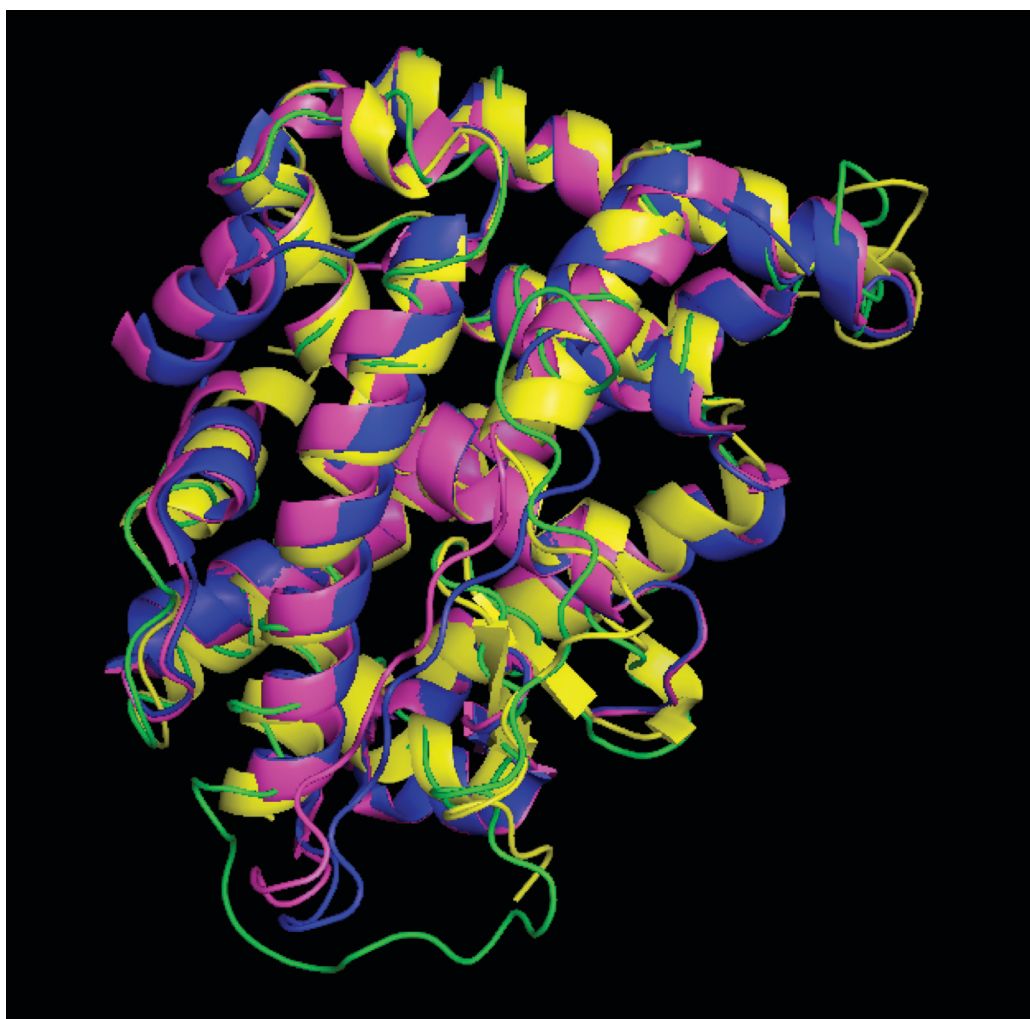


RSC Drug Discovery

Edited by Timothy Clark and Lee Banting

Drug Design Strategies

Computational Techniques and Applications



RSC Publishing

Drug Design Strategies

Computational Techniques and Applications

RSC Drug Discovery Series

Editor-in-Chief:

Professor David Thurston, *London School of Pharmacy, UK*

Series Editors:

Dr David Fox, *Pfizer Global Research and Development, Sandwich, UK*

Professor Salvatore Guccione, *University of Catania, Italy*

Professor Ana Martinez, *Instituto de Quimica Medica-CSIC, Spain*

Dr David Rotella, *Montclair State University, USA*

Advisor to the Board:

Professor Robin Ganellin, *University College London, UK*

Titles in the Series:

- 1: Metabolism, Pharmacokinetics and Toxicity of Functional Groups: Impact of Chemical Building Blocks on ADMET
- 2: Emerging Drugs and Targets for Alzheimer's Disease; Volume 1: Beta-Amyloid, Tau Protein and Glucose Metabolism
- 3: Emerging Drugs and Targets for Alzheimer's Disease; Volume 2: Neuronal Plasticity, Neuronal Protection and Other Miscellaneous Strategies
- 4: Accounts in Drug Discovery: Case Studies in Medicinal Chemistry
- 5: New Frontiers in Chemical Biology: Enabling Drug Discovery
- 6: Animal Models for Neurodegenerative Disease
- 7: Neurodegeneration: Metallostasis and Proteostasis
- 8: G Protein-Coupled Receptors: From Structure to Function
- 9: Pharmaceutical Process Development: Current Chemical and Engineering Challenges
- 10: Extracellular and Intracellular Signaling
- 11: New Synthetic Technologies in Medicinal Chemistry
- 12: New Horizons in Predictive Toxicology: Current Status and Application
- 13: Drug Design Strategies: Quantitative Approaches
- 14: Neglected Diseases and Drug Discovery
- 15: Biomedical Imaging: The Chemistry of Labels, Probes and Contrast Agents
- 16: Pharmaceutical Salts and Cocrystals
- 17: Polyamine Drug Discovery
- 18: Proteinases as Drug Targets
- 19: Kinase Drug Discovery
- 20: Drug Design Strategies: Computational Techniques and Applications

How to obtain future titles on publication:

A standing order plan is available for this series. A standing order will bring delivery of each new volume immediately on publication.

For further information please contact:

Book Sales Department, Royal Society of Chemistry, Thomas Graham House, Science Park, Milton Road, Cambridge, CB4 0WF, UK

Telephone: +44 (0)1223 420066, Fax: +44 (0)1223 420247, Email: books@rsc.org

Visit our website at <http://www.rsc.org/Shop/Books/>

Drug Design Strategies

Computational Techniques and Applications

Edited by

Lee Banting

*Centre of Molecular Design, School of Pharmacy and Biomedical Science,
University of Portsmouth, Portsmouth, UK*

Timothy Clark

Friedrich-Alexander Universitat, Erlangen, Germany

RSC Publishing

RSC Drug Discovery Series No. 20

ISBN: 978-1-84973-167-6

ISSN: 2041-3203

A catalogue record for this book is available from the British Library

© Royal Society of Chemistry 2012

All rights reserved

Apart from fair dealing for the purposes of research for non-commercial purposes or for private study, criticism or review, as permitted under the Copyright, Designs and Patents Act 1988 and the Copyright and Related Rights Regulations 2003, this publication may not be reproduced, stored or transmitted, in any form or by any means, without the prior permission in writing of The Royal Society of Chemistry or the copyright owner, or in the case of reproduction in accordance with the terms of licences issued by the Copyright Licensing Agency in the UK, or in accordance with the terms of the licences issued by the appropriate Reproduction Rights Organization outside the UK. Enquiries concerning reproduction outside the terms stated here should be sent to The Royal Society of Chemistry at the address printed on this page.

The RSC is not responsible for individual opinions expressed in this work.

Published by The Royal Society of Chemistry,
Thomas Graham House, Science Park, Milton Road,
Cambridge CB4 0WF, UK

Registered Charity Number 207890

For further information see our web site at www.rsc.org

Preface

Successful drug design requires a multidisciplinary approach. Understanding the nature of a drug and its intended target requires knowledge of their chemistries, biophysical characteristics and, of increasing importance, their sub-cellular biological context. For this to happen, science and technology must be pushed passed their existing limits. To satisfy this need requires chemists to be highly confident of the valency and covalency between the drug and target, biophysicists to appreciate the dynamic nature of any biomolecular target and potential partners fully, mathematicians to help model molecular systems efficiently and computer scientists to optimise architectures so that complex calculations can be performed practically. Often, as a component of team-based approaches, ‘linchpin’ individuals exist that help catalyse successful collaborations. As with the, now indispensable, hyphenated analytical techniques *e.g.* GC-MS (Gas Chromatography-Mass Spectrometry) multi-disciplinarians such as bio-mathematicians and experts in chemical-computing luckily do exist. These individuals act as the multi-linguists of the scientific community and are often a product of a global educational environment that has facilitated their development. This book has been written in the hope of demonstrating the latest skills of these modern day polymaths and as a showcase for cutting-edge science as applied to drug design problems.

The book starts with excellent accounts of how drug design problems have required the mustering of nearly all the aspects mentioned above. Firstly, the problem of exactly how the naturally occurring compounds known as latrunculins bind to globular actin to exert their biological action and how an apparently simple synthetic substitution can lead to a complex outcome in terms of bioactivity and, secondly, the rationalization of binding of hydroxamate inhibitors to a zinc-based metalloprotein target. Tell Tuttle has had to forge new approaches that merge quantum and molecular mechanical

simulations to describe molecular level interactions more fully that are not treated adequately if a more expedient simplified approach is taken. Rob Deeth follows up this chapter by throwing further light on how transition-metal systems can be modelled successfully by appropriate deployment of molecular mechanical techniques supported by density-functional theory (DFT) approaches. Of particular interest here is the unusual application toward QSPR/QSAR of platinum and copper-based bioactive compounds.

David Ritchie takes us into the biophysical world of protein-protein interactions in recognition of the fact that this aspect of the drug design processes is becoming increasingly important. Here is a prime example of the extreme application of multi-processor computer architecture and programming to make rapid calculation of protein-to-protein orientation and docking for database mining for protein partners readily accessible.

Tim Clark and Alex Voityuk address the modelling, *via* quantum mechanical approaches, of predominantly, but not solely, small molecules. Accurate QSPR and QSAR of ligands often requires the consideration of novel chemistries, subtle dispersion forces and weak hydrogen-bonding situations that the more expedient molecular mechanical methods, which more often than not include simplified treatment of partial charge, fail to model well. Paul Popelier carefully guides us through the state of art for DFT and lets us glimpse at what might be possible due to advances in computational power.

Once information has been derived for small molecule datasets, there is need for accurate, appropriate handling of their structural characteristics, stereochemistry and associated physicochemical and biological properties that can facilitate rapid retrieval of any required information. Chemoinformatics, comprehensively reviewed by Thomas Engel, demonstrates the current 'state of play'. David Whitley then considers the role of topology in characterising the properties of molecules in 3D and their comparison across molecular sets of interest.

A neglected area is that of extended molecular systems such as the giant structures that the inorganic materials can present. Paul Cox addresses the techniques required for effective modelling of alumino-silicates and their relatives acting as novel drug delivery systems and demonstrates that there is wide interest in their applications.

Tim Clark and Lee Banting

Contents

Chapter 1	Quantum Mechanical/Molecular Mechanical Approaches in Drug Design	1
	<i>Tell Tuttle</i>	
1.1	Introduction	1
1.2	QM/MM Methods	4
1.2.1	Additive QM/MM Methods	5
1.2.2	Subtractive QM/MM Methods	7
1.3	Applications of QM/MM Methods to Structure-based Drug Design	8
1.3.1	QM/MM Methods to Aid the Understanding of Ligand–Receptor Interactions	9
1.3.2	QM/MM Methods in Scoring Refinement	16
1.4	Summary	18
	Acknowledgements	19
	References	19
Chapter 2	Transition Metal Systems	27
	<i>Robert J. Deeth</i>	
2.1	Introduction	27
2.2	Methods for Modelling Transition Metal Centres	29
2.2.1	Density Functional Theory	30
2.2.2	Molecular Mechanics	33
2.2.3	Ligand Field Effects in Molecular Mechanics	35
2.3	Scenarios	39
2.3.1	Predicting log <i>P</i> for Pt Complexes	40
2.3.2	Predicting log <i>P</i> for Cu ^{II} Complexes	42

2.3.3	Interactions Involving Explicit Receptors	44
2.3.4	Scenario 1: Metallodrug–DNA Intercalation	45
2.3.5	Platinum–DNA Binding	47
2.3.6	Drug Binding in the Cytochrome P450 Active Site Cavity	48
2.3.7	Drug–Zinc Binding and Substrate Oxidation by P450	50
2.4	Conclusions	53
	References	53
Chapter 3	Modeling Protein–Protein Interactions by Rigid-body Docking	56
	<i>David W. Ritchie</i>	
3.1	Protein–Protein Interactions	56
3.2	Predicting Protein Interaction Surfaces	57
3.3	Protein–Protein Interaction Databases	58
3.4	The Protein Docking Problem	59
3.5	Rigid Body Docking	60
3.6	Docking By Geometric Hashing	61
3.7	Cartesian Grid-based FFT Docking	61
3.8	Spherical Polar Fourier Basis Functions	64
3.9	Spherical Polar Fourier Correlations	67
3.10	Multi-dimensional Spherical Polar FFTs	70
3.11	Exploiting Graphics Processor Units	70
3.12	The ‘Hex’ Docking Program	75
3.13	Conclusions and Future Prospects	79
	Acknowledgements	80
	References	80
Chapter 4	QM Based Modelling	87
	<i>Timothy Clark</i>	
4.1	Introduction	87
4.2	Computational Demands	88
4.3	The Molecular Electrostatic Potential	89
4.4	Local Properties and Intermolecular Interactions	91
4.5	Local Properties and Molecular Surfaces	93
4.5.1	Spherical Harmonic Expansions	93
4.5.2	Surface Autocorrelations	98
4.5.3	Surface Integral Models	99
4.6	Local Properties on Grids	100
4.7	Summary and Conclusions	102
	Acknowledgments	102
	References	103

Chapter 5	Semi-empirical Methods: Current Status and Future Directions	107
	<i>Alexander A. Voityuk</i>	
5.1	Introduction	107
5.2	MNDO and MNDO/d Based Methods	108
5.2.1	MNDO/d	109
5.3	An Overview of Other Semi-empirical Approaches	111
5.3.1	OM x Schemes	111
5.3.2	DFTB	111
5.3.3	MSINDO	111
5.4	Hydrogen Bonds, Non-covalent Interaction and Effects of Environment	112
5.5	Excited States	113
5.6	Electron and Excitation Energy Transfer	114
5.7	Linear Scaling	114
5.8	Concluding Remarks and Perspectives	115
	References	116
Chapter 6	Quantum Chemical Topology: on Descriptors, Potentials and Fragments	120
	<i>Paul L. A. Popelier</i>	
6.1	Introduction	120
6.2	Quantum Chemical Topology (QCT)	121
6.3	Quantum Topological Molecular Similarity (QTMS)	126
6.4	Force Field Development: QCTFF	143
6.5	Bioisosterism	154
6.6	Conclusions	156
	Acknowledgement	157
	References	157
Chapter 7	Cheminformatics in Diverse Dimensions	164
	<i>Thomas Engel</i>	
7.1	Introduction	164
7.2	Zero-dimensional Code Representations	165
7.2.1	CAS Registry Numbers	165
7.2.2	Hash Structural Codes	166
7.2.3	Molecular Descriptors	167
7.3	One Dimensional Code Representations	169
7.3.1	Nomenclature of Chemical and Biochemical Compounds	169
7.3.2	Notation of Chemical Compounds	170
7.3.3	Bit String Representations	173

7.4	Two-dimensional Code Representations	174
7.4.1	Matrices	175
7.4.2	Connection Table	175
7.4.3	Standard Structure Exchange Formats	176
7.5	Conclusion	179
	References	179
Chapter 8	Analysing Molecular Surface Properties	184
	<i>David C. Whitley</i>	
8.1	Introduction	184
8.2	Molecular Surface Models	185
8.3	Molecular Surface Properties	192
8.4	Molecular Surface Similarity	194
8.5	Molecular Surface Property Graphs	198
8.6	Concluding Remarks	204
	References	204
Chapter 9	Atomistic Modelling of Drug Delivery Systems	210
	<i>Paul Cox</i>	
9.1	Introduction	210
9.2	Molecular Mechanics – Energy Minimisation	211
9.2.1	Encapsulation of Multinuclear Platinum Complexes	211
9.2.2	Carbon Nanotubes	213
9.2.3	Fullerenes	215
9.3	Molecular Dynamics	220
9.3.1	Zeolites as Drug Delivery Agents	220
9.3.2	Other Examples of MD Simulations	225
9.4	Quantum Mechanical Methods	226
9.4.1	Delivery of Vancomycin using Mesoporous Materials	226
9.4.2	Design of Conducting Polymers for Drug Delivery	227
9.5	Summary	230
	References	230
	Subject Index	232

CHAPTER 1

Quantum Mechanical/Molecular Mechanical Approaches in Drug Design

TELL TUTTLE

WestCHEM, Department of Pure and Applied Chemistry, University of Strathclyde, 295 Cathedral Street, Glasgow G1 1XL, UK

1.1 Introduction

Quantum mechanical/molecular mechanical (QM/MM) hybrid methods have their origins in the 1970s in the pioneering work of Warshell and Levitt.¹ However, until the last two decades, the uptake of this methodology and its application to solving problems of biochemical interest was essentially non-existent. In the beginning of the 1990s the situation changed dramatically and since this time there has been an explosion in the number of publications that use QM/MM methods to study systems of biochemical interest. Several recent reviews in the literature have provided excellent coverage of the recent developments with a strong focus on the usefulness of QM/MM methods to studying enzyme reactions.^{2–25}

The dramatic uptake of QM/MM methodology has been driven by significant developments in the methodology accompanied by a rapid increase in computer power and decrease in computer cost. Moreover, along with the development in QM/MM methodology, the development of QM methods to deal with larger systems with ever increasing accuracy has allowed these high-level QM methods to be applied to biochemical systems. One can now find examples in the literature where enzyme-catalysed reactions have been

modelled computationally to within chemical accuracy (*ca.* 1 kcal/mol) – a feat unimaginable only a few years previously.²⁶ Despite these impressive strides in accuracy and efficiency, the application of QM/MM methods to the field of drug design has received far less attention until recently.^{8,13,18}

Computational drug discovery is a well-established discipline within the field of computational chemistry.^{27–56} Computational chemistry can be applied to the development of new drugs from both the perspective of the small molecule^{34,47,57–59} (drug-like compound) and the biological target (receptor).^{27,29,39,45,55} In cases where the drug target is unknown or a structure of the target receptor is unavailable, a quantitative–structure activity relationship (QSAR) can be developed by comparison of several ligands with varying biological responses (activities).^{34,47,57–59} One of the oldest and most popular 3D-QSAR methods is a Comparative Molecular Field Analysis (CoMFA), which uses an sp^3 -hybridized carbon with a charge of +1 to evaluate steric and electrostatic features of the compounds in the training set.^{60–64} By comparing the steric and electrostatic features of the compounds in the training set as a function of their activity a relationship between regions that benefit from having certain properties in relative molecular positions (e.g. decreased steric bulk next to a negatively charged region of the molecule) can be identified. While both ligand-based and receptor-based methods have seen significant successes throughout their application, in the current review the focus will be on receptor (structure) based methods.

Structure-based drug design (SBDD) stems from the lock-and-key principle originally proposed by Emil Fischer over 100 years ago. In the lock-and-key model the ligand and the receptor are both considered as rigid objects and the activity of a ligand results from the complementary matching of the steric and electrostatic features of the ligand (key) into the receptor's binding pocket (lock). This principle was subsequently revised in 1958 by Koshland with the proposal of the induced-fit model whereby both the ligand and the receptor are considered to behave dynamically and as such the 'lock' and the 'key' are both able to adapt to some extent to provide a better match where possible during the binding process. However, modelling the dynamic flexibility of a complete macromolecule (receptor) is still not computationally feasible when one wishes to screen a large database of potential drug-like molecules. As such, within modern docking methods a compromise between the lock-and-key method and the induced fit model is generally applied. These methods treat the ligand as a flexible molecule (or a series of different conformations of the ligand is used) and is docked into the rigid receptor,^{65–74} or semi rigid receptor^{75–80} (where flexibility of the side chains in the binding pocket is allowed). Irrespective of the approach used, the key requirement for SBDD is a reliable structure of the target macromolecule.

The SBDD approach usually involves multiple stages. These can be broadly classified as follows.

1. Generation of poses/conformations of the ligand in the active site – this process determines whether the ligand is able, and if so in which orientation, to physically fit into the active site.

2. Scoring and ranking the individual poses – in this process the static poses generated in the first step are given a score, usually based on electrostatic fit between the ligand and the residues that compose the binding site.
3. Refinement of the docked structures and their subsequent ranking – this third step is the most complicated and various different approaches exist to introduce ligand and binding site flexibility into the refinement of the structure (*e.g.* molecular dynamics simulations, random search methods, *etc.*).⁵⁵

Quantum mechanical/molecular mechanical (QM/MM) scoring refinement focuses on the third stage of the SBDD approach – the structural refinement and determination of the final ranking. This refinement process has several problems at a computational level. The methodology employed in the refinement needs to be efficient in order to deal with the large numbers of ligands and poses that need to be refined. Because of this the use of molecular mechanics (*i.e.* empirical force field based approaches) has been the method of choice. However, empirical methods are inherently limited in their accuracy to the quality/specificity of the parameterization.^{81,82} As a result, in repeated regular structures, such as proteins, which are combinations of only 20 amino acids, force field methods are remarkably successful (*e.g.* the CHARMM,^{83,84} AMBER,^{85,86} GROMOS,^{87,88} *etc.* force fields).⁸⁹ However, in the case of ligands, which contain an infinite variety of chemical motifs, the parametrization of effective force fields is much more difficult. As such, docking and scoring programs generally rely on universal force fields for modelling both the ligand and the protein, which are much less successful, but offer a common level of accuracy between the protein and the ligand. In principle, the use of quantum mechanical (QM) methods in the place of molecular mechanical methods would solve these problems, offering a common and accurate description of both the ligand and peptide. However, the main drawback for QM methods is that their computational cost *prohibits their application to such problems*.

The middle ground between purely *ab initio* and purely empirical methods are the set of methods known as semi-empirical methods. These methods offer the chemical flexibility of the *ab initio* methods (though not necessarily the accuracy) but are orders of magnitude more efficient. Throughout the literature one is able to find several examples of the applications of semi-empirical methods in the context of QM/MM biochemical modelling.^{8,13,90–100} The latest semi-empirical methods (OMx,^{101–103} SCC-DFTB,^{104–106} PDDG/PM3,¹⁰⁷ PM6-DH2,¹⁰⁸ *etc.*) offer an accurate and efficient description of the structure and energetics of small molecules. Moreover, these methods have been successfully combined with biochemical force fields to provide an accurate description of the interactions in the binding site, between a ligand – treated at the QM level of theory – and the protein – treated at the MM level of theory.^{8,13,98,99,102}

The general strategy in QM/MM scoring refinement is to run a QM/MM minimization of the binding site (ligand + protein) to refine the structure and binding energies of the various poses obtained from the first two stages of the

virtual screening. This approach simultaneously provides a more accurate modelling of both the ligand and binding site and, as such, should provide enhanced differentiation between the various ligands, as well as a more detailed understanding of the interactions that are present in the binding site. As such, this method should provide considerable benefit in lead optimization activities.

In Section 1.2 an outline of the main QM/MM methodology employed in QM/MM scoring refinement is provided. In Section 1.3 two illustrative examples of the application of this methodology to challenging scoring problems are provided. In the final section (1.4) the main difficulties associated with QM/MM scoring are addressed and the potential future directions of the field are discussed.

1.2 QM/MM Methods

The basic concept of a QM/MM method is to describe the chemically interesting region of the system (in this case the ligand and potentially the binding pocket) using a QM method and the remainder of the system, the environment, at the MM level of theory. In the most common case the ligand binds to the receptor non-covalently (Figure 1.1) and as such only the ligand is treated at the QM level of theory as the MM description of the protein (receptor) is reliable when appropriate biochemical force fields are chosen. The description of the ligand region using a QM method and the receptor using an MM method is trivial. However, the goal of any QM/MM method is to describe the interaction between the two regions – in other words, how do we best describe the coupling between QM and MM parts of the system? The coupling terms between the ligand and the receptor are exactly the interactions that will dictate the binding ability of the ligand and can be quite strong. Indeed, a balanced treatment of the three components, QM, MM and the coupling terms are equally important for any QM/MM scoring function to be successful.

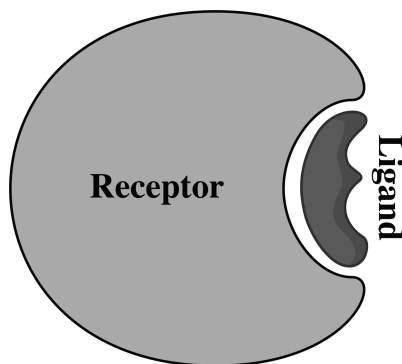


Figure 1.1 Partitioning of the full system into the QM region (Ligand – dark grey) and the MM region (Receptor – light grey).

1.2.1 Additive QM/MM Methods

There exist two families of QM/MM approaches: those based on the additive scheme and those based on the subtractive scheme (see Section 1.2.2). The additive scheme, initially proposed by Warshel and Levitt in 1976,¹ is employed in the majority of QM/MM programs presently in use and generally contains three terms:

$$E_{\text{QM/MM}}(\text{Tot}) = E_{\text{QM}}(\text{Lig}) + E_{\text{MM}}(\text{Rec}) + E_{\text{QM-MM}}(\text{Lig,Rec}) \quad (1.1)$$

where $E_{\text{QM/MM}}(\text{Tot})$ is the total energy of the full system calculated at the QM/MM level of theory, $E_{\text{QM}}(\text{Lig})$ is the energy of the ligand calculated at the QM level of theory, $E_{\text{MM}}(\text{Rec})$ is the energy of the receptor calculated at the MM level of theory and $E_{\text{QM-MM}}(\text{Lig,Rec})$ is the explicit coupling term between the ligand and the receptor. As we shall see in the subtractive scheme there is no explicit coupling term present; however, the variation in how the coupling term is defined allows one to further differentiate between various additive QM/MM approaches. These approaches were described in detail by Bakowies and Thiel in 1996,¹⁰⁹ and have recently been reviewed in the literature; as such, an account will not be provided here.²¹ The QM/MM approach discussed herein refers to the electrostatic embedding approach,¹⁰⁹ unless otherwise stated. For details of other methods the reader is referred to the literature.^{2,4,5,9,10,14,19–21,110–112}

Where there are no bonds that cross the QM/MM boundary, as is the case for non-covalently bound ligands, the coupling term, in the electrostatic embedding approach, has two components:

$$E_{\text{QM-MM}}(\text{Lig,Rec}) = E_{\text{QM-MM}}^{\text{vdW}} + E_{\text{QM-MM}}^{\text{elec}} \quad (1.2)$$

where $E_{\text{QM-MM}}^{\text{vdW}}$ refers to the van der Waals (vdW) interactions between the ligand atoms and the atoms in the receptor; and $E_{\text{QM-MM}}^{\text{elec}}$ refers to the electrostatic interactions between the ligand and receptor atoms. If a covalent bond exists between the QM and MM regions a boundary scheme needs to be chosen in order to properly describe the forces acting across the boundary and also to ensure appropriate ‘capping’ of the QM region. The various schemes for dealing with QM–MM boundary regions have been recently reviewed and the interested reader can find details of these from the reviews.^{10,20,25}

The vdW interactions are treated at the MM level of theory and as such the atoms in the ligand require vdW parameters to be assigned to them. This assignment is usually done by analogy with ‘similar’ atom types that exist in the force field. However, alternative, and more rigorous, approaches for assigning vdW parameters to the atoms treated at the QM level of theory have also been presented.^{16,113} The error introduced by the MM treatment of the vdW interactions between the ligand and the receptor regions is expected to be minimal due to the short-range nature of these interactions.¹¹⁴

Due to its long-range nature, the predominant coupling term between the ligand and the receptor regions is the electrostatic term ($E_{\text{QM-MM}}^{\text{elec}}$). Within the

electronic embedding approach, the electrostatic coupling between the ligand and receptor descriptions is achieved by performing the QM calculation in the presence of the point charges (taken as the partial charges assigned within the force field) of the receptor. This is achieved by including the partial charges as one-electron terms in the QM Hamiltonian, resulting in the additional term ($\hat{H}_{\text{QM-MM}}^{\text{elec}}$), defined in Equation (1.3).

$$\hat{H}_{\text{QM-MM}}^{\text{elec}} = -\sum_i^N \sum_j^L \frac{q_j}{|r_i - R_j|} + \sum_k^M \sum_j^L \frac{q_j Q_k}{|R_k - R_j|} \quad (1.3)$$

The electron positions are given by r_i , the point charges, as defined in the force field, for the receptor are denoted q_j and are located at R_j and Q_k are the charges on the nuclei in the ligand located at R_k . The indices i , j and k correspond to the N electrons, the L point charges in the receptor and M nuclei in the ligand, respectively.

The inclusion of the point charges from the receptor in the Hamiltonian results in the polarization of the electronic structure of the ligand. If the charges in receptor change their positions during a geometry optimization or a molecular dynamics simulation, then the electronic structure is able to respond to these positional changes realistically as the electrostatic coupling is treated at the QM level of theory.

1.2.1.1 Extended Solvation

The long-range nature of the electrostatic interactions imbues them with a special significance in QM/MM calculations. The long-range nature implies that a significantly large proportion of the environment needs to be included in a calculation in order to correctly reproduce the electrostatic effects on the chemically interesting region of the system.^{7,115,116} However, increasing the size of the environment (receptor and solvent) carries two important pitfalls; (1) the larger the number of degrees of freedom in a system, the greater the chance of falling into localized, disconnected, minima; and (2) the larger the system, the more computationally intensive the calculation becomes. During the initial development of QM/MM methods this second factor was considered essentially irrelevant as the calculation of the QM component was expected to be the rate-limiting step. However, as modern QM methods have increased in both speed and accuracy this is no longer necessarily true, particularly when semi-empirical methods are employed for the QM calculation.

The concern of localized, disconnected minima is particularly troublesome in SBDD where the goal is to determine the relative binding energy of different ligands in the binding site and often the variation can be as little as several kcal/mol across a series of competitive ligands. Given this small window of energy differences, spurious variations in ranking that can occur from changes in the environment (*e.g.* alterations in the H-bonding network of the solvent) or in distant parts of the receptor (*e.g.* rotation of an amino acid on the protein surface causing a stronger or weaker salt-bridge to be formed) are

unacceptable. One approach that has been commonly employed to deal with this problem has been to freeze atoms in the solvent and receptor that are farther than a pre-determined distance (typically 10–15 Å) away from the ligand atoms. This procedure reduces the number of degrees of freedom in the QM/MM simulation and treats the ‘frozen’ part of the system as a fixed external potential that interacts with the mobile inner ‘core’ of the system. Despite the fixed nature of the outer region, this approach still requires the computation of all interactions between the frozen atoms in the system in addition to the interaction between every atom in the frozen region with every atom within the inner core region. Thus, freezing the more distant components of a system does not alleviate the second pitfall – *i.e.* the computational cost associated with extended, solvated systems.

This problem of increasing system size has, of course, also plagued those in the MM community where there has been significant effort in developing accurate methods to treat long-range electrostatics. For large systems, where periodic boundary conditions are employed, the use of the Ewald summation is well established and has been employed in the MM community for some time.^{117–119} This approach has also been extended and implemented in a number of packages for QM/MM simulations as well.^{96,120,121} However, the majority of QM/MM simulations are performed on finite (non-periodic) systems where the Ewald summation is no longer valid.^{116,122–124} Therefore the use of solvent boundary potentials^{125–129} has become increasingly prevalent.

The generalized solvent boundary potential¹²⁶ (GSBP) is of particular interest and has already been implemented by a number of QM/MM developers.^{130–134} Within this approach the total system (Tot) is divided into three components – the receptor, which has the same definition as stated previously and is treated at the QM level of theory, an outer region, which includes parts of the receptor and solvent that is within a defined distance from the receptor and is again treated atomistically at the MM level of theory, and finally an extended outer region for all atoms that are outwith the defined boundary. Within the GSBP approach this extended outer region can include irregularly shaped parts of the macromolecule as well as bulk solvent.^{130,134} The solvent molecules are described by a continuous polarizable dielectric, whilst the irregular charge distribution is described by a fixed set of point charges. This approach has been implemented and tested for a number of biomolecular simulations and it is a promising approach for the treatment of long-range electrostatic interactions in QM/MM simulations, which can easily be extended to a range of QM/MM applications, including the scoring refinement step in SBDD.^{130–134}

1.2.2 Subtractive QM/MM Methods

The IMOMM (integrated molecular orbital/molecular mechanics) method by Morokuma and co-workers is the prototypical example of a subtractive QM/MM method.¹³⁵ The IMOMM method was subsequently extended by the authors to allow for the combination of two QM methods in the IMOMO¹³⁶ and finally to the popular ONIOM (our n-layered integrated molecular orbital

and molecular mechanics) method.^{137,138} However, the later ONIOM and IMOMO methods are not, strictly speaking, QM/MM approaches given that the lower level of theory is often a semi-empirical or ‘cheaper’ QM method such as Hartree-Fock, rather than an MM method (a two-level ONIOM method, where the lower level is an MM method, corresponds to the IMOMM approach, which is most commonly used in SBDD studies).^{139–141} The strength of the subtractive QM/MM method is primarily in its ease of implementation. There is *no* explicit coupling between the regions calculated at the QM and MM levels of theory (the $E_{\text{QM-MM}}(\text{Lig}, \text{Rec})$ term in Equation 1.1). Rather, the coupling between the ligand and its environment is handled at the lower (MM) level of theory.

$$E_{\text{QM/MM}}(\text{Tot}) = E_{\text{QM}}(\text{Lig}) + E_{\text{MM}}(\text{Tot}) - E_{\text{MM}}(\text{Lig}) \quad (1.4)$$

The subtractive scheme requires three separate calculations in order to construct the total energy of the system. The first calculation is at the MM level of theory on the entire system. The ligand is then calculated at the QM level of theory and finally, independently, at the MM level of theory. By subtracting the MM energy of the ligand the subtractive scheme removes all ‘internal’ MM contributions of the ligand from the MM energy and thus results in an MM energy that contains only the energy of the receptor and solvent and the MM energy of the interaction between the ligand and its environment. Thus, within the subtractive method the coupling of the ligand to the receptor is simply contained within the initial MM calculation of the full system.

The simplicity of the subtractive method, unfortunately, does come at a cost. As the coupling interaction between the ligand and its environment is handled completely at the MM level of theory, each ligand needs to be well parameterized within the force field representation that is being used. While this is possible for small databases of ligands that are being screened,^{139–141} one of the driving concepts behind QM/MM SBDD is the ability to accurately represent the vast chemical space that ligand compounds occupy using QM methods. Finding or developing adequate parameters for a full library of compounds can be quite challenging and will generally be non-transferrable between different systems or different force fields. The accuracy of the subtractive method is therefore limited by the accuracy of the MM representation of the ligand.

This type of coupling, at the molecular mechanical level of theory, is often referred to as mechanical embedding.¹⁰⁹ The QM calculation of the ligand is generally not performed in the presence of the point charges of the environment and as such the QM density is not polarized by the environment. Alternative subtractive schemes have recently been introduced that now allow electrostatic embedding.^{142–145}

1.3 Applications of QM/MM Methods to Structure-based Drug Design

The overwhelming application of QM/MM methods within structure-based drug design can be divided into two main groups. Firstly, there are several

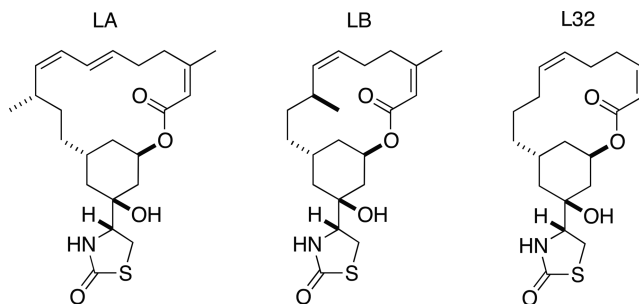
applications of QM/MM methods to gain a detailed understanding of the binding site interactions that play a role in discriminating between closely related ligands. In this type of application there is often only a small number of closely related ligands, which nonetheless demonstrate a large degree of diversity in their binding ability. From an empirical perspective the variation in activity of closely related ligands is often difficult to explain and, as such, detailed investigations into the binding modes of such ligands can shed further light onto the mechanism by which the ligands are stabilized in the binding site and consequently can aid in the further development of new lead compounds. The second main application of QM/MM methods is in the scoring refinement of a set of compounds based on an initial screen. The scoring refinement typically will make use of semi-empirical methods and focus less on the details of the binding mechanism, but rather uses the QM/MM method to gain an accurate representation of the various ligands, relative to the initial force field based screening protocol. In the following we consider two case studies that highlight the important role of QM/MM methods in both of these application areas.

1.3.1 QM/MM Methods to Aid the Understanding of Ligand–Receptor Interactions

The use of QM/MM methods to understand the various factors that help to differentiate the binding ability of closely related ligands is exemplified by two recent publications by Tuttle *et al.*^{98,102} The first of these demonstrates how both the non-covalent interactions between the binding site and the effects of strain energy on ligands, imposed by the binding site, affect the binding ability of Latrunculin A and two of its analogues to G-actin.⁹⁸ In the second study, the role of dispersion in helping to stabilize ligands in the binding site was explored through dispersion augmented (-D) QM-D/MM methods.¹⁰²

1.3.1.1 Latrunculin Analogues Binding to Globular Actin

Latrunculins form 1:1 complexes with globular actin (G-actin), which interferes with the ability of the protein to polymerize and form actin filaments. The use of these compounds in chemical biology as probe molecules is widespread and as such a ‘diverted synthesis’ campaign was carried out by Fürstner *et al.*^{98,146} in order to develop a library of closely related analogues that could shed light onto the mechanism by which the compounds achieved their remarkable biological activity. As a result of this study, one compound (L32, Scheme 1.1) was identified as having biological activity that was as good as, or even better, than the naturally occurring latrunculins. The aim of the computational study was therefore to elucidate the binding mode of Latrunculin A (LA, Scheme 1.1) and two of its analogues – the naturally occurring compound Latrunculin B (LB, Scheme 1.1) and the synthetic analogue L32 – to G-actin. Experimentally, the binding strength was measured by the ability of the compound to inhibit the biological response of actin towards fibroblasts. The experimental study



Scheme 1.1 Naturally occurring compounds Latrunculin A (LA), Latrunculin B (LB), and the synthetic analogue L32.

revealed that the synthetic analog (L32) was as effective an inhibitor as LA ($LA \geq L32 > LB$).⁹⁸

Methodological Approach. The ‘standard’ approach for preparing biological systems for modern QM/MM studies was applied. Following the preparation of the receptor and environment, the ligands were constructed and a conformational search was carried out in order to determine low-energy conformers of the relatively flexible ligands. The resulting low-energy conformers were then docked into the receptor, the resulting complexes were then minimized and finally optimized at the QM/MM level of theory.

Structure Preparation. The structure of actin with LA bound is available from the protein data bank (PDB ID: 1ESV).¹⁴⁷ There are several unresolved residues in the protein structure, although as these residues are remote from the binding site, the backbone was capped with neutral residues at these positions. The protonation state of the ionizable residues was checked with PROPKA¹⁴⁸ and REDUCE,^{149,150} which was also used to determine the orientation of the ambiguous residues (His, Asn, Gln). The crystal waters, three Ca^{2+} ions and ATP were retained; however, no additional hydration was performed.

A conformational search was performed for LA, LB and L32 using the PM3 Hamiltonian as implemented in Spartan.¹⁵¹ The ten lowest energy conformers of each molecule were selected and subsequently optimized within MNDO99¹⁵² using the MNDO/H Hamiltonian,^{153,154} which provides reasonable geometries and relative energies for hydrogen bonded systems.^{153,155} The reliability of the semi-empirical structures was checked via re-optimization of the conformers at the BLYP^{156,157} level of theory with the def2-SVP basis set¹⁵⁸ available in TURBOMOLE.^{159–163}

Docking. For each compound, the ten conformers (optimized with BLYP) were then docked into the actin-binding site using AutoDock 3.0 (AD3).⁷⁰ The AD3 grid was constructed such that it encompassed and was centred on

the binding site of LA (available from the original PDB). AD3 is not parameterized for substrates, such as ATP or Ca^{2+} , and as such these residues are ignored during the docking procedure. As ten different conformational variants of each ligand were tested, the ligand was kept rigid during the docking process. The orientation resulting in the best binding energy for each conformer was selected from each docking run for optimization.

Force Field Minimization. The complex with the best binding energy from each docking simulation was pre-minimized in CHARMM.^{84,164,165} The standard CHARMM force field^{166–168} used in this work does not contain parameters for the LA, LB or L32. However, given the preparative nature of the force field minimization (*i.e.* optimization of the full complex would be performed at the QM/MM level of theory), the parameters and topology of the ligands were automatically generated using the commercial CHARMM force field¹⁶⁹ in the Accelrys suite of programs. The ligands were kept rigid and the remainder of the system was then relaxed during a 10 000-step ABNR minimization.

QM/MM Optimization. The structures resulting from the force field minimization were optimized at the hybrid quantum mechanical/molecular mechanical (QM/MM) level of theory with the modular program package CHEMSHELL.^{170,171} The QM energy and gradients were provided by MNDO99,¹⁵² with MNDO/H^{153,154} as the QM level of theory. Chemshell's internal force field driver, using the CHARMM parameter and topology data,^{166–168} supplied the MM energy and gradients. No electrostatic cut-offs were employed in the QM/MM calculations. Electrostatic embedding¹⁰⁹ was used to couple the QM and MM regions. The QM/MM geometry optimizations were performed with the HDLCOpt algorithm,¹⁷² as implemented in CHEMSHELL. Only the most stable orientation of the ligand in the binding site, as judged by the complex with the lowest QM/MM energy after optimization was used in the analysis of the ligands' binding abilities.

Hydrogen Bond Networks. The binding sites of the three ligands form the QM/MM optimized structures (see Figure 1.2). The binding site of the LA complex (see Figure 1.2a) revealed a H-bond network similar to that observed in the PDB crystal structure that the complex was derived from.¹⁴⁷ This similarity both in the orientation and the H-bonding network between the docked and optimized structure (see Figure 1.2a) and the crystal structure indicates that the methods applied were able to accurately reproduce the correct binding mode. The complex exhibits strong H-bonds between the head group and the surrounding residues, while the macrocycle has no obvious H-bond pairing with the protein. In the head group the NH functionality forms a strong H-bond to Asp107 and the carbonyl forms a strong H-bond to

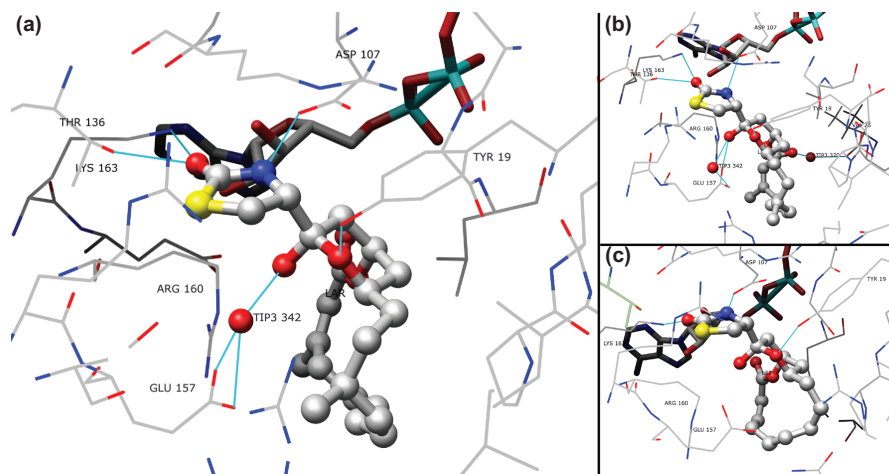


Figure 1.2 H-bond network of (a) LA (b) LB and (c) L32 in the actin binding site after the QM/MM optimization.

Thr136, while there exists only a weak interaction with Lys163. The O atom of the 6-membered ring (O1) forms a strong H-bond with Tyr19.⁹⁸

Analysis of the H-bonding between analogues (LB and L32) and the receptor revealed some changes in the H-bonding network. In the case of LB, these changes led to a slightly weaker H-bond network with the two strong and one average H-bond in the LA complex being exchanged for one strong and two average H-bonds in the LB complex (see Figure 1.2b). However, for the L32 ligand, although there is a shift in the binding pattern, the qualitative strength of the H-bonds (two strong and one average) remains similar to the LA complex. These minor changes in the hydrogen bonding network did offer some insight into the relative binding ability and, consequently, the biological activity of the three compounds (*i.e.* $LA \geq L32 > LB$) although it was unlikely that such minor changes in the binding enthalpy, associated with the redistribution in H-bonding, could fully explain the observed effect on the overall binding ability.⁹⁸

Hydrophobic Contributions. The changes in the H-bond network of the binding site offer a qualitative explanation for the observed differences in the biological activity of the three ligands. However, the contribution from the hydrophobic effect will aid in stabilizing the ligands in the binding site. In Figure 1.3, the hydrophobicity of the binding pocket in the LA complex is mapped onto a Connolly surface^{173,174} of the receptor. The hydrophobicity is determined by the partial charges on the amino acid residues (assigned in the CHARMM force field). The majority of the macrocycle prefers a hydrophobic (non-polar) environment (green regions in Figure 1.4). This is achieved by a small part of the macrocycle (lower left region in Figure 1.4), although the majority of the system is solvent exposed (no surface) or surrounded by polar (red) residues.

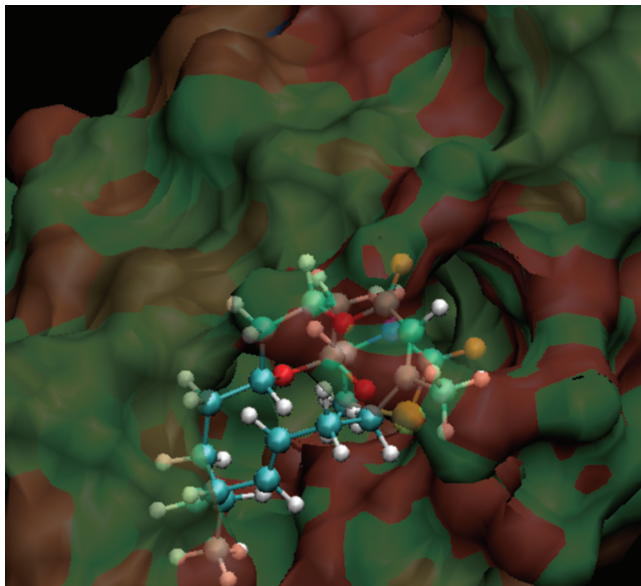


Figure 1.3 Hydrophobicity mapped onto a Connolly surface of the receptor in the LA complex.

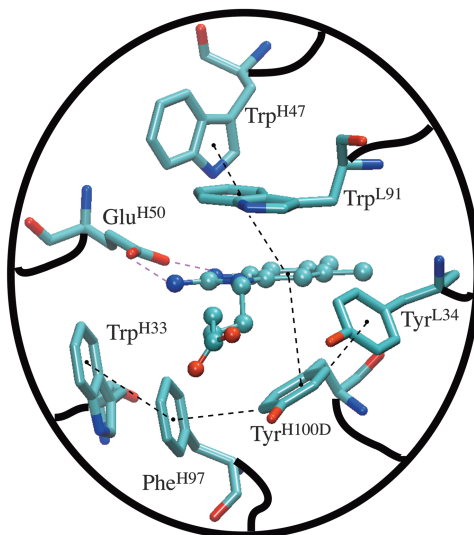


Figure 1.4 The dispersive interactions (black dashed lines) stabilizing the haptent in the binding site are described with the same accuracy as the ionic H-bonds between GluH50 and the haptent (pink dashed lines).

The hydrophobic contribution to the stability of the complex could be estimated using an empirical function based on the change in the solvent accessible surface area (ΔSASA) in the bound and unbound states. The relationship proposed by Spolar and Record:¹⁷⁵

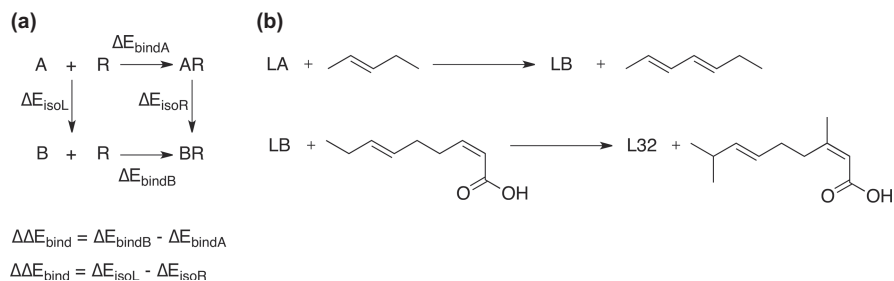
$$\Delta G_{\text{hyd}} = -22(\pm 5)\Delta\text{SASA} \quad (1.5)$$

results in a stabilizing free energy contribution of -20 ± 4.5 kcal/mol when the LA complex is formed. For the smaller ligands LB and L32 the stabilizing effect is slightly smaller (-19 ± 4.3 kcal/mol and -17 ± 4.3 kcal/mol) as a greater percentage of the ligand remains solvent exposed in the optimized structures. The calculated values clearly indicate that the binding of the ligand is favourable from a hydrophobic perspective; however, given that the error bars for this empirical approach the criteria offers no clear distinction between the three ligands.

Relative Binding Energies. The individual binding energy of the substrates can in principle be calculated directly. However, binding potential energies generally over-estimate the binding free energy of a ligand to the receptor, which is measured experimentally, due to the neglect of the dynamic factors that influence the binding ability of ligands. Nonetheless, given the structural similarity of the three compounds it was reasonable to assume that entropic and desolvation effects are approximately equivalent for the three ligands. Therefore, the *relative* binding potential energies ($\Delta\Delta E_{\text{bind}}$) may be reasonable.

The lowest energy structures of the complexes and the ligands optimized in the gas phase (QM = MNDO/H) can be used to calculate binding energies. For LA the binding potential energy is -72.1 kcal/mol. The relative stabilities of the ligands and the complexes can be determined from the isodesmic reactions defined in Scheme 1.1. The partners in the isodesmic reactions (pent-2-ene, hept-2,4-ene, nona-2,6-dienoic acid and 3,8-dimethyl-2,6-dienoic acid) were optimized in the gas phase at the MNDO/H level of theory. The isodesmic reactions indicate that the LB complex is destabilized by 12.5 kcal/mol and the L32 complex by 7.5 kcal/mol, relative to the LA complex. Applying the same isodesmic reactions to the isolated ligands, we find that LB and L32 are destabilized by 2.3 and 1.1 kcal/mol relative to LA in the gas phase.⁹⁸

The cycle depicted in Scheme 1.2 demonstrates that the relative binding energy ($\Delta\Delta E_{\text{bind}}$) of the ligands is the difference of their relative stability in the enzyme (ΔE_{isoR}) and in the gas phase (ΔE_{isoL}). Using the values obtained from the isodesmic reactions, the relative binding energies decrease in the order $\text{LA} > \text{L32} > \text{LB}$ (0 *versus* 6.4 *versus* 10.2 kcal/mol). In all cases the binding of the ligand to the receptor causes a destabilization of the ligand due to the steric constraints imposed by the binding site. The LA and L32 ligands are destabilized by similar amounts (15.2 kcal/mol and 16.8 kcal/mol, respectively), but much less so than the LB ligand (21.5 kcal/mol). A similar distortion results for the receptor upon binding which is consistent with the ligand size, *i.e.* as the



Scheme 1.2 (a) Thermodynamic cycle for calculating relative binding energies. (b) Iso-desmic reactions for comparing the stability of LAR to LB and LB to L32.

ligand size decreases, the distortion of the receptor structure from the unbound state decreases.⁹⁸

Through the QM/MM optimization of the ligand–receptor complexes this study was able to determine that the distortion energies of the ligands, in concert with the differences in the H-bond network, resulted in the large variation in the biological activity for the ligands. The LA and L32 ligands both have H-bond networks that are similar in strength, equally, the ligand distortion of L32 and LA are similar. Thus, LA and L32 have similar binding abilities to the receptor. In contrast, the H-bond network is slightly weaker for LB and the distortion of the ligand from its optimum geometry is larger in this case.

1.3.1.2 Dispersion Interactions in the Binding Site Described by QM/MM Methods

Dispersion interactions, such as π -stacking, play an important role in biochemical structure and stability. As such there are often a number of π -stacking partners in a binding site that a ligand is able to interact with in order to create a stable complex. However, describing dispersion interactions using DFT or semi-empirical methods is notoriously difficult and as such the quality of the methods in describing such interactions within a QM/MM context has been unclear. However, a recent study on the binding of a hapten to the 34E4 enzyme has illustrated the utility of dispersion-augmented QM methods in obtaining an accurate description of the role of both dispersion and hydrogen bonding interactions in receptor–ligand binding.

Dispersion-augmented QM methods (usually denoted by the suffix -D to the QM method, *e.g.* OM3-D) contain an additional parameterized term, with the familiar r^{-6} scaling.^{176,177} Despite the empirical nature of this term, several studies at both the DFT and semi-empirical level have shown that with the appropriate parameterization the inclusion of an empirical dispersion term greatly increases the accuracy of these QM-based methods.^{95,102,176–180}

Within a QM/MM context, mechanistic studies of enzyme-catalysed reactions¹⁸¹ have revealed that including an empirical dispersion term with DFT-based methods for modelling the QM region improves the accuracy of the

calculated potential energy landscape. However, the use of dispersion-augmented methods within a SBDD format is equally important if selected residues of the biomolecule are also included in the QM region. If the receptor is modelled completely at the MM level of theory, the dispersion interactions between the ligand and receptor are handled at the MM level of theory, as described in Section 1.2.1, and this is generally adequate. However, if selected binding site residues are included into the QM region, or if the ligand will potentially have competitive intramolecular dispersion interactions that affect the orientation and conformation of the ligand in the binding site, then a dispersion-augmented QM method is required.

A recent study by Tuttle *et al.* on the role of dispersion forces in the binding site of the antibody 34E4 demonstrated the importance of dispersion forces in stabilizing the binding site and binding the associated hapten (see Figure 1.4).¹⁰² Using the OM3-D semi-empirical method in combination with the CHARMM force field the authors equilibrated the hapten–34E4 complex at the QM/MM level of theory. The OM3-D dispersion terms were found to contribute 26 kcal/mol to the binding of the hapten. The use of dispersion-corrected QM methods is therefore necessary to obtain a consistent and unbiased description of the non-covalent interactions that contribute to the binding energy and consequently the ranking of relative ligands in their ability to bind to a receptor.

1.3.2 QM/MM Methods in Scoring Refinement

QM/MM methods are currently being used in rescoring (scoring refinement) protocols within the SBDD strategy. One of the most important and outstanding applications of these methods is to the field of metalloprotein receptors.^{8,13,99,182} The necessity of using a QM/MM method in the scoring of ligands that bind to metalloprotein systems is the requirement of obtaining a realistic representation of the charge transfer that occurs when ligands bind to the metal centre in the protein. Moreover, a recent study has also highlighted the deficiency of standard force fields when metal ions are incorporated into the binding site, which can lead to the necessity of including proximal residues from the protein into the QM region as well.¹⁸²

Proteins that contain metal ions, and as such a large charge, in the binding site can result in the polarization of proximal residues that alters significantly their charge distribution away from the generic charge distribution that is parameterized in the force field description of the amino acids. However, rather than recalculating the charges for these specific amino acids and fixing them at new potentials in the unbound state, the QM/MM methodology allows for an ‘on-the-fly’ recalculation of the charge distribution simply through the inclusion of these susceptible amino acids into the QM region as well. Clearly, the difficulty with increasing the size of the QM region lies in the additional computational cost associated with this expansion. However, with efficient and accurate semi-empirical QM methods this computational cost is not necessarily prohibitive on modern computing architectures.

In the study by Cho and Rinaldo, they used an ‘extended QM/MM rescoring method’ that uses single point QM/MM calculations on structures obtained from the docking program Glide to recalculate the charge distribution in both the ligand atoms, the metal ion and the atoms in surrounding protein amino acids.¹⁸² The recalculated charges from the QM/MM calculation are then fed back into the docking program, creating a modified force field, which is used to re-dock the ligands into the receptor-binding site. This iterative process can conceivably be continued until the fluctuations between charges, in successive docking runs, have converged. In the study by Cho and Rinaldo, their method was tested on a set of eight matrix metalloproteinases, which had associated ligands with known binding orientations and energies.¹⁸² In seven of the eight cases the correct binding mode was identified with excellent agreement between the calculated and experimental binding orientation. Moreover, the predicted binding energy was also significantly improved, relative to the binding energy predicted with the standard Glide force field. Thus, although this approach does not directly use the QM/MM potential energy in the scoring function, the use of QM/MM methods to refine the docking force field in a self-consistent manner is a further example of the utility of QM/MM methods in SBDD.

A more direct application to the use of QM/MM methods in refining the scoring of ligand binding to a receptor is exemplified by the Balaz *et al.* where 28 hydroxamate inhibitors were bound to the zinc-dependent matrix metalloproteinase-9 (MM9).⁹⁹ In this study the authors used a four-step process to obtain a refined scoring for the ligands to the MM9 receptor, as follows.

1. This involved an initial docking of the ligand to the receptor using a standard docking force field (FlexX) with the best poses from this study chosen based on (i) a structural criterion such that the oxygen atoms in the ligands were aligned with the catalytic zinc atom; and (ii) if (i) was satisfied for multiple poses the best FlexX ranking was then employed.
2. After the docked poses for each of the ligands was obtained the structures were then optimized at the QM/MM level of theory in order to obtain an accurate description of the interaction of the ligand with the metal atom. The QM region for each system involved the ligand, the metal atom, three histidine residues that were covalently linked to the zinc ion and a proximal glutamate that was expected to closely interact with the ligand. The atoms involved in the QM region were used in this study (see Figure 1.5).
3. After the initial QM optimization, the bonds and angles between the zinc ion and the ligand were constrained and a MM MD simulation was carried out on the complex in order to allow the binding site of the complex to relax and to obtain an ensemble of structures that was consistent with the bound state of the system.
4. Finally, structures at 100 fs intervals along the 5 ps MD simulation trajectory were extracted and single point QM/MM calculations were carried out on these structures in order to obtain QM/MM binding energy for the time-averaged structures.

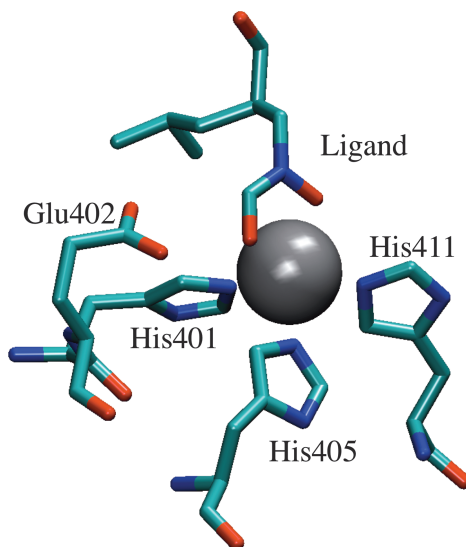


Figure 1.5 QM region included employed in the QM/MM study of 28 hydroxamate inhibitors binding to matrix metalloproteinase-9.⁹⁹ Zinc atom (grey) displayed using the VDW radius, while all other residues are displayed in licorice style.

The inclusion of the MD simulation and QM/MM energy on the time-averaged structures (steps 3 and 4) resulted in a significant improvement in the correlation between the experimental and predicted binding energies. In comparison with the predicted binding from the FlexX simulation, the correlation coefficient is increased from 0.044 (for FlexX) to 0.900 when the QM/MM method described above was used.⁹⁹ It should be noted that this excellent correlation was only possible after the use of a Δ SASA (change in solvent accessible surface area) term in order to include an estimation of desolvation effects. Nonetheless, with this term taken into account, there is still a marked improvement in the accuracy of the correlation when changing from the single optimized QM/MM energy (*i.e.* obtained after step 2) with a Δ SASA term included, to the energy of the time-averaged structures (*i.e.* after step 4).⁹⁹ This result clearly indicates the importance of including both the effect of sampling and obtaining accurate energies for the ensemble of structures in order to discriminate between the binding abilities of structurally similar ligands.

1.4 Summary

The use of QM/MM methods within a drug discovery context is still within its infancy. To date the majority of applications focus on providing a more refined, detailed understanding of specific interactions within the binding site. There exist a few text cases in the literature which exemplify the use of QM/MM methods as part of a screening protocol; however, the application of this

approach to screening libraries of compounds is still a development area and has not yet been taken up significantly by the drug-discovery community.

While QM/MM methods often provide an improved description of ligand–receptor binding this description is often not sufficient to independently provide the correct ranking of ligands in terms of their binding energies. As was observed in the case of the matrix metalloproteinase study,⁹⁹ assessing the hydrophobic (entropic) contribution to the binding ability of a ligand to a receptor is often of equal or even greater importance than the stability gained through individual intermolecular interactions between the ligand and the binding site residues in determining the relative binding ability of ligands. Thus while an improved understanding and description of the intermolecular interactions in the binding site is important, the current inability of QM/MM methods to accurately account for the hydrophobic effect still minimizes their impact in the field.

The area where QM/MM methods have made the most significant impact and appear likely to continue to do so, is in the binding of ligands to metalloproteins and other receptors with highly charged binding sites. The role of quantum effects in these systems (particularly effects such as charge transfer and polarization) for stabilizing the ligand binding are of such importance that the standard empirical force field approaches, which are incapable of describing these effects, are completely inadequate in. In these cases, QM/MM methods offer a significant advantage and as these methods develop further to allow more rapid and diverse screening protocols, their application to metalloprotein receptors will remain a priority.

Acknowledgements

The author acknowledges the Royal Society of Edinburgh for support through a Scottish Executive Personal Research Fellowship, the EPSRC (EP/F031769) and the Glasgow Centre for Physical Organic Chemistry for funding.

References

1. A. Warshel and M. Levitt, *J. Mol. Biol.*, 1976, **103**, 227–249.
2. T. C. Bruice and K. Kahn, *Curr. Opin. Chem. Biol.*, 2000, **4**, 540–544.
3. A. E. Cho, V. Guallar, B. J. Berne and R. Friesner, *J. Comput. Chem.*, 2005, **26**, 915–931.
4. M. J. Field, P. A. Bash and M. Karplus, *J. Comput. Chem.*, 1990, **11**, 700–733.
5. R. A. Friesner and M. D. Beachy, *Curr. Opin. Struct. Biol.*, 1998, **8**, 257–262.
6. R. A. Friesner and V. Guallar, *Annu. Rev. Phys. Chem.*, 2005, **56**, 389–427.
7. M. Garcia-Viloca, J. Gao, M. Karplus and D. G. Truhlar, *Science*, 2004, **303**, 186–195.

8. M. P. Gleeson and D. Gleeson, *J. Chem Inf. Model.*, 2009, **49**, 670–677.
9. I. H. Hillier, in *2nd European Conference on Computational Chemistry (EUCCO-CC2)*, Lisbon, Portugal, 1997, pp. 45–52.
10. H. Lin and D. G. Truhlar, in *10th Electronic Computational Chemistry Conference*, West Long Branch, NJ, 2005, pp. 185–199.
11. R. Lonsdale, K. E. Ranaghan and A. J. Mulholland, *Chem. Commun.*, 2010, **46**, 2354–2372.
12. S. Marti, M. Roca, J. Andres, V. Moliner, E. Silla, I. Tunon and J. Bertran, *Chem. Soc. Rev.*, 2004, **33**, 98–107.
13. L. C. Menikarachchi and J. A. Gascon, *Curr. Top. Med. Chem.*, 2010, **10**, 46–54.
14. G. Monard and K. M. Merz, *Acc. Chem. Res.*, 1999, **32**, 904–911.
15. A. J. Mulholland, *Drug Discov. Today*, 2005, **10**, 1393–1402.
16. R. B. Murphy, D. M. Philipp and R. A. Friesner, *J. Comput. Chem.*, 2000, **21**, 1442–1457.
17. M. B. Peters, K. Raha and K. M. Merz, *Curr. Opin. Drug Discov. Dev.*, 2006, **9**, 370–379.
18. K. Raha, M. B. Peters, B. Wang, N. Yu, A. M. WollaCott, L. M. Westerhoff and K. M. Merz, *Drug Discov. Today*, 2007, **12**, 725–731.
19. H. M. Senn and W. Thiel, *Curr. Opin. Chem. Biol.*, 2007, **11**, 182–187.
20. H. M. Senn and W. Thiel, in *Atomistic Approaches in Modern Biology: From Quantum Chemistry to Molecular Simulations*, 2007, pp. 173–290.
21. H. M. Senn and W. Thiel, *Angew. Chem., Int. Ed. Engl.*, 2009, **48**, 1198–1229.
22. A. Shurki and A. Warshel, in *Protein Simulations*, 2003, pp. 249–313.
23. M. W. van der Kamp, K. E. Shaw, C. J. Woods and A. J. Mulholland, *J. R. Soc. Interface*, 2008, **5**, S173–S190.
24. A. Warshel, *Annu. Rev. Biophys. Biomolec. Struct.*, 2003, **32**, 425–443.
25. T. Tuttle, in *Spectroscopic Properties of Inorganic and Organometallic Compounds*, ed. J. Yarwood, R. Douthwaite and S. B. Duckett, RSC Publishing, Cambridge, 2010.
26. F. Claeysens, J. N. Harvey, F. R. Manby, R. A. Mata, A. J. Mulholland, K. E. Ranaghan, M. Schutz, S. Thiel, W. Thiel and H. J. Werner, *Angew. Chem., Int. Ed. Engl.*, 2006, **45**, 6856–6859.
27. R. Abagyan and M. Totrov, *Curr. Opin. Chem. Biol.*, 2001, **5**, 375–382.
28. C. B-Rao, J. Subramanian and S. D. Sharma, *Drug Discov. Today*, 2009, **14**, 394–400.
29. N. Brooijmans and I. D. Kuntz, *Annu. Rev. Biophys. Biomolec. Struct.*, 2003, **32**, 335–373.
30. C. N. Cavasotto and N. Singh, *Curr. Comput.-Aided Drug Des.*, 2008, **4**, 221–234.
31. S. Cosconati, S. Forli, A. L. Perryman, R. Harris, D. S. Goodsell and A. J. Olson, *Expert. Opin. Drug Discov.*, 2010, **5**, 597–607.
32. W. F. de Azevedo and R. Dias, *Curr. Drug Targets*, 2008, **9**, 1031–1039.
33. P. G. De Benedetti and F. Fanelli, *Drug Discov. Today*, 2010, **15**, 859–866.
34. D. Douguet, *Curr. Comput.-Aided Drug Des.*, 2008, **4**, 180–190.

35. S. Grosdidier and J. Fernandez-Recio, *Expert. Opin. Drug Discov.*, 2009, **4**, 673–686.
36. D. Hecht and G. B. Fogel, *Curr. Comput.-Aided Drug Des.*, 2009, **5**, 56–68.
37. W. L. Jorgensen, *Science*, 2004, **303**, 1813–1818.
38. J. Kirchmair, S. Distinto, D. Schuster, G. Spitzer, T. Langer and G. Wolber, *Curr. Med. Chem.*, 2008, **15**, 2040–2053.
39. D. B. Kitchen, H. Decornez, J. R. Furr and J. Bajorath, *Nat. Rev. Drug Discov.*, 2004, **3**, 935–949.
40. G. Klebe, *J. Mol. Med.*, 2000, **78**, 269–281.
41. Z. D. Konteatis, *Expert. Opin. Drug Discov.*, 2010, **5**, 1047–1065.
42. C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Adv. Drug Deliv. Rev.*, 1997, **23**, 3–25.
43. C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Adv. Drug Deliv. Rev.*, 2001, **46**, 3–26.
44. K. Loving, I. Alberts and W. Sherman, *Curr. Top. Med. Chem.*, 2010, **10**, 14–32.
45. P. D. Lyne, *Drug Discov. Today*, 2002, **7**, 1047–1055.
46. T. J. Marrone, J. M. Briggs and J. A. McCammon, *Annu. Rev. Pharmacol. Toxicol.*, 1997, **37**, 71–90.
47. J. S. Mason, A. C. Good and E. J. Martin, *Curr. Pharm. Design*, 2001, **7**, 567–597.
48. V. Mohan, A. C. Gibbs, M. D. Cummings, E. P. Jaeger and R. L. DesJarlais, *Curr. Pharm. Design*, 2005, **11**, 323–333.
49. C. R. Munteanu, E. Fernandez-Blanco, J. A. Seoane, P. Izquierdo-Novo, J. A. Rodriguez-Fernandez, J. M. Prieto-Gonzalez, J. R. Rabunal and A. Pazos, *Curr. Pharm. Design*, 2010, **16**, 2640–2655.
50. C. Nantasenamat, C. Isarankura-Na-Ayudhya and V. Prachayasittikul, *Expert. Opin. Drug Discov.*, 2010, **5**, 633–654.
51. J. W. Pitera, *Curr. Opin. Drug Discov. Dev.*, 2009, **12**, 388–396.
52. T. T. Talele, S. A. Khedkar and A. C. Rigby, *Curr. Top. Med. Chem.*, 2010, **10**, 127–141.
53. A. Tropsha, *Mol. Inf.*, 2010, **29**, 476–488.
54. S. Vilar, G. Cozza and S. Moro, *Curr. Top. Med. Chem.*, 2008, **8**, 1555–1572.
55. W. P. Walters, M. T. Stahl and M. A. Murcko, *Drug Discov. Today*, 1998, **3**, 160–178.
56. V. Zoete, A. Grosdidier and O. Michielin, *J. Cell. Mol. Med.*, 2009, **13**, 238–248.
57. G. H. Loew, H. O. Villar and I. Alkorta, *Pharm. Res.*, 1993, **10**, 475–486.
58. Y. Marrero-Ponce, M. Iyarreta-Veitia, A. Montero-Torres, C. Romero-Zaldivar, C. A. Brandt, P. E. Avila, K. Kirchgatter and Y. Machado, *J. Chem Inf. Model.*, 2005, **45**, 1082–1100.
59. I. Wallach, *Drug Dev. Res.*, 2011, **72**, 17–25.
60. M. Baroni, G. Costantino, G. Cruciani, D. Riganelli, R. Valigi and S. Clementi, *Quant. Struct.-Act. Relat.*, 1993, **12**, 9–20.

61. R. D. Cramer, D. E. Patterson and J. D. Bunce, *J. Am. Chem. Soc.*, 1988, **110**, 5959–5967.
62. M. Karelson, V. S. Lobanov and A. R. Katritzky, *Chem. Rev.*, 1996, **96**, 1027–1043.
63. G. E. Kellogg, S. F. Semus and D. J. Abraham, *J. Comput.-Aided Mol. Des.*, 1991, **5**, 545–552.
64. G. Klebe, U. Abraham and T. Mietzner, *J. Med. Chem.*, 1994, **37**, 4130–4146.
65. C. Bissantz, G. Folkers and D. Rognan, *J. Med. Chem.*, 2000, **43**, 4759–4767.
66. T. J. A. Ewing, S. Makino, A. G. Skillman and I. D. Kuntz, *J. Comput.-Aided Mol. Des.*, 2001, **15**, 411–428.
67. R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, D. E. Shaw, P. Francis and P. S. Shenkin, *J. Med. Chem.*, 2004, **47**, 1739–1749.
68. T. A. Halgren, R. B. Murphy, R. A. Friesner, H. S. Beard, L. L. Frye, W. T. Pollard and J. L. Banks, *J. Med. Chem.*, 2004, **47**, 1750–1759.
69. I. Halperin, B. Y. Ma, H. Wolfson and R. Nussinov, *Proteins: Struct., Funct. Genet.*, 2002, **47**, 409–443.
70. G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew and A. J. Olson, *J. Comput. Chem.*, 1998, **19**, 1639–1662.
71. I. Muegge and Y. C. Martin, *J. Med. Chem.*, 1999, **42**, 791–804.
72. M. Rarey, B. Kramer, T. Lengauer and G. Klebe, *J. Mol. Biol.*, 1996, **261**, 470–489.
73. M. L. Verdonk, J. C. Cole, M. J. Hartshorn, C. W. Murray and R. D. Taylor, *Proteins: Struct., Funct. Genet.*, 2003, **52**, 609–623.
74. R. X. Wang, Y. P. Lu and S. M. Wang, *J. Med. Chem.*, 2003, **46**, 2287–2303.
75. H. Claussen, C. Buning, M. Rarey and T. Lengauer, *J. Mol. Biol.*, 2001, **308**, 377–395.
76. G. Jones, P. Willett, R. C. Glen, A. R. Leach and R. Taylor, *J. Mol. Biol.*, 1997, **267**, 727–748.
77. A. R. Leach, *J. Mol. Biol.*, 1994, **235**, 345–356.
78. G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell and A. J. Olson, *J. Comput. Chem.*, 2009, **30**, 2785–2791.
79. C. W. Murray, C. A. Baxter and A. D. Frenkel, *J. Comput.-Aided Mol. Des.*, 1999, **13**, 547–562.
80. G. L. Warren, C. W. Andrews, A. M. Capelli, B. Clarke, J. LaLonde, M. H. Lambert, M. Lindvall, N. Nevins, S. F. Semus, S. Senger, G. Tedesco, I. D. Wall, J. M. Woolven, C. E. Peishoff and M. S. Head, *J. Med. Chem.*, 2006, **49**, 5912–5931.
81. N. Foloppe and R. Hubbard, *Curr. Med. Chem.*, 2006, **13**, 3583–3608.
82. N. Huang and M. P. Jacobson, *Curr. Opin. Drug Discov. Dev.*, 2007, **10**, 325–331.
83. B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus, *J. Comput. Chem.*, 1983, **4**, 187–217.

84. A. D. MacKerell, B. R. Brooks, C. L. Brooks, III, L. Nilsson, B. Roux, Y. Won and M. Karplus, in *Encyclopedia of Computational Chemistry*, ed. P. v. R. Schleyer, Wiley, Chichester, 1998, pp. 271–277.
85. D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang and R. J. Woods, *J. Comput. Chem.*, 2005, **26**, 1668–1688.
86. D. A. Pearlman, D. A. Case, J. W. Caldwell, W. S. Ross, T. E. Cheatham, S. Debolt, D. Ferguson, G. Seibel and P. Kollman, *Comput. Phys. Commun.*, 1995, **91**, 1–41.
87. M. Christen, P. H. Hunenberger, D. Bakowies, R. Baron, R. Burgi, D. P. Geerke, T. N. Heinz, M. A. Kastenholtz, V. Krautler, C. Oostenbrink, C. Peter, D. Trzesniak and W. F. Van Gunsteren, *J. Comput. Chem.*, 2005, **26**, 1719–1751.
88. W. R. P. Scott, P. H. Hunenberger, I. G. Tironi, A. E. Mark, S. R. Biller, J. Fennen, A. E. Torda, T. Huber, P. Kruger and W. F. van Gunsteren, *J. Phys. Chem. A*, 1999, **103**, 3596–3607.
89. J. W. Ponder and D. A. Case, in *Protein Simulations*, 2003, pp. 27–86.
90. O. Acevedo and W. L. Jorgensen, *Acc. Chem. Res.*, 2010, **43**, 142–151.
91. P. L. Cummins and J. E. Gready, *Theochem. J. Mol. Struct.*, 2003, **632**, 247–257.
92. K. D. Dubey and R. P. Ojha, *J. Biol. Phys.*, 2011, **37**, 69–78.
93. B. M. Sattelle and A. Almond, *J. Comput. Chem.*, 2010, **31**, 2932–2947.
94. H. M. Senn, J. Kastner, J. Breidung and W. Thiel, *Can. J. Chem.-Rev. Can. Chim.*, 2009, **87**, 1322–1337.
95. T. Tuttle, *J. Phys. Chem. A*, 2009, **113**, 11723–11733.
96. R. C. Walker, M. F. Crowley and D. A. Case, *J. Comput. Chem.*, 2008, **29**, 1019–1031.
97. J. Fanfrlik, A. K. Bronowska, J. Rezac, O. Prenosil, J. Konvalinka and P. Hobza, *J. Phys. Chem. B*, 2010, **114**, 12666–12678.
98. A. Furstner, D. Kirk, M. B. Fenster, C. Aissa, D. De Souza, C. Nevado, T. Tuttle, W. Thiel and O. Muller, *Chem.-Eur. J.*, 2007, **13**, 135–149.
99. A. Khandelwal, V. Lukacova, D. Comez, D. M. Kroll, S. Raha and S. Balaz, *J. Med. Chem.*, 2005, **48**, 5437–5447.
100. T. Tuttle and W. Thiel, *J. Phys. Chem. B*, 2007, **111**, 7665–7674.
101. W. Kohn and L. J. Sham, *Phys. Rev.*, 1965, **140**, 1133–1138.
102. T. Tuttle and W. Thiel, *Phys. Chem. Chem. Phys.*, 2008, **10**, 2159–2166.
103. W. Weber and W. Thiel, *Theor. Chem. Acc.*, 2000, **103**, 495–506.
104. Q. Cui, M. Elstner, E. Kaxiras, T. Frauenheim and M. Karplus, *J. Phys. Chem. B*, 2001, **105**, 569–585.
105. M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai and G. Seifert, *Phys. Rev. B*, 1998, **58**, 7260–7268.
106. G. D. Seabra, R. C. Walker, M. Elstner, D. A. Case and A. E. Roitberg, *J. Phys. Chem. A*, 2007, **111**, 5655–5664.
107. M. P. Repasky, J. Chandrasekhar and W. L. Jorgensen, *J. Comput. Chem.*, 2002, **23**, 1601–1622.

108. M. Korth, M. Pitonak, J. Rezac and P. Hobza, *J. Chem. Theory Comput.*, 2010, **6**, 344–352.
109. D. Bakowies and W. Thiel, *J. Phys. Chem.*, 1996, **100**, 10580–10594.
110. M. J. Field, *J. Comput. Chem.*, 2002, **23**, 48–58.
111. L. Ridder and A. J. Mulholland, *Curr. Top. Med. Chem.*, 2003, **3**, 1241–1256.
112. U. Ryde, *Curr. Opin. Chem. Biol.*, 2003, **7**, 136–142.
113. M. Freindorf, Y. H. Shao, T. R. Furlani and J. Kong, *J. Comput. Chem.*, 2005, **26**, 1270–1278.
114. D. Riccardi, G. H. Li and Q. Cui, *J. Phys. Chem. B*, 2004, **108**, 6467–6478.
115. C. Sagui and T. A. Darden, *Annu. Rev. Biophys. Biomolec. Struct.*, 1999, **28**, 155–179.
116. A. Warshel and A. Papazyan, *Curr. Opin. Struct. Biol.*, 1998, **8**, 211–217.
117. T. Darden, D. York and L. Pedersen, *J. Chem. Phys.*, 1993, **98**, 10089–10092.
118. U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee and L. G. Pedersen, *J. Chem. Phys.*, 1995, **103**, 8577–8593.
119. P. Ewald, *Ann. Phys.*, 1921, **369**, 253–287.
120. J. L. Gao and C. Alhambra, *J. Chem. Phys.*, 1997, **107**, 1212–1217.
121. K. Nam, J. L. Gao and D. M. York, *J. Chem. Theory Comput.*, 2005, **1**, 2–13.
122. P. H. Hunenberger and J. A. McCammon in *CECAM Workshop on Implicit Solvent Models for Biomolecular Simulations*, Lyon, France, 1998, pp. 69–88.
123. P. H. Hunenberger and J. A. McCammon, *J. Chem. Phys.*, 1999, **110**, 1856–1872.
124. W. Weber, P. H. Hunenberger and J. A. McCammon, *J. Phys. Chem. B*, 2000, **104**, 3668–3675.
125. H. Alper and R. M. Levy, *J. Chem. Phys.*, 1993, **99**, 9847–9852.
126. D. Beglov and B. Roux, *J. Chem. Phys.*, 1994, **100**, 9050–9063.
127. J. W. Essex and W. L. Jorgensen, *J. Comput. Chem.*, 1995, **16**, 951–972.
128. F. S. Lee and A. Warshel, *J. Chem. Phys.*, 1992, **97**, 3100–3107.
129. I. G. Tironi, R. Sperb, P. E. Smith and W. F. Vangunsteren, *J. Chem. Phys.*, 1995, **102**, 5451–5459.
130. T. Benighaus and W. Thiel, *J. Chem. Theory Comput.*, 2008, **4**, 1600–1609.
131. D. Riccardi and Q. Cui, in *232nd National Meeting of the American Chemical Society*, San Francisco, CA, 2006, pp. 5703–5711.
132. D. Riccardi, P. Schaefer and Q. Cui, *J. Phys. Chem. B*, 2005, **109**, 17715–17733.
133. D. Riccardi, P. Schaefer, Y. Yang, H. B. Yu, N. Ghosh, X. Prat-Resina, P. Konig, G. H. Li, D. G. Xu, H. Guo, M. Elstner and Q. Cui, *J. Phys. Chem. B*, 2006, **110**, 6458–6469.
134. P. Schaefer, D. Riccardi and Q. Cui, *J. Chem. Phys.*, 2005, **123**.
135. F. Maseras and K. Morokuma, *J. Comput. Chem.*, 1995, **16**, 1170–1179.
136. S. Humbel, S. Sieber and K. Morokuma, *J. Chem. Phys.*, 1996, **105**, 1959–1967.

137. S. Dapprich, I. Komaromi, K. S. Byun, K. Morokuma and M. J. Frisch, *Theochem. J. Mol. Struct.*, 1999, **461**, 1–21.
138. M. Svensson, S. Humbel, R. D. J. Froese, T. Matsubara, S. Sieber and K. Morokuma, *J. Phys. Chem.*, 1996, **100**, 19357–19363.
139. M. P. Gleeson and D. Gleeson, *J. Chem Inf. Model.*, 2009, **49**, 1437–1448.
140. Y. X. Lu, T. Shi, Y. Wang, H. Y. Yang, X. H. Yan, X. M. Luo, H. L. Jiang and W. L. Zhu, *J. Med. Chem.*, 2009, **52**, 2854–2862.
141. Y. X. Lu, Y. Wang, Z. J. Xu, X. H. Yan, X. M. Luo, H. L. Jiang and W. L. Zhu, *J. Phys. Chem. B*, 2009, **113**, 12615–12621.
142. T. H. Rod and U. Ryde, *J. Chem. Theory Comput.*, 2005, **1**, 1240–1251.
143. U. Ryde, *J. Comput.-Aided Mol. Des.*, 1996, **10**, 153–164.
144. T. Vreven, K. S. Byun, I. Komaromi, S. Dapprich, J. A. Montgomery, K. Morokuma and M. J. Frisch, *J. Chem. Theory Comput.*, 2006, **2**, 815–826.
145. T. Vreven and K. Morokuma, in *Symposium on Methods and Applications of Combined Quantum Mechanical and Molecular Mechanical Potentials held at the 222nd American Chemical Society National Meeting*, Chicago, Illinois, 2001, pp. 125–132.
146. A. Furstner, D. Kirk, M. D. B. Fenster, C. Aissa, D. De Souza and O. Muller, *Proc. Natl. Acad. Sci. USA*, 2005, **102**, 8103–8108.
147. W. M. Morton, K. R. Ayscough and P. J. McLaughlin, *Nature Cell Biol.*, 2000, **2**, 376–378.
148. H. Li, A. D. Robertson and J. H. Jensen, *Proteins: Struct. Func. Bioinf.*, 2005, **61**, 704–721.
149. J. M. Word, in *Reduce*, ed. D. U. Biochemistry Department, Durham, NC, 2003.
150. J. M. Word, S. C. Lovell, J. S. Richardson and D. C. Richardson, *J. Mol. Biol.*, 1999, **285**, 1735–1747.
151. R. G. Parr and W. T. Yang, *Density Functional Theory of Atoms and Molecules*, Oxford University Press, New York, 1989.
152. W. Thiel, Max-Planck-Institut für Kohlenforschung, Mülheim an der Ruhr, Germany, 2004.
153. K. Y. Burstein and A. N. Isaev, *Theo. Chim. Act.*, 1984, **64**, 397–401.
154. M. J. S. Dewar and W. Thiel, *J. Am. Chem. Soc.*, 1977, **99**, 4899–4907.
155. A. Goldblum, *J. Comp. Chem.*, 1987, **8**, 835–849.
156. A. D. Becke, *Phys. Rev. A*, 1988, **38**, 3098–3100.
157. C. T. Lee, W. T. Yang and R. G. Parr, *Phys. Rev. B*, 1988, **37**, 785–789.
158. F. Weigend and R. Ahlrichs, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3297–3305.
159. TURBOMOLE, version 5.7.1, COSMOlogic GmbH & Co. KG, Leverkusen, Germany, 2004.
160. R. Ahlrichs, M. Bär, M. Häser, H. Horn and C. Kölmel, *Chem. Phys. Lett.*, 1989, **162**, 165–169.
161. M. Häser and R. Ahlrichs, *J. Comp. Chem.*, 1989, **10**, 104–111.
162. H. Horn, H. Weiss, M. Häser, M. Ehrig and R. Ahlrichs, *J. Comp. Chem.*, 1991, **12**, 1058–1064.

163. O. Treutler and R. Ahlrichs, *J. Chem. Phys.*, 1995, **102**, 346–354.
164. CHARMM, version c31b1, Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, 2004.
165. C. L. Brooks and M. Karplus, *J. Chem. Phys.*, 1983, **79**, 6312–6325.
166. N. Foloppe and A. D. MacKerell, *J. Comp. Chem.*, 2000, **21**, 86–104.
167. A. D. MacKerell and N. K. Banavali, *J. Comp. Chem.*, 2000, **21**, 105–120.
168. A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin and M. Karplus, *J. Phys. Chem. B*, 1998, **102**, 3586–3616.
169. The commercial and academic versions of the CHARMM force fields are not strictly compatible. However, the parameters for the non-bonding terms (van der Waals and coulombic interactions) are comparable.
170. ChemShell, version 3.0a3, CCLRC Daresbury Laboratory, Cheshire, UK, 2004.
171. P. Sherwood, A. H. de Vries, M. F. Guest, G. Schreckenbach, C. R. A. Catlow, S. A. French, A. A. Sokol, S. T. Bromley, W. Thiel, A. J. Turner, S. Billeter, F. Terstegen, S. Thiel, J. Kendrick, S. C. Rogers, J. Casci, M. Watson, F. King, E. Karlsen, M. Sjøvoll, A. Fahmi, A. Schafer and C. Lennartz, *Theochem: J. Mol. Struct.*, 2003, **632**, 1–28.
172. S. R. Billeter, A. J. Turner and W. Thiel, *Phys. Chem. Chem. Phys.*, 2000, **2**, 2177–2186.
173. M. L. Connolly, *J. App. Cryst.*, 1983, **16**, 548–558.
174. M. L. Connolly, *Science*, 1983, **221**, 709–713.
175. R. S. Spolar and M. T. Record, *Science*, 1994, **263**, 777–784.
176. S. Grimme, *J. Comput. Chem.*, 2004, **25**, 1463–1473.
177. S. Grimme, *J. Comput. Chem.*, 2006, **27**, 1787–1799.
178. J. P. McNamara and I. H. Hillier, *Phys. Chem. Chem. Phys.*, 2007, **9**, 2362–2370.
179. P. Jurecka, J. Cerny, P. Hobza and D. R. Salahub, *J. Comput. Chem.*, 2007, **28**, 555–569.
180. P. Jurecka and P. Hobza, *J. Am. Chem. Soc.*, 2003, **125**, 15608–15613.
181. R. Lonsdale, J. N. Harvey and A. J. Mulholland, *J. Phys. Chem. Lett.*, 2010, **1**, 3232–3237.
182. A. E. Cho and D. Rinaldo, *J. Comput. Chem.*, 2009, **30**, 2609–2616.

CHAPTER 2

Transition Metal Systems

ROBERT J. DEETH

Department of Chemistry, University of Warwick, Coventry CV4 7AL, UK

2.1 Introduction

Transition metals (TMs) are ubiquitous in natural systems and mediate a plethora of catalytic, redox and structural processes. The most common ‘naturally occurring’ elements are iron, copper, zinc, nickel, cobalt and molybdenum – all of which are essential to life but present in only trace amounts. Nevertheless, the concentrations of metal ions must be very carefully controlled since excess, or deficiency, can lead to serious illness. Living organisms have thus developed sophisticated mechanisms for sequestering, storing and utilising TMs.

Despite the importance of these ‘inorganic’ elements in biochemical processes, TMs do not yet figure heavily as therapeutic agents. TM compounds are generally considered to be, and often are, highly toxic. Nevertheless, a variety of TMs find use in therapeutic or diagnostic compounds, a selection of which is shown in Figure 2.1. For example, there are the platinum and ruthenium anti-cancer drugs, gold compounds for treating rheumatoid arthritis, and copper and $^{99\text{m}}\text{Tc}$ imaging agents. The platinum complexes such as cisplatin, carboplatin and oxaliplatin are the most successful TM-based anti-cancer therapeutics and, in their own right, are major tools in cancer chemotherapy, while there are of the order 25 million nuclear medicine procedures per annum worldwide based on technetium derived from ^{99}Mo . Thus, not only is there is a precedent for using compounds containing metal centres in medicine but given the richness and diversity of TM chemistry, a bright future is predicted.¹

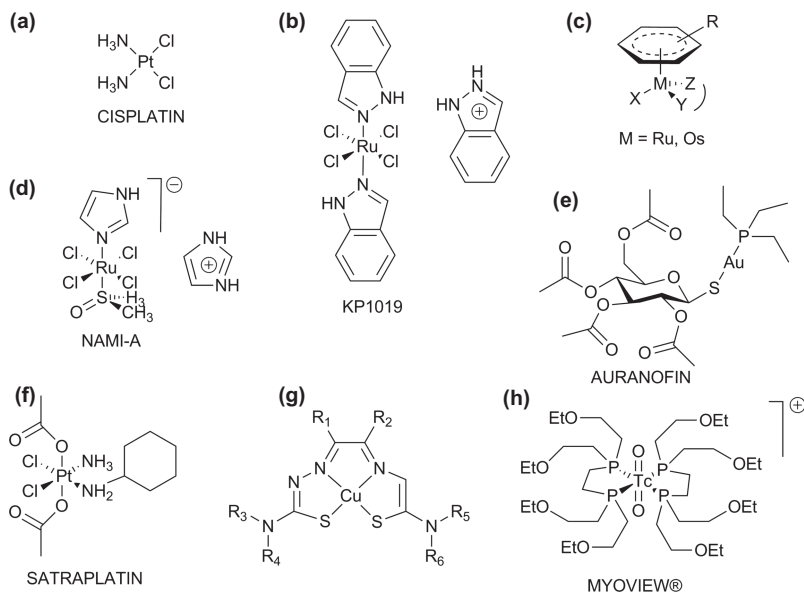


Figure 2.1 Some illustrative examples of TM-based systems of actual or potential therapeutic or clinical benefit. Parts (a)–(d) and (f) show potential or actual anti-cancer drugs; part (e) shows an antiarthritic; and parts (g) and (h) show diagnostic imaging agents.

While drug molecules are critically important, so too are their targets and here again, we find relatively few examples of pharmaceuticals aimed at binding to a TM-containing active site in a protein. This seems curious given that about one third of all proteins are believed to have an absolute requirement for metal ions. Perhaps the best known systems with a TM-containing active site are the zinc centres of matrix metalloproteinases (MMPs) and the iron-heme centres in the many forms of cytochrome P450. P450 is a constituent of the immune system and is often responsible for metabolising drug molecules before they can exert their desired activity, so developing molecules that can temporarily suppress P450 activity is of some interest. As described below, there has also been some work on binding drugs to the active site of methionine aminopeptidase (MetAP), although there is a long history surrounding the identification of both the identity and number of metal centres in the physiologically relevant form.

Another factor which, in the author's opinion, also limits the role of TM systems in pharmaceuticals is that medicinal chemistry is largely, and quite rightly, mostly about synthetic organic chemistry. TM species often provide excellent catalysts for building novel organic molecules for instance but, at the end of the day, the metal must be purged due to toxicity issues. The situation is even worse for the computational chemist working in drug discovery. The tools available to deal theoretically with TM systems are either not available or too computationally expensive to be practical. The advent of Density Functional

Theory (DFT) has gone a long way towards alleviating some of the problems but it will be some time yet before we can afford, for example, to use DFT to carry out a virtual high throughput screen (vHTC) on the million or more compounds found in a typical virtual library, irrespective of whether or not that library holds of any molecules containing a TM centre.

However, the increasing interest in metallodrugs and the continuously growing numbers of metalloenzymes, must inevitably increase the pressure to develop computational methods capable of delivering results which can feed into the discovery process. In this chapter, I will attempt to survey briefly the kinds of metal-containing systems which have been or are currently of some pharmaceutical significance, highlight the properties which challenge the computational chemist and suggest some solutions.

2.2 Methods for Modelling Transition Metal Centres

A wide range of TM centres is encountered in biological and medicinal chemistry. ‘Naturally occurring’ TMs include Fe, Cu, V, Ni, Mo, W, Co, Mn and Zn, to which one can add artificial therapeutic and diagnostic agents involving some of the centres already mentioned above plus Ti, Pt, Au, Ru, Os, Rh, Re and Tc. This large set of metals also displays a diversity of structure and bonding ranging from the relatively ionic classical Werner-type coordination complexes such as *cis*-[PtCl₂(NH₃)₂] (cisplatin, see Figure 2.1a) through to the relatively covalent organometallic species such as the ruthenium and osmium arene complexes (see Figure 2.1f).

Transition metal species possess a number of key chemical differences compared to carbon compounds, and these differences directly influence our ability to model them theoretically. The main potential source of complexity is the presence of an open d shell since the d electrons can exert strong structural effects. For example, certain combinations of d configuration and spin state are associated with specific geometries. Thus, low-spin d⁸ species like Pt^{II} and Au^{III} are square planar while low-spin d⁶ complexes of Pt^{IV} and Ru^{II} are octahedral. In these cases, the d electrons are ‘non-innocent’ and play an important structural role. The ‘stereochemical activity’ of d electrons is widespread but can also be a relatively subtle phenomenon. The *trans* influence in planar species is a case in point.² The strength of a given M-L bond varies as a function of the ligand opposite and, in Pt^{II} complexes for example, can lead to bond length changes of several tenths of an Angstrom.

A more pronounced stereochemical activity is the Jahn–Teller effect commonly seen for d⁹ Cu^{II} complexes but also manifest in the high-spin d⁴ Mn^{III} centres implicated in the oxygen evolving centre of photosystem II. Six-coordinate Cu^{II} complexes are not octahedral. Instead, they generally display a tetragonally elongated geometry although, occasionally, the design of a ligand may favour a tetragonal compression. In either case, the difference between equatorial and axial bond distances can be quite pronounced – of the order of several tenths of an Angstrom.

Metal–ligand bonds also display a phenomenal range of lability. The water exchange rates for simple metal cations in aqueous solution span over 16 orders of magnitude. The dominant chemistry is substitution. The fundamental nature of the dative covalent M–L bond supports heterolytic dissociation since the ligand itself is capable of existing independently. This contrasts strongly with carbon chemistry where breaking a bond to a carbon centre generates extremely reactive species.

The dynamic nature of TM chemistry also often leads to complex behaviour with respect to the coordination number and the ability to support six or seven bonds, even temporarily, exacerbates the problem.

Another important feature of TM chemistry is the ease with which the spin state may change. In several heme systems, for example, the spin state does not remain constant over the course of a multi-step reaction. The switch from one spin state to another and back can have a profound effect on the products of the reaction and has obvious connotations for theoretical modelling.

Given the preceding comments, computational chemists naturally turn to quantum chemistry in order to model TM systems. It is certainly the case that the advent of Density Functional Theory (DFT) has completely revolutionised our ability to carry out good quality calculations on reasonably large molecular systems containing one or more metal centres. Indeed, quantum chemistry has made and continues to make many important contributions in bioinorganic chemistry with the contribution of Siegbahn and co-workers being especially noteworthy.³

2.2.1 Density Functional Theory

DFT is founded in the theorems of Hohenberg and Kohn, who demonstrated a unique relationship between the electron density, ρ , and the total energy of the ground state, E_0 :

$$E_0 = E[\rho] \quad (2.1)$$

This deceptively simple equation implies that if ρ is known, then all the complicated quantum mechanical exchange and correlation contributions to E_0 are known. The latter – *i.e.* correlation – is a particular problem for wave function methods like Hartree–Fock (HF) theory which only provides an averaged correlation energy. Of course, it is possible to recover the missing correlation energy via post-HF methods but these methods do not scale well and rapidly become prohibitively expensive. In contrast, DFT promises a much cheaper but still accurate alternative.

However, the reality is not perfect. Most practical implementations of DFT are based on the Kohn–Sham (KS) formulation that expresses the density in terms of one-electron functions – orbitals. The ‘trick’ of KS-DFT is to compute the exact kinetic energy of a set of hypothetical non-interacting electrons, which has the same density as the true ρ , and fold away the difference between this model kinetic energy and the true kinetic energy into the exchange–correlation term, E_{XC} . Unfortunately, while an inspired and elegant solution to the

problem of a practical formulation of DFT, the exact form of E_{XC} is unknown in general so while DFT is exact in principle, in practice we need to formulate approximate functionals to describe the exchange and correlation energies.

However, an exact solution of the KS-DFT equations exists for the special case of the uniform electron gas. In this case, exchange and correlation can be separated and the value of each at a given point in space depends only on the local value of the density at that point. This construct is therefore referred to as the local density approximation (LDA). The LDA exchange energy E_X^{LDA} , for example, depends simply on the $4/3$ root of ρ :

$$E_X^{LDA} = \frac{-3}{2} \left(\frac{3}{4\pi} \right)^{1/3} \int \rho(r)^{4/3} dr \quad (2.2)$$

The expression for the correlation energy is more complex and is replaced by an interpolation scheme developed by Vosko, Wilk and Nusair (VWN) who computed correlation energies numerically for a range of ρ values.

The LDA is simple, relatively efficient and remarkably accurate for molecular systems where the density is far from uniform. However, the LDA typically overbinds. In response, corrections that depend on the gradient of ρ were added to the base LDA that engendered a whole slew of new Generalised Gradient Approximation (GGA) functionals. The problem here is that systematic mathematical improvement of functionals is extremely challenging. Thus, many have resorted to a healthy dose of empiricism in an attempt to ‘fit’ functionals to experimental and/or numerical data. Becke, for example, developed a gradient corrected exchange functional based on reproducing the exact exchange energies of noble gas atoms.⁴

The Becke exchange energy E_X^B is given by:

$$E_X^B = E_X^{LDA} - \beta \sum_{\sigma} \int \rho_{\sigma}^{4/3} \frac{x_{\sigma}^2}{(1 + 6\beta x_{\sigma} \sinh^{-1} x_{\sigma})} d^3r \quad (2.3)$$

where σ refers to the spin (either up or down), β is an empirically determined constant (0.0042 a.u.), and x_{σ} is a dimensionless ratio which depends on the gradient of the spin density, $\nabla\rho_{\sigma}$, viz.:

$$x_{\sigma} = \frac{|\nabla\rho_{\sigma}|}{\rho_{\sigma}^{4/3}} \quad (2.4)$$

Becke’s continuing quest for yet better functionals soon led him to develop what has become the most popular form of DFT – the so-called hybrid functional. Becke reasoned that the shortcomings of LDA exchange could be corrected by mixing local density exchange with ‘exact’ HF exchange. By empirically adjusting the amount of LDA exchange, his earlier gradient corrected exchange, E_X^B , and HF exchange plus adding LDA and gradient

corrected correlation, a generally superior functional could be developed.⁵⁻⁷ Although Becke used a different correlation functional, the general notion was adopted by the developers of the Gaussian program system who, as described by Stephens *et al.*,⁸ implemented the B3LYP hybrid functional with the following form:

$$E_{XC}^{B3LYP} = (1 - a)E_X^{LDA} + aE_X^{HF} + b\Delta E_X^B + (1 - c)E_C^{LDA} + cE_C^{LYP} \quad (2.5)$$

where a , b and c are empirical constants.

The B3LYP functional has proven to be extremely versatile and is widely employed to calculate a range of properties of TM systems. However, there are some potential limitations. For example, B3LYP does not provide a correct description of dispersion forces and can sometime give large errors for spin state energy differences.

The lack of dispersion is a common theme in quantum chemistry since it is absent from both HF theory and most DFT functionals but it has often been seen as of secondary importance in TM chemistry where the focus is usually on the making and breaking of metal–ligand bonds. However, if one wishes to model the interaction of a TM compound with some biological target, for example, dispersive interactions may be important.

On the one hand, we could use a more sophisticated wave function method such as MP2 but at some cost in terms of additional computational resource. Moreover, MP2 tends to over bind stacked aromatic systems. Alternatively, we can simply add an empirical dispersion correction to an existing (gradient corrected) DFT functional to generate, for example, BLYP-D. Dispersion forces are treated by an inter-atomic potential of the form C_6R^{-6} which is both strongly damped and scaled.⁹ The latter smooths out any dependence on the choice of functional.

The issue of using DFT to compute relative spin state energies is more problematic, which is particularly pronounced in TM chemistry where variable spin states are common. Of course, if the reaction of interest takes place on a single spin surface, the problem may go unnoticed. However, increasingly we are finding TM-mediated reactions in which more than one spin state plays a role. Reactions at the Fe centre in cytochrome P450 often involve a two-state pathway employing two different spin surfaces.¹⁰ Calculations which assume a single spin surface are unable to explain the experimental data satisfactorily.

In such cases it is clearly desirable that the chosen computational method not only predicts the correct ground spin state but also the energy of the excited spin state. Here, the performance of various functionals is patchier. ‘Standard’ B3LYP with 20% HF exchange does not perform well for first-row TM species. However, Salomon *et al.*¹¹ find a linear relationship between the amount of exact (*i.e.* HF) exchange and the high-spin/low-spin energy difference for a series of Fe-sulfur complexes prompting them to propose the B3LYP* functional with only 15% HF exchange as a reasonable choice for first-row complexes. Conversely, Sorkin *et al.*¹² report that ‘standard’ B3LYP performs the best out of a set of GGA and hybrid functionals for Fe_2 , Fe_2^- and FeO^+ .

A recent study by Ye and Neese¹³ concludes that still more complex, so-called, double-hybrid functionals are necessary just for qualitatively correct results although, in the author's opinion, this is something of an exaggeration in that most functionals give the correct ranking of complexes as a function of spin state and simply differ as to the position where the spin state changes.¹⁴

In summary, modern DFT is an extremely useful tool for modelling TM systems. However, *any* quantum chemical method, even DFT, is relatively compute-intensive and thus the type of drug discovery problem that is amenable to a quantum treatment is limited. Comprehensive conformational searching and vHTS, which may require many millions of individual calculations, are challenging for QC and faster methods are generally required.

Semi-empirical molecular orbital theory offers a possible alternative. It is certainly significantly more efficient than, say, 'full' DFT but developing suitable parameters for TM centres is very challenging and the resulting accuracy can be a little variable.¹⁵

2.2.2 Molecular Mechanics

From an efficiency perspective, the 'method of choice' is empirical molecular mechanics (MM). MM offers the necessary speed to enable millions of computations to be completed in a reasonable time frame. However, the simple potential energy expressions employed in 'conventional' MM are not well suited to TM chemistry. One of the easiest issues to appreciate, and thus solve, is the unique labelling problem.¹⁶

A simple force field equation is given in Equation 2.6:

$$E_{\text{tot}} = \sum_{i,j} k_{ij} (r_{ij} - r_{0,ij})^2 + \sum_{i,j,k} k_{ijk} (\theta_{ijk} - \theta_{0,ijk})^2 + \sum_{i,j,k,l} k_{ijkl} [1 + \cos(n_{ijkl}\tau - \phi_{ijkl})] \\ + \left[\sum_{i < j} \frac{\rho_i \rho_j}{\epsilon d_{ij}} + \sum_{i < j} \left(\frac{A_{ij}}{d_{ij}^{12}} - \frac{B_{ij}}{d_{ij}^6} \right) \right] \quad (2.6)$$

where the k are appropriate force constants, r are bond lengths, θ are bond angles, τ are torsion angles, n is the torsional periodicity parameter, ϕ is the torsion offset, ρ are partial atomic charge, ϵ is the dielectric constant, A and B are Lennard-Jones vdW parameters and the summations run over bonded atom pairs (ij), angle triples (ijk) and torsional quadruples ($ijkl$). The non-bonded terms are summed over the distances, d_{ij} , between unique atom pairs excluding bonded pairs and the atoms at either end of an angle triple. For the atoms at the ends of a torsion quadruple, the non-bonded term may be omitted or scaled. The 0 subscript of r or θ refers to 'reference' values for each bond or angle.

Such an expression can only specify a single reference angle for a given A–B–C angle triple. As shown (see Figure 2.2), this presents no difficulty in

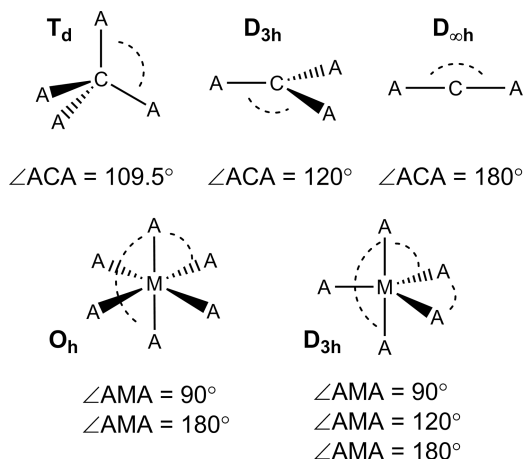


Figure 2.2 Valence angles at central atom for various regular geometries.

carbon chemistry but cannot work in TM chemistry where, for example, an octahedral complex requires two ‘reference’ A–M–A bond angles, 90° and 180° .

Various solutions to this problem have been proposed ranging from replacing the harmonic expressions for angle bending with a Fourier expansion which has multiple minima¹⁶ at the desired reference values, to using a simple ligand–ligand, points-on-a-sphere (POS) repulsion term such that the angular geometry around the TM centre is determined without needing to specify explicit A–M–A angle bending terms.¹⁷

Given a suitable angle bend potential, the most commonly encountered coordination geometries – tetrahedron, square plane and octahedron – can be handled by ‘conventional’ MM. Hence, those metal centres (and their attendant oxidation states) that favour these geometries are amenable to a classical treatment.

Table 2.1 lists some examples of metal ions which are often associated with particular coordination geometries. The preference for a particular geometry is generally a function of the d electron ligand field stabilisation energy (LFSE). In octahedral symmetry, d^3 , low-spin d^6 and high-spin d^8 are associated with large LFSEs while the planar geometry associated with low-spin d^8 configurations can be rationalised in terms of the LFSE.¹⁸

In contrast, those configurations associated with small or zero LFSEs usually give the regular ‘Platonic’ structure associated with their coordination number – a tetrahedron for tetra-coordination, a trigonal bipyramid for penta-coordination and an octahedron for hexa-coordination. Hence, quite a bit of TM coordination chemistry of biological and pharmaceutical relevance can be treated, at least in principle, by MM.

There are three major obstacles.

1. Just because a metal centre can be treated by MM does not mean suitable force field parameters are actually available.

Table 2.1 Combinations of d configuration and spin state that are strongly associated with particular coordination geometries.

<i>Spin state and d configuration</i>	<i>Geometry</i>	<i>Mⁿ⁺ examples</i>
HS d ³	Octahedral	Cr ³⁺ , V ²⁺
LS d ⁶	Octahedron	Co ³⁺ , Ru ²⁺ , Pt ⁴⁺
HS d ⁸	Octahedron	Ni ^{2+a}
LS d ⁸	Planar	Ni ²⁺ , Pd ²⁺ , Pt ²⁺ , Au ³⁺
d ¹ , d ² , HS d ⁵ , d ¹⁰	No strong preference ^b	Fe ³⁺ (d ⁵), Mn ²⁺ (d ⁵), Zn ²⁺ (d ¹⁰)

^awith relatively weak field ligands;^bstructure generally the Platonic structure corresponding to the coordination number.

2. Due to the nature of M–L bonding, the coordination number may be ambiguous so it is difficult to know precisely what to model. This applies particularly to metal centres which have no strong electronic preference for a particular geometry, notably Zn²⁺.
3. Certain combinations of d configuration and spin state do not yield regular geometries. Perhaps the most important example is d⁹ Cu²⁺ which is strongly influenced by Jahn–Teller distortions which can cause a significant problem for conventional MM.

Obstacle 1 is probably a reflection of the perceived need. If more workers needed to be able to deal with more metal centres, then presumably more parameters would have been developed.

Obstacle 2 is simply a challenging problem. It is difficult enough to estimate the binding affinity of a drug for an active site when only non-covalent interactions need to be taken into account. If, in addition, the binding event leads to the rupture of one or more M–L bonds, the level of complexity increases considerably. Does the ligand displace a coordinated solvent water or a protein side chain residue? If the ligand binds via a carboxylate group, for example, is the latter mono- or bidentate; symmetric or asymmetric; bridging or bound to a single metal centre? If the ligand has multiple potential donor sites, which one(s) does it use?

Obstacle 3 is about the electronic effects arising from open-shell d configurations, a topic which has occupied my group for nearly two decades.

2.2.3 Ligand Field Effects in Molecular Mechanics

The energies of the d electrons can exert a significant effect on the structure and energy of TM complexes. Figure 2.3 displays the experimental hydration enthalpies (ΔH_{hyd}) for the first transition series ions Ca²⁺ through Zn²⁺ which illustrates the typical ‘double hump’ variation. The difference between the observed ΔH_{hyd} and the estimated values based on simply interpolating between the points corresponding to d⁰, d⁵ and d¹⁰ configurations correlates very well with the LFSE which can, in turn, be computed from the d-orbital splitting diagram (see Figure 2.4) using values of Δ_{oct} derived from spectroscopy.

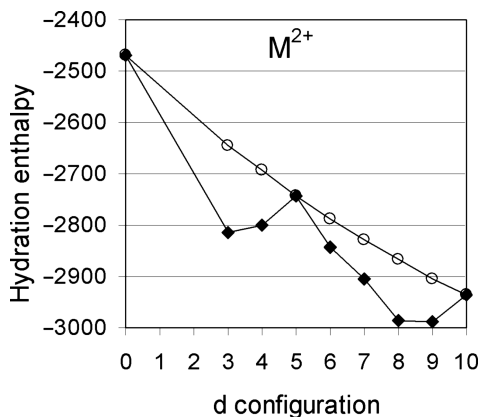


Figure 2.3 Variation of experimental hydration enthalpies (solid diamonds, kJ mol^{-1}) from $d^0 \text{ Ca}^{2+}$ through to $d^{10} \text{ Zn}^{2+}$. The open circles represent the values once d electron stabilisation energy effects are removed (after Johnson and Nelson¹⁹).

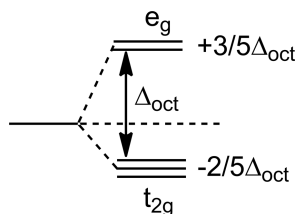


Figure 2.4 Octahedral d-orbital splitting diagram from which the one-electron contribution to the LFSE can be computed.

Hence, superimposed on the general trend of progressively more negative values of ΔH_{hyd} on crossing the series from left to right are fluctuations due to d electron effects. The interpolated values (see Figure 2.3, open circles) represent zero LFSE – hence the intersection of experimental and interpolated points at d^0 , high-spin d^5 and d^{10} configurations.

The LFSE also influences the structure. In some cases – *e.g.* d^3 – the *symmetry* is unaffected and the LFSE simply shortens the M–L bond distance over that which would arise if the LFSE were zero. In other cases, the LFSE leads to a pronounced change in symmetry. The best-known example of this is the Jahn–Teller tetragonal elongation of six-coordinate $d^9 \text{ Cu}^{\text{II}}$ complexes.

As illustrated in Figure 2.5, the asymmetric occupation of the e_g orbitals d_{z^2} and $d_{x^2-y^2}$ leads to a net electronic stabilisation as the axial bonds are lengthened while the equatorial bonds shorten. This motion is resisted by the vibrational potential which, for an ML_6 system, is a minimum for octahedral symmetry. The final tetragonally distorted geometry obtains when these two competing forces are in equilibrium.

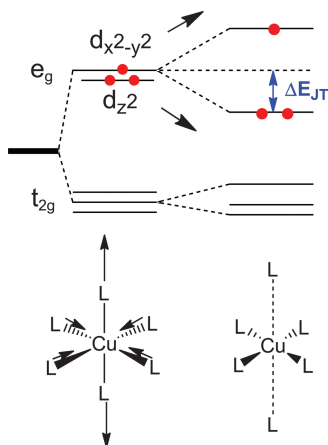


Figure 2.5 Schematic energy level diagram illustrating the Jahn–Teller stabilisation energy accompanying a tetragonal elongation of an octahedral Cu^{II} complex.

If we wish to provide a general computational scheme for modelling TM systems, these d electron effects must be taken into account.²⁰ Conventional MM attempts to do this implicitly. This is not too much of a problem for ‘well-behaved’ geometries alluded to above such as tetrahedral, square planar and octahedral. In these cases, the d electron effects mainly influence the M–L distance and can thus be subsumed within a conventional bond stretch term. However, the difference between equatorial and axial bond lengths in a Jahn–Teller distorted system can be large (commonly 0.3 to 0.5 Å) plus, to first order, any of the three octahedral axes can be the axis of elongation. The problem with conventional MM is that separate bond stretch terms are required for short and long bonds and hence the user must make a choice that may be neither obvious nor desirable.

Alternatively, and the path we have followed, d electron effects can be explicitly added to MM. Since the simplest way to calculate d orbital energies is via ligand field theory (LFT), we have added LF energies and gradients to conventional MM to generate ligand field molecular mechanics (LFMM).²¹

The ligand field contribution is confined to the metal and its immediately bonded neighbours (Figure 2.6). We take the original ‘organic’ force field which handles the ligand region and add to it M–L bond stretch terms (we employ a Morse function better to accommodate the large changes observed for Jahn–Teller distortions), M–L–A angle bend terms, M–L–A–B torsion terms, a purely repulsive L–L term to help determine the angular geometry around the metal, and an explicit LFSE term. L–M–L–A torsion potentials are disabled to avoid problems when the L–M–L angle is 180°.

Since MM is intrinsically bond centred, the most convenient way to express the LFSE is based on the angular overlap model of Schaeffer and Jorgensen.²² The total ligand field potential, V_{LF} , is assumed to be the sum of contributions

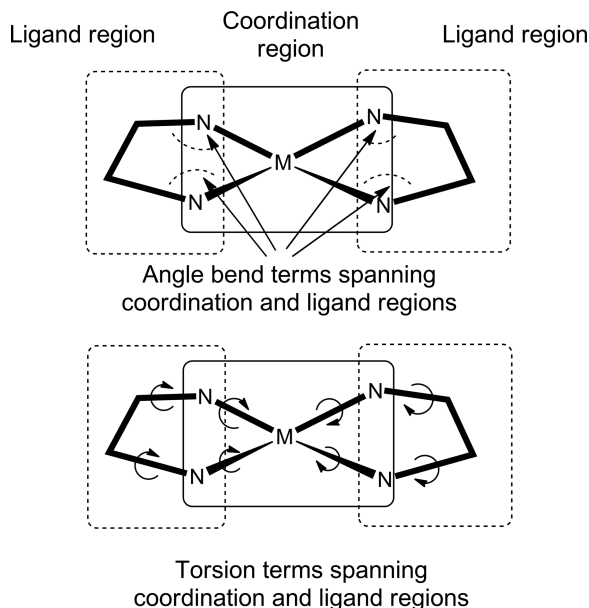


Figure 2.6 Schematic representation of division into coordination and ligand regions and force field terms which span the two.

from the M–L bonds each modelled by AOM parameters describing separate σ and π interactions as illustrated (Figure 2.7).

This construction provides a more physically realistic description of the M–L bond.

However, if the LFSE is based solely on the d orbital energies, it will always favour low-spin configurations. In reality, low-spin states have higher d–d inter electron repulsion energies so the observed spin state is a balance. To model this, we add a bond-length-dependent penalty energy to the low-spin LFSE of the form shown in Equation (2.7). This is also the functional form used to represent the AOM parameters e_σ , $e_{\pi x}$, $e_{\pi y}$ and e_{ds} although, in the interests of simplicity, we seldom use more than a single a_i^j value for a given parameter.

$$e_\lambda = a_0^j + a_1^j r + a_2^j r^{-2} + a_3^j r^{-3} + a_4^j r^{-4} + a_5^j r^{-5} + a_6^j r^{-6} \quad (2.7)$$

The advantage of LFMM is that it is sufficiently sophisticated to mimic far more expensive quantum chemical methods. For example, we get an automatic, unbiased description of Jahn–Teller effects in Cu^{II} complexes,^{23,24} a good treatment of the subtle *trans* influence in planar, low-spin d^8 Pt^{II} complexes² and can even design a single set of LFMM FF parameters capable of reproducing the structures and relative energies of both high-spin and low-spin d^6 Fe^{II} amine complexes.²⁵ Moreover, since the LFSE calculation is parametric, the computational overhead for LFMM is small which facilitates, for example,

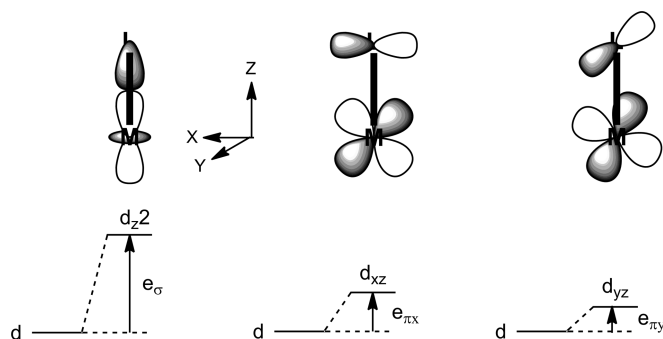


Figure 2.7 Definition of AOM e parameters in terms of local M-L bonding. The σ interaction is always destabilising ($e_\sigma > 0$) while π bonding in the two mutually perpendicular directions to the M-L vector (the local z axis) is shown for a π donor ($e_\pi > 0$).

the relatively long (*i.e.* 16 ns) molecular dynamics simulation of the dicopper metalloprotein tyrosinase.²⁶

Another advantage of LFMM is that the fundamental nature of metal–ligand bonding is the same in small molecule coordination complexes and for metals bound to proteins. Thus, LFMM parameters designed for metal complexes, where experimental crystallographic data are very accurate and high level quantum calculations are feasible, work equally well for metalloproteins such as the highly distorted centre in Type I copper proteins where crystallographic uncertainties are relatively large and high level quantum calculations are not feasible.²⁷

2.3 Scenarios

Having discussed some of the issues surrounding modelling TM species and how these issues might be tackled theoretically, the remainder of this chapter considers a selection of case studies aimed at illustrating the present state of play. These can be divided into two groups. Group 1 comprises investigations where the biological target is implicit. Typically, these are quantitative structure–activity relationships (QSARs) where some known property (or properties) of a set of individual molecules is processed in order to be able to predict the corresponding data for new systems and correlate this with any biological activity. Group 2 comprises investigations where the biological target is explicit. These include docking studies of drug molecules to protein active sites through to mechanistic studies of catalytic reaction pathways.

However, before we get to the modelling studies, it is worth highlighting the critical role played by protein crystallography as well as noting some potential problems. Most bioinorganic quantum chemical studies begin with the structure and new structures are often accompanied by a flurry of computational studies. Docking calculations presuppose you know the structure of the active

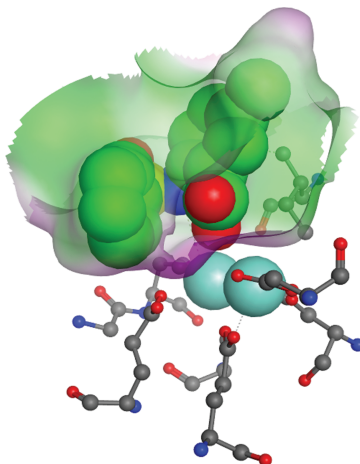


Figure 2.8 Detail from X-ray structure of anthranilic acid sulfonamide (green carbon atoms) bound to the methionine aminopeptidase active site (PDB code: 1YW7). The two Mn centres are shown as magenta spheres.

site. However, single-crystal X-ray diffraction studies of metal-containing proteins are potentially fraught. Firstly, compared to ‘small molecule’ diffraction studies, the resolution of protein crystallography is relatively poor and care must be exercised when interpreting, for example, the metal–ligand distances. Proteins are essentially just giant ligands – exquisitely varied and beautiful – but the basic nature of the coordinate bond remains the same. One should always refer back to the Cambridge Structural Database if one wishes to establish ‘standard’ metal–ligand bond lengths. Photoreduction is also a common issue that might not always be recognised so we must take care that the oxidation state is what we think it is. And we need to be aware of the crystallisation conditions. Consider, for example, methionine aminopeptidase (MetAP).

MetAP removes the N-terminal initiator methionine of nascent proteins that makes MetAP a potential target for cancer therapy. Accordingly, Shepard *et al.*²⁸ set about optimising a series of anthranilic acid sulfonamides guided by X-ray structural studies of substrates bound at the dimanganese active site (see Figure 2.8). However, in the same year, Ye *et al.*²⁹ claimed full kinetic competence for a monometallated active site and obtained the structure, with only the M1 site occupied, by limiting the amount of metal available during crystal growth (see Figure 2.9). Care must therefore be exercised to ensure that the structure derived from X-ray crystallography continues to refer to a catalytically and/or chemically competent form.

2.3.1 Predicting log *P* for Pt Complexes

The platinum anti-cancer drugs are probably the most successful ‘inorganic’ drugs in clinical use today. The ‘parent’ compound, *cis*-diamminodichloroplatinum(II), was discovered accidentally over 40 years ago and received

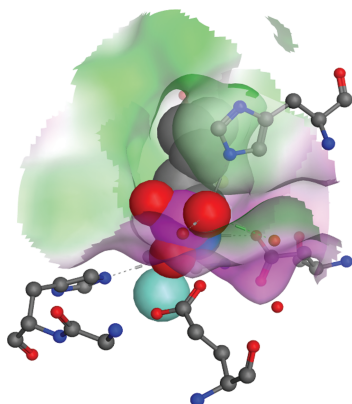


Figure 2.9 Active site detail of MetAP (PDB code 2GTX) with bound norleucinephosphonate (space filling). The single Mn centre is shown as the magenta sphere.

Federal Drug Administration approval about a decade later at the end of 1978. In common with most pharmaceuticals, its discovery engendered a frenzied search for new variations with improved activity and/or reduced side effects. Thousands of platinum complexes were synthesised and tested but only a few have survived the rigours of the approval process. Clearly, discovering and developing new pharmaceuticals is an incredibly difficult and complex business. The correct balance of absorption, distribution, metabolism, excretion and toxicity (ADME/tox) must be achieved. Lipophilicity is an important factor in ADME/toxicity since it relates to absorption, membrane transport, drug–receptor interactions and toxicity, and since the *n*-octanol/water partition coefficient, *P*, is a well accepted physicochemical model of lipophilicity, $\log P$ is widely used parameter in medicinal chemistry. However, for platinum compounds, experimental methods for measuring $\log P$ are expensive, slow and require substantial quantities of material. Hence, a computational method for predicting $\log P$ values is desirable.

Two methods, each employing 24 platinum complexes of the type illustrated (see Figure 2.10), correlated $\log P$ values with various molecular descriptors computed using QC methods.³⁰ The first method employed structures computed by HF theory to obtain the total surface area (TSA), the polar surface area (PSA) and the exposed surface area of platinum atoms (PtSA). A reasonable fit was obtained ($R^2 = 0.894$) with a root mean squared error (RMSE) of 0.354 using Equation (2.8):

$$\log P = -3.39 + 0.0099\text{TSA} - 0.0171\text{PSA} + 0.0543\text{PtSA} \quad (2.8)$$

The second method employed DFT-optimised structures and their associated properties in a linear regression which provides a better fit ($R^2 = 0.952$, RMSE = 0.21) as per Equation (2.9):

$$\log P = 3.06 + 0.0121\alpha - 0.1351\mu - 9.890I_{S,\max} + 1.268\Sigma V_S \quad (2.9)$$

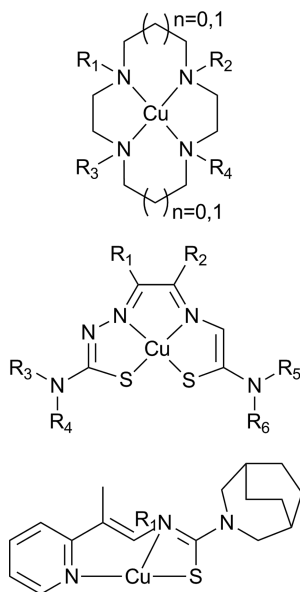


Figure 2.11 Copper chelate radiopharmaceuticals. Top: Cu-azamacrocycles; Middle: Cu-bis(thiosemicarbazones); Bottom: Cu- N -azabicyclo[3.2.2]nonanethiosemicarbazones.

respectively, seven and two further examples for validation. Log P data for 43 compounds were extracted from the literature, with the new MM3* parameters used to compute structures for which no X-ray geometry was available. The data were then subjected to a QSAR analysis using the Molecular Operating Environment (MOE) program³² initially employing 170 molecular descriptors which were eventually ‘pruned’ to just six. Eleven compounds were selected at random for the test set leaving 32 structures in the training set. The results of the QSAR modelling are reproduced in Table 2.2.

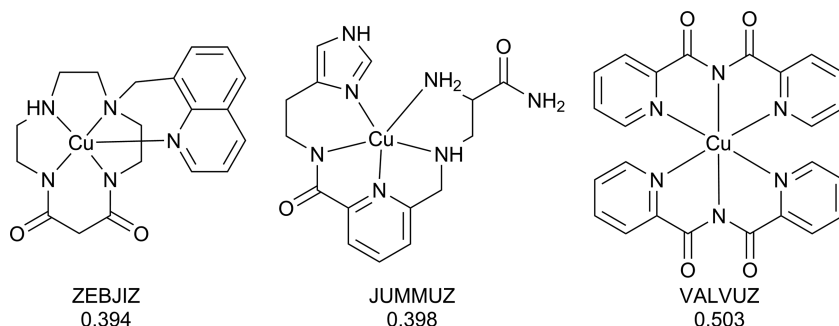
The results in Table 2.2 suggest it is possible to obtain good QSAR treatments for TM species if one is prepared to invest the effort in parameter development. At the same time, these types of complex do not present a very exacting test for MM since they are almost all square planar species, which is one of the ‘regular’ geometries that conventional MM can deal with. Moreover, the majority of the six-coordinate species have axial carboxylate donors with equatorial amines. The carboxylates have independent parameters that need only model long axial contacts while the amines are constrained in a macrocyclic ring and so there is never any ambiguity about the direction of the Jahn–Teller distortion.

The exceptions are five-coordinate species and complexes where a particular ligand type is found in both axial and equatorial positions. Three examples are shown (Figure 2.12), along with the RMSD fit that, for molecules in the training set, was set a limiting value of 0.4 Å as defining acceptable accuracy. Both ZEBJIZ and JUMMUZ were in the training set and are at the 0.4 Å

Table 2.2 Results from QSAR model of the lipophilicity ($\log P$) of the copper(II) complexes.³¹

<i>QSAR</i>	r^2	q^2	<i>RMSE</i>	<i>RMSE_{cv}</i>	n_{mols}	$n_{\text{descriptors}}$
$\log P$ model (training set)	0.95	0.92	0.35	0.35	32	6
$\log P$ model (test set)	0.93	0.91	0.41	0.47	11	6

r^2 refers to non-validated correlation, q^2 to the cross-validated correlation, RMSE to the root mean square error and RMSE_{cv} to the cross-validated root mean square error.

**Figure 2.12** Structural diagrams for copper complexes with their CSD refcodes and RMS deviations (Å) between experimental X-ray and computed MM3* structures.

threshold while VALVUZ was in the test set and has an RMSD somewhat larger than the threshold. Given the strong d electronic effects in d^9 copper(II) species described above, these systems are exactly the ones which conventional MM is expected to have difficulty with and it does.

2.3.3 Interactions Involving Explicit Receptors

When considering modelling TM-containing systems where the interaction between the drug molecule and its target is considered explicitly, we can distinguish five major scenarios:

1. the drug molecule contains one or more TM centres and binds non-covalently to a metal-free active site;
2. the drug molecule contains one or more TM centres and forms one or more dative covalent bonds to the active site;
3. the metal-free drug molecule binds non-covalently to a TM-containing active site;
4. the metal-free drug molecule forms one or more dative covalent bonds to the TM-containing active site;
5. both the drug and the active site contain TM centres.

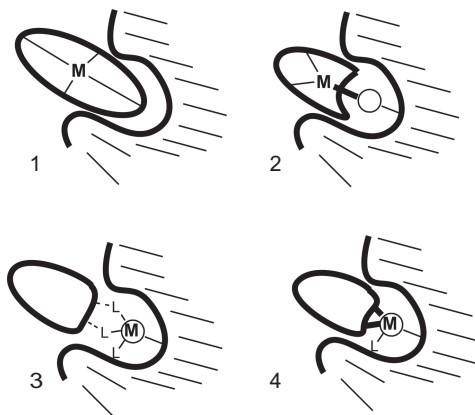


Figure 2.13 Schematic representations of some possible drug–active site interactions involving TM centres.

The author is not aware of any examples of scenario 5, which will thus not be considered further, while scenarios 1 to 4 are sketched schematically in Figure 2.13.

Within each scenario, various levels of sophistication can be applied to modelling studies. These range from docking studies to classical MD simulations through to QM/MM. Typically, docking is used to generate a number of likely poses which form the basis for the subsequent, more sophisticated (and more demanding) MD computations of the structural consequences of binding. QM/MM is usually reserved for addressing those questions for which the classical MD is either unsuited or would require significant re-parameterisation. To illustrate the kind of modelling which can be done and the particular issues surrounding the presence of TM centres, some examples spanning the four scenarios (see Figure 2.13) have been selected. This is in no way intended to be a comprehensive review of the current literature but rather a subjective selection that I hope will nevertheless cover the salient points.

2.3.4 Scenario 1: Metallodrug–DNA Intercalation

Copper–phenanthroline complexes are known to cleave DNA.^{33,34} A putative mechanism is shown in Figure 2.14 and comprises the following steps:

1. $\text{Cu}(\text{phen})_2^{2+}$ is reduced in solution to $\text{Cu}(\text{phen})_2^+$;
2. $\text{Cu}(\text{phen})_2^+$ binds DNA non-coordinatively;
3. $\text{Cu}(\text{phen})_2^+$ is oxidised to $\text{Cu}(\text{phen})_2^{2+}$ by H_2O_2 generating Cu-‘oxo’ and/or Cu-‘hydroxyl’ species, the exact nature of which is still unknown;
4. an oxidative attack mediated by these copper compounds leads to DNA cleavage.

This study illustrates the typical interplay between DFT and classical simulations. One of the principal goals is to explore how the copper complexes

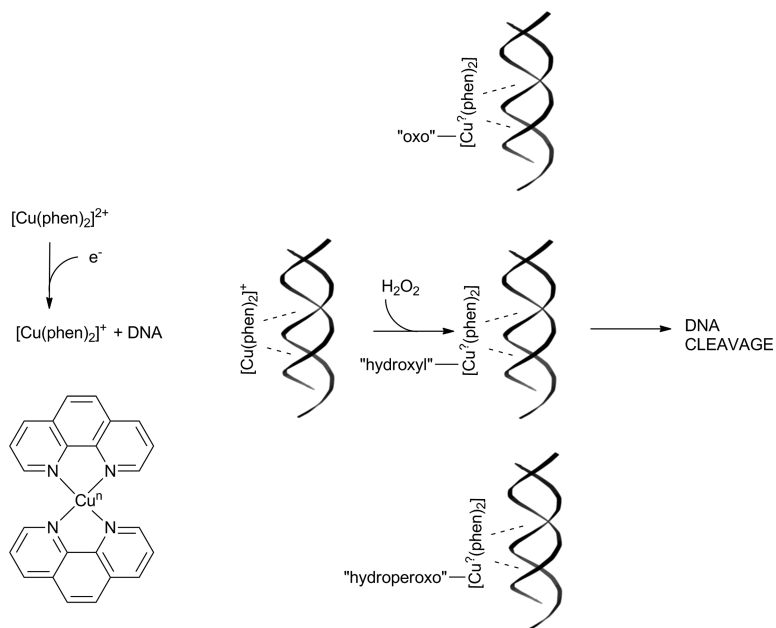


Figure 2.14 Schematic representation of the interaction between $[\text{Cu}(\text{phen})_2]^{2+}$ and DNA. (After Scheme 1 from Ref. 33.)

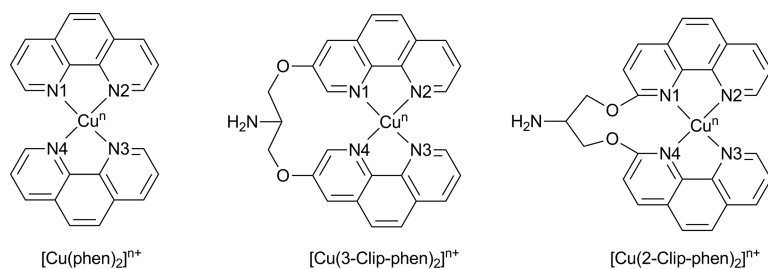


Figure 2.15 Copper complexes considered for DNA docking studies.

shown (see Figure 2.15) might interact with DNA. This involves preliminary docking calculations followed by MD simulations.

The closed shell d^{10} configuration of Cu^{I} makes these four-coordinate complexes relatively well behaved and amenable to conventional MM in that the geometry at the metal centre will, other things being equal, be essentially tetrahedral. Initially, DFT is used to develop parameters for the Cu–N bond stretch terms and for partial atomic charges. In addition, geometry optimisation shows that the N1–N2–N3–N4 dihedral angle between the two phenanthroline planes varies greatly. Since this is likely to have important consequences for any subsequent interaction with DNA, the authors use

further DFT calculations to map out the energy profile as a function of this dihedral and then capture this via a dihedral potential energy term. Having tailored the force field for their particular TM systems, MD simulations can be carried out to examine the binding preferences of the three complexes. The rather flatter structure of $[\text{Cu}(\text{3-Clip-phen})_2]^+$ in comparison to the other two complexes (*i.e.* it has a smaller N1–N2–N3–N4 dihedral) appears to better facilitate binding in the minor groove of DNA which may contribute to its enhanced cleavage activity.

2.3.5 Platinum–DNA Binding

The target for anti-cancer drugs such as cisplatin, *cis*-diaminodichloroplatinum(II), is believed to be DNA. Successive aquation steps replace the chloride ligands with water molecules which, ultimately, are further substituted by donors on the DNA, notably the N7 atom of guanine bases, often on the same DNA strand (see Figure 2.16).

The platinum lesion causes a sharp bend in the DNA that can interfere with the DNA repair machinery to the point where apoptosis occurs.

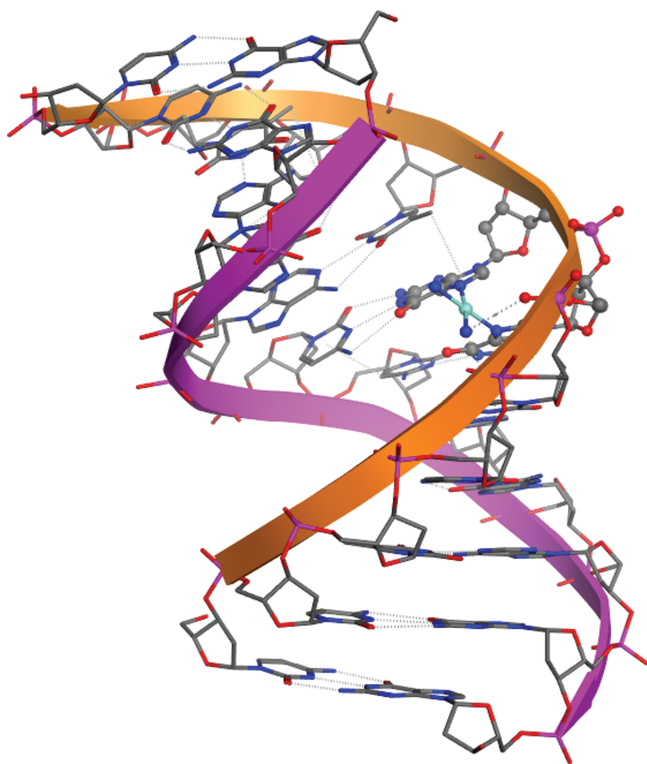


Figure 2.16 X-Ray crystal structure of duplex DNA with a cisplatin (shown in ball-and-stick mode) 1,2-GpG intrastrand crosslink (PDB code 3LPV).

From a modelling perspective, the planar structures of Pt^{II} complexes arise from the low-spin d^8 configuration. However, inasmuch as planarity may be enforced by, for example, an out-of-plane potential energy term, the basic structures of Pt^{II} complexes can be handled via conventional MM.³⁵ However, Pt^{II} complexes also display subtle structural variations due to the *trans* influence. Here, the nature of a given Pt–L bond is affected by the ligand *trans* to it. The effect on bond lengths, for example, can be large so, in general, it is important to be able to account for the *trans* influence. DFT will capture it more or less automatically and we have shown that LFMM can also handle Pt^{II} systems very well.² However, given the relatively few and fairly similar ligand types encountered for Pt–DNA interactions, *trans* influences probably do not vary significantly enough from system to system. Nevertheless, some extra force field development would be required but this can be avoided by taking advantage of a relatively new methodology, QM/MM.

In QM/MM, the ‘active site’ is treated at a relatively high level (*i.e.* DFT) while the remainder is treated with MM. A QM/MM treatment of the system shown (see Figure 2.16), for example, might take as its active site the platinum centre and its four ligands – *i.e.* two ammonia and two guanines – and model these with DFT with the rest treated by MM. The advantages of QM/MM are that no force field parameterisation is necessary, so one can begin calculations immediately, plus the quantum chemical calculation can be further analysed and probed. The disadvantages are that one has to exercise care joining the quantum and classical regions together, and that despite most of the system being in the classical region, QM/MM computations are still relatively expensive.

A recent study by Gkionis and Platts³⁶ looked at binding the $[\text{Pt}(\text{en})]^{2+}$ fragment, and its methylated derivatives, to models of single and double-stranded DNA comprising a pair of bases per strand. Three or four methyls on the nitrogen donors lead to geometry changes and binding energy reductions and a strong correlation between the measured cytotoxicity and the Pt–en binding energy was found.

2.3.6 Drug Binding in the Cytochrome P450 Active Site Cavity

The cytochromes P450 are a class of enzyme responsible for oxidising a wide range of xenobiotic compounds including drug molecules. In humans, there are 55 P450s although ~90% of therapeutic drug metabolism is carried out by just five isoforms. Inhibition and/or induction of these isoforms during the course of treatment has serious clinical implications for undesirable drug–drug interactions especially in cancer patients who take a range of drugs both to combat the disease and to relieve the ensuing symptoms and side effects.

Computer modelling has been used to provide structural insights into how drug molecules orient themselves in the active site cavity prior to being oxidised.³⁷ In this scenario, the metal-containing species, compound I, comprising a heme group with a central Fe bound to a cysteine and a terminal oxo ligand

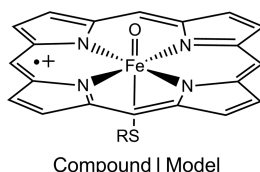


Figure 2.17 Structure of model for compound I radical cation.

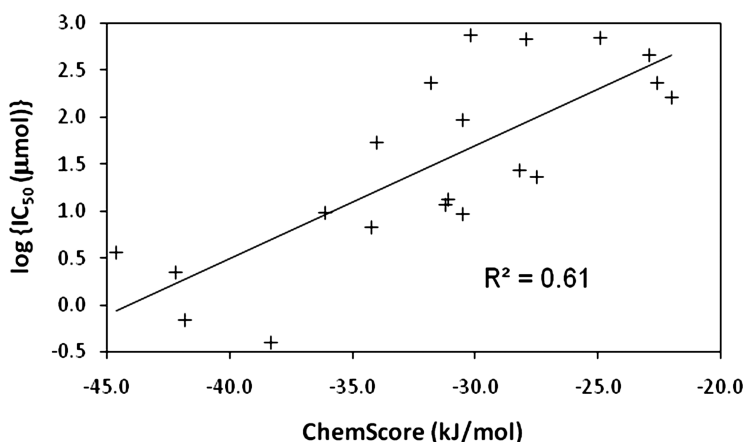


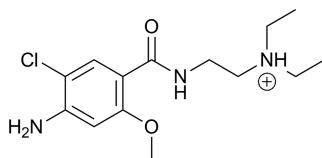
Figure 2.18 Comparison of experimental log (IC₅₀) and calculated ChemScore values (data from Table 1 of Ref. 37).

(see Figure 2.17), contributes to the general active site environment. So, once the structure of the active site is determined, either *via* X-ray diffraction of the particular isoform or from homology modelling studies, the heme group becomes part of the 'background' for docking calculations.

The IC₅₀ values for 20 drugs commonly co-administered to cancer patients were compared with ChemScore values from the best of ten orientations for each compound determined using the GOLD program. The data from reference 37 are reproduced (see Figure 2.18). The regression coefficient is $r^2 = 0.61$ ($q^2 = 0.59$).

The docking experiments highlighted a particular issue with the orientation of the antiemetic drug metoclopramide (see Figure 2.19). The active oxygen is coordinated to the Fe centre of the heme so that oxidation of the drug molecule is expected at positions spatially close to the Fe.

Two orientations of metoclopramide were found with very similar ChemScores, one consistent with the previously reported de-ethylation at the protonated nitrogen site and a second where now it is the basic amine which is in close proximity to the iron centre in the heme and thus likely to be the site of



Metoclopramide

Figure 2.19 Structure of the antiemetic drug metoclopramide.

oxidation. The observation from the modelling prompted the authors to look experimentally for metabolites consistent with amine oxidation, which were duly found.

2.3.7 Drug–Zinc Binding and Substrate Oxidation by P450

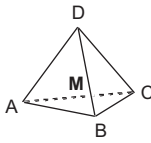
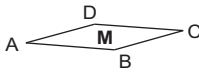
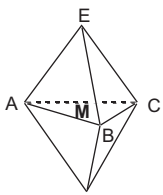
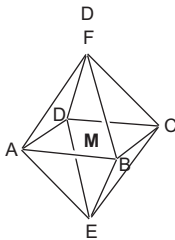
The preceding scenario successfully employed molecular docking but no covalent link between the drug and the metal centre was considered plus the geometry of the iron-heme moiety is well defined, quite rigid and thus relatively easy to handle computationally. In contrast, the zinc centres in various zinc-dependent enzymes are rather more flexible and variable in terms of both the detailed geometry and coordination number, plus drug molecules bind directly to the metal centre. Conventional docking technology treats the metal as a spherical particle and the extra complexity of directional M–L interactions requires enhancement. One such study reports the extension of the FlexX code to deal with both the varied coordination environments found for zinc complexes in general and with the multiple possible binding modes of carboxylates which are themselves commonly associated with zinc binding.³⁸

The authors developed a set of matching rules based on ideal coordination polyhedra, how many vertices are occupied by protein atoms and how many are free for ligand interactions (Table 2.3).

Compared to the previous ‘standard’ (*i.e.* spherical) metal model, the new version performs better, with 42% of those structures tested having an improved RMSD comparison with experiment, especially for zinc species. This is a good result considering all the important zinc-containing proteins, but was fairly predictable given that the assumption of regular coordination polyhedra applies reasonably well to the closed shell d^{10} Zn^{II} configuration. In contrast, d^9 Cu^{II} with its very strong d electronic effects is treated poorly.

The final scenario in this chapter concerns taking the modelling studies to the next level. Since we are often concerned with matters of selectivity that have an important kinetic component, a full characterisation of the system includes locating the transition states (TSs) on the reaction path. At first sight, the bond breaking/bond making processes occurring at the TS appear to demand some form of QM method. So, using P450 as our example, let us explore the possibilities and the pitfalls.

Table 2.3 Extended FlexX binding geometries for metal–ligand interactions.

<i>Geometry</i>	<i>Protein sites</i>	<i>Ligand sites</i>
	A, B, C A, B, C, D	D None
	A, B, C A, B, C, D	D None
	A, B, D A, D, E A, B, C A, B, C, D A, B, D, E A, B, C, D, E	C, E B, C D, E E C None
	A, B, E A, E, F A, B, E, F A, B, C, D A, B, C, D, E A, B, C, D, E, F	C, E, F B, C, D C, D E, F F None

The important step in alkene oxidation by P450 is the transfer of an oxygen atom from an activated iron-heme complex. Depending on the enzyme and the substrate, this could be an insertion into a C–H bond leading to hydroxylation or an addition to a C=C bond leading to epoxidation. The active iron species in P450s is the ferryl-porphyrin-cysteine complex compound I, which contains an Fe=O bond which may formally be formulated as Fe^V=O but is usually considered as Fe^{IV}=O with the porphyrin in a radical cation state (see Figure 2.17). Even though Compound I has quite a complicated electronic structure, gas phase DFT calculations on model complexes using the B3LYP functional have been shown to give a good description. However, different P450 isoforms can oxidise the same substrate in different places, so clearly DFT calculations on model systems will not be adequate and the effect of the surrounding protein must be included. QM/MM is the obvious choice and many groups have applied QM/MM to study the mechanism of substrate oxidation. However, while locating a TS using QM/MM is technically feasible, it is still a demanding calculation. Unfortunately, the calculations depend fairly strongly on conformation and some initial mechanistic conclusions, which were at odds with experiment but based on a single QM/MM path, were subsequently shown to be incorrect by computing the average barriers to hydroxylation and epoxidation for a series of pathways.³⁹ This more comprehensive QM/MM study

was consistent with experiment and highlights the importance of carrying out the necessary conformational sampling. However, with QM/MM this remains a very demanding computational task.

The alternative is to replace QM with a more efficient method like MM. MM is usually considered appropriate only for modelling ground state properties but if one is prepared to do the necessary parameter development, MM is not restricted. Indeed, with his Empirical Valence Bond (EVB) method, Warshel⁴⁰ has for decades shown how a properly constructed force field can be used to model the complete free energy surface from reactants through true first-order saddle points to products while Jensen's SEAM method⁴¹ can be used to optimise to the minimum energy crossing point between the reactant and product potential energy surfaces which, as illustrated (see Figure 2.20) can be quite a good approximation to the true TS.

While this type of calculation is very powerful, issues of selectivity may be decided using a simpler approach. The Q2MM method developed by Norrby⁴² is designed to compute the energy difference between two very similar (often diastereomeric) TSs rather than their absolute barriers to reaction. The starting point is a QM-optimised TS of the reaction of interest using simple, model substrates. The single negative Hessian eigenvalue, corresponding to the TS vector, is replaced by an arbitrarily large positive number and the whole Hessian matrix, plus the TS structure, are used as training data for a complete reparametrisation of the FF. The new FF is essentially guaranteed to reproduce the correct structure of the (model) TS and also captures the correct curvature of the PE surface for all vibrational degrees of freedom except the TS vector. In addition, locating the 'TS' has been converted to an energy minimisation so all the conventional conformational search techniques can be employed.

In their study of aliphatic hydroxylation of substrates by P450, Rydberg *et al.*⁴³ first constructed a Q2MM TS FF using DFT-optimised

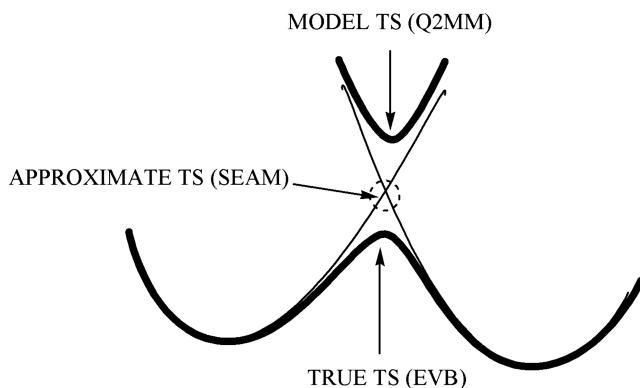


Figure 2.20 Schematic of different ways of representing the potential energy surface. EVB considers both reactant and product surfaces and their coupling to generate a 'true' PE surface with a first order saddle point, while the SEAM intersection and the Q2MM minimum are approximations.

structures and Hessian matrices for 14 complexes of formula $[\text{Fe}(\text{O}-\text{H}-\text{CR})(\text{SCH}_3)(\text{porphyrin})]^\ddagger$. This corresponds to the first (and rate-determining) step wherein the ferryl oxygen abstracts a hydrogen atom to generate $\text{Fe}-\text{OH}$ prior to the OH ‘rebounding’ to attack the aliphatic radical generated in the first step. The new TS force field reproduces DFT very well, with average bond length deviations of $\sim 0.02 \text{ \AA}$ and average bond angle deviations of $2-4^\circ$, and captures the roughly $0.2 \text{ \AA}$ extensions of the $\text{Fe}-\text{O}$, $\text{O}-\text{H}$ and $\text{H}-\text{C}$ distances of the TS.

The new FF can then be used for TS docking which enables a systematic search of conformational space of potential P450 inhibitors at their TS geometries. Excellent results are obtained for flunitrazepam and progesterone each docked into the active sites of the 2C9 and 3A4 isoforms of human P450. However, despite the enhanced sophistication over conventional docking methods and a concomitant improvement in performance, the authors conclude ‘the method works reasonably, but not perfectly’. Of course P450s are naturally ‘promiscuous’ and bind many different substrates that probably explains why the method can predict which substrate sites will be attacked but cannot rank them. They suggest that relaxing the Q2MM constraint of ‘locking’ the TS vector by using an alternative approach such as SEAM searching or EVB may help to alleviate the problem.

2.4 Conclusions

Transition metals carry out diverse functions in biological systems and are finding increasing use as therapeutic and diagnostic agents. Consequently, we can anticipate a growing need for fast, accurate computational methods to aid the discovery process. However, the additional complexity of electronic and geometric structure makes theoretical treatment of TM centres challenging.

At the quantum mechanical level, DFT is the most popular choice and has excellent accuracy even for TM systems. However, in common with all QM techniques, DFT is too slow for high throughput screening or systematic conformational searching where millions of individual calculations are required. Here, empirical methods rule although TM centres present some significant problems. Nevertheless, many metal centres are amenable to molecular mechanics methods although one may have to invest some time and effort in developing parameters. For those more complicated metal centres such as $d^9 \text{ Cu}^{\text{II}}$, extensions to ‘conventional’ MM such as the author’s ligand field molecular mechanics method offer the accuracy and flexibility of quantum methods like DFT but at the computational speed of MM.

References

1. S. H. van Rijt and P. J. Sadler, *Drug Discovery Today*, 2009, **14**, 1089–1097.
2. A. Anastasi and R. J. Deeth, *J. Chem. Theor. Comp.*, 2009, **5**, 2339–2352.
3. P. E. M. Siegbahn, *Faraday Discuss.*, 2003, **124**, 289–296.

4. A. D. Becke, *Phys. Rev. A*, 1988, **38**, 3098–3100.
5. A. D. Becke, *J. Chem. Phys.*, 1993, **98**, 5648–5652.
6. A. D. Becke, *J. Chem. Phys.*, 1993, **98**, 1372–1377.
7. A. D. Becke, *J. Chem. Phys.*, 1992, **97**, 9173.
8. P. J. Stephens, F. J. Devlin, C. F. Chabalowski and M. J. Frisch, *J. Phys. Chem.*, 1994, **98**, 11623–11627.
9. S. J. Grimme, *J. Comput. Chem.*, 2004, **25**, 1463–1473.
10. F. Ogliaro, N. Harris, S. Cohen, M. Filatov, S. P. de Visser and S. Shaik, *J. Am. Chem. Soc.*, 2000, **122**, 8977–8989.
11. O. Salomon, M. Reiher and B. A. Hess, *J. Chem. Phys.*, 2002, **117**, 4729–4737.
12. A. Sorkin, M. A. Iron and D. G. Truhlar, *J. Chem. Theor. Comput.*, 2008, **4**, 307–315.
13. S. F. Ye and F. Neese, *Inorg. Chem.*, 2010, **49**, 772–774.
14. R. J. Deeth and N. Fey, *J. Comp. Chem.*, 2004, **25**, 1840–1848.
15. H. Kayi and T. J. Clark, *J. Mol. Model.*, 2007, **13**, 965–979.
16. V. S. Allured, C. M. Kelly and C. R. Landis, *J. Am. Chem. Soc.*, 1991, **113**, 1–12.
17. P. Comba and T. W. Hambley, *Molecular Modeling of Inorganic Compounds*, 2nd edn., VCH, Weinheim, 2001.
18. R. J. Deeth and M. A. Hitchman, *Inorg. Chem.*, 1986, **25**, 1225–1233.
19. D. A. Johnson and P. G. Nelson, *Inorg. Chem.*, 1995, **34**, 5666–5671.
20. V. J. Burton, R. J. Deeth, C. M. Kemp and P. J. Gilbert, *J. Am. Chem. Soc.*, 1995, **117**, 8407–8415.
21. R. J. Deeth, *Coord. Chem. Rev.*, 2001, **212**, 11–34.
22. C. E. Schaeffer and C. K. Jorgensen, *Mol. Phys.*, 1965, **9**, 401.
23. R. J. Deeth and L. J. A. Hearnshaw, *Dalton Trans.*, 2006, 1092–1100.
24. A. Bentz, P. Comba, R. J. Deeth, M. Kerscher, B. Seibold and H. Wadepohl, *Inorg. Chem.*, 2008, **47**, 9518–9527.
25. R. J. Deeth, A. E. Anastasi and M. J. Wilcockson, *J. Am. Chem. Soc.*, 2010, **132**, 6876–6877.
26. R. J. Deeth and C. Diedrich, *J. Biol. Inorg. Chem.*, 2010, **15**, 117.
27. R. J. Deeth, *Inorg. Chem.*, 2007, **46**, 4492–4503.
28. G. S. Sheppard, J. Y. Wang, M. Kawai, S. D. Fidanze, N. Y. BaMaung, S. A. Erickson, D. M. Barnes, J. S. Tedrow, L. Kolaczowski, A. Vasudevan, D. C. Park, G. T. Wang, W. J. Sanders, R. A. Mantei, F. Palazzo, L. Tucker-Garcia, P. Lou, Q. Zhang, C. H. Park, K. H. Kim, A. Petros, E. Olejniczak, D. Nettesheim, P. Hajduk, J. Henkin, R. Lesniewski, S. K. Davidsen and R. L. Bell, *J. Med. Chem.*, 2006, **49**, 3832–3849.
29. Q. Z. Ye, S. X. Xie, Z. Q. Ma, M. Huang and R. P. Hanzlik, *Proc. Nat. Acad. Sci.*, 2006, **103**, 9470–9475.
30. I. V. Tetko, I. Jaroszewicz, J. A. Platts and J. Kuduk-Jaworska, *J. Inorg. Biochem.*, 2008, **102**, 1424–1437.
31. P. Wolohan, J. Yoo, M. J. Welch and D. E. Reichert, *J. Med. Chem.*, 2005, **48**, 5561–5569.
32. MOE *Molecular Operating Environment 2007.09*, 2007; Chemical Computing Group, Montreal: Montreal, 2007.

33. A. Robertazzi, A. V. Vargiu, A. Magistrato, P. Ruggerone, P. Carloni, P. de Hoog and J. Reedijk, *J. Phys. Chem. B*, 2009, **113**, 10881–10890.
34. A. Robertazzi, A. Magistrato, P. de Hoog, P. Carloni and J. Reedijk, *Inorg. Chem.*, 2007, **46**, 5873–5881.
35. T. R. Cundari, W. Fu, E. W. Moody, L. L. Slavin, L. A. Snyder, S. O. Sommerer and T. R. Klinckman, *J. Phys. Chem.*, 1996, **100**, 18057–18064.
36. K. Gkionis and J. A. Platts, *J. Biol. Inorg. Chem.*, 2009, **14**, 1165–1174.
37. J. L. Yu, M. J. I. Paine, J. D. Marechal, C. A. Kemp, C. J. Ward, S. Brown, M. J. Sutcliffe, G. C. K. Roberts, E. M. Rankin and C. R. Wolf, *Drug Metab. Dispos.*, 2006, **34**, 1386–1392.
38. B. Seebeck, I. Reulecke, A. Kamper and M. Rarey, *Proteins: Struct., Funct., Bioinform.*, 2008, **71**, 1237–1254.
39. R. Lonsdale, J. N. Harvey and A. J. Mulholland, *J. Phys. Chem. B*, 2010, **114**, 1156–1162.
40. A. Warshel, *Annu. Rev. Biophys. Biomol. Struct.*, 2003, **32**, 425–443.
41. F. Jensen and P. O. Norrby, *Theor. Chem. Acc.*, 2003, **109**, 1–7.
42. P. O. Norrby, *Theochem. J. Mol. Struct.*, 2000, **506**, 9–16.
43. P. Rydberg, L. Olsen, P. O. Norrby and U. Ryde, *J. Chem. Theor., Comput.*, 2007, **3**, 1765–1773.

CHAPTER 3

Modeling Protein–Protein Interactions by Rigid-body Docking

DAVID W. RITCHIE

Orpailleur Team, INRIA Nancy – Grand Est, 615 Rue du Jardin Botanique,
54506 Vandoeuvre-lès-Nancy, France

3.1 Protein–Protein Interactions

Protein–protein interactions (PPIs) are central to many biological processes such as metabolism, photosynthesis, transport, immune response and gene expression, for example. At the molecular level the aforementioned processes are often orchestrated by non-covalent associations between two or more proteins to form protein complexes. These complexes may exist as short-lived transitory associations, as in *e.g.* enzyme catalysis, or as long-lived multimeric systems such as the ribosome, transcription factors, cell surface and ion channel proteins, for example. Thus if DNA represents the biological blueprint for life, then proteins make up the molecular machinery. In the current post-genomic era, we now know the blueprints for many proteins, but we still know very little about how they operate at the molecular level. Genome-wide proteomics studies^{1–4} are providing growing lists of putative protein–protein interactions (PPIs), but many of these predicted interactions correspond to completely unknown functions. For example, yeast is known to have around 6000 proteins, giving rise to between about 38 000 and 75 000 PPIs, and it has been estimated that about 50% of these PPIs have been observed experimentally. On the other

hand, the human genome encodes around 30 000 proteins, giving from 154 000 to 370 000 PPIs, of which only around 10% are known to date.^{5,6}

The ‘gold standard’ techniques for determining the three-dimensional (3D) structures of proteins and other macromolecules are undoubtedly X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. However, cryo-electron microscopy (cryo-EM) is becoming increasingly useful for the solution of very large protein structures, and the resolution that can be achieved in cryo-EM is beginning to approach that of crystallography.⁷ Furthermore, fitting high-resolution X-ray protein structures into low-resolution cryo-EM density maps is becoming an increasingly powerful approach to determine the atomic structures of very large protein complexes.^{8–13} Today, around 12 000 distinct 3D protein structures have been deposited in the Protein Data Bank (PDB).¹⁴ However, only a very small proportion of these structures correspond to protein–protein complexes. Furthermore, due to a number of practical difficulties, it seems unlikely that it will become possible to solve the structures of protein complexes using high-throughput structural genomics techniques in the foreseeable future.¹⁵ On the other hand, techniques such as yeast two-hybrid (Y2H) and tandem-affinity-purification mass spectrometry (TAP-MS) can now identify many thousands of PPIs simultaneously in order to build large-scale maps of PPI networks.^{16,17} However, these approaches cannot determine the molecular structures of individual protein complexes, and many of the detected interactions are in fact false-positives. Hence, there is growing interest in using computational techniques to help understand PPIs at the molecular level.^{18,19}

3.2 Predicting Protein Interaction Surfaces

Although the properties of protein–protein interfaces have been studied extensively,^{20–26} it remains a significant challenge to predict reliably the locations of protein–protein interaction surfaces using computational techniques alone.^{21,25} On the other hand, using a database of some 2300 alanine mutations of heterodimeric protein complexes for which affinity measurements and structural information was available, Bogan and Thorn²² showed that often only a small set of core interface residues provides the majority of the binding free energy of a complex. They proposed a ‘hot-spot/O-ring’ model of protein binding sites in which the core hot spot residues are surrounded by a ring of energetically unimportant residues whose main role is to occlude bulk solvent from the hot spot. This study found that occlusion of solvent from the protein–protein interface is a necessary condition for binding, but there is no direct correlation between the experimentally determined binding free energy and buried surface area. Interestingly, this study also found that the bulky side chains of Trp, Arg and Tyr appear in hot spots with high frequencies (21%, 13% and 12%, respectively), whereas Leu, Met, Ser, Thr and Val residues are rarely observed in hot spots (3% frequency or less).²² By analyzing the nature of existing protein complexes, Janin and colleagues^{23,24} independently proposed a similar ‘core-rim’ model of protein binding sites, in which the core

residues resemble the composition of protein interiors and the rim residues, which remain to some extent solvent-accessible in the complex, resemble more the general composition of surface residues.

Although the core-rim interface model seems compelling, it is difficult to define simple rules with which unknown binding sites may be identified. Therefore, a range of machine learning techniques have been developed to predict protein–protein interface sites automatically.^{25,27–35} These systems are typically trained using various combinations of *e.g.* buried surface areas, desolvation and electrostatic interaction energies, hydrophobicity scores and residue conservation scores. Cleft desolvation is often found to be one of the strongest predictive properties,^{25,33} whereas electrostatic complementarity has been found to have little if any predictive capability, and residue conservation has less predictive power than might be expected.²⁵ Broadly speaking, current interface prediction algorithms can generally predict the locations of protein interfaces with around 50% overlap between predicted and native interface residues in up to around 70% of complexes.³⁶ A recent analysis of current binding site prediction tools by de Vries and Bonvin³⁷ confirms that several machine learning based approaches are becoming useful predictive tools. On the other hand, as part of their analysis of blind docking results for the community-wide CAPRI (Critical Assessment of PRedicted Interactions) blind docking experiment,³⁸ Lensink and Wodak³⁹ recently showed that the quality of the interfaces predicted by protein–protein docking algorithms is significantly higher than that obtained by other interface prediction methods.

3.3 Protein–Protein Interaction Databases

Structural PPI databases are becoming important assets to analyze and predict the structures of protein complexes. Some recent examples include PQS,⁴⁰ 3DID,⁴¹ PIBASE,⁴² iPFam,⁴³ InterDom,⁴⁴ Interpare,⁴⁵ 3D Complex,⁴⁶ SCOPPI⁴⁷ and PROTCOM.⁴⁸ Such databases are often derived from protein–protein complexes found in the PDB, and they usually provide additional information to annotate the nature of the interface. For example, many databases distinguish permanent dimers, which often have interfaces composed of mostly hydrophobic residues, from hetero-complexes which tend to have more polar residues at the interface. The PQS system also distinguishes interfaces between biologically active subunits from crystal packing contacts, and hence provides a significantly richer source of protein–protein interface data than the original experimentally determined structures.⁴⁰ Many databases provide the buried surface area of the interface and some (*e.g.* 3DID, PIBASE) list the identities and secondary structure types of the residues at each interface. Several databases (*e.g.* InterDom, iPFAM, Interpare, PIBASE, SCOPPI) describe interfaces at the protein chain and protein domain levels, and link out to the SCOP⁴⁹ and/or Pfam⁵⁰ protein domain classifications. The 3D Complex database may be used to compute new classifications of protein complexes using quaternary structure symmetry and topology criteria, for example. Other protein structure resources, such as the Protein Docking Benchmark,⁵¹ list the

separate unbound partners of known protein complexes and provide large numbers of decoy complexes for testing and developing new docking algorithms.^{52,53}

Structural PPI databases are being used increasingly to help predict the structures of unknown protein complexes. For example, if a given pair of proteins are homologous to the components of a known complex, it is reasonable to expect that they might interact in a similar way. Obviously, current databases contain only a very small fraction of all possible protein–protein complexes. Nonetheless, as structural genomics initiatives continue to populate the space of protein 3D structures, it seems clear that using structural database systems to predict PPIs by homology will become an increasingly powerful approach.³⁶

3.4 The Protein Docking Problem

Protein docking is the task of calculating the 3D structure of a protein complex starting from the individual structures of the constituent proteins. Since proteins have dynamical structures that may change conformation on binding, predicting how proteins bind and form complexes is rather like trying to assemble the pieces of a complex 3D jigsaw puzzle in which the given parts do not fit together perfectly.

In one of the earliest computer simulations of protein–protein interactions, Wodak and Janin⁵⁴ modeled the docking of bovine pancreatic trypsin inhibitor (BPTI) onto the active site of trypsin. For this calculation, BPTI was initially placed near the trypsin binding site. It was then rotated in about 2300 angular steps, and at each orientation it was translated towards trypsin until the calculated energy function exceeded a given threshold, suggesting a close contact had been achieved. For each such contact, the surface area buried in the interface was estimated by counting the number of pairwise residue contacts, and 47 orientations with the most contacts were selected for further refinement by minimizing the energy over all six rigid body degrees of freedom. To speed the energy calculations, the binding free energy was estimated using residue pair potentials.⁵⁵ The whole calculation took 40 min on an IBM 370 mainframe computer. Subsequent analysis of the results showed that a near-native orientation for the complex appeared at rank 12 in the list of minimized poses. This early calculation demonstrated that it was possible to predict the docking mode of a pair of proteins by using reduced representations of each protein and by minimizing a simple energy-based scoring function.

Following the first computational experiment of Wodak and Janin, protein docking has since become a distinct computational field that combines knowledge and techniques from several disciplines including physics, chemistry, biology, mathematics and computing. There have been several reviews of protein docking algorithms^{56–65} and the state of the art is regularly assessed and reviewed in the CAPRI blind docking experiment^{66–71} and its partner conference, Modeling of Protein Interactions in Genomes.⁷² Although it appears that current docking algorithms still cannot reliably predict binding affinity,⁷³

protein docking techniques are helping to provide a better physical picture of PPIs,³⁶ and they are beginning to be used as an *in silico* technique to help predict PPI networks.^{74,75}

Recent docking and molecular dynamics (MD) simulation studies support a picture of protein complexes being formed in at least a two-step process.^{76–80} In the initial collision encounter complex, recognition takes place through desolvation and burial of key hot spot anchor residues at the center of the nascent interface, the conformations of which do not significantly change on binding.^{77,79} This is followed by a latching phase in which peripheral interface residues may adjust their rotameric conformations into complementary arrangements.⁷⁷ Arguably, there is a final induced fit step⁷⁸ in which interface side chains adjust their torsion angles to adopt off-rotamer conformations⁸¹ and interfacial waters become frozen into their crystallographically observable positions.⁸² Although this is clearly an idealized picture of a complex dynamical process, it is broadly compatible with the experimental and statistical studies of known protein interfaces.^{22–24} Nonetheless, it is not necessary to simulate precisely the physics of protein complex formation in order to make useful predictions about PPIs. For example, even a relatively low quality protein docking model can help to identify potentially druggable PPIs.^{83–86}

3.5 Rigid Body Docking

As most proteins consist of many thousands of atoms, it is generally not practical to predict putative docking orientations using conventional molecular mechanics based approaches because of the enormous computational cost of calculating MD trajectories for multiple trial starting positions, and because of the difficulty of distinguishing multiple local minima within those trajectories. Hence, current docking algorithms employ a variety of efficient search and energy-based scoring techniques, such as fast Fourier transform (FFT) correlations, geometric hashing and Monte Carlo (MC) sampling, for example. In order to make the calculation tractable, such algorithms typically begin by assuming that the structures to be docked are rigid. This essentially reduces the problem to a six-dimensional (6D) rotational–translational search space. Rigid body docking approaches can often produce a relatively small list of up to a few hundred orientations, amongst which a near-native binding mode is often observed.

However, despite the use of improved scoring functions that typically include models of desolvation, hydrophobicity and electrostatics, current algorithms still have difficulty in identifying the best solution from the list of false positives, or decoys. Therefore, most current docking algorithms use a two-step searching and scoring procedure, in which *ab initio* techniques are used to generate an initial list of putative complexes which are then re-scored using available biophysical information and knowledge-based potentials derived from analyses of existing protein–protein interfaces.^{87–89} Nonetheless, it can still take several hours or even days to complete a docking calculation using high performance workstations.³⁶ Hence several docking algorithms have been implemented

using batch queuing systems on computational clusters.^{90–94} However, thanks to some striking recent improvements in the computational performance of graphics processor units (GPUs), the ‘Hex’ docking algorithm can now perform the rigid-body part of a docking calculation in a matter of seconds on a modern desktop GPU.⁹⁵ This is described in further detail below.

3.6 Docking By Geometric Hashing

The geometric hashing algorithm of Fischer *et al.*⁹⁶ is a docking method adapted from a computer vision object recognition technique.⁹⁷ In this approach, the molecular surfaces of the proteins to be docked, often called a ‘ligand’ and a ‘receptor’, are first processed to identify local critical points, or ‘caps’ and ‘pits’, and triplets of such points are stored in a hash table.⁹⁸ The search for docking orientations then consists of matching cliques of ligand caps to receptor pits in order to define candidate ligand poses with respect to a fixed receptor. The poses generated in this way are then evaluated by projecting the volumes of the two proteins onto a 3D grid and by evaluating a complementarity score based on the number of overlapping grid cells.⁹⁹ Although the geometric hashing technique is combinatorial in the number of critical points used, and the grid-scoring stage is relatively expensive, the overall approach is very fast if a small number of well-chosen critical points can be identified. On the other hand, if the proteins exhibit relatively large conformational changes on binding, there is a risk that near-native poses might not be generated from the initial surface analysis of the unbound starting structures. Hence, most current rigid body docking algorithms use FFT techniques in order to cover the rigid body search space exhaustively.

3.7 Cartesian Grid-based FFT Docking

The first FFT-based approach to the protein docking problem was described by Katchalski-Katzir *et al.*¹⁰⁰ In this technique, the two proteins to be docked, A and B, are placed in a 3D Cartesian grid and each volume element, or ‘voxel’, of the grid is assigned a value according to:

$$A[\mathbf{x}] = \begin{cases} \sigma & \text{if } A[\mathbf{x}] \text{ is a surface voxel} \\ \tau & \text{if } A[\mathbf{x}] \text{ is a surface voxel} \\ 0 & \text{otherwise} \end{cases} \quad (3.1)$$

and

$$B[\mathbf{x}] = \begin{cases} \sigma & \text{if } B[\mathbf{x}] \text{ is a surface voxel} \\ \delta & \text{if } B[\mathbf{x}] \text{ is a surface voxel} \\ 0 & \text{otherwise,} \end{cases} \quad (3.2)$$

where $\mathbf{x} = (x, y, z)$. This is illustrated schematically in Figure 3.1. The goal is then to find the translation that maximizes the degree of overlap, or correlation,

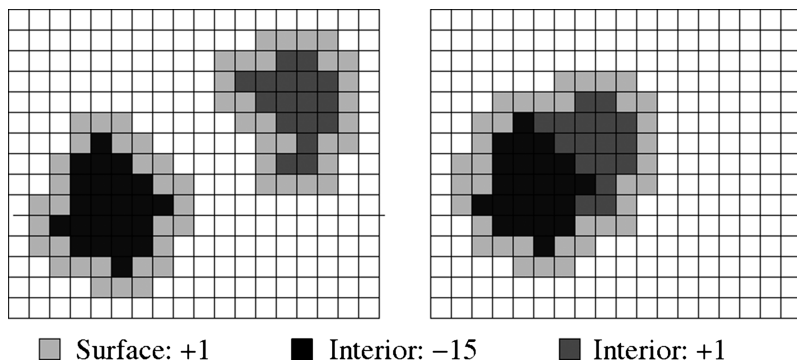


Figure 3.1 A two-dimensional illustration of Cartesian grid-based docking. The interior volume occupied by each protein is represented as a set of discrete voxels (dark shades) surrounded by a layer of surface voxels (light shade). The goal of a docking search is to find the orientation which maximises the overall overlap between the surface and interior regions of each protein.

between the interior and surface voxels of the two proteins whilst avoiding significant overlap between their interior voxels. In their original algorithm, Katchalski-Katzir *et al.* put $\tau = -15$ and $\sigma = \delta = +1$. The correlation is calculated for all possible translations of the ligand voxels with respect to a fixed receptor. In other words, the goal is to find the translation ($\Delta \mathbf{x}$) which maximizes the complementarity score, C :

$$C[\Delta \mathbf{x}] = \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} \sum_{z=0}^{N-1} A[\mathbf{x}] B[\mathbf{x} - \Delta \mathbf{x}]. \quad (3.3)$$

In practice, the correlation may be calculated very efficiently using 3D FFTs, although multiple FFTs must be calculated over a range of rotated orientations of one of the proteins in order to cover the 6D search space.¹⁰⁰

For simple shape-matching calculations, 3D grid correlations may also be calculated efficiently using integer bit-wise operations.¹⁰¹ However, it is standard practice nowadays to use FFT techniques to calculate the correlations because the $O(N^3 \ln^3 N)$ algorithmic complexity of the 3D FFT is much more favorable for large N than the $O(N^6)$ cost of a direct calculation. Furthermore, floating point FFTs allow general scalar quantities to be correlated. For example, from Poisson's equation, the classical *in vacuo* electrostatic energy, E , of a charge density distribution, $\phi(\mathbf{x})$, in the presence of an electrostatic potential, $\rho(\mathbf{x})$, is given by¹⁰²

$$E = \frac{1}{2} \int \phi(\mathbf{x}) \rho(\mathbf{x}) d\mathbf{x}. \quad (3.4)$$

Hence the *in vacuo* electrostatic interaction energy between a pair of proteins, A and B, may be calculated as a function of translations of protein B using:

$$E(\Delta\mathbf{x}) = \frac{1}{2} \int (\phi_A(\mathbf{x})\phi_B(\mathbf{x} - \Delta\mathbf{x}) + \rho_A(\mathbf{x})\phi_B(\mathbf{x} - \Delta\mathbf{x}))d\mathbf{x}. \quad (3.5)$$

It is worth noting at this stage that conventional Cartesian FFT docking approaches essentially represent each property to be correlated as a Fourier expansion of the form

$$\phi_A(\mathbf{x}) = \frac{1}{N^{\frac{3}{2}}} \sum_{s=0}^{N-1} \sum_{t=0}^{N-1} \sum_{u=0}^{N-1} A_{stu}^{\phi} e^{-\frac{2\pi i(sx+ty+uz)}{N}}, \quad (3.6)$$

where the A_{stu}^{ϕ} represent the expansion coefficients of the electrostatic potential, and the complex exponential terms provide the basis functions for the expansion. In other words, performing a forward FFT on a 3D grid of electrostatic potential values implicitly calculates all of the expansion coefficients simultaneously. Then, thanks to the orthogonality of the basis functions, the correlation between a pair of scalar properties may be calculated using *e.g.*

$$\int \phi_A^*(\mathbf{x})\rho_B(\mathbf{x} - \Delta\mathbf{x})d\mathbf{x} = \frac{1}{N^3} \sum_{s=0}^{N-1} \sum_{t=0}^{N-1} \sum_{u=0}^{N-1} A_{stu}^{\phi*} A_{stu}^{\rho} e^{\frac{2\pi i(s\Delta x+t\Delta y+u\Delta z)}{N}}, \quad (3.7)$$

where the asterisk denotes complex conjugation. In other words, the desired overlap integral as a function of the translational shifts may be calculated as the inverse Fourier transform of the product of two forward transforms. Equation (3.7) is essentially a statement of the well-known convolution theorem.

Several Cartesian grid-based FFT docking approaches (for example, FTDOCK,¹⁰³ DOT,¹⁰⁴ MolFit¹⁰⁵ and ZDOCK¹⁰⁶) use 3D FFT libraries to calculate shape and electrostatic correlations in this way. They differ mainly according to the way they estimate the potential and charge density from partial atom charges and how these properties are discretized onto FFT grid points. They also differ in the manner in which they estimate the non-uniform dielectric nature of the medium. Nonetheless, by following the above pattern, other types of potentials may be incorporated in FFT correlation calculations. For example, ZDOCK 2.3 uses the ‘ACE’ potential of Zhang *et al.*¹⁰⁷ to estimate desolvation effects,¹⁰⁶ and the MolFit program of Berchanski *et al.*¹⁰⁸ implements a simple model of hydrophobicity by using the complex part of each grid cell to assign additional weight to contacts between hydrophobic residues. It is often found that including such additional scoring terms helps to improve the rank of near-native docking poses, although it remains a considerable challenge to be able to predict good poses in the absence of prior knowledge about the location of the binding site(s). However, a major practical disadvantage of the Cartesian grid-based FFT approaches is that the translational FFT correlations must be repeated for many thousands of rotational samples of one of the molecules, and this computational overhead seems unavoidable in the translational FFT model.

Thus, despite the rigid-body assumption, Cartesian grid-based FFT docking algorithms are inherently computationally expensive.

3.8 Spherical Polar Fourier Basis Functions

The spherical polar Fourier (SPF) correlation approach was developed in order to address the main limitations of the Cartesian FFT-based methods.¹⁰⁹ Instead of encoding protein shapes and properties in a Cartesian grid using complex exponential basis functions, in the SPF approach each protein is represented as an expansion of orthonormal SPF basis functions using a spherical polar coordinate system located at its center of mass. The relationship between Cartesian and spherical polar coordinate systems is illustrated in Figure 3.2A. Using a pair of spherical polar coordinate systems, the 6D search space may be defined in terms of one intermolecular distance and five Euler rotation angles (see Figure 3.2B). This allows docking correlations to be calculated primarily using *rotational* FFT correlations. In the context of a 6D search space, this is much more efficient than using translational FFTs because only one forward Fourier transform is required for each property.

The SPF approach begins with a similar voxel-based representation of protein shape to that of Katchalski-Katzir *et al.*, but instead of calculating Cartesian FFTs the voxel samples are used only to encode the protein shapes as expansions of SPF basis functions. For example, the volume of the receptor molecule which is interior to the van der Waals surface is encoded as a ‘shape density’ function, $\tau_A(\mathbf{r})$, to order N using

$$\tau_A(\mathbf{r}) = \sum_{n=1}^N \sum_{l=0}^{n-1} \sum_{m=-l}^{m=l} a_{nlm}^{\tau} R_{nl}(r) y_{lm}(\theta, \phi), \quad (3.8)$$

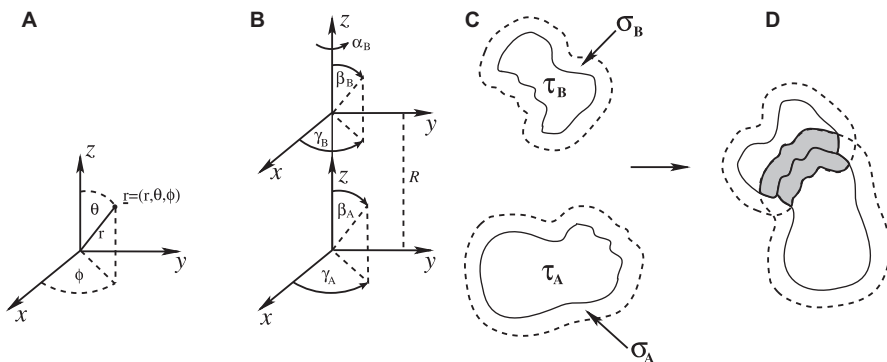


Figure 3.2 (A) The relationship between spherical polar (r, θ, ϕ) and Cartesian (x, y, z) coordinates. (B) An illustration of the 6D docking search space in terms of one intermolecular distance, R , and five Euler rotation angles ($\beta_A, \gamma_A, \alpha_B, \beta_A, \gamma_A$). (C) A sketch of the interior (τ) and surface skin (σ) density functions of a pair of proteins to be docked. (D) The final predicted docking orientation in which the degree of overlap between opposite pairs of density functions is maximized (shaded region).

where $\mathbf{r} = (r, \theta, \phi)$ are 3D spherical polar coordinates, a_{nlm}^τ are the expansion coefficients, $y_{lm}(\theta, \phi)$ are normalized real spherical harmonic (SH) functions¹¹⁰ and $R_n(r)$ are orthonormal radial basis functions. Since the radial index, n , counts from unity, the highest harmonic order and highest polynomial power in each coordinate is therefore $L = N - 1$.^{109,111} Thanks to the orthogonality of the basis functions, each expansion coefficient may be isolated by multiplying each side of Equation 3.8 by $R_{kj}(r)y_{js}(\theta, \phi)$ and integrating to obtain:

$$a_{kjs}^\tau = \int \tau_A(\mathbf{r}) R_{kj}(r) y_{js}(\theta, \phi) d\mathbf{r}. \quad (3.9)$$

Hence it is straightforward to calculate the expansion coefficients by numerical integration over the voxels in a Cartesian grid.¹⁰⁹ By defining and calculating the volumes of a surface skin region for each protein, $\sigma_A(\mathbf{r})$ and $\sigma_B(\mathbf{r})$, in a similar way, and by considering the overlap volumes shown in Figures 3.2C and 3.2D, protein shape complementarity may be written as a sum of overlap integrals between opposing pairs of surface and interior shape density functions:

$$C = \int (\sigma_A(\mathbf{r})\tau_B(\mathbf{r}) + \sigma_B(\mathbf{r})\tau_A(\mathbf{r}) - Q\tau_A(\mathbf{r})\tau_B(\mathbf{r}))d\mathbf{r}, \quad (3.10)$$

where Q is a penalty factor which penalizes interior–interior overlaps.

It should be noted that the subscript ranges given explicitly in the above expressions ensure that only linearly independent polynomial powers are enumerated. However, for notational clarity, such ranges will be suppressed in subsequent expressions on the understanding that all terms will vanish unless $|m| \leq l < n \leq N$. As it is necessary to use expansions to at least $N = 25$ (or $L = 24$) to encode sufficient detail for docking, and because the summation in Equation (3.8) to order N involves $\frac{N(N+1)(2N+1)}{6}$ coefficients, the shape of each protein is therefore encoded by a pair of coefficient vectors of at least 5525 elements.

The radial basis functions used here for protein shape representations are orthonormal Gauss–Laguerre (GL) functions, defined as:^{111,112}

$$R_{nl}(r) = \left[\frac{2(n-l-1)!}{\lambda^{\frac{3}{2}} \Gamma(n+\frac{1}{2})} \right]^{\frac{1}{2}} e^{-\frac{\rho^2}{2}} \rho^l L_{n-l-1}^{(l+\frac{1}{2})}(\rho^2), \quad (3.11)$$

where $L_k^z(x)$ are generalized Laguerre polynomials,¹¹² ρ is a scaled distance, $\rho = \frac{r}{\lambda^{\frac{1}{2}}}$, and λ is a radial scale factor. These functions are orthonormal in the sense that

$$\int_0^\infty R_{nl}(r) R_{n'l'}(r) r^2 dr = \delta_{nn'}. \quad (3.12)$$

As these functions have a Gaussian pre-factor products of the GL and SH basis functions are often called Gaussian-type orbitals (GTOs) in the quantum

chemistry literature. The Gaussian factor ensures that these functions rapidly tend to zero at large radial distances. For typical protein domains consisting of around 100 to 150 residues, setting $\lambda = 20$ causes most of the zeros of the basis functions to appear within about $30 \text{ \AA}$ from the origin. It is also possible to construct orthonormal radial functions which decay only exponentially at large distances. These more diffuse exponential-type (ETO) functions are more suitable for representing long-range functions such as the electrostatic potential, for example. Figure 3.3 shows some real SPF shape density representations of the HyHel-5 antibody and hen egg lysozyme proteins (PDB code 3HFL) using GTO basis functions at various polynomial expansion orders.

The complex SHs, $Y_{lm}(\theta, \phi)$, are given by:

$$Y_{lm}(\theta, \phi) = \left[\frac{(2l+1)(l-m)!}{4\pi(l+m)!} \right]^{\frac{1}{2}} P_{lm}(\cos \theta) e^{im\phi} \quad (3.13)$$

where $P_{lm}(\cos \theta)$ are the Legendre polynomials. These functions are orthonormal in the sense that:

$$\int_0^{2\pi} \int_0^\pi Y_{lm}(\theta, \phi) Y_{l'm'}(\theta, \phi) \sin \theta d\theta d\phi = \delta_{ll'} \delta_{mm'}. \quad (3.14)$$

The real SHs, $y_{lm}(\theta, \phi)$, can be formed by taking linear combinations of pairs of complex functions having the same absolute value of the azimuthal index, m . This can be written as

$$y_{lm}(\theta, \phi) = \sum_{m'} U_{mm'}^{(l)} Y_{lm'}(\theta, \phi), \quad (3.15)$$

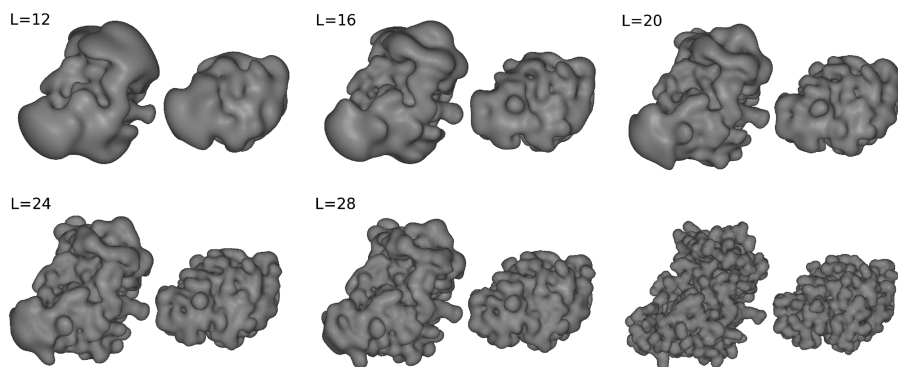


Figure 3.3 SPF shape density isosurfaces at various GTO expansion orders for the complex between the HyHel-5 antibody Fv domain (left) and hen egg lysozyme (right). Here, the two proteins are separated by $15 \text{ \AA}$ for clarity. The bottom right pair shows atomic Gaussian representations of the van der Waals surfaces from which the SPF expansions are derived. This figure is reproduced from Ritchie *et al.*¹⁴⁴ by permission of Oxford University Press.

where $U^{(l)}$ is a unitary $(2l+1) \times (2l+1)$ matrix. For a given value of the latitudinal index, l , this equation may be written as an explicit matrix equation of the form

$$\begin{pmatrix} y_{ll} \\ y_{lm} \\ y_{l0} \\ y_{l\bar{m}} \\ y_{l\bar{l}} \end{pmatrix} = \begin{pmatrix} 1/\sqrt{2} & 0 & 0 & 0 & (-1)^l/\sqrt{2} \\ 0 & 1/\sqrt{2} & 0 & (-1)^m/\sqrt{2} & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & i(-1)^m/\sqrt{2} & 0 & -i/\sqrt{2} & 0 \\ i(-1)^l/\sqrt{2} & 0 & 0 & 0 & -i/\sqrt{2} \end{pmatrix} \begin{pmatrix} Y_{ll} \\ Y_{lm} \\ Y_{l0} \\ Y_{l\bar{m}} \\ Y_{l\bar{l}} \end{pmatrix}. \quad (3.16)$$

The unitary matrices are useful here because they allow real scalar quantities to be expressed as complex polynomial expansions. For example, Equation (3.8) may be written in complex form as

$$\sigma_A(\mathbf{r}) = \sum_{nlm} A_{nlm}^\sigma R_{nl}(r) Y_{lm}(\theta, \phi), \quad (3.17)$$

where the complex coefficients, A_{nlm}^σ , are related to the real expansion coefficients by

$$A_{nlm}^\sigma = \sum_{m'} U_{m'm}^{(l)} a_{nlm'}^\sigma. \quad (3.18)$$

3.9 Spherical Polar Fourier Correlations

By substituting $\sigma'_B(\mathbf{r}) = \sigma_B(\mathbf{r}) - Q\tau_B(\mathbf{r})$ into Equation (3.10), the shape complementarity expression may be written as a two-term overlap integral:

$$C = \int (\sigma_A(\mathbf{r})\tau_B(\mathbf{r}) + \sigma'_B(\mathbf{r})\tau_A(\mathbf{r}))d\mathbf{r}. \quad (3.19)$$

We normally put $Q = 11$. Furthermore, by taking complex linear combinations of expansion coefficients,

$$A_{nlm} = \sum_{m'} (a_{nlm'}^\tau + ia_{nlm'}^\sigma) U_{m'm}^{(l)} \quad (3.20)$$

and

$$B_{nlm} = \sum_{m'} (b_{nlm'}^{\sigma'} + ib_{nlm'}^\tau) U_{m'm}^{(l)}, \quad (3.21)$$

and thanks to the orthogonality of the basis functions, Equation (3.19) may be calculated very easily as a scalar product:

$$C = \text{Re} \left(\sum_{nlm} A_{nlm}^* B_{nlm} \right). \quad (3.22)$$

Although it is relatively expensive to calculate the shape complementarity or the interaction energy of a single orientation in this way, the main benefit of the

SPF approach is that it allows multiple rotational orientations to be evaluated very rapidly using FFT techniques. For example, it can be shown that the SPF coefficients transform amongst themselves under rotation according to:

$$A_{nlm}(\alpha, \beta, \gamma) = \sum_{m'} D_{mm'}^{(l)}(\alpha, \beta, \gamma) A_{nlm'}, \quad (3.23)$$

where the $D_{mm'}^{(l)}(\alpha, \beta, \gamma)$ are matrix elements of the Wigner rotation matrices.¹¹⁰ In other words, the effect of rotating a protein may be simulated by transforming only its expansion coefficients according to Equation (3.23). The Wigner rotation matrices may be defined as

$$D_{mm'}^{(l)}(\alpha, \beta, \gamma) = e^{-im\alpha} d_{mm'}^l(\beta) e^{-im'\gamma}, \quad (3.24)$$

where the real $d_{mm'}^l(\beta)$ are often expressed in terms of Jacobi polynomials.¹¹⁰ In order to calculate rotational FFTs, it is convenient to expand the $d_{mm'}^l(\beta)$ as a product of complex exponentials¹¹³ to obtain

$$D_{mm'}^{(l)}(\alpha, \beta, \gamma) = \sum_t i^{m-m'} \Delta_{lm}^l \Delta_{lm'}^l e^{-im\alpha} e^{-it\beta} e^{-im'\gamma}, \quad (3.25)$$

where

$$\Delta_{lm}^l = d_{lm}^l\left(\frac{\pi}{2}\right). \quad (3.26)$$

However, because the β rotation angle ranges from 0 to 2π , an additional scaling factor must be applied to map this angle onto the natural range of the FFT, as described previously.¹¹⁴

It can also be shown that the effect of translating a protein by a distance R along the z axis may be simulated by transforming its expansion coefficients according to:

$$A_{nlm}(R) = \sum_{kj} T_{nl,kj}^{(|m|)}(R) A_{kjm}, \quad (3.27)$$

where the $T_{nl,kj}^{(|m|)}(R)$ are the matrix elements for translations of the GTO basis functions. The translation matrix elements for both the GTO and ETO radial bases may be calculated analytically using Fourier–Bessel transform theory.¹¹¹ However, due to the many high order factorials that appear in the analytic expressions it is necessary to calculate high order translation matrix elements using the ‘GMP’ multi-precision math library.¹⁴⁵ Nonetheless, because the translation matrices are completely independent of the proteins to be docked, they can be calculated just once for a range of translations, and stored on disc for future use. Using pre-calculated translation matrix elements greatly speeds up subsequent docking calculations.

Given that the operations of rotating and translating SPF expansions may be implemented by applying the corresponding matrix operations to the expansion

coefficients, our strategy for calculating docking correlations is to calculate lists of rotated and translated coefficient vectors for the receptor and ligand proteins, and to evaluate Equation (3.22) for all possible pairs of such vectors. For example, by casting the coefficient vectors in complex form,¹¹⁴ and by letting $A_{nlm}(R, \beta_A, \gamma_A)$ and $B_{nlm}(\alpha_B, \beta_B, \gamma_B)$ represent rotated and translated complex coefficients, the shape complementarity score as a function of the remaining twist angle degree of freedom, α_B , may be calculated as

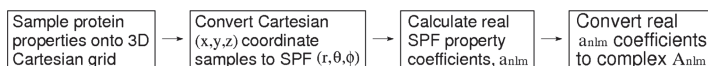
$$C(\alpha_B) = \sum_m S_m(R, \beta_A, \gamma_A, \beta_B, \gamma_B) e^{-im\alpha_B}, \quad (3.28)$$

where the intermediate S_m coefficients are given by

$$S_m(R, \beta_A, \gamma_A, \beta_B, \gamma_B) = \sum_{nl} A_{nlm}^*(R, \beta_A, \gamma_A) B_{nlm}(\alpha_B, \beta_B, \gamma_B). \quad (3.29)$$

As Equation (3.28) has the form of a one-dimensional (1D) Fourier series, the calculation over multiple rotational samples for α_B may be performed efficiently using a FFT. Using a FFT grid dimension of 64 gives angular increments of $\frac{360^\circ}{64} = 5.625^\circ$. This gives a good compromise between computational speed and angular resolution. For the remaining rotation angles, near-regular patterns of (β, γ) angles are generated from icosahedral tessellations of the sphere.¹⁰⁹ For example, an icosahedral tessellation of 812 vertices gives angular samples with an average separation of about 7.5° . In order to complete a docking search, Equation (3.28) is evaluated over multiple pairs (812×812) of (β, γ) molecular rotations, and the entire calculation is repeated over a range of intermolecular separations, R . In *Hex*, the default is to use 50 translational steps of $0.8 \text{ \AA}$. This generates in the order of two billion (2×10^9) trial docking orientations. The overall approach is summarized in Figure 3.4.

A Sampling protein properties



B Docking pairs of proteins

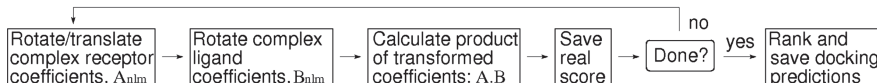


Figure 3.4 Conceptual flowcharts of the main processing steps in the SPF protein docking algorithm. In practice, rotated lists of coefficients for the receptor and ligand calculated just once and saved for subsequent use in a series of 1D FFTs, or the ligand rotations are calculated *en masse* using 3D rotational FFTs. This figure is reproduced from Ritchie *et al.*¹⁴⁴ by permission of Oxford University Press.

3.10 Multi-dimensional Spherical Polar FFTs

By using Equation (3.25) to expose different Euler rotational angles as complex exponentials, it is possible to express the overall calculation as a list of 3D or even 5D FFTs.¹¹⁴ For example, a 3D rotational FFT can be calculated using

$$C(\alpha_B, \beta_B, \gamma_B) = \sum_{muv} e^{-i(m\alpha_B + 2u\beta_B + v\gamma_B)} S_{muv}(R, \beta_A, \gamma_A), \quad (3.30)$$

where the S_m coefficients are now given by

$$S_{muv}(R, \beta_A, \gamma_A) = \sum_{nl} A_{nlm}^*(R, \beta_A, \gamma_A) B_{nlv} A_{lv}^{um}, \quad (3.31)$$

and where A_{lv}^{um} is a rotational scaling factor.¹¹⁴ Equation (3.30) is normally evaluated using a 3D FFT grid of $48 \times 24 \times 48$ elements, which gives rotational steps of 7.5° for each of the three ligand rotation angles ($\alpha_B, \beta_B, \gamma_B$). In each case, outer iterations over the remaining degrees of freedom must be performed to cover the 6D search space. 5D rotational FFTs may be calculated in a similar way. Each 3D FFT correlation corresponds to ‘spinning’ the ligand through an array of ($\alpha_B, \beta_B, \gamma_B$) rotation angles with respect to each fixed receptor orientation, whereas the 5D FFT calculation corresponds to spinning both molecules simultaneously around all five ($\beta_A, \gamma_A, \alpha_B, \beta_B, \gamma_B$) rotation angles. Although it might be expected that 5D FFTs should be faster to calculate than 3D and 1D FFTs, this is often not the case in practice due to the very large memory requirements of the 5D FFT grid.

As an SPF calculation may be expressed as a series of matrix multiplications and FFTs, it is straightforward to use standard thread programming techniques to spread the calculation over as many processors as are available on a compute node. Hence, on machines with one or two quad-core central processor units (CPUs), for example, the calculation may be spread over four or eight cores, respectively. On Linux systems, the POSIX ‘pthreads’ library was used to implement a simple ‘producer/consumer’ thread model in which multiple producer threads perform the calculation, and a single consumer thread collects the results. Synchronization between threads is achieved using mutexes. On Windows machines, synchronization is achieved in a similar way using semaphores. On a quad-core machine, this gives almost a four-fold speed-up compared to using a single core. However, only about a six-fold speed-up is achieved on dual quad-core systems. Hence, spreading a docking calculation over more CPUs in a multi-node compute cluster would be expected to give only a relatively small additional speed-up.

3.11 Exploiting Graphics Processor Units

In recent years, many scientific calculations have benefited from the very high arithmetic capabilities of modern GPUs.¹¹⁵ Initially, it required considerable skill and knowledge of graphics programming techniques to transform a

scientific calculation into a form that could be executed by dedicated pixel processing hardware on a GPU. However, with the advent of programmable GPUs and software development tools such as Brook¹¹⁶ and the CUDA (Common Unified Device Architecture) toolkit,¹⁴⁶ it is now much easier to deploy scientific software on GPUs. For example, using GPUs to calculate protein and DNA sequence alignments can give speed-ups from at least a factor of 10 (ref. 117 and 118) to over 100 (ref. 119) compared to the same calculation on conventional CPUs. Similarly, GPUs can give speed-ups from around 10 to 100 for MD simulations,^{120,121} and up to a factor of 130 for quantum chemistry calculations.¹²² GPUs have also been used to accelerate the calculation of protein accessible surface areas, for example, giving speed-up factors from around 100 to over 300, depending on the size of the protein.¹²³ Encouraged by these earlier successes, we recently implemented SPF docking correlations on Nvidia GPUs using the CUDA development libraries.

Nvidia GPUs contain a number of programmable multi-processors (MPs) each of which comprises multiple scalar processors (SPs), or ‘cores’, and a modest amount of very fast on-chip memory (‘shared memory’) that may be accessed simultaneously by the SPs. Each GPU also has a large amount of relatively slow off-chip ‘global’ memory (from a few hundred megabytes to several gigabytes) which is accessible by each SP. For example, a contemporary GeForce GTX 285 device has 30 MPs, 240 SPs, 16 Kb of shared memory per MP, 1 Gb of global memory and a clock speed of 1.48 GHz. Given that each SP can calculate at least one single precision floating point arithmetic operation (or ‘flop’) per clock cycle, this device is capable of at least 355 Gflops. However, in contrast to conventional CPUs that often have several megabytes of fast cache memory, a considerable draw-back of current GPUs is that they do not have any cache memory at all. This is significant because access to global memory is about 80 times slower than accessing shared memory. Hence it is important to ensure that data traffic between the MPs and global memory is kept to a minimum, and that as much work as possible is done on data in fast shared memory.

The Nvidia device architecture promotes a very fine-grained ‘SIMT’ (Simultaneous Instructions Multiple Threads) programming model in which individual threads of execution are responsible for manipulating a small number of closely related data elements. This model is well suited for performing simple operations with high arithmetic intensity, such as those found in matrix multiplications and FFTs. For such calculations, it is natural to let one SIMT thread be responsible for calculating one element of an array, for example. In the CUDA SIMT model, threads are given numerical index identifiers and are grouped into ‘blocks’ with consecutive indices. Blocks of threads may further be grouped into ‘grids’ of thread blocks. Thus, threads may be indexed using 1D, 2D or 3D indexing schemes, respectively. CUDA devices schedule and execute blocks of threads by dividing them into ‘warps’ of 32 threads. Each warp executes one instruction at a time, so maximum efficiency is achieved when all of the threads of a warp execute the same sequence of instructions. Similarly, access to global memory is most efficient when the data is aligned in memory on even word boundaries, and when all threads in a half-warp access consecutive

memory elements simultaneously, for example. When this occurs, the MPs can coalesce multiple memory accesses into a single transaction, and they can hide the latency of global memory by running several warps concurrently.

In order to exploit the above hardware characteristics, SPF docking calculations have been implemented using a small number of CUDA ‘kernel’ functions. These functions are written using a similar syntax to the C programming language, but they are executed in parallel by the MPs. Each kernel performs a basic operation such as matrix–matrix multiplication, for example. To exploit the limited amount of shared memory and to optimize access to global memory, each matrix is divided into tiles of 16×16 data elements, and pairs of tiles are multiplied using blocks of 256 threads. Together with the use of coalesced memory transactions, this basic technique provides sufficient arithmetic intensity to hide the latency of global memory. Since a 6D docking search mainly entails rotating, translating and cross-multiplying multiple receptor and ligand coefficient vectors, it is natural to express the calculation on the GPU as a sequence of simple matrix operations. The main strategy to achieve this in the SPF approach presented above is to map all pairs of n_l subscripts to single index, p , and to round up the total number of elements, P , to a multiple of 16. This scatters the elements of the of a compact coefficient vector into a sparse 2D array of dimension $P \times (2L + 1)$ elements, as illustrated (see the upper part of Figure 3.5). Re-indexing M rotated coefficient vectors in this way gives a 3D block of $P \times M \times (2L + 1)$ coefficients in sparse format. Similarly, the SPF translation matrix elements are re-indexed to give a sparse 3D block of $P \times P \times (L + 1)$ elements. This allows all M coefficient vectors to be translated together by performing $2L + 1$ tiled matrix–matrix multiplications, as illustrated (see the lower part of Figure 3.5).

In order to complete the 1D FFT docking scheme, two further GPU kernels were implemented. The first of these cross-multiplies and zero-pads pairs of receptor and ligand coefficient vectors according to Equation (3.29). This gives a list of $N \times M$ 64-element twist angle coefficient vectors which can be evaluated as a batch of 1D FFTs using Nvidia’s cuFFT library. The output from the FFT is an array of docking scores, or pseudo-energies, as a function of the intermolecular twist angle α_B . This array is scanned by the second kernel to identify the lowest energy for each twist angle, and the resulting list of poses and energies is copied back to main CPU memory. These steps are illustrated (see Figure 3.6). The above sequence of operations is repeated for each translational step of the 6D search.

The 3D FFT scheme generally follows the same execution path as the 1D FFT scheme, but only a single un-rotated ligand coefficient vector has to be copied to the GPU to calculate the array of 3D FFT coefficients (Equation 3.31). This array is then passed to the 3D cuFFT function, and the best poses and energies are copied back to CPU main memory, as before.

Figure 3.7 shows the overall docking correlation rates at different polynomial expansion orders for both the 1D and 3D FFT calculation schemes described above. This shows that a GPU can calculate the 1D FFT docking scheme significantly faster than the same calculation on the CPU, and indeed also

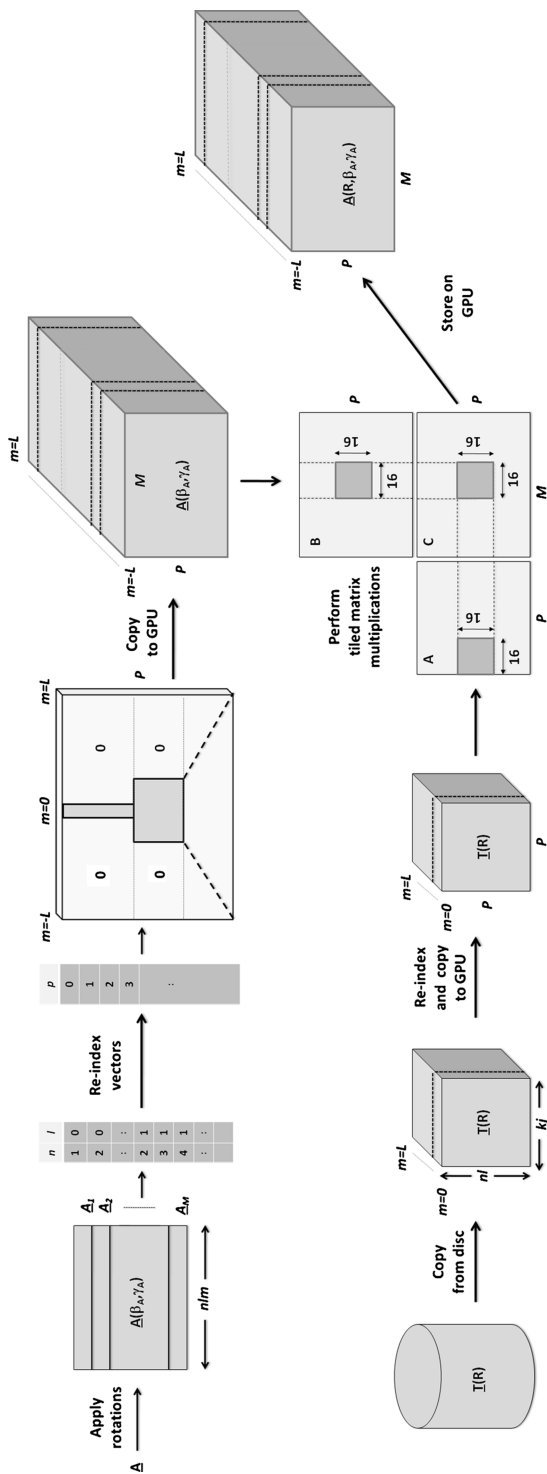


Figure 3.5 Re-ordering the expansion coefficient vectors into sparse arrays suitable for tiled matrix multiplication on the GPU. Here, M rotated receptor coefficient vectors are re-indexed to give a $2L + 1$ planes of dimension $M \times P$. Similarly, the translation matrix elements are re-indexed to give $L + 1$ planes of dimension $P \times P$. This allows multiple coefficient vectors to be translated efficiently using tile-based matrix multiplications. This figure is reproduced from Ritchie and Venkatraman⁹⁵ by permission of Oxford University Press.

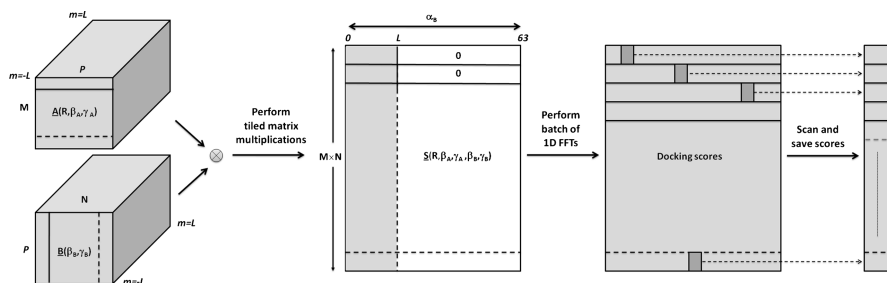


Figure 3.6 Calculating multiple 1D FFTs on the GPU for multiple pairs of transformed receptor and ligand coefficient vectors. For each translation step R , M translated and rotated receptor coefficient vectors and N rotated ligand coefficient vectors are combined using Equation 3.29 to give $M \times N$ vectors of twist angle coefficients (Equation 3.28). The $S_m(R, \beta_A, \gamma_A, \beta_B, \gamma_B)$ coefficients are then transformed into an array of docking scores using a batch of 1D FFTs (Equation 3.28), and the result is scanned to find the best score for each twist angle, α_B . This figure is reproduced from Ritchie and Venkatraman⁹⁵ by permission of Oxford University Press.

significantly faster than the 3D FFT on both the GPU and CPU. For example, the 1D FFT scheme can evaluate 236 million poses per second using $N=18$ on the GPU, and 104 million poses per second using $N=25$. Compared to using a single CPU core, these rates correspond to speed-up factors of 100 and 130, respectively. As might be expected, 3D FFT calculation rates are less sensitive to the polynomial order than the 1D FFTs because they require less explicit matrix arithmetic and they can benefit more from the $O(N \ln N)$ nature of the FFT. On the other hand, Figure 3.7 also shows that compared to the CPU, a GPU does not give any significant speed-up for 3D FFT calculations. This is because 3D FFT algorithms require multiple passes through the data volume, and on the GPU this exposes the latency of repeatedly accessing global memory without the benefit of fast cache memory. On the other hand, because multiple 1D FFTs can be processed in a single pass over global GPU memory, it follows that the SPF 1D FFT scheme is especially well suited to exploit current GPU architectures. As *Hex* performs an initial scan of the search space using $N=18$, and because only a small fraction of the remaining poses are re-scored using $N=25$, the overall benefit of using a GPU to calculate the 1D FFT docking scheme is dramatic. Using this two-stage approach, a global rigid body-docking search over all six degrees of freedom can be completed in around 15 seconds on a modern GPU. This is, at least, two orders of magnitude faster than conventional Cartesian grid-based FFT docking approaches.⁹⁵

It should be noted that not all stages of the SPF docking calculation have been implemented on the GPU. For example, the translation matrix elements must be calculated using the GMP math library, and this library is not available for GPUs. On the other hand, because the overall calculations have been structured to use coarse-grained CPU threads, it is straightforward to use similar threading techniques to control multiple GPUs simultaneously.

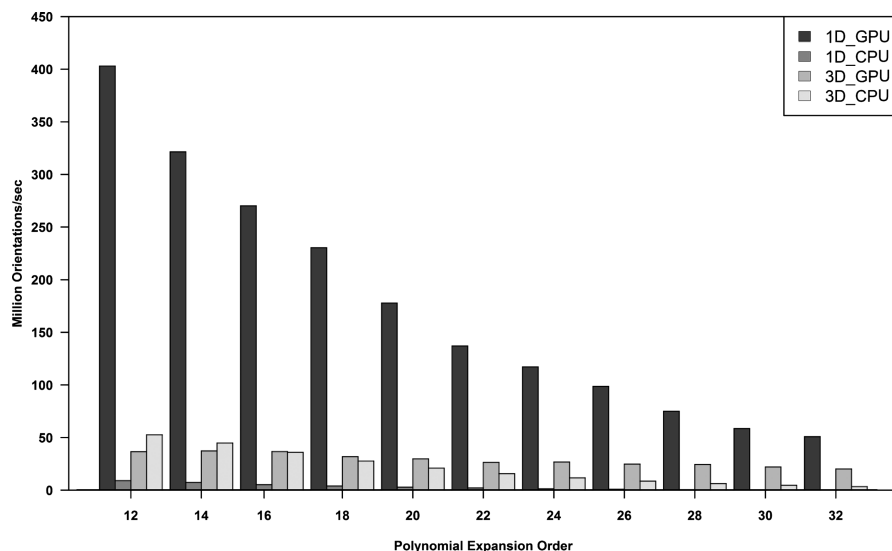


Figure 3.7 Docking correlation rates (millions of poses per second) using the 1D and 3D correlation schemes at various polynomial expansion orders. The GPU rates include all data transfers between the GPU and CPU, the time required to re-index and translate the coefficients, and the time spent in the cuFFT library. All rates exclude the costs of calculating the initial forward transform and reading the translation matrices from disc. This figure is reproduced from Ritchie and Venkatraman⁹⁵ by permission of Oxford University Press.

To date, the above calculations have been tested successfully using up to eight Intel CPU cores and two Nvidia GPUs simultaneously.

3.12 The ‘Hex’ Docking Program

The program *Hex* was developed to explore the use of the SPF approach for accelerating the first rigid body part of a protein docking calculation. *Hex* is an interactive program that lets a user load a pair of proteins, view them in a variety of 3D graphical modes, perform docking calculations, and visualize the results in a single easy-to-use environment. *Hex* runs equally on Linux, Windows and Mac OS X systems, and it can use as many CPUs and GPUs as are available for its docking calculations. *Hex* may be downloaded at no cost for academic use.¹⁴⁷ As high performance GPUs are still relatively expensive, a GPU-powered version of *Hex* has also been made available as a public web service called ‘HexServer.’¹²⁴ This has been implemented using a simple form-based interface through which users may up-load a pair of protein structures to be docked by *Hex*, and from which the user may download a set of predicted binding modes.¹⁴⁸

In order to evaluate *Hex* exhaustively, the SPF approach was tested on the complexes of version 2 of the Protein Docking Benchmark.⁵¹ To provide a consistent pseudo-random starting orientation, all proteins were initially fitted onto the coordinates of the corresponding native complex, and a small off-grid rotation was then applied to the ligand. The orientations calculated in each docking run were clustered using a greedy algorithm with a 9 Å clustering threshold,¹²⁵ and the lowest energy member of each cluster was selected as the ‘solution’ for that cluster. All other members of each cluster were discarded. Seven different docking runs were performed for each complex to assess the shape-based and electrostatic components of the scoring function, and to investigate the difference between blind docking and the use of prior knowledge of one or both protein’s binding sites. The results obtained for the first 50 benchmark complexes are shown (see Table 3.1), which lists the target complexes in approximate order of difficulty. The first set of figures in this table gives the results for blind shape-only docking of bound subunits, presented as the rank and deviation of the first solution found within 10 Å RMS of the complex (here called a ‘hit’), along with the total number of such hits found within the top 2000 solutions. This threshold broadly corresponds to the definition of an ‘acceptable’ prediction under the CAPRI assessment criteria.⁶⁶ Encouragingly, acceptable solutions are found within the top 10 in many cases, of which several are in fact ranked first out of the billions of orientations evaluated. This shows that the *Hex* shape-based scoring function can often identify near-native crystallographic orientations.

The rest of Table 3.1 presents results for docking unbound structures. As might be expected, the rank of the best shape-only solution is often considerably worse than that obtained when docking bound components. On the other hand, including the ETO electrostatic interaction term in the correlation often improves the rank of the best solution, placing several complexes within the top 20, for example. However, using electrostatic correlations can also worsen the prediction in some cases, but it is not clear how to predict *ab initio* which those cases might be.

In practice, it is becoming increasingly rare that completely blind docking is necessary because biochemical or biophysical knowledge is often available to indicate the identities of key interaction residues. In general, *Hex* allows a given receptor and ligand residue to be rotated onto the *z* axis before each docking run, and by setting small values for the β_A and β_B angular ranges, it is then straightforward to focus a docking calculation around a given pair of residues in a known or hypothesized interface. Hence, four further constrained docking runs were performed for each complex to simulate such data-driven docking scenarios. In these runs, the range of the FFT searches were constrained by applying the restriction $\beta_A \leq 45^\circ$ to simulate using knowledge of the receptor binding site (tabulated as ‘One Constraint’), and additionally by setting $\beta_B \leq 45^\circ$ to simulate using knowledge of both the receptor and ligand binding sites (‘Two Constraints’). Each such constraint reduces the size of the search space by a factor of about four, and speeds up the calculation by a corresponding amount. As can be seen from the final columns of Table 3.1, providing two

Table 3.1 Hex docking results for the first 50 “rigid body” complexes of the Protein Docking Benchmark (version 2).

Code	B-B Shape-Only Blind Search		U-U Shape-Only Blind Search		U-U Shape + Elec Blind Search		U-U Shape-Only One Constraint		U-U Shape + Elec One Constraint		U-U Shape-Only Two Constraints		U-U Shape + Elec Two Constraints	
	Rank (RMS)	Hits	Rank (RMS)	Hits	Rank (RMS)	Hits	Rank (RMS)	Hits	Rank (RMS)	Hits	Rank (RMS)	Hits	Rank (RMS)	Hits
1AVX	46 (4.8)	20	108 (8.9)	7	111 (8.9)	4	40 (8.9)	12	75 (9.0)	14	18 (9.0)	43	12 (9.0)	45
1AY7	40 (8.9)	16	645 (9.9)	4	—	—	99 (3.5)	20	234 (9.8)	1	17 (6.7)	39	17 (9.7)	18
1BVN	1 (1.1)	29	63 (9.1)	20	389 (9.6)	7	29 (9.6)	35	3 (6.6)	36	4 (5.1)	49	2 (9.6)	39
1CGI	1 (0.7)	24	42 (9.4)	17	47 (4.6)	9	20 (9.4)	14	42 (9.8)	11	4 (9.4)	31	4 (4.6)	24
1D6R	273 (1.3)	24	447 (7.7)	1	119 (7.6)	4	49 (7.7)	8	31 (7.7)	8	8 (7.7)	37	5 (7.7)	31
1DFJ	167 (4.2)	14	17 (9.5)	14	1 (4.2)	30	3 (9.5)	24	1 (4.2)	30	2 (9.5)	32	1 (4.2)	35
1E6E	1 (2.1)	14	109 (5.6)	10	5 (2.2)	24	24 (5.6)	19	3 (1.5)	29	5 (5.6)	38	1 (7.7)	49
1EAW	1 (1.0)	17	9 (5.0)	20	1 (4.0)	37	7 (5.0)	25	1 (4.0)	35	1 (5.0)	42	1 (4.0)	42
1EWY	19 (7.7)	16	76 (9.1)	12	24 (9.7)	14	114 (8.1)	12	103 (6.8)	7	9 (8.1)	37	9 (7.6)	23
1EZU	2 (0.9)	13	—	—	—	—	—	—	—	—	86 (6.7)	10	287 (6.2)	4
1F34	1 (1.4)	25	124 (6.7)	11	—	—	48 (7.1)	15	—	—	11 (5.4)	22	26 (6.5)	11
1HIA	3 (1.2)	30	51 (8.7)	6	8 (8.9)	15	72 (8.7)	21	15 (9.9)	22	15 (6.7)	33	6 (8.3)	32
1MAH	1 (0.9)	16	2 (1.2)	20	1 (1.1)	28	1 (1.2)	27	1 (1.2)	30	1 (1.2)	33	1 (1.2)	30
1PPE	1 (1.0)	42	2 (9.7)	47	4 (3.0)	31	1 (9.7)	49	1 (3.0)	46	1 (3.0)	43	1 (3.0)	45
1TMQ	1 (2.1)	19	356 (5.9)	9	427 (6.0)	6	45 (5.9)	21	264 (2.3)	7	7 (5.9)	39	10 (6.6)	38
1UDI	1 (1.6)	17	8 (6.2)	9	20 (6.2)	10	4 (6.2)	22	7 (6.2)	25	1 (6.2)	32	5 (6.2)	37
2MTA	11 (1.4)	18	136 (9.0)	4	79 (9.8)	20	38 (9.0)	17	12 (8.4)	24	15 (7.7)	33	15 (8.7)	31
2PCC	1007 (9.1)	1	—	—	18 (6.9)	33	14 (9.3)	20	12 (5.1)	31	5 (9.3)	37	14 (6.3)	44
2SIC	3 (0.7)	10	57 (8.8)	8	—	—	21 (8.9)	10	44 (1.0)	9	4 (8.9)	31	4 (1.0)	35
2SNI	1 (1.5)	18	256 (9.6)	7	101 (9.6)	6	39 (7.1)	15	40 (4.4)	11	5 (7.1)	31	5 (4.4)	25
7CEI	5 (1.3)	17	61 (8.7)	5	4 (8.4)	19	11 (8.7)	17	3 (8.4)	22	2 (8.7)	29	1 (8.4)	35
1AHW	6 (1.9)	10	234 (8.0)	3	7 (8.0)	12	31 (8.0)	12	5 (8.0)	40	3 (8.0)	42	5 (8.0)	38
1BVK	44 (1.5)	6	—	—	508 (6.7)	7	134 (9.4)	7	184 (6.8)	10	71 (9.9)	23	22 (6.8)	24
1DQJ	—	—	—	—	—	—	216 (8.6)	6	440 (9.9)	2	22 (8.6)	24	73 (8.1)	11
1E6J	—	—	—	—	—	—	26 (8.9)	12	16 (8.4)	22	2 (8.9)	37	4 (8.4)	41
1JPS	24 (1.3)	5	—	—	36 (8.8)	11	170 (6.6)	9	14 (6.6)	27	15 (6.6)	29	1 (8.8)	30
1MLC	62 (1.2)	5	408 (3.6)	2	—	—	25 (3.6)	13	22 (3.7)	28	3 (3.6)	29	2 (3.7)	23
1VFB	23 (1.1)	3	—	—	—	—	97 (9.1)	14	51 (7.1)	10	14 (9.1)	36	12 (7.1)	35

Table 3.1 (Continued)

Code	B-B Shape-Only		U-U Shape-Only		U-U Shape + Elec		U-U Shape-Only		U-U Shape + Elec		U-U Shape-Only		U-U Shape + Elec	
	Rank (RMS)	Hits	Blind Search	Hits	Rank (RMS)	Hits	Rank (RMS)	Hits	One Constraint	Rank (RMS)	Hits	Rank (RMS)	Hits	Two Constraints
1WEJ	—	—	—	—	—	—	26 (1.7)	13	—	2 (1.7)	20	8 (1.7)	29	1 (1.7)
2VIS	—	—	—	—	—	—	—	—	—	—	—	—	—	—
1A2K	29 (5.4)	12	—	—	—	—	—	—	—	—	—	186 (9.3)	5	274 (9.1)
1AK4	—	—	—	—	—	—	—	—	—	—	—	—	—	—
1AKJ	30 (8.4)	25	209 (9.6)	10	17 (9.4)	27	110 (6.3)	15	23 (2.7)	35	35	23 (9.6)	36	5 (9.6)
1B6C	3 (1.8)	19	593 (9.0)	2	755 (8.9)	2	88 (9.0)	5	133 (8.5)	5	5	19 (9.0)	27	7 (9.7)
1BUH	28 (1.0)	9	743 (7.7)	2	289 (7.8)	4	52 (7.7)	14	19 (7.7)	13	13	28 (7.7)	19	8 (7.7)
1E96	133 (1.1)	5	—	—	302 (8.6)	2	246 (9.4)	6	119 (8.6)	8	8	37 (9.7)	13	43 (8.5)
1F51	3 (1.4)	21	371 (9.6)	5	—	—	149 (9.6)	12	58 (9.3)	3	3	9 (7.6)	19	8 (7.5)
1FC2	605 (6.5)	2	—	—	—	—	—	—	—	—	—	5 (7.8)	31	297 (7.7)
1FQJ	7 (1.0)	14	41 (8.0)	12	7 (7.9)	14	14 (8.0)	21	7 (7.7)	28	28	—	6	4 (7.7)
1GCO	1 (1.0)	16	—	—	—	—	—	—	—	—	—	92 (6.2)	6	—
1GHQ	—	—	—	—	—	—	828 (8.9)	2	—	—	—	30 (8.9)	13	175 (6.7)
1HE1	1 (1.5)	24	37 (6.4)	18	88 (6.3)	15	10 (6.4)	26	28 (7.2)	25	25	2 (7.6)	39	9 (7.2)
1I4D	31 (1.5)	19	—	—	—	—	—	—	—	—	—	505 (8.1)	1	481 (9.4)
1KAC	36 (1.2)	7	687 (8.7)	1	271 (8.9)	5	7 (4.4)	19	4 (4.4)	26	26	4 (4.4)	33	2 (4.4)
1KLU	—	—	—	—	—	—	—	—	—	—	—	591 (9.7)	2	—
1KTZ	—	—	—	—	—	—	—	—	—	—	—	238 (9.4)	4	25 (6.0)
1KXP	1 (1.1)	22	36 (9.4)	13	1 (7.5)	13	15 (9.4)	19	1 (6.9)	30	30	7 (9.4)	24	1 (6.9)
1ML0	—	—	—	—	—	—	7 (9.1)	8	33 (7.0)	11	11	1 (9.1)	22	3 (5.6)
1QA9	86 (5.9)	7	—	—	161 (9.9)	3	587 (7.5)	8	481 (6.8)	4	4	25 (5.3)	28	23 (4.5)
1RLB	409 (8.8)	2	—	—	—	—	—	—	—	—	—	305 (6.3)	7	384 (6.3)

For each of seven docking runs with different initial conditions, this table lists the rank, RMS deviation, and total number of “hits” found for each case. Here, any prediction for which the ligand C α coordinates are within 10 Å RMS of the crystallographic complex structure is considered as a “hit.” Here, Code refers to the PDB code of the crystallographic complex structure; Rank is the rank of the first hit found within 10 Å RMS of the crystallographic complex; Hits is the total number of hits found within the top 2,000 predictions. A hyphen denotes no acceptable solution within the top 2,000 predictions. B-B and U-U indicate whether the starting structures were taken from the crystallographic complex (bound-bound) or were crystallised separately (unbound-unbound). For the antibody/antigen complexes (1AHW, 1BVK, 1DQJ, 1E6J, 1DQJ, 1JPS, 1MLC, 1VFB, 1WEJ, and 2VIS), the C α coordinates of heavy chain residue 37 were used as the the antibody coordinate origin. For all other structures, the centre of mass was used as the coordinate origin. This table is reproduced from Ritchie *et al.* 114 by permission of Oxford University Press.

loose constraints is often sufficient to improve considerably the rank of near-native solutions. With just one constraint, adding the electrostatic contribution to the scoring function helps considerably to identify the general orientation of the binding mode. With two search constraints, the shape-only scoring function is often able to rank a near-native orientation within the first 10 predictions, and adding electrostatics can give a further improvement. Overall, Table 3.1 shows that useful docking predictions can be made under the rigid body assumption if the conformational change on binding is relatively small and if some prior knowledge is available to constrain the search space.

3.13 Conclusions and Future Prospects

The *Hex* SPF rigid body docking algorithm is now sufficiently fast that covering the 6D rigid body search space is no longer a rate-limiting step in protein–protein docking protocols. However, one weakness of the *Hex* algorithm is its relatively simple shape plus electrostatic scoring function. Several groups have developed more sophisticated knowledge-based docking potentials derived from statistical analyses of known protein–protein interfaces. For example, Zhang *et al.*¹⁰⁷ developed an ‘atomic contact potential’ (ACP) using a small number of atom types, and also the more recent DFIRE (distance-scale finite ideal gas reference state)¹²⁶ potential using 19 atom types. Kozakov *et al.*¹²⁷ developed their ‘decoys as reference state’ (DARS) potential by counting the frequency of pair-wise atom occurrences in actual complexes compared to the corresponding frequencies found in a large number of realistic but incorrect decoy complexes. The DARS potential has been very successful in the recent rounds of the CAPRI blind docking experiment.^{128,129} Therefore, a natural development would be to implement SPF correlations of the DFIRE or DARS potentials, for example.

Recent CAPRI results also show that using prior experimental information to focus the docking search or to re-score *ab initio* decoys can significantly improve the quality of the predictions, and has therefore become an integral component of many docking protocols.⁶⁹ Hence, another useful enhancement to *Hex* would be to promote the scores of putative docking orientations that involve known or hypothesized interface residues. This approach is nicely exemplified in the data-driven HADDOCK system¹³⁰ that uses ambiguous interaction restraints (AIRs)¹³¹ to define interaction constraints from a wide range of experimental sources.

Of course, a fundamental weakness of any rigid body docking algorithm is the assumption that the conformational changes that occur on binding will be small. However, even if there are no major changes in backbone conformations, induced fit effects can sometimes involve substantial changes of side chain torsion angles, particularly for flexible residues such as Lys and Arg. However, it is difficult to predict which specific side chains might change conformation in a given complex. Kimura *et al.*⁷⁶ argue that, from a dynamical point of view, there is insufficient time available during a collision encounter for extensive conformational rearrangements to take place. Using short time-scale MD

simulations with explicit solvent, they showed that certain key interface residues generally adopt the same conformation in the unbound and bound structures whereas peripheral interface residues are more likely to adopt a range of rotameric states. In other words, specific residues seem to act as ready-made recognition motifs or ‘anchors’ for docking.⁷⁶ By using short MD simulations of unbound structures, Camacho was able to identify successfully key anchor residues for several CAPRI targets.¹³² However, even relatively short MD simulations are still computationally expensive, and they require considerable expertise to set up and to analyze the results.

Several groups have used principal component analysis (PCA) of MD trajectories to generate protein conformations for docking.^{80,133–137} The result of a PCA is a matrix of eigenvectors and a list of associated eigenvalues that together describe the principal components and amplitudes, respectively, of the internal motions within a protein. Often, most of the internal motions within a protein can be adequately described by the first few eigenvectors.^{138,139} Thus a PCA analysis may be considered as a form of dimensionality reduction. PCA-based approaches have been shown to provide a promising way to generate candidate conformations for docking.^{135,136,140} However, structures generated from PCA eigenvectors can violate standard bond length and torsion angle ranges, and hence need to be energy minimized.¹⁴¹

Instead of starting from an expensive MD calculation, a promising alternative approach is to construct a simple Gaussian network model (GNM) of protein flexibility,¹⁴² and to derive the eigenvectors directly from the associated Hessian matrix.¹⁴³ However, GNM eigenvectors do not fully span the conformational space between the unbound and bound forms,¹⁴⁴ and the structures generated in this way also need to be energy minimized. Hence, it is not yet clear whether PCA-based approaches can offer an efficient alternative to performing explicit MD simulations. Nonetheless, it seems inevitable that incorporating more sophisticated models of protein flexibility into protein docking simulations will require considerable computational resources. The SPF correlation approach presented here suggests that if the calculations required by these new models can be formulated as basic matrix operations, they could be implemented very profitably on modern GPUs.

Acknowledgements

Part of this work was funded by the Agence Nationale de la Recherche, grant number ANR-08-CEXC-017-01. The Nvidia Professor Partner Programme is acknowledged for the gift of a graphics card.

References

1. P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. J. Yang, M. Johnston, S. Fields and J. M. Rothberg, *Nature*, 2000, **403**, 623.

2. T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori and Y. Sakaki, *Proc. Natl. Acad. Sci.*, 2001, **98**, 4569.
3. A. C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. M. Michon and C. M. Cruciat, *Nature*, 2002, **415**, 141.
4. Y. Ho, A. G. A. Heilbut, G. D. Bader, L. Moore, S. L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Sørensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. V. Hogue, D. Figey and M. Tyers, *Nature*, 2002, **415**, 180.
5. P. Aloy and R. B. Russell, *Nature Biotech.*, 2004, **22**, 1317.
6. G. T. Hart, A. K. Ramani and E. M. Marcotte, *Genome Biol.*, 2006, **7**, 120.
7. J. Frank, *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*, Oxford University Press, New York, 2006.
8. W. Wriggers, R. A. Milligan and J. A. McCammon, *J. Struct. Biol.*, 1999, **125**, 185.
9. A. M. Roseman, *Acta Crystallogr., Sect. D*, 2000, **D56**, 1332.
10. M. G. Rossmann, *Acta Crystallogr., Sect. D*, 2000, **D56**, 1341.
11. W. Wriggers and P. Chacón, *J. Appl. Crystallogr.*, 2001, **34**, 773.
12. J. A. Kovacs, P. Chacon, Y. Cong, E. Metwally and W. Wriggers, *Acta Crystallogr., Sect. D*, 2003, **D59**, 1371.
13. M. G. Rossmann, F. Arisaka, A. J. Battisti, V. D. Bowman, P. R. Chipman, A. Fokine, S. Halfstein, S. Kanamura, V. A. Kostyuchemko, V. V. Mesyanzhinov, M. M. Schneider, M. C. Morais, P. G. Leiman, L. M. Palermo, C. R. Parrish and C. Xiao, *Acta Crystallogr., Sect. D*, 2007, **D63**, 9.
14. H. M. Berman, T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J. D. Westbrook and C. Zardecki, *Acta Crystallogr., Sect. D*, 2002, **D58**, 899.
15. R. B. Russell, F. Alber, P. Aloy, F. P. Davis, D. Korkin, M. Pichaud, M. Topf and A. Sali, *Curr. Opin. Struct. Biol.*, 2004, **14**, 313.
16. E. M. Marcotte, P. Pellegrini, M. J. Thompson, T. O. Yeates and D. Eisenberg, *Nature*, 1999, **402**, 83.
17. D. Eisenberg, E. M. Marcotte, I. Xenarios and T. O. Yeates, *Nature*, 2000, **405**, 823.
18. P. Aloy, M. Pichaud and R. B. Russell, *Curr. Opin. Struct. Biol.*, 2005, **15**, 15.
19. P. Aloy and R. B. Russell, *Nat. Rev. Mol. Cell Biol.*, 2006, **7**, 188.
20. S. Jones and J. M. Thornton, *J. Mol. Biol.*, 1997, **272**, 121.
21. S. Jones and J. M. Thornton, *J. Mol. Biol.*, 1997, **272**, 133.
22. A. A. Bogan and K. S. Thorn, *J. Mol. Biol.*, 1998, **280**, 1.
23. P. Chakrabarti and J. Janin, *Proteins: Struct., Funct., Genet*, 2002, **47**, 334.

24. R. P. Bahadur, P. Chakrabarti, F. Rodier and J. Janin, *Proteins: Struct., Funct., Genet.*, 2003, **53**, 708.
25. N. J. Burgoyne and R. M. Jackson, *Bioinformatics*, 2006, **22**, 1335.
26. J. Janin, F. Rodier, P. Chakrabarti and R. P. Bahadur, *Acta Crystallogr., Sect. D*, 2007, **D63**, 1.
27. A. Koike and T. Takagi, *Protein Eng. Des. Selec.*, 2004, **17**, 165.
28. H. Neuvirth, R. Raz and G. Schreiber, *J. Mol. Biol.*, 2004, **338**, 181.
29. A. J. Bordner and R. Abagyan, *Proteins: Struct., Funct., Bioinf*, 2005, **60**, 353.
30. J. R. Bradford and D. R. Westhead, *Bioinformatics*, 2005, **21**, 1487.
31. H. Chen and H. X. Zhou, *Proteins: Struct., Funct., Bioinf*, 2005, **61**, 21.
32. A. del Sol and P. O'Meara, *Proteins: Struct., Funct., Bioinf*, 2005, **58**, 672.
33. J. Fernández-Recio, M. Totrov, C. Skorodumov and R. Abagyan, *Proteins: Struct., Funct., Bioinf*, 2005, **58**, 134.
34. J. L. Chung, W. Wang and P. E. Bourne, *Proteins: Struct., Funct., Bioinf*, 2006, **62**, 630.
35. A. Porollo and J. Meller, *Proteins: Struct., Funct., Bioinf*, 2007, **66**, 630.
36. D. W. Ritchie, *Curr. Protein Pep. Sci.*, 2008, **9**, 1.
37. S. J. de Vries and A. M. J. J. Bonvin, *Curr. Protein Pep. Sci.*, 2008, **9**, 394.
38. J. Janin, K. Henrick, J. Moult, L. Ten Eyck, M. J. E. Sternberg, S. Vajda, I. Vakser and S. J. Wodak, *Proteins: Struct., Funct., Genet.*, 2003, **52**, 2.
39. M. F. Lensink and S. J. Wodak, *Proteins: Struct., Funct., Bioinf*, 2010, **78**, 3073–3084.
40. E. R. Jefferson, T. P. Walsh and G. J. Barton, *J. Mol. Biol.*, 2006, **23**, 1118.
41. A. Stein, R. B. Russell and P. Aloy, *Nucleic Acids Res.*, 2005, **33**, D413.
42. F. P. Davis and A. Sali, *Bioinformatics*, 2005, **21**, 1901.
43. R. D. Finn, M. Marshall and A. Bateman, *Bioinformatics*, 2005, **21**, 410.
44. P. Heuser, D. Baù and D. Schomburg, *Proteins: Struct., Funct., Bioinf*, 2005, **61**, 1059.
45. S. Gong, C. Park, H. Choi, J. Ko, I. Jang, J. Lee, D. M. Molser, D. Oh, D. S. Kim and J. Bhak, *BMC Bioinformatics*, 2005, **6**, 207.
46. E. D. Levy, J. B. Pereira-Leal, C. Chothia and S. A. Teichmann, *PLoS Comp. Biol.*, 2006, **2**, e155.
47. C. Winter, A. Henschel, W. K. Kim and M. Schroeder, *Nucleic Acids Res.*, 2006, **34**, D310.
48. P. J. Kundrotas and E. Alexov, *Nucleic Acids Res.*, 2007, **35**, D575.
49. A. G. Murzin, S. E. Brenner, T. Hubbard and C. Chothia, *J. Mol. Biol.*, 1995, **247**, 536.
50. A. Bateman, E. Birney, R. Durbin, S. R. Eddy, K. L. Howe and E. L. Sonnhammer, *Nucleic Acids Res.*, 2000, **28**, 263.
51. J. Mintseris, K. Wiehe, B. Pierce, R. Anderson, R. Chen, J. Janin and Z. Weng, *Proteins: Struct., Funct., Genet.*, 2005, **60**, 214.
52. J. J. Gray, S. Moughan, C. Wang, O. Schueler-Furman, B. Kuhlman, C. A. Rohl and D. Baker, *J. Mol. Biol.*, 2003, **331**, 281.

53. D. Douguet, H. C. Chen, A. Tovchigrechko and I. A. Vakser, *Bioinformatics*, 2006, **22**, 2612.
54. S. J. Wodak and J. Janin, *J. Mol. Biol.*, 1978, **124**, 323.
55. M. Levitt, *J. Mol. Biol.*, 1976, **104**, 59.
56. M. J. E. Sternberg, H. A. Gabb and R. M. Jackson, *Curr. Opin. Struct. Biol.*, 1998, **8**, 250.
57. A. H. Elcock, D. Sept and J. A. McCammon, *J. Phys. Chem.*, 2001, **105**, 1504.
58. C. J. Camacho and S. Vajda, *Curr. Opin. Struct. Biol.*, 2002, **12**, 36.
59. I. Halperin, B. Ma, H. Wolfson and R. Nussinov, *Proteins: Struct., Funct., Genet*, 2002, **47**, 409.
60. G. R. Smith and M. J. E. Sternberg, *Curr. Opin. Struct. Biol.*, 2002, **12**, 28.
61. C. Deremble and R. Lavery, *Curr. Opin. Struct. Biol.*, 2005, **15**, 171.
62. A. D. J. van Dijk, R. Boelens and A. Bonvin, *FEBS J.*, 2005, **272**, 293.
63. A. Bonvin, *Curr. Opin. Struct. Biol.*, 2006, **16**, 194.
64. J. J. Gray, *Curr. Opin. Struct. Biol.*, 2006, **16**, 183.
65. B. Pierce, Z. Weng, Y. Xu, D. Xu and J. Liang, eds, *Computational Methods for Protein Structure Prediction and Modeling, Volume 2*, Springer, New York, 2007, pp. 109–134.
66. R. Méndez, R. Leplae, L. De Maria and S. J. Wodak, *Proteins: Struct., Funct., Genet*, 2003, **52**, 51.
67. S. Vajda and C. J. Camacho, *Trends Biotech.*, 2004, **22**, 110.
68. S. J. Wodak and R. Méndez, *Curr. Opin. Struct. Biol.*, 2004, **14**, 242.
69. R. Méndez, R. Leplae, M. F. Lensink and S. J. Wodak, *Proteins: Struct., Funct., Bioinf*, 2005, **60**, 150.
70. J. Janin, *Protein Sci.*, 2005, **14**, 278.
71. S. Vajda, *Proteins: Struct., Funct., Bioinf*, 2005, **60**, 176.
72. S. Vajda, I. A. Vakser, M. J. E. Sternberg and J. Janin, *Proteins: Struct., Funct., Genet*, 2002, **47**, 444.
73. P. L. Kastritis and A. M. J. J. Bonvin, *J. Proteome Res.*, 2010, **9**, 2216.
74. R. Mosca, C. Pons, J. Fernandez-Recio and P. Aloy, *PLoS Comp. Biol.*, 2009, **5**, e1000490.
75. T. Yoshikawa, K. Tsukamoto, Y. Hourai and K. Fukui, 2009, **49**, 693.
76. S. Kimura, R. C. Brower, S. Vajda and C. J. Camacho, *Biophys. J.*, 2001, **80**, 635.
77. D. Rajamani, S. Thiel, S. Vajda and C. J. Camacho, *Proc. Natl. Acad. Sci.*, 2004, **101**, 11287.
78. R. Grünberg, J. Leckner and M. Nilges, *Structure*, 2004, **12**, 2125.
79. L. P. Ehrlich, M. Nilges and R. C. Wade, *Proteins: Struct., Funct., Bioinf*, 2005, **58**, 126.
80. G. R. Smith, M. J. E. Sternberg and P. A. Bates, *J. Mol. Biol.*, 2005, **347**, 1077.
81. C. Wang, O. Schueler-Furman and D. Baker, *Protein Sci.*, 2005, **14**, 1328.
82. A. D. J. van Dijk and A. M. J. J. Bonvin, *Bioinformatics*, 2006, **22**, 2340.
83. K. Dev, *Nat. Rev. Drug Discov.*, 2004, **3**, 1047.

84. J. C. Fuller, N. J. Burgoyne and R. M. Jackson, *Drug Discovery Today*, 2008, **14**, 155.
85. J. Seco, F. J. Luque and X. Barril, *J. Med. Chem.*, 2009, **15**, 2363.
86. C. Pons, S. Grosdidier, A. Solernou, L. Pérez-Cano and J. Fernández-Recio, *Proteins: Struct., Funct., Bioinf*, 2010, **78**, 95.
87. C. J. Camacho, D. W. Gatchell, S. R. Kimura and S. Vajda, *Proteins: Struct., Funct., Genet*, 2000, **40**, 525.
88. L. Li, R. Chen and Z. Weng, *Proteins: Struct., Funct., Genet*, 2003, **53**, 693.
89. C. Zhang, S. Liu and Y. Zhou, *Proteins: Struct., Funct., Bioinf*, 2005, **60**, 314.
90. R. Chen, L. Li and Z. Weng, *Proteins: Struct., Funct., Bioinf*, 2003, **52**, 80.
91. S. Comeau, D. Gatchell, S. Vajda and C. Camacho, *Nucleic Acids Res.*, 2004, **32**, W96.
92. D. Schneidman-Duhovny, Y. Inbar, R. Nussinov and H. Wolfson, *Nucleic Acids Res.*, 2005, **33**, W363.
93. A. Tovchigrechko and I. Vakser, *Nucleic Acids Res.*, 2006, **34**, W310.
94. S. Lyskov and J. Gray, *Nucleic Acids Res.*, 2008, **36**, W233.
95. D. W. Ritchie and V. Venkatraman, *Bioinformatics*, 2010, **26**, 2398–2405.
96. D. Fischer, R. Norel, H. Wolfson and R. Nussinov, *Proteins: Struct., Funct., Genet*, 1993, **16**, 278.
97. O. Bachar, D. Fischer, R. Nussinov and H. J. Wolfson, *Protein Eng.*, 1993, **6**, 279.
98. S. L. Lin, R. Nussinov, D. Fischer and H. J. Wolfson, *Proteins: Struct., Funct., Genet*, 1994, **18**, 94.
99. D. Fischer, S. L. Lin, H. L. Wolfson and R. Nussinov, *J. Mol. Biol.*, 1995, **248**, 459.
100. E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A. A. Friesem and C. Aflalo, *Proc. Natl. Acad. Sci.*, 1992, **89**, 2195.
101. P. N. Palma, L. Krippahl, J. E. Wampler and J. J. G. Moura, *Proteins: Struct., Funct., Genet*, 2000, **39**, 372.
102. J. D. Jackson, *Classical Electrodynamics*, Wiley, New York, 1975.
103. H. A. Gabb, R. M. Jackson and M. J. E. Sternberg, *J. Mol. Biol.*, 1997, **272**, 106.
104. J. G. Mandell, V. A. Roberts, M. E. Pique, V. Kotlovyyi, J. C. Mitchell, E. Nelson, I. Tsigelny and L. F. Ten Eyck, *Protein Eng.*, 2001, **14**, 105.
105. A. Heifetz, E. Katchalski-Katzir and M. Eisenstein, *Protein Sci.*, 2002, **11**, 571.
106. R. Chen and Z. Weng, *Proteins: Struct., Funct., Genet*, 2003, **51**, 397.
107. C. Zhang, G. Vasmatzis, J. Cornette and C. DeLisi, *J. Mol. Biol.*, 1997, **267**, 707.
108. A. Berchanski, B. Shapira and M. Eisenstein, *Proteins: Struct., Funct., Bioinf*, 2004, **56**, 130.
109. D. W. Ritchie and G. J. L. Kemp, *Proteins: Struct., Funct., Genet*, 2000, **39**, 178.
110. L. C. Biedenharn and J. C. Louck, *Angular Momentum in Quantum Physics*, Addison-Wesley, Reading, MA, 1981.

111. D. W. Ritchie, *J. Appl. Crystallogr.*, 2005, **38**, 808.
112. A. Erdelyi, W. Magnus, F. Oberhettinger and F. G. Tricomi, *Higher Transcendental Functions*, Vol. 2, McGraw-Hill, New York, 1953.
113. A. R. Edmonds, *Angular Momentum in Quantum Physics*, Princeton University Press, New Jersey, 1957.
114. D. W. Ritchie, D. Kozakov and S. Vajda, *Bioinformatics*, 2008, **24**, 810.
115. J. D. Owens, D. Luebke, N. Govindaraju, M. Harris, J. Krüger, A. E. Lefohn and T. J. Purcell, *Comp. Graph. Forum*, 2007, **26**, 80.
116. I. Buck, T. Foley, D. Horn, J. Sugerman, K. Fatahalian, M. Houston and P. Hanrahan, *ACM Trans. Graph.*, 2004, **23**, 777.
117. M. C. Schatz, C. Trapnell, A. L. Delcher and A. Varshney, *BMC Bioinformatics*, 2007, **8**, 474.
118. S. A. Manavski and G. Valle, *BMC Bioinformatics*, 2008, **9**, S10.
119. M. A. Suchard and A. Rambaut, *Bioinformatics*, 2009, **25**, 1370.
120. J. E. Stone, J. C. Phillips, P. L. Freddolino, D. J. Hardy, L. G. Trabuco and K. Schulten, *J. Comp. Chem.*, 2007, **28**, 2618.
121. J. A. van Meel, A. Arnold, D. Frenkel, S. F. Portegies Zwart and R. G. Belleman, *Mol. Sim.*, 2008, **34**, 259.
122. I. S. Ufimtsev and T. J. Martínez, *J. Chem. Theory. Comput.*, 2008, **4**, 222.
123. D. Dynerman, E. Butzlaff and J. C. Mitchell, *J. Comp. Biol.*, 2009, **16**, 523.
124. G. Macindoe, L. Mavridis, V. Venkatraman, M.-D. Devignes and D. W. Ritchie, *Nucleic Acids Res.*, 2010, **38**, W445.
125. D. Kozakov, R. Brenke, S. R. Comeau and S. Vajda, *Biophys. J.*, 2005, **89**, 867.
126. C. Zhang, S. Liu, Q. Zhu and Y. Zhou, *J. Med. Chem.*, 2005, **48**, 2325.
127. D. Kozakov, R. Brenke, S. R. Comeau and S. Vajda, *Proteins: Struct., Funct., Bioinf*, 2006, **65**, 392.
128. Y. Shen, R. Brenke, D. Kozakov, S. R. Comeau, D. Beglov and S. Vajda, *Proteins: Struct., Funct., Bioinf*, 2006, **69**, 734.
129. D. Kozakov, D. R. Hall, D. Beglov, R. Brenke, S. R. Comeau, Y. Shen, K. Li, J. Zheng, P. Vakili, I. Ch Paschalidis and S. Vajda, *Proteins: Struct., Funct., Bioinf*, 2010, **78**, 3124–3130.
130. C. Dominguez, R. Boelens and A. M. J. J. Bonvin, *J. Am. Chem. Soc.*, 2003, **125**, 1731.
131. M. Nilges, *J. Mol. Biol.*, 1995, **245**, 645.
132. C. J. Camacho, *Proteins: Struct., Funct., Bioinf*, 2005, **60**, 245.
133. M. Zacharias and H. Sklenar, *J. Comp. Chem.*, 1999, **20**, 287.
134. M. Zacharias, *Proteins: Struct., Funct., Bioinf*, 2004, **54**, 759.
135. C. N. Cavasotto, J. A. Kovacs and R. A. Abagyan, *J. Am. Chem. Soc.*, 2005, **127**, 9632.
136. M. Zacharias, *Proteins: Struct., Funct., Bioinf*, 2005, **60**, 252.
137. G. R. Smith, P. W. Fitzjohn, C. S. Page and P. A. Bates, *Proteins: Struct., Funct., Bioinf*, 2005, **60**, 263.
138. A. Amadei, A. B. M. Linssen and H. J. C. Berendsen, *Proteins: Struct., Funct., Genet*, 1993, **17**, 412.

- 139. A. May and M. Zacharias, *Biochim. Biophys. Acta*, 2005, **1754**, 225.
- 140. D. Mustard and D. W. Ritchie, *Proteins: Struct., Funct., Bioinf*, 2005, **60**, 269.
- 141. M. Lei, M. I. Zavodoszsky, L. A. Kuhn and M. F. Thorpe, *J. Comp. Chem*, 2004, **25**, 1133.
- 142. M. Tirion, *Phys. Rev. Lett.*, 1996, **77**, 1905.
- 143. K. Hinsen, *Proteins: Struct., Funct., Genet*, 1998, **33**, 417.
- 144. A. May and M. Zacharias, *Proteins: Struct., Funct., Bioinf*, 2007, **70**, 794.
- 145. see <http://gmplib.org/>
- 146. see <http://www.nvidia.com/>
- 147. Download from <http://hex.loria.fr/>
- 148. HexServer is available at <http://hexserver.loria.fr/>

CHAPTER 4

QM Based Modelling

TIMOTHY CLARK

Centre for Molecular Design, School of Pharmacy and Biomedical Sciences,
Institute of Biomedical and Biomolecular Sciences, University of
Portsmouth, Mercantile House, Hampshire Terrace, Portsmouth, PO1 2 EG,
UK and Computer-Chemie-Centrum, Friedrich-Alexander-Universität
Erlangen-Nürnberg, Nögelsbachstrasse 25, 91052 Erlangen, Germany

4.1 Introduction

Biological and pharmaceutical modelling uses predominantly classical mechanical techniques (force fields and related techniques, such as classical scoring functions). This has the advantages of computational efficiency and beneficial scaling with system size, but is also limited in accuracy by common (but not strictly necessary) approximations used in the classical mechanical models. Other chapters in this book deal with semiempirical MO-Theory,¹ density-functional theory (DFT),² quantum mechanical descriptions of hydrogen bonding³ and the quantum-mechanics based atoms in molecules (AIM) technique.⁴ This chapter will therefore serve as an introduction and to describe some specific applications of quantum mechanical techniques in biological and pharmaceutical modelling.

The ‘missing’ effects in most classical mechanical models are well known; most force fields are not polarisable (*i.e.* the charge distributions do not react to external electrostatic fields) and the common net atomic monopole (charge) model cannot reproduce important features of the molecular electrostatic potential around molecules. However, common classical mechanical models also have their strengths. They can, for instance, generally describe weak

interactions (dispersion) better than many quantum mechanical techniques. This has led to a flurry of ‘dispersion-augmented’ QM techniques that will be described below. High-quality force fields are also likely to be able to describe the relative energies of competing conformations of a molecule more accurately than all but the best quantum mechanical techniques. In particular the last generation of DFT functionals suffered from inconsistencies such as an inability to deal with branching properly^{5–9} that would not be expected from a well-parameterised force field. Thus, quantum mechanical techniques are not always automatically better than force fields, and indeed, ‘expensive’ quantum mechanical methods are not always better than ‘cheap’ ones. The following sections will give an outline of some of the considerations that make quantum mechanical techniques realistic and viable options for certain types of modelling. The selection is necessarily selective, but serves to illustrate some important points.

4.2 Computational Demands

Many routine quantum mechanical calculations can now be carried out on low-level hardware in a few seconds or less. This was not always the case and so some resistance to using quantum mechanics for large-scale applications remains. As recently as 1998, treating a complete database of just over 50 000 compounds with AM1 semi-empirical molecular orbital (MO) theory¹⁰ was a task suitable for the current generation of supercomputer.¹¹ Seven years later, processing the entire NCI database of approximately 250 000 compounds on distributed hardware was easily possible.¹² Interestingly, the technical reliability of the calculations (*i.e.* whether the calculations ran successfully) and the handling of tens of thousands of structures presented more problems in both studies than the CPU time required. Quantum mechanical calculations are less tolerant than their force-field equivalents. Thus, protonation states and charges must be correct (no unpaired electrons), counterions should be removed and starting geometries should be as strain-free as possible to avoid unwanted bond shifts caused by the spurious proximity of two unbound atoms. Quantum mechanical geometry optimisations also require an extra check that is not necessary for force-field optimisations. As the bonds are neither defined nor necessarily constant within a quantum mechanical geometry optimisation, an analysis of the starting and optimised structures must be used to detect possible rearrangements during the optimisation. These are typically proton shifts to give a different tautomer or to convert an input zwitterion into the form with no charged groups. Relatively few zwitterionic structures represent minima in the gas phase, so that the latter type of rearrangement is particularly common. It can be avoided by performing the geometry optimisation with a continuum solvation model, but quantum mechanical geometry optimisations in cheminformatics applications are generally performed in the gas phase.

Quantum mechanical calculations for very large systems are now also becoming routine. Linear scaling (a linear increase of the computational

requirements with the number of atoms) quantum mechanical techniques are becoming common, and in some cases standard. Linear scaling DFT has a long history¹³ and resolution of the identity (RI) techniques¹⁴ have increased the size of structures treatable by DFT significantly. Nonetheless, DFT geometry optimisations of even a moderate sized protein remain a daunting task. The faster, more approximate semi-empirical MO techniques¹⁵ and charge-consistent tight-binding DFT techniques¹⁶ are more realistic for the complete treatment of biopolymers by quantum mechanical methods, whereas more expensive quantum mechanical techniques are usually combined with a force field environment in hybrid QM/MM calculations.^{17,18}

Semi-empirical MO techniques have been used most for treating complete proteins because two separate linear scaling techniques have been implemented for these methods; the localised molecular orbital technique introduced by Stewart in MOZYME¹⁹ and the more conventional divide and conquer (D&C) methodology.^{20,21} The former has been used to optimise the geometries of proteins up to approximately 17 000 atoms,²² whereas the latter has been used primarily for single-point calculations on experimental protein structures.²³ More recently, a highly parallel implementation of a semi-empirical MO program has been used to perform a single-point calculation on 76 000 atoms using 1024 CPUs.²⁴

4.3 The Molecular Electrostatic Potential

The molecular electrostatic potential, MEP, $V(\mathbf{r})$,²⁵ has long been used as a local property in biological and pharmaceutical modelling and in other quantitative structure–property (QSPR) relationships. It is defined at any given point in space as the energy of interaction of a positive unit point charge (in atomic units) with the unpolarised molecule. The MEP is defined as

$$V(\mathbf{r}) = \sum_A \frac{Z_A}{|\mathbf{R}_A - \mathbf{r}|} - \int \frac{\rho(\mathbf{r}')}{|\mathbf{r}' - \mathbf{r}|} d\mathbf{r}' \quad (4.1)$$

where Z_A is the charge on nucleus A , located at $\mathbf{R}_A$, and $\rho(\mathbf{r}')$ is the electron density at point $\mathbf{r}'$. The two contributions to the MEP at any point are a repulsion arising from the interaction of the probe charge with the nuclei of the molecule and a negative one that results from the interaction of the probe charge with the molecular electrons. The positive term is the simple summation over all atoms of the molecule, whereas the negative one involves integrating the electron density over space and is therefore far more expensive to calculate. However, this integration is not significant in comparison to the total compute task of carrying out a DFT or *ab initio* MO calculation, but amounts to more effort than a complete semi-empirical MO self-consistent field calculation. Luckily, the two-electron integrals in MNDO-like semi-empirical MO techniques are expressed in terms of atom-centred multipoles up to quadrupole,²⁶ so that the MEP can be expressed in terms of interactions of the probe monopole

with the atom-centred monopole, dipole and quadrupole.²⁷ This reduces the calculation of the MEP to a sum over atoms and gives remarkably accurate results compared with DFT or MP2 MEP values.²⁸

One of the early uses of the MEP was paradoxically to derive atom-centred point charges for use in force fields and other classical mechanical techniques.^{29–32} The MEP-fitting errors indicate that atomic monopoles represent a severe approximation as a technique for calculating the MEP. Figure 4.1 shows a comparison of the MEP at the standard isodensity surface³³ of the DNA base guanine. The differences, principally the lack of resolution in the MEP calculated from the potential-derived charges, are manifest.

The situation for the molecule shown in Figure 4.1 is relatively favourable for the atomic monopole description. In more extreme cases, one and the same atom may have both positive and negative areas of the MEP on its surface. It is clearly impossible to describe this situation with a single charge on the atom. Politzer *et al.*³⁴ have called this situation a ‘fallacy of atomic charges’. It is particularly important for σ -hole bonding,³⁵ which is a special case of halogen bonding,³⁶ and plays an important role in many biological systems,³⁷ including many drugs that contain chlorophenyl groups. Figure 4.2 shows the calculated MEP at the standard isodensity surface of chlorpromazine with a prominent σ -hole on the chlorine atom collinear with and opposite the C–Cl bond.

Nonetheless, 3D-QSAR techniques such as CoMFA^{TM38} almost always use a classical charge model to calculate the interaction with a charged probe, which is nothing other than the MEP. However, recent work^{38,40} suggests that the MEP may not be as dominant as has been believed in determining interactions between ligands and their receptors and that other local electronic properties may be equally or more important.

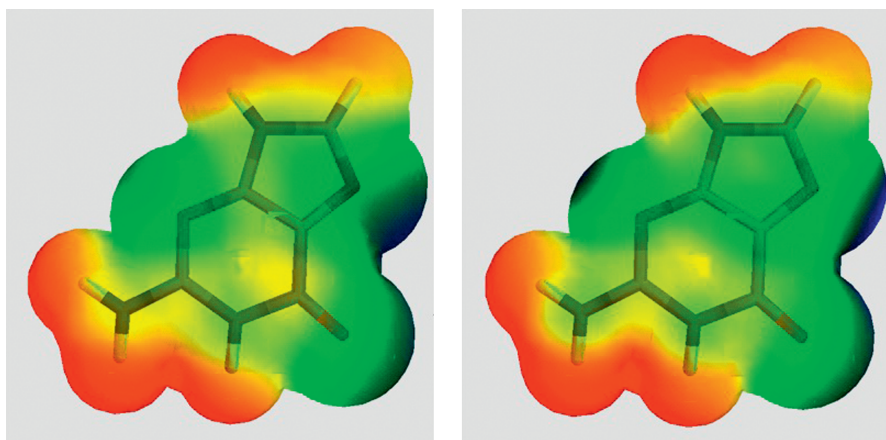


Figure 4.1 Comparison of the molecular electrostatic potentials at the standard isodensity surface of guanine calculated with (a) MEP-derived (AM1*) atomic monopoles and (b) AM1*.

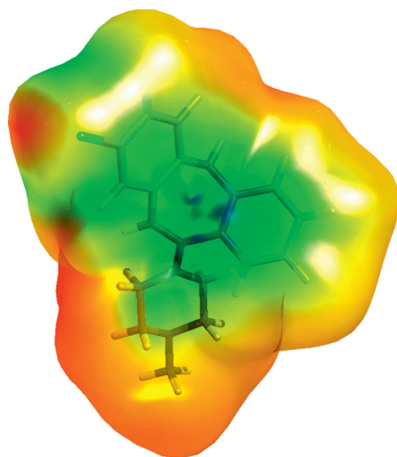


Figure 4.2 The molecular electrostatic potential at the standard isodensity surface of chlorpromazine showing the prominent σ -hole collinear with and opposite to the C–Cl bond.

4.4 Local Properties and Intermolecular Interactions

The theory of intermolecular interactions has been known for a long time,⁴¹ but recently more emphasis has been placed on defining local properties that are directly related to the factors that contribute to intermolecular interactions.⁴² Basically, the major components are the direct Coulomb interaction, which depends on the MEP, Pauli repulsion at short ranges, dispersion and electron donor–acceptor (Lewis acid–base) interactions. The Coulomb interaction has traditionally been modelled in biological and pharmaceutical applications by a simple interaction between atom-centred multipoles, as outlined above. Clearly, the molecular electrostatic potential is the local property directly related to the simple Coulomb interaction, although the molecule's electrostatic field has the advantage that it depends less strongly on the molecular charge than the MEP and that it provides more spatial definition. However, the field is a vector property, so that it has not been used extensively in QSAR or QSPR applications. A variation that may prove useful is the field normal to the molecular surface. This field is a key property in polarised continuum models (PCM) of solvation effects⁴³ and has been suggested to be related to the electrostatic solvation energy by a simple surface integral.⁴⁴

Electron donor–acceptor interactions are usually neglected in classical mechanical modelling techniques. The earliest local properties used to describe the donor and acceptor properties of molecules were different variations of the so-called Fukui function, which was proposed by Parr and Yang⁴⁵ based on considerations derived from Fukui's frontier molecular orbital (FMO) theory.⁴⁶ The major weakness of FMO theory is that it may not consider donor or acceptor orbitals that lie close in energy to the highest occupied MO

(HOMO) or lowest unoccupied MO (LUMO), respectively. This is often the case for molecules with π -systems, including many drugs and drug candidates. The average local ionisation energy, $I(\mathbf{r})$ proposed by Sjöberg *et al.*⁴⁷ represents a more complete description of the electron donor properties in the vicinity of a molecule. It is defined as

$$I(\mathbf{r}) = - \frac{\sum_{i=1}^{HOMO} \varepsilon_i \rho_i(\mathbf{r})}{\sum_{i=1}^{HOMO} \rho_i(\mathbf{r})} \quad (4.2)$$

Where ε_i is the eigenvalue of molecular orbital I , and $\rho_i(\mathbf{r})$ is the electron density assignable to orbital i at point $\mathbf{r}$. Thus the average local ionisation energy is simply a Koopmans theorem ionisation potential weighted by the contribution of the individual molecular orbitals to the total electron density at the point in question. $I(\mathbf{r})$ has been suggested to be related to the polarisability^{48–52} and its properties and usefulness have been described in a recent review.⁵³

The concept of the average local ionisation energy can be extended to the electron-accepting properties of molecules by defining the analogous local electron affinity,⁵⁴ $A(\mathbf{r})$, as

$$A(\mathbf{r}) = - \frac{\sum_{i=HOMO+1}^{NORBS} \varepsilon_i \rho_i(\mathbf{r})}{\sum_{i=HOMO+1}^{NORBS} \rho_i(\mathbf{r})} \quad (4.3)$$

in which $NORBS$ is the total number of molecular orbitals. Equation 4.3 is exactly analogous to equation 4.2 except that the summation now occurs over the virtual, rather than the occupied, orbitals. However, equation 4.3 is only applicable to techniques such as the older semi-empirical MO methods that use only a minimal s and p orbital basis set. It can be extended to techniques that use polarisation functions by a simple intensity filter.⁵⁵

Together, the average local ionisation energy and the local electron affinity provide an extremely useful description of the donor–acceptor properties of molecules. These appear prominently in a variety of QSPR models^{56–62} and for predicting chemical reactivity in models for metabolism.⁶³ The connection between donor and acceptor properties and chemical reactivity is particularly important for qualitative reactivity theories.⁵⁴

Two further properties⁵⁴ can be derived from the local ionisation energy and electron affinity in analogy with the Mulliken electronegativity⁶⁴ and Pearson's hardness.⁶⁵ The local electronegativity, $\chi(\mathbf{r})$, is given by

$$\chi(\mathbf{r}) = \frac{I(\mathbf{r}) + A(\mathbf{r})}{2} \quad (4.4)$$

and the local hardness, $\eta(\mathbf{r})$, by

$$\eta(\mathbf{r}) = \frac{I(\mathbf{r}) - A(\mathbf{r})}{2} \quad (4.5)$$

Together, these properties define the element- and binding-specific features that are not covered by the MEP. Traditionally, for instance, surface-based models use not only the electrostatics at the surface, but also the identity of the atom to which the surface element is assigned.^{66–69} Such additional qualifiers are only necessary if the non-electrostatic electronic local properties of the molecule at the point in question are neglected. Given the ‘extreme quantum mechanical’ view that molecules are clouds of indistinguishable electrons moving in the external field of static (within the Born–Oppenheimer approximation) point-charge nuclei, neither atoms nor bonds exist within molecules, so that we should not need descriptors or qualifiers derived from the concept of atoms. This extreme view contrasts strongly with the atoms in molecules (AIM) approach,⁴⁴ which, however, relies on the powerful but approximate concept of the transferability of the electron density assignable to specific types (element and hybridisation) of atoms. In many ways, AIM provides a bridge between the atomistic and the extreme quantum mechanical views of modelling molecules.

The ‘missing’ local property not described above is the local polarisability, $\alpha(\mathbf{r})$.⁵⁴ This is based on a parameterised version⁷⁰ of Rivail’s variational scheme for calculating the molecular electronic polarisability^{71–73} partitioned in an arbitrary fashion analogously to a population analysis into atomic or even atomic orbital contributions to an additive polarisability model.⁷⁴ The local polarisability serves to describe dispersion interactions.

Figure 4.3 shows projections of the six local properties onto the surface of a drug-like molecule.

As Figure 4.3 suggests, there is little correlation⁵⁴ between the different local properties except between $I(\mathbf{r})$ and $A(\mathbf{r})$ and the properties derived from them ($\chi(\mathbf{r})$ and $\eta(\mathbf{r})$). The ‘additional’ local properties therefore add information to the descriptions of molecules.⁶²

4.5 Local Properties and Molecular Surfaces

4.5.1 Spherical Harmonic Expansions

Once the local properties have been defined, techniques must be found to use the information contained in them to construct QSAR and QSPR models. One way to do this is to define the surface itself analytically, so that properties projected onto it can also be described by analytical functions. Perhaps the best known technique for fitting molecular surfaces is to use spherical harmonic expansions.^{75–77} Spherical harmonics are familiar to most chemists as functions that define the angular dependence of atomic orbitals.⁷⁸ Real spherical

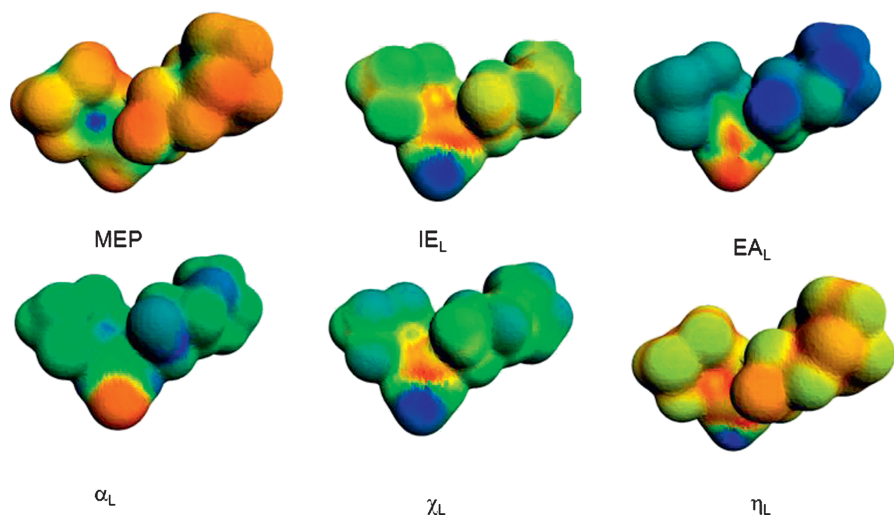


Figure 4.3 The six local properties projected onto the standard isodensity surface of a drug-like molecule.

harmonic functions $y(\theta, \phi)$ are defined as

$$y(\theta, \phi) = \{N_{lm}P_{lm}(\cos \theta) \cos|m|\phi\}, m \geq 0$$

and

$$y(\theta, \phi) = \{N_{lm}P_{lm}(\cos \theta) \sin|m|\phi\}, m < 0$$

where θ and ϕ are the Euler angles that define the function, N_{lm} are normalisation factors and $P_{lm}(\cos \theta)$ are associated Legendre functions.^{75,79,80} The molecular surface can then be fitted to the following equation:

$$r(\theta, \phi) = \sum_{l=0}^N \sum_{m=-l}^l a_{lm} y_{lm}(\theta, \phi) \quad (4.7)$$

where $r(\theta, \phi)$ is the distance from the molecular centre to the surface along the vector defined by the angles θ and ϕ , N is the order of the expansion and a_{lm} are fitting coefficients, which can be determined easily.⁸¹ A spherical harmonic expansion of order N is defined by $(N + 1)^2$ coefficients a_{lm} , so that spherical harmonic expansions of molecular surfaces represent extremely compact descriptions of molecules. However, there are two potential difficulties. The first is that the molecular surface must be single valued along any radial vector extending from the centre. This is shown schematically in Figure 4.4.

In Figure 4.4a, vectors a and b are single valued (*i.e.* they cross the molecular surface only once), whereas vectors c and d are not because they cross the surface three times each. An appropriate surface for fitting to spherical

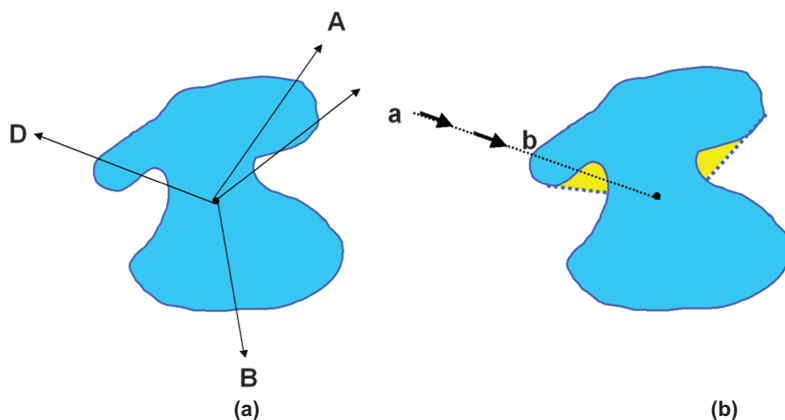


Figure 4.4 (a) 2D-schematic diagram of radial vectors cutting the molecular surface. Vectors **a** and **b** are single-valued, whereas **c** and **d** cross the surface three times (b) the 'bay' areas (yellow) not included in the molecular shrink-wrap surface. Reproduced with permission of Cepos InSilico Ltd.

harmonic expansions is therefore the so-called shrink-wrap surface,^{82–84} which is constructed by 'shrinking' the surface of a sphere larger than the molecule along radial vectors until it touches the molecule for the first time. This results in some excluded areas as shown in yellow in Figure 4.4b.

The second technical difficulty in determining spherical harmonic surfaces is to define the most suitable centre for the expansion. The first choice, the centre of mass, is usually appropriate, but may actually lie outside the molecular surface for U- or L-shaped molecules. In these more difficult cases, more complex techniques can be used to determine an appropriate centre for the spherical harmonic expansion.⁸⁵ Using such techniques, almost all drug-like molecules can be fitted except for O-shaped ones with holes in the centre.

Figure 4.5 shows spherical harmonic surfaces fitted to an original shrink-wrapped isodensity surface for clozapine.

The surface for $N=0$ is simply a sphere that is described by a single coefficient (the radius). Increasing the order to $N=1$ corresponds to adding the possibility of translating the centre for the expansion and so changes the surface little in this example, for which the centre of mass used for the expansions is close to the ideal position. Increasing the order of the spherical harmonic expansion from zero to one is the exact equivalent of adding p functions to an s basis set to provide polarisation. Increasing the order to $N=2$ (equivalent to adding d orbitals to a basis set) causes a far more visible change in the molecular shape, making it ellipsoid. At values of N of five and larger, the overall shape of the molecule is easily recognisable. $N=5$, which is defined by 36 coefficients, is sufficient for many high-throughput applications and by $N=15$ (256 coefficients) the atomistic details of the surface become visible. Thus, spherical harmonic expansions represent extremely compact and efficient representations of molecular shapes that can also be manipulated analytically (see below).

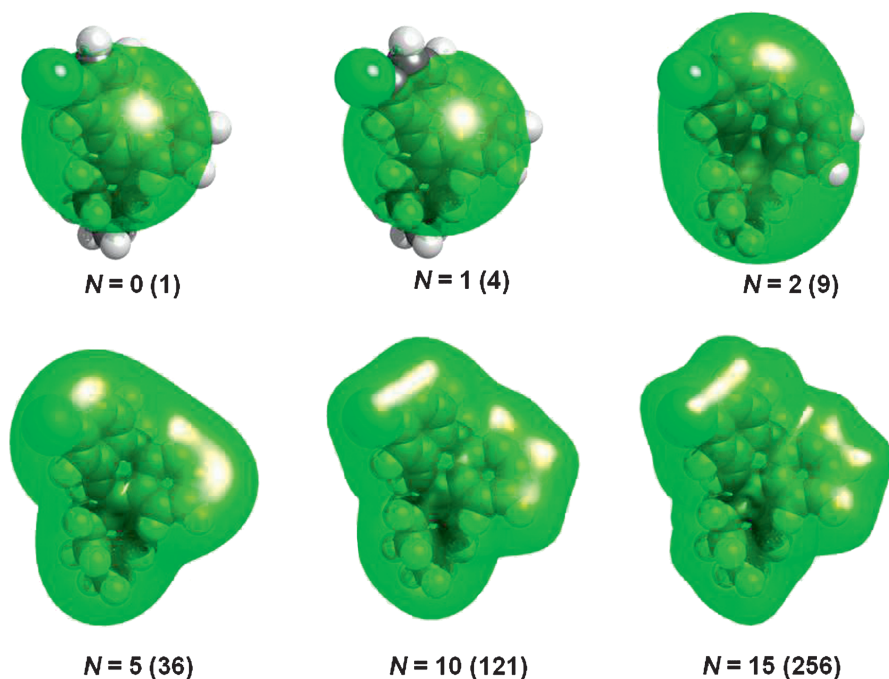


Figure 4.5 Spherical-harmonic fits of different order N to the isodensity shrink-wrap surface of clozapine. The numbers in brackets give the number of coefficients for each surface.

Shape descriptions are clearly important, but a complete analytical description of the local properties that govern intermolecular interactions is of even more use. To this end, the six local properties projected onto the molecular surface can also be represented as spherical harmonic expansions. Figure 4.6 shows the MEP of clozapine expressed as spherical harmonic expansions of different order N projected onto the $N=15$ fitted spherical harmonic surface.

As for the shape, $N=0$ gives an average value for the MEP over the entire surface. In the case of the MEP, however, $N=1$ corresponds to adding the molecular dipole and $N=3$ to the MEP arising from molecular monopole, dipole and quadrupole moments. Note that the spherical harmonic approach is in this case mathematically equivalent to a monocentric multipole expansion. The MEP is a highly resolved property that requires quite high orders in the spherical harmonic expansion before all important features are reproduced. Note in Figure 4.6, for instance, that the area of positive MEP collinear with the C–Cl bond (the σ -hole) first appears weakly for $N=10$ and is only fully developed at $N=15$.

Since spherical harmonic expansions can be rotated analytically, they are ideal objects for fast overlaying of molecules. Ritchie *et al.*⁸⁶ extended their work on proteins⁷⁷ to small molecules to enable fast overlaying and similarity

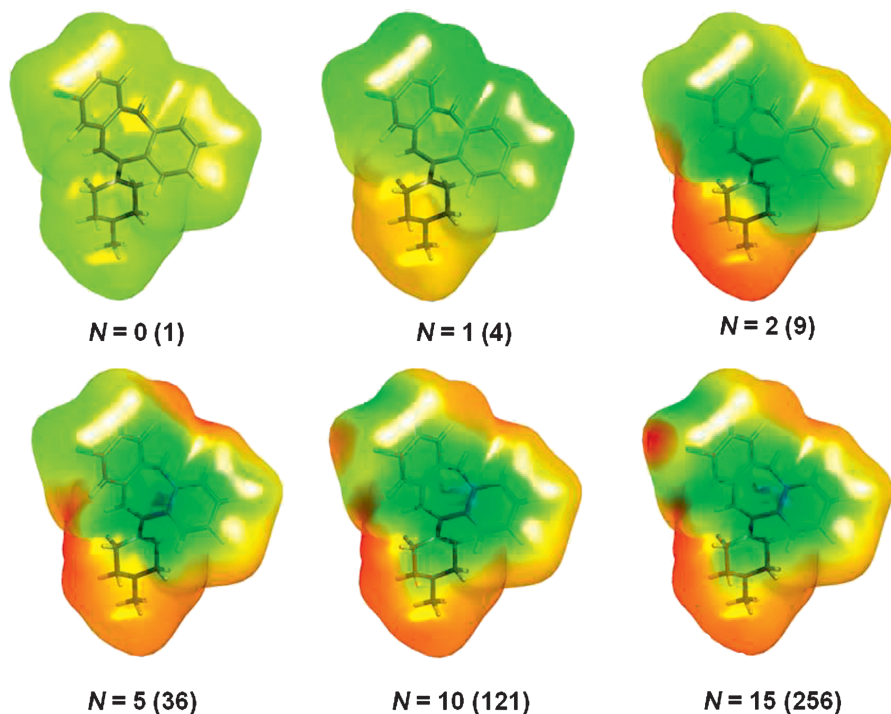


Figure 4.6 Spherical-harmonic fits of different order N to the MEP projected onto the isodensity shrink-wrap surface of clozapine of order 15. The numbers in brackets give the number of coefficients for each surface.

calculations on large databases. Comparisons can be made using the molecular shape or any one of the local properties. Surprisingly, the MEP generally does not play a significant role in QSARs derived using this approach.^{39,40} An alternative to overlaying molecules is to use the so-called canonical orientation, which is derived from coefficients of the spherical harmonic expansion using fixed rules.⁸⁶ The canonical orientation allows very fast shape-based screening or even screening based on the local properties.

A further very fast screening technique based on spherical harmonic expansions are the so-called rotationally invariant fingerprints (RIFs).⁸⁶ These are simply a vector of ‘hybridisation coefficients’, H_l for each of the orders l of the spherical harmonic expansion defined in equation 4.7 (or the equivalent expression for any local property other than the shape). The hybridisation coefficients are defined as

$$H_l = \sum_{m=-l}^l a_{lm}^2 \quad (4.8)$$

where m , l and a_{lm} have the same meanings as in equation (4.7). Because the coefficients a_{lm} are normalised,

$$\sum_{l=1}^N H_l = \sum_{l=1}^N \sum_{M=-l}^l a_{lm}^2 = 1 \quad (4.9)$$

RIFs are simply the hybridisation coefficients for up to defined orders N for the molecular shape and local properties at the surface in a predefined order. They provide a fixed length rotationally invariant description of the molecular shape and intermolecular binding properties. They can be used for virtual screening by calculating, for instance Euclidean distances between the RIFs of the molecules to be compared. Properties can also be selected by using only the elements of the RIFs that correspond to the hybridisation coefficients of the required properties.

4.5.2 Surface Autocorrelations

Autocorrelations have long been used in pharmaceutical modelling. Gasteiger *et al.*⁸⁷ introduced the idea of surface-based autocorrelations based on classical atomistic representations of molecules. However, surface autocorrelations acquire extra predictive power when used together with the local properties described above. A Gaussian-weighted autocorrelation function is a vector of L values, $R(d)$, is defined as:

$$R(d) = \sum_{x=1}^N \sum_{y>x}^N (p(\mathbf{x}) \cdot p(\mathbf{y})) e^{-\alpha(r_{xy}-d)^2} \quad (4.10)$$

where L is the number of distances, d , calculated, $p(\mathbf{x})$ and $p(\mathbf{y})$ are the property in question at points x and y , N is the total number of surface points, r_{xy} is the distance between points x and y and α is a smoothing function that determines the width of the Gaussian distance weighting.

Figure 4.7 shows a series of autocorrelations for trimethoprim using some of the local properties. Note that for the properties that can be both positive and negative (*i.e.* MEP, $A(\mathbf{r})$), three different autocorrelation functions are required to differentiate between $+/+$ and $-/-$ peaks (because the product is positive in both cases).

Note that, in contrast to the techniques described above, calculating autocorrelations is expensive because it formally involves LN^2 evaluations of the weighted product of the properties. As a finely triangulated surface for a drug-like molecule typically consists of 5000 to 10 000 surface points, the calculational task can become significant. Although promising because of its high information content, the surface autocorrelation technique has not yet found wide application.

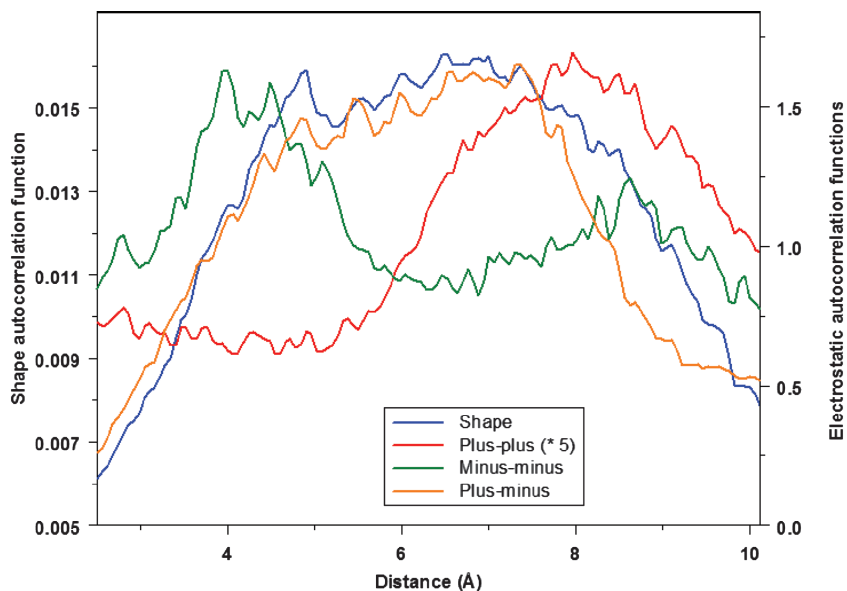


Figure 4.7 Surface autocorrelation functions for trimethoprim based on the AM1 standard isodensity surface.

4.5.3 Surface Integral Models

Surface integral models are a particularly attractive way to use local properties at a molecular surface for QSPR modelling. The idea was originally proposed by Richards, Williams and Tute,^{88,89} but were first implemented as a usable model by Brickman *et al.*^{66,67} The idea of surface integral models has some justification because volume integrals can be represented as surface integrals.⁴⁴ In our context, a simple surface integral model could be defined as

$$P = \int_{\text{surface}} f(z(\mathbf{r}), y(\mathbf{r}), x(\mathbf{r}), \dots) \simeq \sum_{i=1}^{ntri} f(z(\mathbf{r}_i), y(\mathbf{r}_i), x(\mathbf{r}_i), \dots) \cdot A_i \quad (4.11)$$

where P is the target property, $f(z(\mathbf{r}), y(\mathbf{r}), x(\mathbf{r}), \dots)$ is a function of the local properties $z(\mathbf{r}), y(\mathbf{r}), x(\mathbf{r}), \text{etc.}$ The triangulated molecular surface consists of $ntri$ tesserae with the area of tessera i being A_i . The points $\mathbf{r}_i$ correspond, for instance, to the midpoints of the tesserae i , although many definitions are possible for the values of the local properties assigned to triangle i . Models of this type can be constructed quite successfully using linear regression to construct the target function from the integrals of polynomial functions of the local properties,⁶⁰ but models constructed this way tend to have quite large constant terms, so that the function obtained does not represent a real local property, but rather a correction term to a mean (and therefore constant) value

for the training dataset. Nonetheless, such models can reproduce demonstrably non-local properties such as solvation free energies quite successfully.

A more successful way to construct practical surface integral models is to use a binning technique.⁵⁶ In this case, a range of bins are defined for the local properties themselves and, in the published example, products of pairs of local properties. Descriptors are then defined as:

$$D(P(\mathbf{r}), i) = \sum_{j=1}^{ntri} A_j \cdot \delta(P(\mathbf{r}_j), i) \quad (4.12)$$

where the descriptor $D(P(\mathbf{r}), i)$ for bin i of property $P(\mathbf{r})$ is simply the sum of the areas A_j of all the triangles for which the value of $P(\mathbf{r}_j)$ falls within the range of bin i , which is defined by the function $\delta(P(\mathbf{r}_j), i)$. Surface integral models derived using such descriptors tend to have small constant terms and are therefore closer to representing real local properties. The disadvantage of such models is that the range and partitioning for the bins must be defined before the model is constructed, so that the values of the local properties for the entire dataset must be calculated before the binned descriptors can be determined. In principle, this could also mean that unknown compounds may lie outside the range described by the binning scheme. This can be avoided by using first and last bins that are open on their low and high sides, respectively, but this only masks the fact that the compound probably lies outside the applicability domain of the model.

4.6 Local Properties on Grids

Some of the most successful 3D-techniques in computer-aided drug design rely on calculating interaction energies between the target molecule and specific probes at the points of a regular grid around the molecule. Perhaps the most prominent of these are Goodford's GRID program⁹⁰⁻⁹² and comparative molecular field analysis (ComFATM).⁹³ In both cases, interaction energies between the molecule (usually represented as an atomistic force field) and the probe (simple models such as methane for hydrophobic interaction, etc.) are calculated classically.

Calculating the local properties on a grid around the molecule is a useful and fast alternative procedure. The electron density gives ample information about steric repulsion and the MEP encodes the information usually obtained by polar probe molecules, but does not suffer from the deficits of atomistic point charge models. Hydrogen bonding properties are encoded by $I(\mathbf{r})$ and $A(\mathbf{r})$ in addition to the MEP, but the two donor-acceptor properties provide considerable extra information. Dispersion interactions are encoded in $\alpha(\mathbf{r})$.

Using the local properties in place of probe interaction energies for ComFA usually yields more predictive models than the classical model if the same molecular overlays and grids are used,^{94,95} so that local properties derived from economical quantum mechanical techniques appear to be the methods of choice for such applications.

As intermolecular interactions are coded so effectively in the local properties described above, we should be able to characterise the typical binding sites around molecules and receptors simply on the basis of the local properties calculated on a grid. Interaction energies can be calculated accurately in the gas phase using high level *ab initio* theory. Figure 4.8 shows iso-energy diagrams for the interaction of water, ammonia and the hydroxide anion with a water molecule.

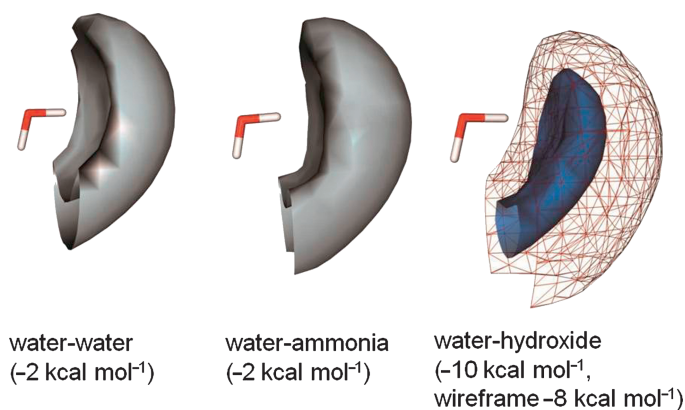


Figure 4.8 Isoenergy contours for the interaction energies (MP2/aug-cc-pVTZ) of water with three different H-bond acceptors.

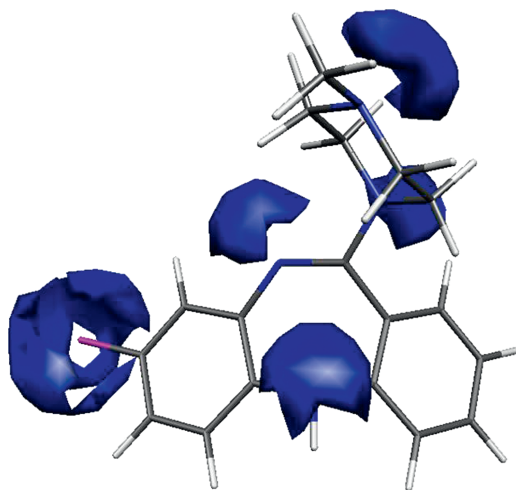


Figure 4.9 H-bond acceptor sites for clozapine as calculated using an artificial neural net trained to reproduce the *ab initio* data using the local properties as input.

Figure 4.8 demonstrates the effect of hydrogen bond acceptors of different strength on the intermolecular interaction energy, but also that the geometries of the binding interaction do not vary very strongly when the acceptor is changed. It is therefore possible to train, for instance, a simple feed-forward artificial neural net to recognise binding sites around molecules from the values of the local properties. Figure 4.9 shows the results of such an analysis for hydrogen bond acceptor sites on clozapine. Note that the non-classical equatorial hydrogen bond acceptor site around the chlorine atom is recognised although no chlorine compounds were used to derive data to train the neural net.

Using different probe molecules for the *ab initio* calculations allows separate neural nets to be trained for different types of binding interaction, so that complete pharmacophore maps can be calculated that take both known, classical binding sites and less expected ones that result from unusual electronic effects into account.⁹⁶

4.7 Summary and Conclusions

Classical atomistic modelling has been a trusty servant of computer-aided drug design for more than two decades and will still render sterling service for years to come. However, it does have its limits. Some of these, such as being able to treat σ -hole interactions correctly, simply require more sophisticated electrostatic models, such as those based on an AIM analysis.⁴ Others, such as the change in electronic character of the faces of aromatic rings on substitution,⁵⁶ require quantum mechanical techniques that do not assume transferability of atomic properties. It is also possible that our current emphasis on electrostatic interactions may not be appropriate and that the hitherto relatively neglected donor–acceptor properties of molecules may in the long run be more important.

Most of the examples described in this chapter are based on semiempirical molecular orbital theory, which does a good job for one-electron properties, but less well for geometries and energies. However, the latter need not be calculated quantum mechanically. Very high quality force fields could provide excellent geometries and relative conformational energies, so that whatever quantum mechanical technique is used can use the force field geometry and relative energies. This opens the field to computationally more expensive quantum mechanical techniques such as density functional theory; although there is little indication at the moment that this would give better results.

Acknowledgments

I should like to thank all my co-workers, past and present, from both Erlangen and Portsmouth, who have contributed to the ideas presented here. They are named in the references. Thanks are also due to countless computational chemists in different pharmaceutical companies who have often functioned as

sounding boards for wild ideas, have contributed some of their own and have generally ensured that some degree of practicality remains.

References

1. Alexander Voityuk, Chapter 5 of this book.
2. Rob Deeth, Chapter 2 of this book.
3. Tell Tuttle, Chapter 1 of this book.
4. Paul Popellier, Chapter 6 of this book.
5. M. D. Wodrich, D. F. Jana, P. V. R. Schleyer and C. Corminboeuf, *J. Phys. Chem A*, 2008, **112**, 11495.
6. M. D. Wodrich, C. Corminboeuf, P. R. Schreiner, A. A. Fokin and P. V. R. Schleyer, *Org. Lett.*, 2007, **9**, 1851.
7. M. D. Wodrich, C. Corminboeuf and P. v. R. Schleyer, *Org. Lett.*, 2006, **8**, 3631.
8. S. Grimme, *Angew. Chem., Int. Ed. Engl.*, 2006, **45**, 4460.
9. S. Grimme, *Org. Lett.*, 2010, **12**, 4670.
10. M. J. S. Dewar, E. G. Zoebisch, E. F. Healy and J. J. P. Stewart, *J. Am. Chem. Soc.*, 1985, **107**, 3902.
11. B. Beck, A. Horn, J. E. Carpenter and T. Clark, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 1214.
12. P. Murray-Rust, H. S. Rzepa, J. J. P. Stewart and Y. Zhang, *J. Mol. Model.*, 2005, **11**, 532.
13. See, for instance C. F. Guerra, J. G. Snijders, G. te Velde and E. J. Baerends, *Theor. Chem. Acc.*, 1998, **99**, 391.
14. M. Sierka, A. Hogeamp and R. Ahlrichs, *J. Chem. Phys.*, 2003, **118**, 9136.
15. T. Clark and J. J. P. Stewart, MNDO-like semiempirical molecular orbital theory and its application to large systems, in *Computational Methods for Large Systems*, ed. J. J. Reimers, Wiley, Chichester, 2011.
16. M. Elstner and M. Gaus, The self-consistent-charge density-functional tight-binding (SCC-DFTB) method: An efficient approximation of density functional theory, in *Computational Methods for Large Systems*, ed. J. J. Reimers, Wiley, Chichester, 2011.
17. H. M. Senn and W. Thiel, *Angew. Chem., Int. Ed. Engl.*, 2009, **48**, 1198.
18. P. L. Cummins and J. E. Gready, *J. Comput. Chem.*, 1997, **18**, 1496.
19. J. J. P. Stewart, *Int. J. Quant. Chem.*, 1996, **58**, 133.
20. T. S. Lee, D. M. York and W. Yang, *J. Chem. Phys.*, 1996, **105**, 2744.
21. A. Van der Vaart, V. Gogonea, S. L. Dixon and K. M. Merz Jr., *J. Comput. Chem.*, 2000, **21**, 1494.
22. J. J. P. Stewart, *J. Mol. Model.*, 2009, **15**, 765.
23. X. He and K. M. Merz Jr., *J. Chem. Theor. Comput.*, 2010, **6**, 405.
24. M. Hennemann and T. Clark, manuscript in preparation.
25. *Chemical Applications of Atomic and Molecular Electrostatic Potentials. Reactivity, Structure, Scattering, and Energetics of Organic, Inorganic, and Biological Systems*, ed. P. Politzer and D. G. Truhlar, Plenum Press, New York, 1981.

26. M. J. S. Dewar and W. Thiel, *J. Am. Chem. Soc.*, 1977, **99**, 4899.
27. D. Bakowies and W. Thiel, *J. Comput. Chem.*, 1997, **17**, 87.
28. A. H. C. Horn, J.-H. Lin and T. Clark, *Theor. Chem. Acc.*, 2005, **114**, 159; Erratum: *Theor. Chem. Acc.*, 2007, **117**, 461.
29. B. H. Besler, K. M. Merz Jr. and P. A. Kollman, *J. Comput. Chem.*, 1990, **11**, 431.
30. L. E. Chirlian and M. M. Francl, *J. Comput. Chem.*, 1987, **8**, 894.
31. C. M. Breneman and K. B. Wiberg, *J. Comput. Chem.*, 1990, **11**, 361.
32. B. Beck, T. Clark and R. C. Glen, *J. Comput. Chem.*, 1997, **18**, 744.
33. R. F. W. Bader, M. T. Carroll, J. R. Cheeseman and C. Chang, *J. Am. Chem. Soc.*, 1987, **109**, 7968.
34. P. Politzer, J. S. Murray and M. C. Concha, *J. Mol. Model.*, 2008, **14**, 659.
35. T. Clark, M. Hennemann, J. S. Murray and P. Politzer, *J. Mol. Model.*, 2007, **13**, 291.
36. P. Politzer, P. Lane, M. C. Concha, Y. Ma and J. S. Murray, *J. Mol. Model.*, 2007, **13**, 305.
37. P. Auffinger, F. A. Hays, E. Westhof and P. Shing Ho, *Proc. Natl. Acad. Sci.*, 2004, **101**, 16789.
38. R. D. Cramer III, D. E. Patterson and J. D. Bunce, *J. Am. Chem. Soc.*, 1988, **110**, 5959.
39. D. T. Manallack, *J. Mol. Model.*, 2008, **14**, 797.
40. Dave Ritchie.
41. Intermolecular interactions, <http://www.teokem.lu.se/~gunnar/material.pdf>, G. Karlström and B. Jönsson, accessed January 16th 2011.
42. T. Clark, K. G. Byler, and M. J. de Groot in *Molecular Interactions - Bringing Chemistry to Life* (Proceedings of the International Beilstein Workshop, Bozen, Italy, May 15–19, 2006), Logos Verlag, Berlin, 2008, pp. 129–146.
43. Tomasi, B. Mennucci and R. Cammi, *Chem. Rev.*, 2005, **105**, 2999.
44. R. J. Abraham, B. D. Hudson, M. W. Kermode and J. R. Mines, *J. Chem. Soc., Faraday Trans. I*, 1988, **84**, 1911.
45. R. G. Parr and W. Yang, *Density Functional Theory of Molecules*, Oxford University Press, 1989.
46. K. Fukui, *J. Chem. Phys.*, 1952, **20**, 722.
47. P. Sjöberg, J. S. Murray, T. Brinck and P. A. Politzer, *Can. J. Chem.*, 1990, **68**, 1440.
48. J. S. Murray and P. Politzer, *J. Mol. Struct. (Theochem)*, 1998, **425**, 107.
49. J. S. Murray, S. Ranganathan and P. Politzer, *J. Org. Chem.*, 1991, **56**, 3734.
50. P. Politzer, P. Lane, J. S. Murray and T. Brinck, *J. Phys. Chem.*, 1992, **96**, 7938.
51. J. S. Murray, P. Lane, T. Brinck, K. Paulsen, M. E. Grice and P. Politzer, *J. Phys. Chem.*, 1993, **97**, 9369.
52. T. Brinck, J. S. Murray and P. Politzer, *Mol. Phys.*, 1992, **76**, 609.
53. P. Politzer, J. S. Murray and F. Bulat, *J. Mol. Model.*, 2010, **16**, 1731.

54. B. Ehresmann, B. Martin, A. H. C. Horn and T. Clark, *J. Mol. Model.*, 2003, **9**, 342.
55. T. Clark, *J. Mol. Model.*, 2010, **16**, 1231.
56. C. Kramer, B. Beck and T. Clark, *J. Chem. Inf. Model.*, 2010, **50**, 429.
57. C. Kramer, B. Beck and T. Clark, *J. Chem. Inf. Model.*, 2010, **50**, 404.
58. D. J. Livingstone, T. Clark, M. G. Ford, B. D. Hudson and D. C Whitley, *SAR and QSAR in Env. Res.*, 2008, **19**, 285.
59. T. Clark, M. Ford, J. Essex, W. G. Richards and D. W. Ritchie, A non-atom-based paradigm for modeling, QSAR and QSPR, in *QSAR and Molecular Modelling in Rational Design of Bioactive Molecules*, EuroQSAR 2004 Proceedings, ed. E. Aki and I Yalcin, CADDDS in Turkey, Ankara, 2006, pp. 536–537.
60. B. Ehresmann, M. De Groot and T. Clark, *J. Chem. Inf. Model.*, 2005, **45**, 1053.
61. T. Clark, *J. Mol. Graph. Model.*, 2004, **22**, 519.
62. B. Ehresmann, M. J. de Groot, A. Alex and T. Clark, *J. Chem. Inf. Comp. Sci.*, 2004, **44**, 658.
63. M. Hennemann, A. Friedl, M. Lobell, J. Keldenich, A. Hillisch, T. Clark and A. H. Göller, *ChemMedChem*, 2009, **4**, 657.
64. R. S. Mulliken, *J. Chem. Phys.*, 1935, **3**, 573.
65. R. G. Pearson, *J. Am. Chem. Soc.*, 1963, **85**, 3533.
66. P. Pixner, W. Heiden, H. Merx, G. Moeckel, A Möller and J. Brickmann, *J. Chem. Inf. Comp. Sci.*, 1994, **34**, 1309.
67. R. Jäger, S. Kast and J. Brickmann, *J. Chem. Inf. Comp. Sci.*, 2003, **43**, 237.
68. C. J. Cramer and D. G. Truhlar, *Acc. Chem. Res.*, 2008, **41**, 760.
69. D. Bashford and D. A. Case, *Annu. Rev. Phys. Chem.*, 2000, **51**, 129.
70. G. Schürer, P. Gedeck, M. Gottschalk and T. Clark, *Int. J. Quant. Chem.*, 1999, **75**, 17.
71. D. Rinaldi and J.-L. Rivail, *Theor. Chim. Acta*, 1974, **32**, 243.
72. J.-L. Rivail and D. Rinaldi, *Comptes Rendus, Serie B: Sciences Physiques*, 1976, 283.
73. D. Rinaldi and J.-L. Rivail, *Theor. Chim. Acta*, 1973, **32**, 57.
74. B. Martin, P. Gedeck and T. Clark, *Int. J. Quant. Chem.*, 2000, **77**, 473.
75. B. S. Duncan and A. J. Olson, *Approximation and Characterization of Molecular Surfaces*, Scripps Insitute, San Diego, California, 1995.
76. D. W. Ritchie and G. J. L. Kemp, *J. Comput. Chem.*, 1999, **20**, 383.
77. Dave Ritchie, Chapter 3 of this book.
78. D. L. Beveridge and J. A. Pople, *Approximate Molecular Orbital Theory*, McGraw-Hill, New York, 1970.
79. N. L. Max and E. D. Getzhoff, *IEEE Comput. Graphics Appl.*, 1988, **8**, 42.
80. D. W. Ritchie, *Parametric Protein Shape Recognition*, Ph.D. Thesis, University of Aberdeen, 1998.
81. J.-H. Lin and T. Clark, *J. Chem. Inf. Model.*, 2005, **45**, 1010.
82. J. H. Van Drie, *J. Chem. Inf. Comp. Sci.*, 1997, **37**, 38.
83. J. H. Van Drie and R. A. Nugent, *SAR and QSAR in Env. Res.*, 1998, **9**, 1.

84. J. Erickson, D. J. Neidhart, J. Van Drie, D. J. Kempf, X. C. Wang, D. W. Norbeck, J. J. Plattner, J. W. Rittenhouse, M. Turon, N. Wideburg, W. E. Kohlbrenner, R. Simmer, R. Helfrich, D. A. Paul and M. Knigge, *Science*, 1990, **249**, 527.
85. ParaSurf¹⁰ User Manual, Cepos InSilico Ltd., Bedford, 2010, pp. 12–13 (<http://www.ceposinsilico.de/pdf/ParaSurf10.pdf>, accessed 16th February 2011).
86. L. Mavridis, B. D. Hudson and D. W. Ritchie, *J. Chem. Inf. Model.*, 2007, **47**, 1787.
87. M. Wagener, J. Sadowski and J. Gasteiger, *J. Am. Chem. Soc.*, 1995, **117**, 7769.
88. N. G. J. Richards, P. B. Williams and M. Tute, P. Löwdin, *Int. J. Quant. Chem.: Quantum Biology Symposium*, 1991, 299.
89. N. G. J. Richards, P. B. Williams and M. Tute, *Int. J. Quant. Chem.*, 1992, **44**, 219.
90. P. J. Goodford, *J. Med. Chem.*, 1985, **28**, 849.
91. R. C. Wade, K. J. Clark and P. J. Goodford, *J. Med. Chem.*, 1993, **36**, 140.
92. R. C. Wade and P. J. Goodford, *J. Med. Chem.*, 1993, **36**, 148.
93. R. D. Cramer, III, D. E. Patterson and J. D. Bunce, *J. Am. Chem. Soc.*, 1988, **110**, 5959.
94. K. G. Byler, *3D-QSAR and Physical Property Modeling Using Quantum-Mechanically-Derived Molecular Surface Properties*, PhD Thesis, Universität Erlangen-Nürnberg, 2007.
95. S. Güßregen, K. G. Byler and T. Clark, manuscript in preparation.
96. A. El Kedawy, M. Hennemann and T. Clark, manuscript in preparation.

CHAPTER 5

Semi-empirical Methods: Current Status and Future Directions

ALEXANDER A. VOITYUK

Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona,
Spain and Universitat de Girona, Institute of Computational
Chemistry, Spain

5.1 Introduction

Quantum mechanical calculations of the electronic structure and properties of biological systems remains a formidable challenge in computational chemistry. High-level *ab initio* and density functional theory (DFT) methods provide insight into the nature of relatively small molecules. The size and conformational flexibility of biomolecules call for much more efficient techniques capable of treating systems up to several thousand atoms. Semi-empirical methods (SMs) allow such systems to be addressed at the quantum mechanical level.^{1–5} The success of SMs lies in having some adjustable parameters. Tuning these parameters one can implicitly include the effects of electron correlation and extended basis sets and thereby achieve chemical accuracy at a fraction of the computational cost of *ab initio* and DFT methods (SMs are typically 3 to 4 orders of magnitude faster than DFT calculations). The role and place of SMs in quantum chemistry is considered by Clark.⁵ It is well known that properties of biomolecules considerably depend on molecular motion, and, therefore, the calculated characteristics should be averaged over many configurations.

In particular, surrounding dynamics can essentially affect chemical transformation in the active site of enzymes.⁶ *Ab initio* and DFT methods are computationally too demanding to allow for sufficient sampling of conformational states even for medium-size models. The speed of SMs makes them amenable to simulations over time scales of tens of nanoseconds.

The development of semi-empirical techniques has a long history;⁷ over the past decades, SMs have been widely used in organic chemistry, biochemistry and drug design.^{1–5,12} The accuracy of SMs has been continuously improved, new algorithms for treatment of very large systems have been suggested. This makes SMs suitable for modeling various systems, where specific electrostatic interactions, hydrogen bonds, and solvent dynamics are of particular importance.¹³ In this review, we consider semi-empirical methods that are of interest for application in biochemistry and related fields.

5.2 MNDO and MNDO/d Based Methods

In 1978, Dewar and Thiel published a seminal paper, in which they described a new semi-empirical approach called MNDO (Modified Neglect of Differential Overlap).¹⁴ This method employs an elegant and efficient scheme to calculate two-center two-electron integrals using point charges.¹⁵ Although MNDO is derived from Hartree–Fock theory, the balanced description of electrostatic interactions as well as carefully defined semi-empirical parameters allowed one to achieve a higher accuracy as compared with earlier SMs based on the CNDO and INDO schemes. Within a short time, the MNDO scheme became very popular and was widely employed for quantum mechanical studies of organic molecules. With the passage of time, various limitations were found. In particular, the MNDO method fails to reproduce the structure and the energy of hydrogen bonded systems. Several schemes based on the MNDO Hamiltonian have been developed, e.g. AM1¹⁶ and PM3,¹⁷ which in general provide more accurate results than MNDO. AM1 and PM3 differ from MNDO by including several Gaussian terms in the core–core repulsion function. For some molecules, however, some artifacts in potential energy surfaces were found. Jorgensen and his co-workers showed that further modification of the core–core repulsion in MNDO and PM3 improves the accuracy of these schemes.^{18,19}

A key step in the development of SMs is parameterization, a procedure to search parameters that determine the semi-empirical Hamiltonian. The MNDO and related methods were parameterized using the ground-state properties: heats of formation, molecular geometries, dipole moments, and ionization potentials, as reference data. Today, the reference data for many molecules can be found in the WebBook²⁰ and the Cambridge Structural Database.²¹ To assess the advantages and limitations of a semi-empirical scheme, many tests have to be performed for different properties and different classes of molecules. The performance is established by comparison of estimated molecular properties with the corresponding experimental data or/and the results of more sophisticated calculations (see Tables 5.1 and 5.2). Systematic assessment of the SMs has been reported.^{1,2,4,14–23}

Table 5.1 Mean absolute errors of SMs for organic compounds.^a

Property	N ^b	MNDO	AM1	PM3	OM3	DFTB
ΔH_f , kcal/mol	140	6.3	5.5	4.2	2.9	7.7
Bond lengths, Å	242	1.4	1.7	1.1	2.0	1.5
Bond angles, deg.	101	2.6	1.9	2.1	1.8	1.3
Ionization potential, eV	52	0.46	0.35	0.42	0.45	3.82
Dipole moment, D	63	0.35	0.26	0.27	0.27	0.37
Vibrational frequencies, cm ⁻¹	112	241	172	151	120	90

^aThe data are taken from Ref. 23.^bNumber of comparisons.**Table 5.2** Mean absolute errors of SMs for organic compounds (all data are in kcal/mol).^a

Compounds	N ^b	G3	B3LYP	MNDO	AM1	PM3	OM2	DFTB
ΔH_f , small species	81	0.7	2.3	7.7	7.4	6.7	3.36	9.19
ΔH_f , larger species	47	0.9	7.1	7.1	6.3	4.4	3.15	4.50
ΔH_f , alkanes C ₁ –C ₁₆	16	0.5 ^c	15.4 ^d	1.8	10.9	2.2	2.03	5.99
Ionization energies	32	1.1	5.1	12.5	12.2	11.9	7.13	9.70
Electron affinities	29	1.0	3.6	15.4	11.8	9.2	8.6	12.30

^aThe data are taken from Ref. 23.^bNumber of comparisons.^cG3 data only up to C₈H₁₈.^dThe error increases with molecular size.

As is well known, DFT calculations with the commonly used exchange-correlation potentials (e.g. B3LYP) provide rather accurate energetics for small-size molecules; their performance for large systems is, however, less satisfactory (see Table 5.2). Unlike DFT, the accuracy of SMs does not change much by increasing the size of molecular systems.²⁴ The MNDO, AM1, PM3 methods and their variants are available in many programs for quantum mechanical calculations and maybe routinely used in computational modeling of biomolecules.

5.2.1 MNDO/d

The MNDO and related schemes were formulated for the sp-AO basis and cannot be applied to compounds of transition metals because of the omission of d orbitals. An extension of the MNDO scheme to the spd basis, MNDO/d, has been suggested by Thiel and Voityuk.^{25,26} This scheme was extensively tested for main group elements; MNDO/d was found to describe hypervalent compounds of P, S, and halogens much more accurately than the methods based on the sp-AO framework. The major advantage of MNDO/d and related schemes is that they can be applied to study biomolecules including transition metals such as Fe, Cu, and Mo. The extension of the AM1 method to AM1/d and the parameters for Mo was described in detail.²⁷ The AM1/d parameterization of

Mo showed that there were systematic deviations for certain types of Mo–X bonds. Inclusion of Gaussian-type functions commonly used in AM1 did not improve the data.²⁷ On modification of the AM1/d scheme, by introducing Mo–X bond specific parameters in the core–core repulsion, the accuracy of the results was significantly increased. The use of the diatomic parameters allows the method to be extended to other elements without modifying the underlying schema. The parameterization procedure for transition metals is not as straightforward as for the main group elements. Unlike organic molecules, experimental heats of formation and metal–ligand bond enthalpies are sparse and prone to errors and, because of this, the reference data for many elements should be augmented with the results of *ab initio* and DFT calculations. High-level calculations of transition metal compounds are extremely time consuming and cannot be considered as a convenient source of the reference energies. In many cases, results from DFT calculations are found to be not accurate enough. For molecular geometries, gas phase structures are available for a number of metallorganics, but X-ray crystallographic data must be used as reference geometries of inorganic compounds. The orbital energies of the transition metal in complexes cannot be used to estimate the corresponding vertical ionization energies (the Koopmans' theorem does not hold for transition metal compounds due to large electronic reorganization). Furthermore, 'smooth' optimization of semi-empirical parameters is limited by problems with SCF convergence for some reference systems.

It has proven difficult, however, to obtain a general set of semi-empirical parameters to describe the diverse range of molecular properties of transition metal compounds. For this reason, specific parameters for particular chemical problems have to be elaborated. This approach is typified by the following examples. Parameterization of AM1/d for a phosphoryl transfer process was described by York *et al.*²⁸ The AM1/d method was also parameterized for accurate estimation of bond energies and structural parameters in complexes of Pt with large π conjugated systems.²⁹ A set of PM3/d parameters for Fe was developed to describe structure and redox properties of the active sites of iron-containing proteins, including iron–sulfur clusters and iron–heme complexes.³⁰ To model hydrogen abstraction catalyzed by iron species, the reaction-specific PM3/d parameters for Fe were suggested.³¹

Two semi-empirical schemes based on the MNDO/d framework have been recently reported. In the AM1* method,^{32,33} the original AM1 parameters are used for H, C, N, O, and F. For other elements (P, S, and Cl,³² Al, Si, Ti, and Zr,³³ Br and I,³⁴ Cu and Zn,³⁵ V and Cr,³⁶ Co and Ni,³⁷ Mn and Fe,³⁸ and Au³⁹), the semi-empirical parameters have been elaborated for the spd basis. The strengths and weaknesses of AM1* have been discussed in detail.^{32–39} Most of the approximations used in PM6⁴⁰ are identical to those in AM1/d,²⁷ while some modifications were made to the core–core repulsion term. The method has been parameterized for some 70 elements. To obtain the final parameter set, *ca.* 9000 species was employed. PM6 is found to give good results for biomolecules. For a representative set of species, its average absolute error in ΔH_f is 4.4 kcal mol^{–1}, whereas B3LYP, PM3, and AM1

show errors of 5.2, 6.3, and 10.0 kcal mol⁻¹, respectively. The method provides reasonable geometries of transition metal complexes and may be considered as a good starting point for further reaction-specific parameterization of 3d and 4d elements.

5.3 An Overview of Other Semi-empirical Approaches

Beside the MNDO and MNDO/d related methods, which play a significant role in practical calculations, several other semi-empirical approaches are of interest for application in biochemistry.

5.3.1 OMx Schemes^{41,42}

OM1, OM2, and OM3 schemes described organic molecules more accurately than the MNDO based approaches (see Tables 5.1 and 5.2) due to inclusion of explicit corrections that mimic the Pauli exchange repulsion. All of the OMx schemes provide similar results. As these methods are still parameterized only for 4 elements (H, C, N, and O), their extensive application for modeling biological systems is somewhat restricted. OMx may be employed to treat excited state properties of molecular systems (see below).

5.3.2 DFTB

The DFTB (density functional tight binding) method is an approximate scheme, derived by a second-order expansion of the total DFT energy expression,⁴³ the formalism, parameterization and applications of this method are considered in detail by Elstner.⁴⁴ This approach appears to perform well for biological systems. In particular, it reproduces accurately conformational energies and geometries for polypeptides.²³ The overall accuracy of DFTB is similar to PM3 (see Tables 5.1 and 5.2). This method is readily recommended for the study of biomolecules,⁴⁴ but apparently is less suited to the study of radicals and electronically excited states.²³

5.3.3 MSINDO

MSINDO was developed by substantial modifications of the SINDO1 formalism.^{10,45} The MSINDO parameters are available now for elements H, Li to F, Na to Cl, K to Zn, and Ga to Br. MSINDO employs the spd basis for Al to Cl in order to account for hypervalent compounds.⁴⁶ To date this method is mostly applied to models in solid-state chemistry and there are only a few studies related to biochemistry. In particular, MSINDO provides good results for electronic properties and geometry of signaling biomolecules, and can be used for the analysis of structure–activity relationships.⁴⁷

5.4 Hydrogen Bonds, Non-covalent Interaction and Effects of Environment

Correct treatment of intra-molecular and inter-molecular hydrogen bonds (HBs) in the QM treatment of biomolecules is vital. As noted previously, the original MNDO scheme fails to properly describe the structure and the energy of HBs, which makes it unsuitable for biological applications. While AM1 and PM3 give better results, the accuracy of these schemes is still not satisfactory. A modified MNDO scheme was suggested to treat the structure and energetics of systems with both weak and strong hydrogen bonds.⁴⁸ A variant of the PM3 method was recently introduced for modeling HB in nucleic acids.⁴⁹ The performance of semi-empirical methods for hydrogen bonded systems has been considered in detail.⁵⁰

In biomolecules dispersive interactions, *e.g.* π - π stacking, can be of equal importance to hydrogen bonding interactions,¹³ and their accurate QM calculation requires computationally expensive high-level *ab initio* methods in combination with large basis sets. Standard DFT and semi-empirical methods often fail to properly describe these interactions.¹³ An important step to address this situation was made by Grimme, who suggested adding empirical corrections to the total energy of the system.⁵¹ Augmenting the AM1, PM3 and OMx schemes with a dispersion term improves significantly the description of hydrogen bonding and van der Waals interactions in biomolecules.^{52,53} The corresponding corrections were also developed for PM6⁵⁴ and DFTB.⁴⁴ Extensive tests showed that a dispersion corrected method reproduced the interaction energy between biomolecules within 1.5 kcal mol⁻¹, which makes them valuable for molecular docking and conformational analysis.

Since experimental studies usually involve molecules in solution, solvent effects should be accounted for in computational models. Existing approaches to treat the environment effects can be divided into two distinct groups: continuum and discrete schemes. In continuum models, the solute molecules form a cavity within the dielectric continuum that represents the solvent. Accurate PCM models developed by Cramer and Truhlar for semi-empirical methods appear to provide a reliable treatment of chemical reactions in solution.⁵⁵ Hybrid QM/MM (quantum mechanics/molecular mechanics) approaches are often applied in the discrete solvent model: the solute molecule is treated by QM and the solvent by a suitable force field.⁵⁶ In these models, averaging over the degrees of freedom of the solvent is necessary. This can be achieved by sampling over a large number of snapshots from classical molecular dynamics (MD).

QM/MM approaches have become the method of choice for modeling biochemical reactions.⁵⁷ QM schemes are required to describe chemical transformations, electronic excitations, charge and electronic energy transfer processes. The size and conformational complexity of biopolymers require methods capable of treating up to several 100 000 atoms and allowing for simulations over time scales of tens of nanoseconds. However, even semi-empirical calculations are usually restricted to systems of up to a few hundred atoms. In the

QM/MM model introduced by Warshel and Levitt,⁵⁸ a QM method is employed for the chemically active region (*e.g.* substrates and co-factors in an enzymatic reaction) whereas the surroundings (*e.g.* protein and solvent) are treated by an MM method. Many variants of QM/MM methods based on semi-empirical treatment of QM region have been described.^{59,63} These and other related schemes enable quite accurate modeling of biochemical reactions at a reasonable computational effort. Many interesting examples of biomolecular QM/MM studies are provided in the literature, see *e.g.* Table 5.2 in the review by Senn and Thiel.⁵⁷ SMs are also used in the so-called ONIOM models developed by Morokuma and co-workers.^{64,65} ONIOM is based on an additivity scheme wherein different QM methods are applied to different regions of a molecular system.

5.5 Excited States

Theoretical modeling of excited states is of potential interest because of photosynthesis and several other biochemical processes related to electronic excitations. The interpretation of experimental data on excited state properties of biomolecules remains challenging and QM methods are often invoked to analyze the excited state properties. Most accurate theoretical approaches, MS-CASPT2,⁶⁶ and coupled cluster methods⁶⁷ are well established and may be successfully applied to relatively small molecules with up to 30 heavy atoms. Reliable models of many biological systems lie, however, beyond these high-level approaches. Recently a MS-CASPT2 benchmark set has been introduced for electronically excited states of small organic molecules with a total of 223 excitations,⁶⁸ which provides a good basis to assess the performance of less sophisticated but more efficient computational approaches. Time-dependent density functional theory (TD-DFT) gives reasonable results for medium-sized molecules,⁶⁹ but has limited use for treating charge transfer transitions.⁷⁰ It should be noted that the bottleneck, due to molecular size, is more restrictive here than for ground state calculations. The CIS method is an affordable technique to compute electronic excitations in larger molecules. This approach, however, is currently not accurate enough and a typical error in the calculated excitation energies is larger than 1 eV.

SMs are computationally much less demanding than high-level *ab initio* methods, and they continue to be very useful in modeling excited state properties of biomolecules. The INDO/S method, developed by Ridley and Zerner 40 years ago,⁷¹ is the most popular semi-empirical approach for calculating electronic excitations in extended organic systems. New parameters for oxygen were suggested to improve the results for carbonyl compounds.⁷² The method was extended for several main-group elements and transition metals,^{73,74} INDO/S appears the most accurate in predicting low-lying π - π^* and n - π^* bands of molecules containing first- and second-row elements.⁷⁵ As the standard method predicts triplet excited states too low in energy, the modified scheme⁷⁶ should be used. Since the INDO/S method is not suitable for geometry optimization, the molecular structure should be taken from experiment or

calculated by other computational approaches. In addition to transition energies, INDO/S may be employed to explore charge distribution in both the ground and excited states. Two methods, NDDO/G⁷⁵ and CNDO/L,⁷⁷ were recently elaborated to calculate excited state properties of organic and biological molecules.

CIS calculations using MNDO, AM1, or PM3 Hamiltonians are shown to underestimate excitation energies of organic chromophores.⁷⁸ The results become more accurate by including pair-doubly excited configurations in CI.⁷⁹ Despite the fact that the OM x methods were parameterized for ground state properties, they show good performance for electronically excited states (the reference data for transition energies, oscillator strengths, and excited-state dipole moments are well reproduced) within the CISDTQ scheme (single, double, triple, and quadruple excitations are accounted for in CI).⁷⁸ This approach, however, appears to be computationally demanding for large chromophores.

5.6 Electron and Excitation Energy Transfer

Electron transfer (ET) and excitation energy transfer (EET) reactions within proteins play a key role in bioenergetic pathways and have been extensively studied using computational methods.^{80,81} In the past decade, long-distance ET mediated by DNA has also received considerable theoretical attention.^{82,83} The ET and EET free energy (ΔG) and the electronic coupling (V) are key parameters that determine the efficiency of these processes. Electronic couplings are difficult to derive from experimental data and QM approaches are widely employed to estimate these matrix elements. The *ab initio* CIS and TD DFT methods seem to be of limited use.⁷⁰ CASSCF and MS-CASPT2 calculations provide more reliable estimates of ΔG and V , but they are tractable only for small systems. As thermal fluctuations of the protein and DNA structure affect the electronic couplings considerably the parameters should be estimated by their averaging over many configurations.^{84,85} To take into account the influence of structural dynamics on the efficiency of ET and EET, very efficient quantum chemical methods should be applied. SMs are widely employed to study intra- and inter-molecular ET.^{86,87} AM1 was successfully used to explore fast ET through a polyglycine α -helix.⁸⁸ Recently, the performance of several semi-empirical methods in modeling ET in π -stacks of nucleobases was assessed.⁸⁹ INDO/S is shown to provide good estimates for V and ΔG ; the MNDO, AM1, and PM3 methods give reasonable ΔG values, but fail to predict the couplings in the π stacks. The DFTB scheme performs well when calculating ET parameters.⁹⁰

5.7 Linear Scaling

While SMs are much more efficient than *ab initio* and DFT approaches, one runs into a serious computational problem on application to systems consisting of thousands of atoms. The most time-consuming step is the diagonalization of

the Fock matrix, which scales as N^3 for N basis functions. This problem may be addressed using linear-scaling methods.^{1,91} The basic concept of these approaches is to divide a molecular system into a set of small sub-systems and to solve the eigenvalue problem for these molecular fragments instead of diagonalization of the global Fock matrix. Then, an approximate total electron density matrix is built up from the contributions of the subsystems. The linear scaling methods allow semi-empirical methods to be applied to real-size biological systems consisting of many thousands of atoms.⁹¹ Recently, the LocalSCF method was proposed and implemented for SMs as a computationally inexpensive tool to treat large systems.¹ It was tested for various polypeptides, up to 264 residues. For the largest system, LocalSCF was found to be 160 times faster than the conventional SCF procedure. The calculations also showed that the time dependency is still nearer $N^{1.5}$ than N . The LocalSCF method was compared with the MOZYME scheme⁹² with respect to computational performance and the ability to reproduce matrix diagonalization results.⁹⁵ MOZYME is found to be more accurate in the prediction of total energy, whereas LocalSCF is more reliable in calculating conformational energies, dipole moments, and atomic charges. For a model system, containing more than 20 000 atoms, LocalSCF is three times faster than MOZYME. Both techniques reproduce satisfactorily the target data for the MNDO, AM1, and PM3. Overall, the linear scaling methods allow the calculations of molecular systems up to 10 000 atoms. They can be employed to calculate Mulliken charges, dipole moments, and some other molecular properties, but they are currently impracticable for geometry optimization.

5.8 Concluding Remarks and Perspectives

The major advantage of semi-empirical MO techniques is their speed and versatility, which make them very useful for modeling biomolecules. In cases where one needs to reveal the main features, to screen electronic properties or to analyze the trends, rather than to accurately describe an individual system, semi-empirical schemes are the method of choice. The combination of semi-empirical methods with molecular dynamics allows one to study the effects of structural motion on the properties and functions of proteins and nucleic acids. Computed ET and EET parameters are especially sensitive to conformational fluctuations in biomolecules, and therefore, reliable models of these systems should account for the structural dynamics in nanosecond and longer time scales. Currently, such simulations are affordable only with semi-empirical approaches.

Modern SMs provide quite accurate data for the ground-state energetics and structure of organic molecules. Dispersion corrections of the total energy have considerably improved the performance of the methods for hydrogen bonds and van der Waals interactions. The calculations of transition metal complexes are still not so accurate and further effort here is required. More reliable results may be obtained using specific parameters for specific chemical problems.

The performance of SMs in prediction of activation barriers and transition state structures seems to be inferior to that of more sophisticated calculations. Since SMs suffer from occasional outliers, more effort is required for their comprehensive assessment and corresponding corrections.

Design of new semi-empirical approaches as well as further extension of the existing methods to other elements is needed to obtain more reliable tools for computational modeling of biomolecules. Electron transfer and excitation energy transfer processes play an important role in bioenergetic pathways, hence special schemes elaborated for accurate treatment of these reactions are of great interest. To this end, both ground-state and excited-state properties should be used as reference data for parameterization of the semi-empirical methods. Photochemistry is another important area, where carefully chosen semi-empirical tools should be very helpful.

References

1. W. Thiel, in *Theory and Applications of Computational Chemistry: The First 40 Years*, ed. C. E. Dykstra, K. S. Kim, G. Frenking and G. E. Scuseria, Elsevier, Amsterdam, 2005, p. 559.
2. J. J. P. Stewart, *Rev. Comput. Chem.*, 1990, **1**, 45.
3. M. C. Zerner, *Rev. Comput. Chem.*, 1991, **2**, 313.
4. W. Thiel, *Adv. Chem. Phys.*, 1996, **93**, 703.
5. T. Clark, *J. Mol. Struct. (Theochem)*, 2000, **530**, 1.
6. H. M. Senn and W. Thiel, *Curr. Opin. Struct. Biol.*, 2007, **11**, 182.
7. J. A. Pople and D. L. Beveridge, *Approximate Molecular Orbital Theory*, McGraw-Hill, New York, 1970.
8. G. P. Ford, *J. Mol. Struct. (Theochem)*, 1997, **401**, 253.
9. B. Beck and T. Clark, in *Drug Discovery and Design*, Kluwer, 1998, p. 131.
10. T. Bredow and K. Jug, *Theor. Chem. Acc.*, 2005, **113**, 1.
11. V. Gogonea, D. Suarez, A. van der Vaart and K. M. Merz, *Curr. Opin. Struct. Biol.*, 2001, **11**, 217.
12. K. Raha, M. B. Peters, B. Wang, N. Yu, A. M. Wolla-Cott, L. M. Westerhoff and K. M. Merz, *Drug Discov. Today*, 2007, **12**, 725.
13. K. E. Riley, M. Pitonak, P. Jurecka and P. Hobza, *Chem. Rev.*, 2010, **110**, 5023.
14. M. J. S. Dewar and W. Thiel, *J. Am. Chem. Soc.*, 1977, **99**, 4899.
15. M. J. S. Dewar and W. Thiel, *Theor. Chim. Acta*, 1977, **46**, 89.
16. M. J. S. Dewar, E. G. Zoebisch, E. F. Healy and J. J. P. Stewart, *J. Am. Chem. Soc.*, 1985, **107**, 3902.
17. J. J. P. Stewart, *J. Comput. Chem.*, 1989, **10**, 209.
18. M. P. Repasky, J. Chandrasekhar and W. L. Jorgensen, *J. Comput. Chem.*, 2002, **23**, 1601.
19. I. Tubert-Brohman, C. R. W. Guimaraes and W. L. Jorgensen, *J. Chem. Theory Comput.*, 2005, **1**, 817.
20. I. Tubert-Brohman, C. R. W. Guimaraes and W. L. Jorgensen, *J. Chem. Theory Comput.*, 2005, **1**, 817.

21. I. Tubert-Brohman, C. R. W. Guimaraes and W. L. Jorgensen, *J. Chem. Theory Comput.*, 2005, **1**, 817.
22. W. Thiel, in *Computational Thermochemistry*, ed. K. K. Ikakura and D. J. Frurip, American Chemical Society, Washington, DC, 1998, p. 142.
23. N. Otte, M. Scholten and W. Thiel, *J. Phys. Chem. A*, 2007, **111**, 5751.
24. J. Tirado-Rives and W. L. Jorgensen, *J. Chem. Theory Comput.*, 2008, **4**, 297.
25. W. Thiel and A. A. Voityuk, *Theor. Chim. Acta*, 1992, **75**, 54.
26. W. Thiel and A. A. Voityuk, *J. Phys. Chem.*, 1996, **100**, 616.
27. A. A. Voityuk and N. Rösch, *J. Phys. Chem.*, 2000, **104**, 4089.
28. K. Nam, Q. Cui, J. Gao and D. M. York, *J. Chem. Theory Comput.*, 2007, **3**, 486.
29. A. A. Voityuk, *J. Phys. Chem. A*, 2009, **113**, 11801.
30. J. P. McNamara, M. Sundararajan, I. H. Hillier, J. Ge, A. Campbell and C. Morgado, *J. Comput. Chem.*, 2006, **27**, 1307.
31. I. Tejero, À. González-Lafont and J. M. Lluch, *J. Comput. Chem.*, 2007, **28**, 997.
32. P. Winget, A. H. C. Horn, C. Selçuki, B. Martin and T. Clark, *J. Mol. Model.*, 2003, **9**, 408.
33. P. Winget and T. Clark, *J. Mol. Model.*, 2005, **11**, 439.
34. H. Kayi and T. Clark, *J. Mol. Model.*, 2009, **15**, 295.
35. H. Kayi and T. Clark, *J. Mol. Model.*, 2007, **13**, 965.
36. H. Kayi and T. Clark, *J. Mol. Model.*, 2009, **15**, 1253.
37. H. Kayi and T. Clark, *J. Mol. Model.*, 2010, **16**, 29.
38. H. Kayi and T. Clark, *J. Mol. Model.*, 2010, **16**, 1109.
39. H. Kayi, *J. Mol. Model.*, 2010, **16**, 1029.
40. J. J. P. Stewart, *J. Mol. Model.*, 2007, **13**, 1173.
41. M. Kolb and W. Thiel, *J. Comput. Chem.*, 1993, **14**, 775.
42. W. Weber and W. Thiel, *Theor. Chem. Acc.*, 2000, **103**, 495.
43. M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai and G. Seifert, *Phys. Rev. B*, 1998, **58**, 7260.
44. M. Elstner, *Theor. Chem. Acc.*, 2006, **116**, 316.
45. D. N. Nanda and K. Jug, *Theor. Chim. Acta*, 1980, **57**, 95.
46. B. Ahlswede and K. Jug, *J. Comput. Chem.*, 1999, **20**, 563.
47. N. Ferro and T. Bredow, *J. Comput. Chem.*, 2010, **31**, 1063.
48. A. A. Voityuk and A. A. Bliznyuk, *Theor. Chim. Acta*, 1987, **72**, 223.
49. T. J. Giese, E. C. Sherer, C. J. Cramer and D. M. York, *J. Chem. Theory Comput.*, 2005, **1**, 1275.
50. P. Hobza, F. Hubálek, M. Kabelác, P. Mejzlík, J. Sponer and J. Vondrášek, *Chem. Phys. Lett.*, 1996, **257**, 31.
51. S. Grimme, *J. Comput. Chem.*, 2004, **25**, 1463.
52. J. P. McNamara and I. H. Hillier, *Phys. Chem. Chem. Phys.*, 2007, **9**, 2362.
53. T. Tuttle and W. Thiel, *Phys. Chem. Chem. Phys.*, 2008, **10**, 2159.
54. J. Rezac, J. Fanfrlik, D. Salahub and P. Hobza, *J. Chem. Theory Comput.*, 2009, **5**, 1749.
55. C. J. Cramer and D. G. Truhlar, *Chem. Rev.*, 1999, **99**, 2161.

56. V. V. Vasiyev, A. A. Bliznyuk and A. A. Voityuk, *Int. J. Quant. Chem.*, 1992, **44**, 897.
57. H. M. Senn and W. Thiel, *Angew. Chem., Int. Ed. Engl.*, 2009, **48**, 1198.
58. A. Warshel and M. Levitt, *J. Mol. Biol.*, 1976, **103**, 227.
59. U. C. Singh and P. A. Kollman, *J. Comput. Chem.*, 1986, **7**, 718.
60. M. J. Field, P. A. Bash and M. Karplus, *J. Comput. Chem.*, 1990, **11**, 700.
61. V. Théry, D. Rinaldi, J. L. Rivail, B. Maigret and G. J. Ferenczy, *J. Comput. Chem.*, 1994, **15**, 269.
62. M. A. Thompson, E. D. Glendening and D. Feller, *J. Phys. Chem.*, 1994, **98**, 10465.
63. D. Bakowies and W. Thiel, *J. Phys. Chem.*, 1996, **100**, 10580.
64. M. Svensson, S. Humbel, R. D. J. Froese, T. Matsubara, S. Sieber and K. Morokuma, *J. Phys. Chem.*, 1996, **100**, 19357.
65. F. Maseras and K. Morokuma, *J. Comput. Chem.*, 1995, **16**, 1170.
66. B. O. Roos, K. Andersson, M. P. Fülscher, P. Å. Malmqvist, L. Serrano-Andres, K. Pierloot and M. Merchán, *Adv. Chem. Phys.*, 1996, **93**, 216.
67. H. Koch, O. Christiansen, P. Jørgensen, A. M. S. de Meras and T. Helgaker, *J. Chem. Phys.*, 1997, **106**, 1808.
68. M. Schreiber, M. R. Silva-Junior, S. P. A. Sauer and W. Thiel, *J. Chem. Phys.*, 2008, **128**(134110).
69. D. Jacquemin, V. Wathelet, E. A. Perpète and C. Adamo, *J. Chem. Theory Comput.*, 2009, **5**, 2420.
70. A. Dreuw and M. Head-Gordon, *Chem. Rev.*, 2005, **105**, 4009.
71. J. Ridley and M. C. Zerner, *Theoret. Chim. Acta*, 1973, **32**, 111.
72. J. Li, T. Zhu, C. J. Cramer and D. G. Truhlar, *J. Phys. Chem. A*, 1999, **104**, 2178.
73. M. C. Zerner, G. H. Loew, R. F. Kirchner and U. T. Mueller-Westerhof, *J. Am. Chem. Soc.*, 1980, **102**, 589.
74. W. P. Anderson, W. D. Edwards and M. C. Zerner, *Inorg. Chem.*, 1986, **25**, 2728.
75. A. A. Voityuk, M. C. Zerner and N. Rösch, *J. Phys. Chem. A*, 1999, **103**, 4553.
76. J. Ridley and M. C. Zerner, *Theoret. Chim. Acta*, 1976, **42**, 223.
77. L. A. Montero-Cabrera, U. Röhrig, J. A. Padron-García, R. Crespo-Otero, A. L. Montero-Alejo, J. M. García de la Vega, M. Chergui and U. Röthlisberger, *J. Chem. Phys.*, 2007, **127**, 145102.
78. M. R. Silva-Junior and W. Thiel, *J. Chem. Theory Comput.*, 2010, **6**, 1546.
79. T. Clark and J. Chandrasekhar, *Isr. J. Chem.*, 1994, **33**, 435.
80. H. B. Gray and J. R. Winkler, *Annu. Rev. Biochem.*, 1996, **65**, 537.
81. J. R. Winkler, *Curr. Op. Chem. Biol.*, 2000, **4**, 192.
82. Y. A. Berlin, I. V. Kurnikov, D. N. Bertran, M. A. Ratner and A. L. Burin, *Top. Curr. Chem.*, 2004, **237**, 1.
83. A. A. Voityuk, in *Computational Studies of RNA and DNA*, ed. J. Šponer and F. Lankas, Springer, Dordrecht, 2006, p. 485.
84. I. A. Balabin, D. N. Beratan and S. S. Skourtis, *Phys. Rev. Lett.*, 2008, **101**, 158102.

85. A. A. Voityuk, *J. Chem. Phys.*, 2008, **128**, 115101.
86. S. F. Nelsen and M. D. Newton, *J. Phys. Chem. A*, 2000, **104**, 10023.
87. S. F. Nelsen, M. N. Weaver, A. E. Konradsson, J. P. Telo and T. Clark, *J. Am. Chem. Soc.*, 2004, **126**, 15431.
88. T. Herz, P. Gedeck and T. Clark, *J. Am. Chem. Soc.*, 1999, **121**, 1379.
89. A. A. Voityuk, *Chem. Phys. Lett.*, 2006, **427**, 177.
90. T. Kubar, P. B. Woiczikowski, G. Cuniberti and M. Elstner, *J. Phys. Chem. B*, 2008, **112**, 7937.
91. W. Yang and T. S. Lee, *J. Chem. Phys.*, 1995, **103**, 5674.
92. A. Van der Vaart, V. Gogonea, S. L. Dixon and K. M. Merz, *J. Comput. Chem.*, 2000, **21**, 1494.
93. J. M. Millam and G. E. Scuseria, *J. Chem. Phys.*, 1997, **106**, 5569.
94. J. J. P. Stewart, *Int. J. Quant. Chem.*, 1996, **58**, 133.
95. V. M. Anisimov, V. L. Bugaenko and V. V. Bobrikov, *J. Chem. Theory Comput.*, 2006, **2**, 1685.

CHAPTER 6

Quantum Chemical Topology: on Descriptors, Potentials and Fragments

PAUL L. A. POPELIER

Manchester Interdisciplinary Biocentre (MIB), 131 Princess Street,
Manchester M1 7DN, UK and School of Chemistry, University of
Manchester, Oxford Road, Manchester M13 9PL, UK

6.1 Introduction

Nowadays *ab initio* quantum chemistry features more in drug design than back in the 1990s, the approximate time of its significant uptake. Applied initially by only a handful of enthused and patient visionaries of medicinal chemistry, *ab initio* chemistry is now part of the armoury of standard tools in the pharmaceutical industry. In fact, for the last five years, computational chemists employed in that industry routinely obtain wave functions using a highly affordable PC tucked away under their work desk.

Carrying out accurate *ab initio* calculations remains an ambitious agenda in drug design. The rather sudden appearance of Density Functional Theory (DFT) in quantum chemistry made it possible to include electron correlation in calculations on molecules with up to 50 second-row atoms already in the late 1990s. Such calculations used not to be feasible when only ‘classical’ post Hartree–Fock methods, such as MP2, were available. As drug molecules typically have a molar mass of less than 500 g mol⁻¹, they could immediately benefit from this increased accuracy. Admittedly, very useful benchmark results are appearing¹ at levels of

theory that are probably ‘as good as they will ever be’ for the non-relativistic electronic molecular Schrödinger equation. For example, the single-point inter-molecular interaction energy for the parallel-displaced benzene dimer is $-11.6 \text{ kJ mol}^{-1}$, obtained using CCSD(T) at the estimated complete basis set limit, but only at the MP2/aug-cc-pVTZ optimised inter-monomer distance. Such accurate values for interaction energies are unlikely to change with the advent of the next available hardware or software and, hence, are here to stay. The significance of such ‘perpetual benchmarks’ cannot be underestimated. However, the CCSD(T) method, which is, sometimes, called the gold standard of quantum chemistry, scales very poorly with basis set size, that is, by a power of seven. By the second half of the 2000s, efficient implementations allowed calculations at the CCSD(T) level of theory with a thousand basis functions and more.² Although impressive, this number of basis functions is not enough to embed the aforementioned benzene dimer into a more realistic environment than the vacuum. Nevertheless, it is those larger systems that one needs to be able to calculate properties for, in order to capture the attention of a wider community of scientists and industrialists. Of course, the challenge is much greater because, even if one could obtain future-proof energies and geometries, the dynamic nature of interactions remains elusive. Molecular simulation can deliver such nature but at the cost of invoking only approximate energy evaluation, typically by means of a force field.

The challenges that still exist today may portray the optimism of the 1990s as misplaced. There was much optimism as to what computational chemistry could deliver for drug design. Ironically, now that computational chemistry is closer to this goal, both the molecular modelling industry and the large-scale pharmaceutical industry appear to be shrinking. The best one can strive for is enhancing the realism of the methods and data that computational chemistry can deliver to the research field of drug design. Over the years, we have attempted to do this on three fronts, all inspired by a common method called quantum chemical topology (QCT),^{3–6} which will be introduced briefly below. These three fronts are force field design, bio-isosterism and QCT’s electronic descriptors in quantitative structure activity/property relationships (QSAR/QSPR). This chapter presents a summary of much of the work on these three fronts of investigation. The intention is not to spend equal space on all three subjects. In fact, recently there have been moves to reduce our effort on the QSAR front, most space will be devoted to it here to underline the state of play. Bio-isosterism is the most recent strand of investigation and will be discussed very briefly. The plan is to intensify our research efforts on force field design, which will be reviewed elsewhere in much more detail in the near future. Bio-isosterism needs to be seen in the context of fragment-based drug design and, therefore due to limitations of space, this activity will be discussed more briefly than the other two.

6.2 Quantum Chemical Topology (QCT)

Quantum chemical topology (QCT)⁵ is a branch of quantum chemistry that uses the concepts and language of *dynamical systems* to retrieve insight into

quantum systems and with the ultimate goal of making them predictive.⁷ A detailed justification for its name was given in a previous footnote⁸ and in another appendix⁹ in addition to Section 2 of ref. 6. Dynamical systems¹⁰ constitute a branch of mathematics that largely originates from advanced classical mechanics. Central to dynamical systems is the notion of a *gradient vector field*, which is a collection of gradient paths. A *gradient path* is a curve of steepest ascent, here appearing in a quantum mechanical function of interest. This function can be, for example, an electron density, its Laplacian, an electron localisation function (ELF),¹¹ a molecular electrostatic potential¹² or a virial field.¹³ The employment of gradient paths is the central feature that defines the approach of QCT. They may be traced by following the gradient vector in space, as it changes direction, point by point. The tracing can be achieved by numerical methods such as Runge-Kutta-Cash-Karp.^{14–16} Gradient paths have an orientation, as they propagate from a low function value to higher. A gradient traverses space searching out the highest function value that it will encounter from its initial point. For the vast majority of gradient paths this point is a maximum in the function, which is coined an *attractor*. For example, if the function is an electron density then the attractor coincides with a nucleus for all practical purposes. The portion of space carved out by all gradient paths terminating at the attractor constitutes a topological atom. Hence, the gradient vector field partitions the system into so-called *basins*, which are subspaces, each housing an attractor.

One should also realise that a gradient path always connects two critical points. A critical point is a point in space where the gradient vanishes. Clearly, a maximum is a critical point; it is one of four types of rank 3, *i.e.* no vanishing local curvatures critical points that are possible in three dimensions. The electron density of an isolated molecule shows an ocean of critical points at infinity, which are all minima. The gradient paths that form a topological atom connect these minima at infinity with the maximum, *i.e.* attractor, that dominates the atom. The partitioning into atoms is accomplished without using any reference electron density. This is an important hallmark of QCT. Indeed, by essentially being a partial differential the gradient exposes an *internal* difference in the function that it operates on. In this way, QCT adheres to the key principle of Occam's razor because it does not invoke a reference function. The notion of a gradient path is *minimal* rather than *simple*. The term 'minimal' refers to an economy of assumption that leads to an approach in which an object under study can express itself in all its intrinsic complexity without being squeezed in an *a priori* 'corset' of unnecessary assumptions and approximations. In contrast, the term 'simple' refers to imposing such a 'corset'. Doing so creates artificial problems that may be avoided. A compelling example, of the difference between *simple* and *minimal*, is that of rationalising planetary motion around the Sun. Whereas Ptolemean astronomy attempted to explain observations by means of imposed circular motion, Kepler disposed of these corrections within corrections that this model employed. He introduced the ellipse as a minimal concept that beautifully explained the observations without artificial corrections. Admittedly, an ellipse is not perceived as elegant as the

circle as it lacks a circle's appealing symmetry. However, an ellipse is elegant in a different and deeper way: it captures the essence of the physics of planetary motion.

Returning to the remaining types of critical points, there are two additional types of saddle points in 3D. One is a maximum in two dimensions and a minimum in one dimension, and is called a $(3,-1)$ point for reasons explained elsewhere.⁴ The second type is called $(3,+1)$ and is a minimum in two dimensions and a maximum in the remaining dimension. Note that a $(3,+1)$ has a dual relationship to the $(3,-1)$ in that maxima and minima have been swapped. A gradient path connecting a minimum and a $(3,-1)$ form a *separatrix* between two basins. Again, this term is couched in the language of dynamical systems and refers to the sharp separation between two non-overlapping basins. In the two-dimensional phase portrait of the pendulum spanning angular displacement and angular velocity, there is a separatrix between the two dynamical regimes: swinging and free rotational motion.

Figure 6.1 illustrates the technical terms introduced so far by means of a van der Waals complex. Depicted is an imidazole dimer partitioned into atoms that do not overlap. Instead, the atoms are delineated by sharp boundaries without interatomic gaps. The atoms also exhaust space in that each point in space belongs to an atom. The algorithm¹⁷ that constructs this picture uses simplices rather than cubes that impose an undesirably high symmetry and which are,

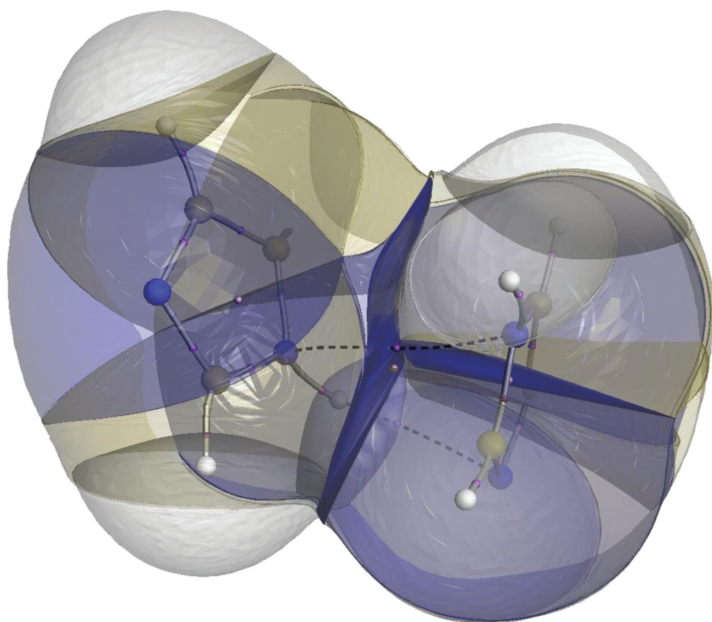


Figure 6.1 A finite-element representation of an imidazole dimer in accordance with quantum chemical topology. The bond critical points are marked with small purple spheres while ring critical points are marked by pink spheres.

therefore, simple rather than minimal. The simplices spatially discretise atoms, in the spirit of finite elements. The whole dimer is surrounded by a constant electron density envelope serving as the outer boundary of atoms constituent of a complex. This boundary is not topological in nature because it is not a separatrix. Instead, it is an arbitrary edge that helps visualising atoms in isolated molecules, and enables them to be allocated a finite volume. However, one can argue that an *isolated* molecule, or van der Waals system, actually does not exist. When travelling away from a molecule one will, sooner or later, encounter another molecule in the Universe. Therefore, one will find a separatrix that separates the original ‘isolated’ molecule from the one just encountered. As a result, all molecules are actually separated from their environment by separatrices, although they may be unreasonably far away. Any molecule, whether in an isolated state or in a condensed matter state, is *completely* bounded by topological boundaries. This is important because it eliminates the artificial need for constant electron density envelopes set at reasonable but actually arbitrary values.

Figure 6.1 shows the atomic volumes over which property densities can be integrated to deliver the corresponding atomic properties. For example, the 3D integration^{18,19} of the electron density over an atomic volume yields the electronic population of that atom. When corrected for the nuclear charge contained within the atom, this population is converted to an atomic charge. If the property density is an electron density multiplied by a vector centred on a nucleus that sweeps a full atomic volume, then 3D integration yields an atomic dipole moment. Higher rank multipole moments will feature in Section 6.4. A (3,−1) critical point in the electron density is called a bond critical point (BCP) and a (3,+1) critical point is called a ring critical point (RCP). Both types of critical points are shown in Figure 6.1. In 2003, a full classification of the nine possible gradient paths that can exist in terms of the two rank 3 critical points they connect was clarified.²⁰ This classification is general and valid for any quantum mechanical function, including an electron density. One type of gradient path originates at a BCP and terminates at a maximum and is marked by a dashed curve in Figure 6.1. The two gradient paths, each originating from the same BCP and terminating at two different nuclei, are collectively called an *atomic interaction line*.

The function of focus determines which aspect of the quantum system is revealed. For example, if the function is the electron density $\rho(\mathbf{r})$, then atoms inside single molecules or molecular complexes, are revealed, as demonstrated above in some detail. If the function is the Laplacian of an electron density, $\nabla^2\rho(\mathbf{r})$, to mention a second example, then one may recover lone pair regions amongst others. This proposal refers to work carried out in the 1980s, which associated critical points in $\nabla^2\rho(\mathbf{r})$ with lone pairs^{21,22} and thereby provided a purely empirical basis for the widely used VSEPR model.²³ However, more recent work^{24–26} focused on the full topology of the Laplacian and went beyond mere critical points. Here, for the first time, basins in the Laplacian were plotted and the electron density integrated over their volume so that their population may be studied.

In 1975, the French mathematician Thom in a monograph²⁷ called *Stability and Morphogenesis* boosted the development of the ‘Quantum Theory of Atoms in Molecules (QTAIM)’,³ which laid the basis of QCT. Thom saw the Universe as a ceaseless creation, evolution and destruction of forms. Inspired by earlier work of another mathematician called Arnold,²⁸ Thom introduced *catastrophe theory* in the 1960s, which became very popular in the 1970s due to the efforts of Zeeman. Catastrophe theory is a branch of bifurcation theory in the study of dynamical systems. Bifurcation theory classifies phenomena characterised by sudden shifts in behaviour arising from small changes in circumstances. Catastrophe theory was applied²⁹ to the topology of $\rho(\mathbf{r})$, which led to universal insight into three-centre systems and the formation of cage structures.

The Silvi group also used catastrophe theory to study³⁰ the evolution of the topology³¹ and rationalised chemical reaction mechanisms by following how electron pairs reorganise. In this case, QCT is manifest in a gradient vector field operating on a three-dimensional function ELF. Therefore, this work, started in the first half of the 1990s,³² resides under the QCT umbrella. So does the work of Joubert *et al.* who combined a QTAIM and ELF analysis³³ to obtain unique insight into the bond-formation process in a prototypical S_N2 reaction.

A further segment of QCT is the topological study of the electrostatic potential,³⁴ another 3D function studied by Gadre’s group.³⁵ Tsirelson *et al.* managed to have the gradient vector field of the electrostatic potential partition space into electrically neutral pseudo-atoms. Blanco *et al.*³⁶ proposed an interacting quantum atoms (IQA) approach, that is a correlated energy decomposition scheme based on QTAIM, and which is preceded by our earlier work.³⁷ IQA offers energetic underpinning for the topological expression of chemical bonding³⁸ by monitoring competition between atom pairs in terms of their exchange energy. Topological energy partitioning also highlights³⁹ how current force fields focus is only on interatomic interaction energies. The importance of self-energy terms were made clear from our energy decomposition of torsional barriers.⁴⁰

The so-called ‘virial field’, defined as the trace of the Schrödinger stress tensor,³ is another 3D function perceived in terms of its gradient vector field¹³ and hence resides under the umbrella of QCT. The penultimate example of a QCT analysis is that of a magnetically induced molecular current distribution.⁴¹ Finally, the gradient vector field of the intracule density reveals correlation cages,⁴² providing new insights into electron–electron interactions in atoms and molecules. This final example concludes the list of QCT analyses mentioned here. Alternative methods of interpretative quantum chemistry^{43–49} do not share the central concept of the gradient vector field. This crucial difference is what draws together the topological analysis of the various functions under the heading of QCT, thus making them distinct from alternative methods.

It should be clear from the above that QCT is more than a population method, although it is often used for this purpose in the wider community. It would be naive to confuse its use for what it is. In fact, those embracing QCT as a now *de facto* method of extracting chemical insight beyond structure are the high-resolution X-ray crystallographers.

6.3 Quantum Topological Molecular Similarity (QTMS)

A full account of the development and achievements of quantum topological molecular similarity (QTMS) can be found in a recent book.⁵⁰ Here is presented a more compact account guided by a number of questions receiving answers from the series of papers our group have published on QTMS over the last decade. The roots of QTMS are in the *ansatz* that the electron density of a molecule is the ultimate source of its properties and that one should therefore compare the electron densities of two different molecules to gauge similarity. If the similarity between two molecules is high in terms of their electron densities then it is expected that they will possess similar high-level properties, such as binding to a protein. This *ansatz* is shared with a method called quantum molecular similarity (QMS), which emerged from the Carbó group in the 1980s. In 1980 they published⁵¹ a seminal paper proposing a measure of molecular similarity based on the electron density. This article was well ahead of its time and became widely cited, probably due to its conceptual boldness and ultimate promise. However, the actual practical implementation of the idea behind this article continued well into the late 1990s, with a reliance on increasingly sophisticated mathematical techniques. The philosophy of the QMS approach, and its various acronyms such as QMSM and MQSM, is supported by Hohenberg and Kohn's existence theorem that states that the non-degenerate ground-state electron density determines the Hamiltonian. If the Hamiltonian of a molecule is known then this molecule's properties may be obtained by virtue of solving the corresponding Schrödinger equation. Carbó *et al.* proposed a similarity index, which is essentially an overlap integral between the respective calculated electron densities of two different molecules normalised for their molecular size. This gauge resembles a correlation index already in use in pattern recognition but Carbó's innovation was the use of, originally semi-empirical and later *ab initio*, electron densities in molecular similarity analysis. This innovation is a product of the driving force of physicalising biological and pharmaceutical science, a general force that has also driven the development of QTMS.

In its original formulation, the practical calculation of the Carbó index required excessive CPU time. Secondly, this approach suffered from multiple maxima in molecular superposition and alignment procedures. Thirdly, it faced the problem of the dominance of the core densities in influencing the numerical value of the similarity index. Finally, the index is a single number and hence obliterates knowledge about the contributions of molecular fragments. The first problem can be tackled by the so-called *atomic shell approximation* (ASA),⁵² which expresses the molecular electron density by a linear combination of Gaussian *s* functions. In order to save even more CPU time, the Carbó group proposed application of the ASA to pro-molecular densities.⁵³ A pro-molecular density is a simple superposition of atomic electron densities. Superficially, molecular and pro-molecular electron densities look extremely similar but this very small difference is exactly what makes a molecule more than a mere

superposition of free unhybridised atoms. If the latter picture was sufficient then it would be pointless to spend many unnecessary CPU hours on a CCSD(T) or more approximate calculation. Therefore, drawing conclusions from an index that operates on pro-molecular electron densities should be met with deep suspicion. The second problem of alignment, in general, cannot be circumvented. This challenge faces the same potential combinatorial explosion that global optimisation algorithms attempt to avoid. However, if a set of molecules has a common skeleton, as in many QSARs, then there is a one-to-one mapping between the atoms or bonds of two different molecules. In this case it is straightforward to superimpose molecules. The third problem of core densities was circumvented by reformulating the similarity index with electrostatic potentials,⁵⁴ valence electrons⁵⁵ or momentum densities.⁵⁶ The fourth problem, namely that of the global nature of the similarity index, was addressed⁵⁷ within the QMS context after QTMS had successfully solved it.⁵⁸ In fact, the way QTMS works precludes this problem, so it is more accurate to state that QTMS actually did not have to solve this problem. Then how does QTMS work and where did it all start?

In 1995, the concept of the *BCP space* was launched⁵⁹ and tested on the drug haloperidol. Before worrying about how to superimpose respective molecular electron densities of two molecules, one should ask if the calculation of their molecular similarity requires the *full* electron density. As explained above (see Section 6.2), BCPs are special points in space, remote from the nuclei that they connect but still fully dependent of their presence, both in terms of location and core density. Locating these BCPs is a fast and robust process.⁶⁰ Now, imagine an abstract space in which each axis represents a different quantum mechanical function that is evaluated at a given BCP. This space can be two-dimensional, for example, and spanned by the electron density ρ and the ellipticity, which is denoted by ϵ . The latter is obtained as a ratio of two local curvatures and serves as a simple shape descriptor of electronic structure, for instance, expressing the π -character of a C–C bond. Phenol conveniently illustrates this 2D BCP space, as shown in Figure 6.2.

Phenol's 13 bonds are represented by 13 different points in BCP space. Since the OH group breaks the symmetry of a mirror plane orthogonal to the molecular plane, no two BCP points coincide: each bond has a unique electron density and ellipticity. The C–C bonds have the greatest ellipticity, which is unsurprising as they are part of the π -system in the aromatic ring. The C–O bond has the greatest electron density of all possible bonds in phenol. The fine structure of this pattern would change if another substituent were added or if the OH group were replaced by another functional group. There are many bonds that share the same degree of ellipticity. For example, there are C–H bonds in phenol that have practically the same ellipticity as that of the C–O bond. If one continuously altered the bond length of C–H then one could surely match the ellipticity of C–O exactly. Therefore, bonds cannot be characterised uniquely in only one dimension. However, in two dimensions, the precise value of two BCP properties, such as the ellipticity and the electron density, already fully determine a bond. A 2D property space 'locks in' a given bond, expressing

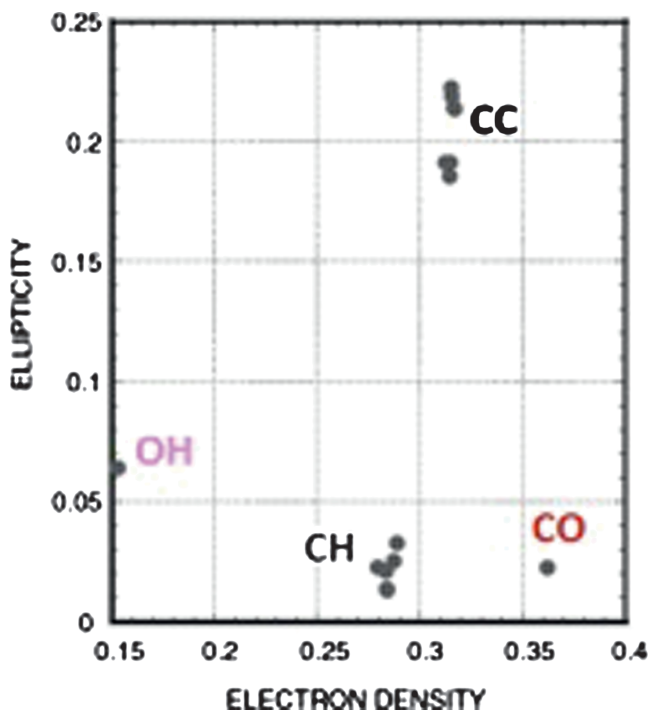


Figure 6.2 Depiction of phenol in a two-dimensional BCP space. The ellipticity ϵ (dimensionless) is plotted against electron density ρ (in a.u.) evaluated at each of the 13 BCPs.

all the subtle substitution effects that a bond may experience. The idea of BCP space can be readily extended to an arbitrarily high dimension. For example, the Laplacian of the electron density, denoted $\nabla^2\rho$, is a useful property to join the two properties already discussed to construct a 3D BCP space. The Laplacian gauges the local concentration of electron density and classifies bonds as closed-shell or shared interactions.^{61,62}

The drug haloperidol can be thought of as essentially consisting of four fragments, that when completed to molecules become fluorobenzene, chlorobenzene, butyraldehyde and 4-hydroxypiperidine. All in, haloperidol may be represented by 51 bonds in this 3D BCP space.^{58,59} The $C_{\text{arom}}-C_{\text{arom}}$ bonds clustered in BCP space give rise to two subclusters, the smaller one representing the two pairs of phenyl C–C bonds adjacent to the C–F or the C–Cl bond. These four bonds exhibit somewhat greater ellipticity, which is due to the fact that halogens are π -donors. The ellipticity even detects that fluorine is a stronger π -donor than chlorine. One should be aware that there is no magic about the very position of a BCP. We noticed early on that properties evaluated in the vicinity of the BCP could also be used to characterise a bond. However, the BCP position does not need parameters to define it. In contrast, a bond midpoint, for example, is determined by a positional parameter arbitrarily set

to one-half. BCPs, in contrast, are self-defining and appear naturally in the gradient vector field, without need for parametrisation. BCPs are minimal while bond midpoints are simple.

An early verification of BCP space attempted to reproduce a ranking according to molecular similarity applied to a simple set of substituted benzenes, containing benzene, aniline, nitrobenzene, *para*-nitroaniline, *meta*-nitroaniline and *ortho*-nitroaniline. The ranking according to BCP space disagreed with that of a study⁶³ using the Carbó molecular similarity index r_{AB} . As no experiment could resolve this discrepancy, verification of BCP space required a case where experiment could.⁵⁸ Attention focused on one of the oldest QSARs and a bedrock of classical physical organic chemistry: Hammett's substituent constant σ . If a similarity index formulated in BCP space can be related to σ , then this would vindicate the concept of BCP space. This indeed proved so, as demonstrated by work⁵⁸ on substituted benzoic acids, acting as a typical set of congeneric molecules sharing a common skeleton. A convenient measure of similarity between molecules A and B starts with the Euclidean distance between two points in BCP space. It is important that the components of this distance are first standardised such that they are dimensionless because it is impossible to add physical quantities that have different units. The distance d_{ij} between two BCPs i and j in 3D BCP space is defined as the square root of the sum of the squares of the differences between the three properties evaluated at BCP i and BCP j . Subsequently, the distance $d(A,B)$ between two molecules A and B is defined as a sum of the distances d_{ij} associated with the two molecules A and B. Of course, the lower the value $d(A,B)$, the more similar the molecules A and B are. Two important questions arise in connection with the exact formulation of the sum leading to $d(A,B)$: Which BCPs of molecule A should be compared to which BCPs of molecule B? and How many BCPs should be included in this sum? An answer to the first question is that a BCP in molecule A should be compared with the corresponding BCP in molecule B, according to the evident one-to-one mapping that exists between two molecules with a common skeleton. For example, the $C_{\text{ipso}}-C_{\text{ortho}}$ BCP in *para*-chlorobenzoic acid, say *syn* with respect to the COOH substituent at C_{ipso} , must be matched with the $C_{\text{ipso}}-C_{\text{ortho}}$ BCP in *para*-nitrobenzoic acid. An answer to the second question led to an interesting feature of QTMS, namely that it highlights the 'locus of action' in the molecule. As explained below in detail, the statistics of regression of distances in BCP space against experimental data highlight the part of the molecule that best relates to experimental data. For example, the σ constants are essentially pK_a values and therefore relate to the dissociation of this acidic group. Therefore, in substituted benzoic acids one expects the COOH group to be the locus of action. The sum resulting in $d(A,B)$ does then only contain three d_{ij} values: one for O-H, one for C=O and one for C-O. In the analysis of a small set of *para*-substituted benzoic acids the NH_2 substituent was taken as a reference because it has the lowest value for σ (it is the most electron-donating group of all). A straightforward plot of σ versus $d(\text{NH}_2, S)$, where S represents an arbitrary substituent, yielded a correlation coefficient r^2 of 0.993. To further test QTMS, five more *para*-substituted benzoic acids were

examined, with S being set to COCH_3 , CHO , phenyl, OH , or O^- . QTMS correctly predicts that COCH_3 and CHO are bracketed by CN and Cl for all experimental values found. Experiment does not provide a single value but a range of values, albeit narrow, as systems are revisited by updated experimental techniques. In this regard it has been pleasing to find that more recent experimental data matched the prediction of BCP space better than older data for the phenyl substituent. Finally, the OH substituent is equally correctly bracketed by OCH_3 and CH_3 , which is not easy to verify experimentally due to experimental difficulties only resolved in 1960, more than two decades after Hammett's pioneering work. The case of O^- is interesting since it offers a prospect beyond interpolation, namely that of extrapolation, a holy grail of QSAR. Indeed, the experimental σ value puts O^- 'left of NH_2 ', *i.e.* this substituent has a σ value lower than that of NH_2 . QTMS correctly predicted that O^- is even more electron-donating than NH_2 . Further success was achieved when studying a small set of *meta*-substituted benzoic acids.

This important case study⁵⁸ showed that quantum similarity measures using the full electron density were unnecessarily cumbersome and biased by chemically unimportant regions. The location of BCPs and the evaluation of their properties are computationally inexpensive. Molecular alignment can be bypassed if a mapping between molecules is clear, as is the case when a common skeleton can be identified. BCPs occur in regions of the electron density that are more sensitive to bonding characteristics than core regions. Their properties depend on the accuracy of the whole molecular electron density. Secondly, QTMS revealed that an experimental activity sequence will only be reproduced if the distance measure is confined to contributions from *BCPs associated with the locus of action only*. This finding remains intact in later statistically more sophisticated treatments, as discussed below. This conclusion is independent of the exact definition of the Euclidean distance because those treatments do not invoke this distance.

It is pertinent here to discuss the role of the bond length in QTMS. It should be emphasised that QTMS and BCP space operate on optimised molecular geometries. According to unpublished evidence, QSAR models built from unoptimised geometries show very poor correlation coefficients. The equilibrium bond lengths, R_e , can themselves be used as descriptors in QTMS. Later, it emerged that various respectable QSARs could be generated with optimised bond lengths alone. This observation gives rise to the question *whether the BCP properties, such as ρ or its Laplacian, actually add any information or value over and above bond lengths*. A subsequent QTMS study⁶⁴ attempted to address this question by investigating possibly linear relationships between BCP properties and bond lengths. For that purpose a data set of 57 molecules was designed, based on single conformations of the 20 natural amino acids and smaller derived molecules, yielding 731 BCPs. It transpired that *local* linear relationships did indeed emerge in a set of bonds provided there was little variation in the chemical surroundings of the bonds. If a wider variety of bonds is allowed then the correlation between the BCP properties and the bond length breaks down dramatically. From this lack of correlation one may safely conclude that

BCP properties cannot be trivially recovered or even predicted by knowledge of bond length alone. So, in general, one should view the equilibrium bond length as another BCP property, even if this is perhaps somewhat artificial at first sight. This is how we discovered the value of *ab initio* optimised bond length in QSAR. Curiously, bond length does not seem to feature much, if at all, as a descriptor in the QSAR literature. Work from our group submitted for publication shows that bond lengths on their own are excellent descriptors for prediction of pK_a .

The simple ranking procedure and linear regression used in the early stage of QTMS was subsequently replaced by partial least squares (PLS).^{65,66} Essentially, PLS is a linear regression method allied with principal component (PC) analysis.⁶⁷ The latter is an unsupervised technique that constructs new variables from the original, raw, variables such that the variation of the data along the dimensions spanned by the new variables may be exposed. This analysis identifies variables that describe the internal data structure compactly by revealing the variables with the largest variation. Similarly, PC analysis facilitates omission of variables associated with little variation. The PCs in PLS are called latent variables (LVs). The dependent variable, an activity of physical property of interest, is then expressed as a linear combination of the LVs in the PLS equation. The first LV will explain most of the variance in the independent set, the second LV the next largest and so on. SIMCA-P prescribes a criterion for the significance of an LV. If $q^2 < 0.097$ then the LV is not significant and no more LVs are computed. The PLS regression is then deemed complete.

QTMS work⁶⁸ adopted PLS methodology to obtain regression equations for benzoic acids in terms of LVs constructed from BCP descriptors. The dependent variables were again Hammett σ constants and the independent variables consisted of ρ , $\nabla^2\rho$, λ_1 , λ_2 , λ_3 , ϵ , K and G . The latter two quantities are kinetic energy densities evaluated at the BCPs. When integrated over a topological atom the values for K and for G become identical and are trivially related to the total atomic energy. The λ_i symbol represents an eigenvalue of the Hessian of the electron density ρ with respect to the Cartesian coordinates describing ρ . The eigenvalues express the coordinate-independent intrinsic curvature of the electron density around a BCP. The signs of these eigenvalues are used to classify the four possible types of critical points that can occur in 3D. The fact that $\nabla^2\rho = \lambda_1 + \lambda_2 + \lambda_3$ is not a problem for PLS. Not only can PLS cope with many variables, but also with variables connected through a linear combination. In summary, the whole molecular set is described by $8n$ variables, where n is the number of bonds present in the common molecular skeleton.

The correlation coefficient r^2 and the cross-validated correlation coefficient⁶⁷ q^2 indicate the quality of the PLS equation. Randomisation of the dependent variables ('Y') determines if the likelihood of the correlation occurred by chance. If good correlations are obtained with incorrect data, the initial regression may be due to chance factors alone. Randomisation is a standard option in the SIMCA-P computer package,⁶⁹ which may also be used for the PLS analysis itself. Details of the randomisation procedure are best referred to

in the first QTMS work⁶⁸ that used PLS, or directly in this program's manual. Assessment of the QSAR's validity is essential before any results can be interpreted in a physical sense.

The more elaborate PLS handling of *para*-benzoic acids recovered the earlier results, discussed above, these being obtained by means of a more elementary statistical treatment. Furthermore, QTMS was also successfully applied⁶⁸ to other carboxylic acid systems, such *para*-phenylacetic acid, 4-X-bicyclo-[2.2.2]octane-1-carboxylic acids and even polysubstituted benzoic acids. All *ab initio* optimisations were carried out at the following levels of theory: AM1//AM1, HF/3-21G(d)//HF/3-21G(d), HF/6-31G(d)//HF/6-31G(d), B3LYP/6-311 + G(2d,p)//HF/6-31G(d) and B3LYP/6-311 + G(2d,p)//B3LYP/6-311 + G(2d,p). In order to generate a consistent series of results throughout all QTMS studies from our group to date, these levels were conveniently labelled from A to E, respectively.

The question arises of how PLS highlights the locus of action in a molecule. It has been shown in the first PLS study⁶⁸ that the locus of action in benzoic acids may be recovered. Hence, *a priori* knowledge of a centre of action for any given activity is unnecessary⁷⁰ with QTMS. *Variable importance in the projection* (VIP) values⁷¹ may also be employed. The VIP value indicates the relative importance of each independent variable in the regression. Hence, factors that contribute substantially to any fit have high VIP scores. A typical profile of descending VIP values will show a sequence of BCP descriptors that jump from one BCP to another. In other words, the ranking of VIP values does not respect locality of the descriptors. For example, the highest VIP value may correspond to the Laplacian of the 5th BCP, while the next highest value corresponds to the ellipticity of the 11th BCP. These jumps, from one location in the molecule to another, make the highlighting of the locus of action difficult. A PC analysis, carried out by the program SPSS,⁷² can help in localising BCP descriptors by reducing their number and condensing them into PCs. Typically only one, two or at maximum three PCs are sufficient to describe the quantum information at a BCP. PLS will then deliver a VIP plot, which highlights the BCPs important to explain the variation in a given, i.e. measured or observed, activity or property. These variables may now be associated directly with a region of electron density that is recognised as a bond. The variable reduction, introduced by a PC analysis, makes it easier to localise the locus of action. In summary, the ('raw') BCP descriptors are used to carry out the regression for the purpose of *prediction*. On the other hand, a reduced set of variables arising from a PC analysis of localised BCP properties is used to *interpret* the PLS analysis.

It happens that the highest VIPs may contain other parts of the molecule that cannot be readily associated with the mode of activity. However, it is believed that QTMS has sufficient suggestive power to highlight the part of the molecule responsible for the measured response and below an illustrative example is covered. In some cases this highlighting occurs with high contrast and that a small number of VIP variables dominate a decreasing sequence of variables with much lower values. Often, however, the VIP plot reveals a rather slowly

decreasing profile throughout. In this case, the diffuse nature of the profile makes it harder to discriminate a locus of action from the rest of the molecule. Finally, it should be pointed out that so-called ‘contaminations’ to the process may occur. For example, some levels of theory highlight the C_{ortho}–H BCP as more important than the C=O BCP in COOH. However, it is satisfying that all five levels of theory, A to E as above, rank the carboxylic O–H bond as the greatest VIP value. This is the bond that is broken on dissociation of the acid, which is expressed by its pK_a value and hence by a Hammett acidity constant.

A regression technique, such as PLS, demands an independent, or ‘X’ variable is sampled in the same relative location in each molecule of the dataset. For example, in a set of substituted anilines the bond between the nitrogen of the amino group and the phenyl carbon to which it is attached is readily identified in each molecule of the full data set. In such a congeneric set of molecules, one bond occurring in one molecule can be unambiguously mapped onto the same bond in another molecule, the reason being that bonds are marked by their position in the molecule and by the elements they connect. For example, in a set of meta-substituted benzoic acids, the carboxylic C–O bond appears in each molecule and is unique, unless one of the substituents is a carboxylic group so that two carboxyl groups are attached to the phenyl ring. Therefore, it is straightforward to map the C–O BCP in one molecule to the C–O BCP in another. In a strict sense, the common skeleton in a series of congeneric molecules is a subset of atoms, with the same atomic number in each molecule and topologically connected to the same atoms, again with the same atomic number.

The requirement of this strict version of a common skeleton, such as in *para*-benzoic acids, was relaxed in a subsequent QTMS study.⁷³ This relaxation enabled the PLS analysis of a larger and more diverse set of aromatic carboxylic acids. For example, with a small modification, *p*-phenylacetic acids can be added to a set of *para*- and *meta*-substituted benzoic acids. The former acids contain a methylene group that cannot be brought into correspondence with any fragment within the benzoic acids. This is why the methylene hydrogens (see Figure 6.3) do not have numerical labels. What is more, the methylene carbon is labelled as C₅ because it takes on the role of the atom to which the carboxyl group is bonded, corresponding to the C₅ atom in the benzoic acids. As a result, it is still possible to map 15 bonds of the *p*-phenylacetic acids to the corresponding bonds in the benzoic acids. Figure 6.3 illustrates a colour-coded plot, ‘rainbow plot’, indicating the influence of the bonds through their VIP value in explaining the observed acidity of a set of benzoic and phenylacetic acids. The active site has the highest influence, red, followed by yellow > green > blue > purple > black. Variables with a VIP value less than unity can be discarded as unimportant and are marked in black. Bonds with a VIP value of more than unity are assigned a colour in accord with the visible spectrum. In other words, if at least one of the PCs that represent a bond has a VIP value larger than unity it is assigned a colour other than black and subsequent linearly spaced intervals of decreasing VIP values are assigned to yellow, green, blue and violet, respectively.

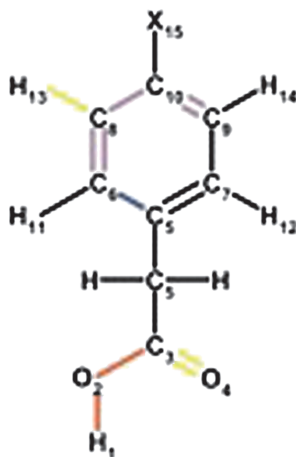


Figure 6.3 Rainbow plot illustrating the influence of the various bonds in the observed acidity of all benzoic and phenylacetic acids. The active site has the highest influence (red). Legend: red > yellow > green > blue > purple > black.

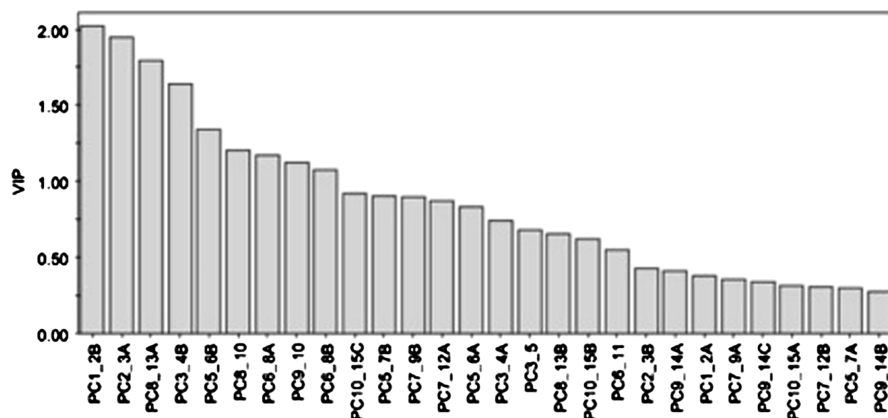


Figure 6.4 Variable importance in the projection (VIP) plot for the PLS regression of *para*-, *meta*-benzoic acids and *para*-phenylacetic acids.

The VIP plot (see Figure 6.4) demonstrates that the PCs associated with three of four of the highest VIP values describe the COOH group. One PC corresponds to C₈–H₁₃, a C–H bond meta to the carboxyl group, which unfortunately contaminates the locus of action. There are 28 PCs and each PC represents topological information furnished by any of the 15 bonds of the common skeleton. Some bonds are characterised by more than one PC. For example, C₉–H₁₄ appears as three different PCs, A, B and C. The large number of carboxylic acids appears to ‘sharpen’ the VIP plot towards the high end, compared to plots for a smaller set of carboxylic acids. It is hypothesised that a

larger data set delivers a more pronounced locus of action. Finally, we point out that it is remarkable that molecular information obtained in the gas phase can predict Hammett acidity constants, which are derived from acid–base equilibria and hence implicitly include solvation effects.

With a locus of action being sufficiently pronounced, it should be possible to investigate if this locus alters its position in the molecule when the observed activity, the ‘Y variable’, changes. This was shown⁷⁴ to be the case early on in the development of QTMS. Full details, only summarised here, may be consulted in Ref. 50. Damborsky and Schultz studied⁷⁵ the experimentally determined biodegradability and toxicity of *para*-substituted phenols. The second-order rate constant of the compound’s oxidation (k_b) yielded values for the biodegradability while toxicity was represented by the 50% growth inhibitory concentration (IGC₅₀). Due to their availability we also added the pK_a values and the logarithm of the octanol–water partitioning coefficient, $\log K_{ow}$, as ‘Y variables’. VIP plots in the style of Figure 6.4, one for each Y variable, may be constructed, as in Ref. 50, but they are not reproduced here. If the Y variable is set to the pK_a value then the locus of action transpires to be the phenolic O–H bond, closely followed by the C–O bond. The former bond is indeed the one that breaks in the deprotonation of the acid–base reaction. Unfortunately, in the VIP plot for pK_a plot the VIP values of C_{ortho} –H and C_{meta} –H are still relatively high, closely following the VIP values of O–H and C–O. The $\log K_{ow}$ VIP plot highlights the C–X bond attaching the substituent X to the phenyl group. The bonds with the next highest VIP values are C–C bonds in the phenyl ring, starting with those adjacent to the C–X bond. These bonds have much reduced VIP influence. The partition coefficient is not a property involving bond breaking or making, so it is unsurprising that the only bond clearly highlighted (C–X) is not in the common skeleton. In the $\log k_b$ VIP plot a C_{ortho} – C_{meta} bond features prominently. Whether the oxygenase enzyme in charge of this biodegradation interacts strongly with this bond while oxygenating the phenols is an open-ended question. It would be rewarding if QTMS predicted this correctly. Finally, in the $\log(1/IGC_{50})$ VIP plot the highlighted bonds are C_{meta} –H and C_{ortho} –H, which again awaits an experimental verification. It is promising that these four observed properties, ‘Y variables’, can all be modelled by the same group of descriptors, all of which arose from only 13 points in the electron density. In summary, as the Y variable against which is regressed changes, different regions of the molecular electron density are being highlighted each time, that is, the locus of action changes. Without *a priori* knowledge of the systems or mechanisms of action, QTMS manages to reproduce the properties of interest.

The success of QTMS in the province of classical physical organic chemistry, in particular the classic Hammett QSARs, inspired an excursion⁷⁶ into medicinal chemistry. Traditional Chinese medicine identified the α,β -unsaturated ketone, (*E*)-1-(4-hydroxyphenyl)but-1-en-3-one, as an anti-tumour agent. Ducki *et al.*⁷⁷ synthesised a series of substituted analogues in the hope to increase the parent compound’s moderate anti-tumour activity. We carried out a PLS analysis of the IC₅₀, the Y variable here, against BCP properties

supplemented with $\log P$ values for 17 substituted ketones. A validated regression emerged with $r^2=0.91$ and $q^2=0.86$. QTMS highlighted five bonds as the locus of action since their corresponding VIP values dominated. These five bonds encompassed the three bonds in the enone system, $C=C-C=O$, in addition to the bond attaching it to the terminal methyl group and the bond attaching the enone system to the phenyl ring. The experimental group of Ducki *et al.* surmised that a Michael addition was responsible for the mode of activity. It is rewarding that QTMS independently supported this hypothesis. An analysis of Mulliken charges, however, provided a QSAR with little predictability and pointed at the terminal methyl group as the locus of action followed by $\log P$.

The ever-increasing movement to use QSAR in toxicology led us to explore QTMS's potential in an environmental context. A set of well-known molecules, the polychlorinated dibenzo-*p*-dioxins (PCDDs),⁷³ provided an attractive test ground given their core molecular rigidity. The influence of conformational flexibility remained an unexplored complication on the influence of the substituents. PCDDs span a wide range of toxic effects most of which involve binding the aromatic hydrocarbon receptor (AhR). QTMS was applied to predict three different activities (pEC_{50}) of a set of 13 ecologically relevant PCDDs. As the activity refers to the ability to bind to a receptor, the locus of action is not involved in bond breaking as in the case of carboxylic acids. In retrospect, this explains why the QTMS regression is mediocre. A more complete analysis should have introduced QCT atomic properties such as monopole, dipole or quadrupole, and higher, moments as they are the generators of atomic electrostatic potentials.⁷⁸ Still, using BCP properties, the locus of action focussed near the lateral C–Cl bonds. This conclusion overlaps with deductions made by Bonati *et al.* in their electrostatic (MEP) study⁷⁹ of the ligand–receptor recognition step. Our results complement the CoMFA work of Waller and McKinney⁸⁰ in which they point out the importance of lateral halogen substitution on overall stereoelectronic structure. They claim that molecular polarisability is most affected by the lateral halogens. However, further research would be welcome into the convergence of various results on these systems from different analyses including QTMS.

A further QTMS study⁸¹ helped restore confidence in the ability of QTMS to work successfully. A very good and statistically validated QSAR was obtained for base-promoted hydrolysis rate constants for a set of 40 esters, including acetates, benzoates and ethyl-substituted esters, *i.e.* $R-(C=O)-O-CH_2CH_3$. Without the need for molecular superposition, a variety of models was constructed, which all point toward the molecular fragment $(O=C)-C-O$ being most significant to explain the range of hydrolysis rate constants. When the model is confined to this locus of action then $r^2=0.93$ and $q^2=0.86$ is obtained for all 40 esters using BCP properties evaluated with bond lengths optimised simply at HF/6-31G* level. This work is relevant for environmental exposure and risk analysis. This successful application of the QTMS method demonstrates the lack of requirement to measure infrared frequencies,⁸² which itself has been introduced as a faster method to circumvent the measurement of rate constants.

The prediction of pK_a values presents an opportune case study to further test QTMS, given its success in predicting Hammett acidity constants and the general importance of pK_a . This physical quantity features in pharmacokinetic properties and more generally in the chemical industry, which has an interest in knowledge of the environmental fate or hazard of compounds. A subsequent QTMS study⁸³ suggested strong QSAR models for a set of aliphatic carboxylic acids, anilines, and phenols in estimating their pK_a values. These results improved on previous attempts at modelling pK_a by others. One should note that pK_a values of the 36 anilines actually refer to their conjugate acids, but pK_a is used as a measure of the amine's basicity as $pK_a + pK_b = pK_w$ (where K_w is the ionisation constant of water). We obtained progressively better correlation statistics with increasing level of theory. For example, the q^2 value obtained was 0.95, just using bond lengths at B3LYP/6-311 + G(2d,p) level.

In a subsequent study the classic test case of the 'steroid set' that Cramer *et al.*⁸⁴ used to launch their CoMFA analysis was investigated. Wagener *et al.*⁸⁵ corrected the multiple errors in topological coding and stereochemistry that Cramer *et al.* introduced. In the QTMS analysis,⁸⁶ AM1 and HF/3-21 G* bond lengths were used in conjunction with PLS and a genetic algorithm (GA) to predict the corticosteroid-binding globulin (CBG) binding activity. First, an initial PLS model was constructed incorporating all descriptors. Then a GA selected the optimum number of descriptors for use in a subsequent PLS analysis. A new PLS model was then obtained using the descriptors selected by the GA. Good r^2 and q^2 values were observed and the notorious steroid 'number 31', 2 α -methyl-9 α -fluorocortisol, encouragingly, was found not to be an outlier. However, a critical eye could question to what extent the *precise* knowledge of bond length is actually required. For example, the C–O bond length appearing in the 6-membered 'A' ring of the common skeleton has a high VIP value but hardly varies between certain steroids. The large difference between double C=O bond lengths and single C–O bond lengths is most likely the actual reason for the correct prediction of activity. Again, this test case involves only binding rather than a reaction, so it is not obvious why an 'interior property' such as bond length or BCP property would actually be predictive. The same QTMS study⁸⁶ also focused on nitrofuran derivatives, which are used as antibacterial and anticancer agents. From the AM1 bond lengths a model with two LVs was generated for antibacterial activity, expressed as $\log(1/IC_{50})$, against *S. aureus*, yielding $r^2 = 0.92$ and $q^2 = 0.78$. Another model, for the *C. crescentus*, also passed the randomisation test and delivered $r^2 = 0.96$ and $q^2 = 0.90$. The rainbow plot, not shown here, displayed a dramatic difference between the locus of action for the two types of bacteria. The researchers who published the experimental data failed to respond to an invitation to comment on this conclusion.

Further studies employing QTMS⁸⁷ involved mutagenic activity. Halogenated hydroxyfuranones, mutagen-X, derivatives and triazenes were examined. QTMS delivered an interesting result in the case of the triazenes. Shusterman *et al.* proposed⁸⁸ two alternatives for the, still obscure, hydroxylation mechanism that is thought to create mutagens. Path A involves a two-step process,

where an electron is first transferred from the triazene to the enzyme cytochrome P-450. Then the resulting radical cation loses a proton to give a triazene-substituted methyl radical. The alternative path B is a one-step process in which this radical is formed by direct abstraction of a hydrogen atom. Previously published QSAR equations⁸⁸ were unable to distinguish paths A and B. Based on analysis of the VIP plot, not shown, QTMS suggests that pathway A may be favoured over pathway B. This example shows again how QTMS engages with experiment and arrives at a conclusion not reached by alternative means.

Given the success of QTMS in predicting Hammett acidity constants, the next QTMS study⁸⁹ explored a wider class of Hammett constants. It transpired that QCT descriptors can also substitute σ^+ and σ^- based on a reported work on *para*-substituted phenols, substituted toluenes and bromophenethylamines. Meaningful loci of action were highlighted. The evidence accumulated in this work suggested that QCT descriptors can most likely provide a reliable model to predict any type of Hammett constant. A valid model can be obtained without invoking Hammett constants, thereby removing concerns about which type of constant to invoke or whether the given substituent's constant is in a database.

At this point hopefully sufficient evidence has been mustered to conclude that QTMS is able to produce good QSARs for only one of the three familiar classes of descriptors; QTMS captures the electronic effects and not the steric nor the hydrophobic effects. As a consequence of this clear discriminatory capacity, QTMS is able to discriminate, based on quantum chemical electron densities, whether a given activity is due to steric effects or electronic effects. Surely, the QTMS descriptors can be combined with externally provided steric and/or hydrophobicity parameters, but QTMS does not generate them independently. Armed with the knowledge that QTMS can tell if a measured quantity, 'Y variable', depends on electronic effects, an unusual set of *ortho* alkyl substituted phenols was tackled in a subsequent QTMS study.⁹⁰ These phenols are known for their cytotoxicity and have previously been investigated by the Hansch group.⁹¹ QTMS results did not support their conclusion that a steric factor was important in the determination of the cytotoxicity. In fact, QTMS deemed any steric contribution or influence to be absent. The addition of the Taft steric parameter or the hydrophobic parameter to the bond length descriptors did not improve the models constructed at any of the three levels of theory used.

Further QTMS work⁹² on phenols followed, this time the focus was on computing a model of the relative substituent effects on the bond dissociation enthalpies (Δ BDEs) for a set of 39 compounds. The data set included a diverse set of substituents with mono-substituted and poly-substituted derivatives that exhibited different electronic and steric effects. For substituents in the *para* position the model required a correction for radical stabilisation enthalpy. The C–O bond adjacent to the phenolic hydroxyl group was selected as the most important bond represented by the VIP plots for both the bond length and BCP models.

Subsequently, seven datasets of medicinal interest were investigated.⁹³ They consisted of: (i) the influenza inhibition constants for a set of benzimidazoles,

(ii) the interaction constants for a set of amides and the enzyme liver alcohol dehydrogenase, (iii) pK_a values for a set of substituted imidazolines and (iv) imidazoles, (v) the ability of a set of indole derivatives to displace [3H]flunitrazepam from binding to bovine cortical membranes, (vi) the natriuretic activity of sulfonamide carbonic anhydrase inhibitors and (vii) the toxicity of a series of benzyl alcohols. A PLS analysis in conjunction with a genetic algorithm delivered excellent models. These case studies, in addition, were able to highlight the locus of action within the molecule whose structure determines the activity, *i.e.* the ligand rather than the associated enzyme.

At the expense of interrupting the QTMS chronology studies, six out of these seven sets were revisited more recently by Hemmateenejad and co-workers⁹⁴ by using the multiway data analysis method called molecular maps (MOLMAP). A 3D array of QTMS descriptors is transferred into new 2D parameters using Kohonen networks, followed by PLS. The method improved the original statistics. Moreover, VIP plots confirmed the original findings⁹³ about loci of action. The same workers revisited⁹⁵ this dataset in the context of their recently proposed substituent electronic descriptors. It should be mentioned that Kohonen networks had featured earlier in the context of QTMS⁹⁶ in an attempt to confront the notoriously difficult QSAR of 1,4-dihydropyridine calcium channel blockers.

Validating a QSAR by means of an external validation set is recognised as a more stringent test than cross-validation. A pair of QTMS studies did exactly that: using external test sets to assess the substituent effects on interaction energies of hydrogen-bonded DNA Watson–Crick base pairs. Figure 6.5 shows the typical geometries of these base pairs. One study⁹⁷ reported on the capacity of BCPs of *monomeric* base pairs, gas phase, to predict BSSE corrected *ab initio* interaction energies of cytosine derivatives substituted in position C₅ and in position C₆. For the C^{6X} dataset, the best model yielded $r^2=0.93$ and a

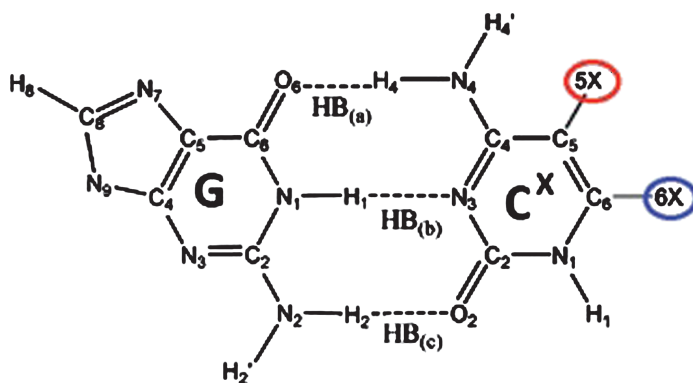


Figure 6.5 Schematic geometry and labelling of the guanine (G) and substituted cytosine (C^X) Watson–Crick base pairs. In the set of van der Waals complexes with unsubstituted cytosine monomers the guanine is substituted in the C₈ position.

root-mean-square (RMS) energy of 0.5 kJ mol^{-1} for the 28 complexes in the training set. This model was evaluated by an external test set (14 complexes), yielding an r^2 value of 0.96 and an RMS error of 0.4 kJ mol^{-1} . A model based on NPA charges does not achieve this high linear correlation for $\text{C}^{\text{6X}}\text{:G}$. For the C^{5X} dataset, the QCT descriptors generated a linear model, with r^2 values of 0.92 and 0.97 and RMS values of 1.7 and 1.2 kJ mol^{-1} for the training set (31 compounds) and the external test set (11 compounds), respectively. A second study⁹⁸ focused on a rich variety of substituents, this time on guanine. Guanine was substituted in the C_8 position, not marked in Figure 6.5 but clear from the context, and again formed van der Waals complexes with cytosine, which is now always unsubstituted. The model, trained on 37 training set complexes, was tested by an external test set of 9 complexes, yielding $r^2 = 0.99$ and an RMS value of 0.2 kJ mol^{-1} . The results indicate that the bonds $\text{C}_6=\text{O}_6$ and N_2-H_2 at the hydrogen-bonded frontier of the guanine derivatives play an important role in transmitting the substituent effects.

Subsequent QTMS studies focused further on *external* predictability but in a statistically more sophisticated manner. Phenols are important in nutrition and medicine, given their cytotoxic potential. In addition they are prevalent in the environment and are likely to elicit often unknown ecotoxic effects. One study⁹⁹ focused on their hepatotoxicity, calculating the QTMS descriptors at five different levels of theory. Preliminary modelling highlighted three compounds as outliers. It was reassuring to find out that they had a mode of action different to that of the non-outliers. New statistical measures were introduced in order to assess external validation. A predictive coefficient R^2_{pred} was calculated via Equation (6.1):

$$R^2_{\text{pred}} = 1 - \frac{\sum (Y_{\text{obs}} - Y_{\text{pred}})^2}{\sum (Y_{\text{obs}} - \bar{Y}_{\text{Training}})^2} \quad (6.1)$$

where Y_{obs} and Y_{pred} respectively represent the observed and predicted toxicity values of the test set compounds, while $\bar{Y}_{\text{Training}}$ is the mean observed value of the training set. Secondly, the squared regression coefficient (r^2) between observed and predicted values of the test set compounds may conceal that the predicted values differ much from the observed activity values, in spite of a good overall intercorrelation. To better indicate the external predictive capacity of a model, a modified r^2 term (r_m^2), defined before¹⁰⁰ is given in Equation (6.2):

$$r_m^2 = r^2 \left(1 - \sqrt{r^2 - r_0^2} \right) \quad (6.2)$$

The coefficient r_0^2 in Equation (6.2) represents the squared correlation coefficient between the observed and predicted values of the test set compounds when setting the intercept to zero.¹⁰¹ Predicted values will be very close to observed activity values for excellent external prediction. In that case, the r^2 value will be very near to the r_0^2 value and r_m^2 will be almost equal to r^2 . The lower boundary of r_m^2 is zero, which is the worst possible scenario. This QTMS

study on phenol hepatotoxicity demonstrated the value of QTMS descriptors because they increased the external predictability over and above that of PLS models using pK_a and Hammett σ^+ .

In view of the widespread industrial use of nitroaromatics and their consequent ecotoxicological hazard potential, QTMS explored¹⁰² predictive models for their toxicity to *S. cerevisiae*. QTMS descriptors were used to complement the deficiency of E_{LUMO} in setting up predictive QSAR models from the viewpoint of external validation. The nitro group was identified as the locus of action.

A third QTMS study¹⁰³ that benefitted from more rigorous external validation focused on pK_a prediction of phenols in seven different solvents, that is, water, dimethyl sulfoxide, methanol, dimethylformamide (DMF), acetonitrile, isopropanol and *t*-butanol. For some solvents the number of test set compounds was considerably smaller, and hence the regression-based external validation parameters R^2_{pred} , r^2 and r_m^2 were less appropriate. Therefore, an additional parameter, the root mean square error of prediction (RMSEP) was introduced and calculated according to Equation (6.3):

$$RMSEP = \sqrt{\frac{(Y_{obs} - Y_{pred})^2}{N_{Test}}} \quad (6.3)$$

where N_{Test} indicates the number of test set compounds. For the development of the PLS models, a hierarchical method was adopted. Good predictive models were obtained, except for the solvent isopropanol. On considering all seven solvents, the equilibrium bond lengths gave consistently good results whereas ellipticity descriptors were consistently of less importance. Moreover, VIP plots for the best models highlighted the importance of the C–O bond connecting the phenolic oxygen to the aromatic ring.

The prediction of pK_a received much more attention from QTMS, given its enormous importance in both life sciences and chemistry. Indeed, the propensity of a compound to donate or accept a proton is fundamental to understanding chemical and biological processes. In a sizeable QTMS study,¹⁰⁴ global and subset models were created for 228 carboxylic acids using PLS, support vector machines (SVM) and radial basis function neural networks (RBFNN). The models were extensively validated using 4-, 7-, and 10-fold cross-validation, with the validation sets selected based on systematic and random sampling. The proposed method compared favourably to ACD/Laboratories, VCCLAB, SPARC and ChemAxon's pK_a prediction software. Out of all the methods, QTMS has the lowest RMSE for the *para*- and *meta*-substituted benzoic acids and aliphatic carboxylic acids but has the highest RMSE for the *ortho*-substituted benzoic acids. Even more elaborate, but as yet unpublished, QTMS studies followed on carboxylic acids, anilines and phenols, using only equilibrium bond lengths. This time, *ortho*-substituted compounds delivered better statistics but at the cost of the globality of the model.

In two recent studies, published back to back, we showed that a single descriptor, *i.e.* an *ab initio* bond length, is able to predict pK_a . The emphasis was placed on model simplicity and a demonstration that more accurate predictions emerge from single bond length models. A data set of 171 phenols was studied.¹⁰⁵ The carbon–oxygen bond length, connecting the OH to the phenyl ring, consistently provided accurate predictions. In this study the pK_a of *meta*- and *para*-substituted phenols was predicted by a single bond length model within 0.50 log units. However, accurate prediction of the pK_a of *ortho*-substituted phenols necessitated their splitting into groups referred to as *high-correlation subsets* in which the pK_a of the compounds strongly correlated with a single bond length. The highly compound-specific single bond length models produced better predictions than models constructed with more compounds and more bond lengths. Outliers were readily identified using single bond length models and in most cases we were able to determine the reason for any outlier discrepancy.

The prediction of pK_a from a single *ab initio* bond length was then extended¹⁰⁶ to provide equations for benzoic acids and anilines. The HF/6-31G(d) level of theory was used for all geometry optimisations. Similarly to phenols, the *meta*-/ *para*-substituted benzoic acid pK_a values could be predicted from a single model constructed from one bond length. This model delivered an impressive RMSEP value of 0.13 pK_a units. The prediction of *ortho*-substituted benzoic acids required the identification of high-correlation subsets, where the compounds in the same subset have at least one of the same, *e.g.* halogens, hydroxy or *ortho*, substituents. Two pK_a equations were provided for *o*-halogen benzoic acids and *o*-hydroxybenzoic acids, where the RMSEP values were 0.19 and 0.15 pK_a units, respectively.

Related to pK_a prediction is a QTMS study on the basicities of pyridines in the gas phase and in aqueous solution.¹⁰⁷ PLS and the machine-learning technique Kriging generated validated models for 125 pyridine derivatives. The *meta* and *para* set was appropriately described by PLS and Kriging, whereas the *ortho* set was only properly described using Kriging. The prediction of the basicities was found to be more problematic in aqueous solution than in the gas phase. In general, the descriptors derived from the neutral or protonated molecules provided predictive models of similar quality.

Finally, a QTMS study¹⁰⁸ on the toxicity of 77 aromatic aldehydes to the unicellular protozoan *T. pyriformis* was divided into a training set ($n = 58$) and a test ($n = 19$) set, and 58 models were constructed using PLS and genetic partial least squares (G/PLS). For both PLS and G/PLS, the models built at the HF/6-31G(d) level show better predictivity, based on overall prediction, than the models developed at any of the other five levels of theory. This identified compound lipophilicity as having the highest VIP value.

A complementary overview⁵⁰ on the development of QTMS was recently published by the present author. This account contains some general reflections on the ‘expensive ways of counting carbon’ and on pro-molecular electron densities speeding up the evaluation of molecular similarity indices. Both pitfalls are discussed at length in that account.

6.4 Force Field Development: QCTFF

Unless quantum computing suddenly becomes a practical reality, there will be a continuing need for force fields for decades to come. Only the use of force fields enables the prediction of the structure and dynamics of proteins in solution with present-day computer technology. Even taking into account the respectable five-fold gain typically associated with emerging GPU technology, *ab initio* algorithms cannot deliver the performance levels required. An otherwise efficiently implemented algorithm to calculate the energy and forces of a simple system such as the imidazole dimer at the MP2/aug-cc-pVTZ level of theory costs more just over a full CPU day when computed in parallel on two processors using a well-known *ab initio* program, ‘that no man shall pronounce’. How could one then perform a molecular dynamics (MD) simulation over 5 nanoseconds of a system of, say, a hundred imidazole molecules surrounded by thousands of water molecules, to mention a modest example? Even if the evaluation of energy and forces for each time step of typically 0.5 fs could be executed in only one CPU hour, then such a simulation would still take more than a thousand CPU years.

The first question is, what is the current state-of-the-art of force fields? This is a question that could drive a whole review but rather than addressing this here it is more useful to convey a broad impression. There is a large community of researchers who only use force fields, often *via* commercial packages with impressive interfaces and stunning graphics. While users of course benefit from such seductive ‘mouse-click environments’ they are shielded from the tough science behind the interaction potential that drives their simulation and the sampling procedures. There is an inevitably much smaller research community that worries about the hard reality behind the glossy pictures and smooth graphical user interface (GUI). What is more, this community is split into a group that focuses on the potential and another distinct group that focuses on sampling. There does not appear to be a group or individual that researches both.

The community that focuses on the interaction potential comprises possibly only a dozen groups worldwide. This is a relatively small research effort given the pivotal role that a force field plays in the reliability of predicted system properties. To this author it is unclear why not more funding is invested into fundamentally improving force fields. In fact, when practically unlimited funding is available, as in the case of the D. E. Shaw venture, then only about three researchers out of a work force of about fifty are concentrating on the potential. The main thrust of their research effort is making longer and larger simulations possible, with either standard or merely tweaked potentials. Curiously, challenging the foundations of current popular force fields is not seen as a priority. Ironically, longer simulations are more prone in exposing fundamental flaws in a force field at the level of the whole system. These flaws could have been exposed already by validating a given force field against *ab initio* calculations on a small part of the whole system. Perhaps referees and funding councils do not find fundamental research on potentials ‘sexy’ probably because it is not an emergent activity with the usual promise, can-do mentality

and excitement caused by surprise? Instead, this type of research, however, aims at getting something right for the right reasons. This may not be ‘sexy’ in the stringent sense of short-lived satisfaction, but perhaps it is ‘sexy’ in a deeper sense that long-term success and satisfaction bring about.

There is, however, enough readily available computer power to enable a fundamental overhaul of the electrostatic contribution to a biomolecular potential. Below, a possible scenario for this aim will be outlined and illustrated with work on-going in our laboratory. With the increasing ‘physicalisation’ of the life sciences comes a responsibility of the biomolecular modelling community to make sure it invests into properly designed potentials. Only then will experimentalists take seriously the output of this community as a trustworthy and independent source of information. One should make no mistake: chemistry is a science of small energy differences. An error of a kilojoule per mole is already a sizeable quantity in the arena of weak non-covalent interactions that hold biological systems together, especially if this error is accumulated adding fragment–fragment interaction energies.

In partial summary then, the vibrant area of molecular simulation continues to make major strides in increasing system size and simulation length, serving an ever growing community of beneficiaries, including experimentalists and industry. While more and more ambitious problems are tackled at accelerated pace, it is a ‘suppressed truth’ that the results ultimately depend on the quality of the potential. Fortunately, some research groups^{109–118} continue to improve potentials.

David Case, one of the developers of the popular force field AMBER, urged *further research into the accuracy of force-field potentials, particularly electrostatic terms*.¹¹⁹ In fact, the basic architecture of force fields has not changed much since the 1980s, or even before. Point charges still dominate the electrostatic interaction in spite of well-known inaccuracies at short range. In his textbook,¹²⁰ Jensen echoes Case’s concern where he writes on p.43 that ‘*Obtaining a good description of the electrostatic interaction between molecules, or between different parts of the same molecule, is one of the big problems in force field work*’. Jensen lists four main deficiencies of modelling the electrostatic energy by fixed atomic charges. A study¹²¹ in 2009 showed that the failure of a simulation of a fast-folding human protein was due to force-field inaccuracy rather than insufficient sampling.

Multipole moments deliver a future-proof strategy to tackle electrostatic interaction. They embody the rigorous expression of the electron density that generates the electrostatic potential. Point charges can only approximate the short-range interaction that occurs in condensed matter. Evidence in favour of multipole moments has been accumulated over the years.^{122–125} In their extensive study¹²⁶ comparing various electrostatic force fields applied in organic crystals in the context of polymorphism, Brodersen *et al.* wrote, ‘*The multipole approach was found to be superior over charge models for more or less rigid molecules . . . , the use of multipole models leads to an improvement in the description of inter-molecular interactions.*’

The second important requirement for a potential to be more realistic is the inclusion of polarisation effects. Polarisation is so important that, in 2007, a

special volume¹²⁷ was dedicated to it in the *Journal of Chemical Theory and Computation*. The multitude of polarisation methods can be categorised in three broad approaches:¹²⁸ (i) Drude oscillators, which describe electronic induction by the movement of a fictitious mass bonded to polarisable atoms via a stiff spring; (ii) fluctuating charges, which consists of changing the charge of the atoms along the simulation by partial equalisation of electronegativities; and (iii) distributed polarisabilities to model induction energies, following a self-consistent procedure to determine the induced dipoles. If forced to categorise our method under any of these three, it resembles most the latter approach but differs in important ways, as made clear below. The three broad approaches previously mentioned are actually based on old ideas, which although at one time enabled a feasible computational route, now impose unnecessary constraints and assumptions. It is our believe that the time is ripe to embark on a more progressive approach, which embraces the increase in computing power by more than five orders of magnitude since the 1980s. This enables one to drastically reshape the electrostatics within an interaction potential from a point charge with a fictitious spring to an atom fully equipped with an electron density and endowed with precise knowledge on how to change its multipole moment, up to high rank, in response to a changing environment.

Conceptually, polarisation is still very much treated within the context of perturbation theory. This is an old and natural approach if the interacting molecules are far apart. The 0th-order energy term is the sum of the monomer energies, the 1st-order energy term is the electrostatic interaction and the 2nd-order energy term consists of mutual induction and dispersion energy terms. The source of the difficulties encountered at short range is the extrapolation of the isolated molecule to short range. In other words, at long range a free molecule is well-defined in practice as an 'isolated' quantum system but at short range the interacting molecule becomes part of an entangled quantum system. At this range molecules stop owning their own electrons for all practical purposes and it is problematic to discern the original isolated molecules in Hilbert space. However, at short range, molecules overlap considerably in terms of their wave functions. This seriously challenges the perturbation approach due to the need for antisymmetrisation between the interacting molecules. As a result, we cannot unambiguously assign an order to a term. A related difficulty is that there can be no 0th-order Hamiltonian that has antisymmetrised wavefunction products as eigenstates. Hence the standard Rayleigh–Schrödinger perturbation theory can no longer be used and a multitude of perturbation treatments have been proposed.¹²⁹ Moreover, electron correlation is not included,¹³⁰ or if it is it leads to an extremely complicated formalism¹³¹ and is found to strongly affect repulsion. The question is then whether one should abandon the concept of a monomeric wavefunction at short range, and instead embrace supermolecular wavefunctions. This is what was carried out about a decade ago, focusing on supermolecular electron densities, which are sufficient to model the Coulomb interaction. When two molecules come close to each other, the boundaries between the molecules become unclear if one holds on to the *image* of isolated molecule. By doing so the weakness in the extrapolation of an isolated molecule to short range becomes apparent. However, QCT offers a

way to delineate individual molecules interacting in a van der Waals complex at any range, short or long. QCT defines a molecule as a collection of topological atoms that never overlap.

Returning to the state-of-the-art of force fields, when considering aqueous ion simulations there are also shortcomings. Although there are more than 50 water potentials,¹³² including polarisable, flexible and dissociable ones, the development of accurate aqueous ion potentials is limited. Many are *effective potentials*, incorporating the effects of polarisation implicitly as part of the parameterisation. These limitations have been noted when inhomogeneous systems are considered, such as interfaces and narrow channels.^{133–135} These ion potentials proposed are limited by the simple *ad hoc* models that have also been developed for water.

Before progressing onto more detail on radically different work on polarisable electrostatics, it is worth summing up the four cornerstones on which this approach is based. The first cornerstone is the *partitioning method* followed to define an atom in a system. There are two possible pitfalls here. First one may not even be aware that a partitioning operates in the definition of atomic charges. Secondly, if one is aware then one must make a decision as to which partitioning method to use. Some research, typically at the user end of biomolecular modelling, can be naive by speaking of ‘the’ atomic charge, blissfully unaware that there are about a dozen ways to define an atomic charge. Moreover, some of these papers do not even cite the paper explaining the method behind the atomic charge. Let us not forget that it is the electron density that generates the electrostatic potential. Hence, if one wants to know the electrostatic potential generated by an atom then one has to define the electron density of that atom. This question distils to how to partition the electron density of the full system into atomic electron densities. In the so-called CHELP¹³⁶ procedure this question is hidden. In this method atomic charges are fitted so as to reproduce a given molecular electrostatic potential in the best least-squares sense, at a grid of several hundred points. The only constraint, through a Lagrangian multiplier, to this non-linear least-squares fit is that the sum of the fitted charges reproduce the total molecular charge. It transpires¹³⁷ that this procedure does not unambiguously define atomic charges. Singular value decompositions of the least-squares matrices show that, in general, the least-squares matrix for this fitting problem will be rank deficient. This means that statistically valid charges cannot be assigned to all the atoms in a given molecule. In a typical small biomolecule such as acetylcholine about half the atomic charges may safely be set to zero. It is clear that this mathematical procedure, even if correctly applied, has no physical or chemical reality. Setting the charge of an electronegative atom to zero clearly violates its electronegative nature. In our work we have chosen to proceed with a QCT partitioning method. The QCT monopole moment is defined (see Equation 6.4) by a 3D volume integral:

$$Q_{oo}(\Omega) = -e \int_{\Omega} \rho(r) d\tau \quad (6.4)$$

where ρ is the electron density of the molecule to which the atom belongs and Ω denotes the *atomic basin* or portion of Cartesian space that the atom adopts. To convert this monopole moment to an atomic charge one simply adds the positive nuclear charge inside the atomic basin to it. Early work¹³⁸ showed that a QCT atomic charge cannot accurately reproduce the exact electrostatic potential, due to the atom involved, on grid points of the water-accessible surface. But then, the QCT atomic charge has not been designed to accomplish this. A QCT charge is not fitted but reflects what it really is: an allocation of electron density to an atom. The atomic charge is essentially an atom's zeroth moment, brutally summarising how much charge it houses, no more no less. At large distance this description is increasingly more accurate and in the limit exact. However, at medium and short range, one needs higher rank multipole moments to define an exact electrostatic potential correctly. This is indeed what happens when QCT, the dipole, quadrupole and octupole moments are added to the multipole series expansion that represents the electrostatic potential. For C_α in alanine the root mean square deviation between the exact and multipole-predicted atomic electrostatic potential is only 0.1 kJ mol^{-1} . The same work¹³⁸ shows that formal convergence of the multipole expansion can be reached if enough high-rank multipole moments are included. With increasing rank, each multipole moment delivers a more detailed description of the electron density of the atom. Ultimately, the description captures enough detail of the shape of the electron density in order to predict the electrostatic potential correctly.

The second cornerstone is *transferability*. Transferability is the idea that information obtained for a small system can be transferred to a large system, without much change. In other words, large systems can be described by piecing together fragments for which accurate information such as geometry, energy and electron density is available. Transferability is an old and very important idea that lies at the foundation of the whole of chemistry. Without it chemistry would be a hopeless science because each aggregate of atoms would have completely different properties, disconnected from the properties of fragments of this aggregate. The transferability of a fragment is embodied and expressed in the concept of a functional group. This concept expresses the locality of a subsystem by allowing it to be defined in a useful and practical way.

The fact that force fields exist in the first place is due to transferability. Therefore, it may appear natural to designate transferability as the first cornerstone. However, a moment of thought makes clear that one cannot define and gauge transferability quantitatively without first having a partitioning method. This is why partitioning is designated as the first cornerstone. Perfect transferability is an unattainable limit but one may quantify it for practical purposes. For the first time, the transferability of topological atoms were quantified from atom-atom (multipolar) electrostatic interaction energies.¹³⁹ The water trimer and pentahydrated serine were used as proof of principle. The atomic multipole moments in serine turned out to be more transferable than those of the water cluster, with the exception of the hydroxyl group. Energy is the ultimate physical quantity that quantifies transferability and hence enables the assessment of the degree of transferability. The latter can be expressed in

terms of an *atom type*. An atom type is essentially a ‘transferable packet of information’. It is important to *detect* transferability rather than *impose* it, which seems to be common practice in the construction up of popular force fields such as AMBER. Early work in our lab computed atom types for the time,^{140,141} based on atomic properties. This work¹⁴² exposed weaknesses in the assignment of atom types in popular force fields such as AMBER. The electrostatic potential generated by an atom can also be used to assess transferability.¹⁴¹ The potential acts as a ‘half-way house’ between atomic multipole moments, used to compute atom types, and interatomic electrostatic interaction. The long-range effect of conjugation could be measured in terms of energetic transferability using retinal as a test system, which contains an eight-carbon linearly conjugated chain.

The third cornerstone is the use of multipole moments. They arise naturally in the Taylor expansion of the exact electrostatic interaction between two atoms. The monopole–monopole interaction is only the first term of this expansion. Each atom is described by a set of multipole moments relative to a local frame $\{x, y, z\}$. The electrostatic energy associated with the interaction of the two atoms a and b is then given by Equation (6.5):

$$U_{elec}^{ab} = \sum_t \sum_u Q_t^a T_{tu}^{ab} Q_u^b \quad (6.5)$$

where T is a purely geometrical interaction tensor and Q_t^a is the t -th multipole moment centred at the nucleus of atom a . Following the notation in Ref. 143 we can define, in the manner of Equation (6.4), the x component of the dipole moment via Equation (6.6):

$$Q_{11c}(\Omega) = -e \int_{\Omega} x \rho(r) d\tau \quad (6.6)$$

or one of the five quadrupole moments via Equation (6.7):

$$Q_{20}(\Omega) = -e \int_{\Omega} \frac{1}{2} (3z^2 - r^2) \rho(r) d\tau \quad (6.7)$$

The spherical tensors introduced here are less intuitive than their Cartesian counterparts but they are more compact because of their irreducibility. For example, there are only five quadrupole moments instead of the six possible Cartesian ones. The discrepancy between number of components becomes more dramatic with increasing multipole rank ℓ . In the spherical tensor formalism there are only $2\ell + 1$ multipole moments, *i.e.* components. Thus, the full atomic electron density can be compressed into only a handful of numbers, *i.e.* multipole components. Admittedly, this handful is more than a single point charge positioned on the atom’s nucleus but the point charge is inherently inaccurate at short and medium range. People have corrected for this well-known inaccuracy by adding more off-nuclear point charges. For example,

whereas once the TIP3P potential modelled water by only three point charges, TIP4P increased this number by one extra charge on oxygen, while TIP5P has two extra point charges representing the oxygen lone pairs in water. What is rarely emphasised is that extra point charges cause a huge computational overhead because their long-range behaviour leads to a large number of charge–charge interactions that scales with the square of the number of point charges in the system.

Equation (6.8) illustrates the complexity of the interaction tensor T , expressing the quadrupole–quadrupole element $T_{20,20}^{ab}$:

$$T_{20,20}^{ab} = \frac{3}{16\pi\epsilon_0} R^{-9} \left\{ 35(\mathbf{z}_a \bullet \mathbf{R})^2 (\mathbf{z}_b \bullet \mathbf{R})^2 - 5R^2 [(\mathbf{z}_a \bullet \mathbf{R})^2 (\mathbf{z}_b \bullet \mathbf{R})^2] \right. \\ \left. - 20R^2 (\mathbf{z}_a \bullet \mathbf{R})(\mathbf{z}_b \bullet \mathbf{R})(\mathbf{z}_a \bullet \mathbf{z}_b) + 2R^4 (\mathbf{z}_a \bullet \mathbf{z}_b)^2 + R^4 \right\} \quad (6.8)$$

where $\mathbf{R}$ is the *inter-site vector* linking the nuclear positions, R is its magnitude and $\mathbf{z}$ is the vector representing the z axis of the local frame (or axis system) cantered on atom a or b . The various elements of the interaction tensor capture the orientation of the axis system of atom a relative to that of atom b , and their inter-site distance. The FLOP count of such an expression is of course larger than that associated with two interacting point charges, but the interaction between two quadrupole moments drops off with a R^{-5} power. This means that this high FLOP count only occurs locally. In general, the interaction rank L equals $\ell_A + \ell_B + 1$. For example, the interaction rank L between a dipole moment ($\ell_A = 1$) and a quadrupole moment ($\ell_B = 2$) is 4, so that this interaction behaves as $R^{-L} = R^{-4}$. Two point charges interact according to rank $L = 0 + 0 + 1 = 1$, explaining their $1/R$ long-range behaviour previously mentioned.

The price paid for the multipolar series expansion, compared to the totally impractical unexpanded and exact energy, is possible lack of convergence. Discussing this important technical matter is outside the scope of this chapter but our group spent much time studying this topic. A companion study⁷⁸ to the one mentioned above showed that the convergence behaviour of the actual electron density inside an atomic basin is due to its decay rather than to the atom's shape. A subsequent article³⁷ studied the convergence behaviour of intramolecular Coulomb interaction between atoms in a set of small molecules, as well as van der Waals complexes described by a single wave function. Another study¹⁴⁴ also addressed this for van der Waals complexes, but now described by monomeric wave functions. Using the shift method¹⁴⁵ we showed very good convergence behaviour for 1,3 and 1,4 interactions¹⁴⁶ and defined pseudo-convergence. This work culminated in a study¹⁴⁷ attacking the convergence problem for 1,2 interactions. Along this path emerged work¹⁴⁸ that allowed the electrostatic potential to be evaluated accurately at short range, by introducing continuous multipole moments called *Bessel moments*. A second distinct idea improved convergence by swapping the allocation of coordinates to the familiar $r_<$ and $r_>$ variables in the Taylor expansion of $1/|\mathbf{r} - \mathbf{r}'|$. This work delivered¹⁴⁹ so-called *inverse moments*. Finally, the

multipole expansion of the exchange energy can also be expanded in a multipolar series.¹⁵⁰ This series converges faster than the Coulomb energy discussed so far, and generates what we called *exchange moments*, which are calculated with explicit reference to molecular orbitals. We have demonstrated that the corresponding expansion converges quickly. A final comment on convergence concerns the use of multipole moments in a protein, which is our ultimate goal, ideally in the presence of a drug ligand. For medium-range interactions we were able to set up a function¹⁵¹ that relates atom type, interatomic distance, interaction rank L and energy error, of the multipole energy compared to the exact one. That paper answered questions such as ‘*How close can two atoms get before their interaction energy diverges?*’ or ‘*Which is the lowest (and therefore cheapest) rank necessary to achieve a certain error in energy?*’ Elaborate work, using the smallest natural protein crambin, addressed these and other questions, concerning the level of theory, for example, and will be submitted for publication shortly.

The fourth and final cornerstone is the use of machine learning to model polarisation as well as charge transfer, both on the same footing. Originally we worked within the context of long-range Rayleigh–Schrödinger perturbation theory and studied charge transfer along the hydrogen bond in the water dimer via QCT distributed polarisabilities.¹⁵² Later the perturbation approach was abandoned in favour of capturing polarisation directly from supermolecular calculations. The main idea is to create a *direct mapping* between a given atomic multipole moment and the coordinates of its neighbouring atoms.

To illustrate this, imagine a water molecule surrounded by four water molecules in liquid water. Figure 6.6 shows an image, produced by an in-house¹⁷ GUI, of this central water molecule in a water pentamer. This particular configuration was taken from a MD, rigid body, simulation of liquid water under ambient conditions using a multipolar electrostatic potential.¹⁵³

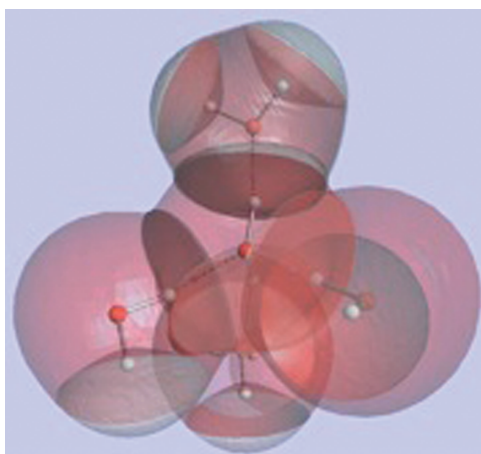


Figure 6.6 QCT partitioning of a water pentamer sampled from a multipolar MD simulation without polarisation. The topological atoms occupy finite volumes.

As the four water molecules neighbouring the central water change position and orientation, the multipole moments of the atoms in the central water molecule change in concert. The machine learning method receives the coordinates of the neighbouring waters as input and is trained to predict, as output, the multipole moments of atoms in the central water. In this approach there is no explicit polarisability, such as α for dipolar polarisability. Instead, the polarisability is implicitly accounted for in the mapping that the machine learning method constructs. The mapping can be a set of trained weights if the machine learning method is a neural network. We began with this method and showed¹⁵⁴ that it is fast and accurate for small water clusters up to a tetramer.

The performance of neural networks starts to deteriorate from the water tetramer onwards because they cannot handle high-dimensional spaces well, especially if combined with large variations in the input variables. The variation in intramolecular coordinates is smaller in the intermolecular case because in the former case atoms are more restricted in their motion. This is why neural networks operate well¹⁵⁵ for intramolecular polarisation, as demonstrated on the test cases glycine and *N*-methylacetamide. A more recent study compared multilayer perceptrons, i.e. the classical artificial neural networks, mentioned above, radial basis function neural networks and Kriging.¹⁵⁶ The latter method, which is also known as Gaussian process regression,¹⁵⁷ proved to be the most accurate but also the most costly in terms of training time. However, since computers are guaranteed to become cheaper in the future, we decided to proceed with Kriging in the development of the QCT force field, QCTFF.

Figure 6.7 schematically illustrates the main idea of how polarisation is tackled in a QCTFF. A number of molecular geometries, typically of the order of a hundred, is obtained by distorting the nuclear skeleton away from a local

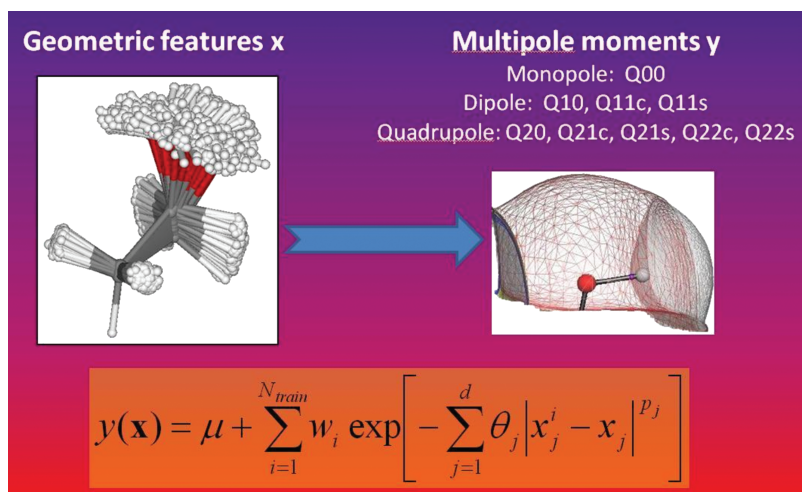


Figure 6.7 Schematic diagram illustrating Kriging modelling intramolecular polarisation.

equilibrium, here one of ethanol. A normal mode analysis is useful to guide this distortion along ‘sensible’ lines. After training, the multipole moments of the oxygen, as an example in this figure, can be predicted by the Kriging equation (marked in orange). In this formula, μ is the mean of the observed values, N_{train} is the number of examples used to build the model, w_i is the Kriging weight, computed from the correlation matrix, for example i , d is the number of (geometric) features (or dimensionality of the interpolation space) and θ_j and p_j are the j -th components of the Kriging vector parameters θ and $\mathbf{p}$. The same approach is used for intermolecular interactions, where the generator of the geometries is an MD simulation rather than a normal mode analysis. After training, Kriging can then predict, for example, the direction and magnitude of the dipole moment on oxygen in water, depending on the positions of all four surrounding waters (see Figure 6.6).

The principle behind Kriging is that it models observations as realisations of an underlying probabilistic model. In our case, the observations are values of a particular atomic multipole moment. This probabilistic model is a *probability density function* that returns the probability of observing a set of values of a particular multipole moment. Two multipole values arising from different water cluster configurations, for example, are correlated through the difference in coordinates of the configurations. The parameters of the probability density function are chosen to be most consistent with the training data, i.e. they are chosen to maximise *the likelihood of the model given the observed data*.

Kriging offers four important advantages.

- The coordinates, i.e. ‘features’ in machine learning language, are ranked according to importance in the construction of the Kriging model. This provides physical insight as well as a secure way to eliminate coordinates, ‘feature selection’.
- The performance of Kriging, in terms of interaction energies, scales linearly with the dimension of feature space.
- Kriging delivers the trained model as an analytical formula, linking the multipole moment with the cluster coordinates. Hence, the interaction energy is analytical and can be differentiated. Hence, forces can be computed quickly and accurately.
- The knowledge of how an environment influences an atom is stored in Kriging parameters. Hence, there is no need to perform iterative calculations to self-consistency during a simulation in which the polarisation potential is used.

Kriging delivers unprecedented accuracy. For example, 55% of the water pentamer configurations in an *external* test set of a 1000 water clusters show an error in the total electrostatic interaction of less than 4.2 kJ mol^{-1} (1 kcal mol^{-1}). The corresponding energy prediction time is 1.4 s (Intel Core 2 Quad Q6600 @2.40 GHz with 3.25 GB of RAM). The performance of Kriging for intramolecular polarisation is even more impressive, where the energy error of electrostatic interaction between C_α and surrounding atoms within capped

alanine is less than 1 kJ mol^{-1} ('chemical accuracy') for 80% of 1000 test configurations of that molecule, which contains 60 features.

Due to space restrictions one can provide here only a flavour of the development work that underpins QCTFF. Many 'technical matters' have been discussed at length in papers that were cited above, as well as in forthcoming ones focusing on intramolecular polarisation.^{158,159} Details on how to calculate the forces on the nuclei in the context of high-rank polarisable multipolar electrostatic interactions will also be divulged in due course. The philosophy of QCTFF's development is to model electrostatics 'up front'. If this contribution is treated rigorously and has the correct rationale then the other terms in the full interaction potential will model the remaining non-electrostatic contribution. The final and important step in QCTFF's construction, as a generic force field, is to combine the polarised atoms with the so-called 'bonded part'. The latter will initially be represented by the classical contributions involving harmonic stretch and valence terms, and possibly torsional terms. Preliminary work suggests that the combination of bonded terms with the (non-bonded) electrostatic part gives rise to a stable optimisation procedure (using Levenberg–Marquardt). Before closing with some general remarks let us summarise the advantages that QCT offers in the context of the development of a force field.

- As a QCT atom is defined from the electron density, which is measurable and theoretically well defined, it can be obtained regardless of the type of basis set used (Gaussian *even if diffuse*, plane wave, Slater).
- As QCT atoms do not overlap there is no penetration (energy) and hence need for damping functions.
- QCT atoms are rooted in quantum mechanics, thereby enhancing their physical meaning.
- The multipolar expansion of atoms still converges at very short range.
- QCT atoms can be visually represented due to their finite shape, providing an intuitive understanding of interactions in complex situations. We use the spherical tensor formalism¹²⁹ as it is more compact than the Cartesian one.

One should realise that the Kriging approach could eventually also operate on intra-atomic energies. Equation (6.9) shows how the total Hartree–Fock energy of any (quantum) system, whether it is a single molecule or a system of molecules, can be part be partitioned into intra- and interatomic contributions:³⁹

$$E_{tot} = \sum_A E_{kin}^A + \frac{1}{2} \sum_A E_{coul}^{AA} + \frac{1}{4} \sum_A E_X^{AA} + \frac{1}{2} \sum_{\substack{A,B \\ B \neq A}} E_{coul}^{AB} + \frac{1}{4} \sum_{\substack{A,B \\ B \neq A}} E_X^{AB} \quad (6.9)$$

Here the first three terms are intra-atomic, running over atomic index A , describe the kinetic, Coulomb and exchange contributions. The latter two terms describe the Coulomb and exchange interaction terms, which are sums running

over distinct pairs of atoms. One can then study how these five types of terms, that govern all interactions, cooperate to form a torsion potential.⁴⁰ Torsion barriers arise from a subtle interplay between the different types of physical terms. A torsional rotation around a central C–C bond is caused by a completely different interplay between terms than one around an O–O bond. For this total energy partitioning approach to be practical, the absolute energy error of the Kriging model has to be low enough, again below the oft-quoted chemical accuracy of 1 kJ mol^{-1} . The atomic energies can be of the order of a thousand kJ mol^{-1} , so the relative accuracy has to be very high. This whole approach is drastically different to that of existing force fields, which all lack kinetic energy terms, for example. However, if successful it would eliminate the need for bonded parameters appearing in classical harmonic, cubic and coupling terms.

Finally, MD simulations already accomplished with QCT multipole moments with interaction $L = 5$ could not be discussed due to space restriction. Secondly, QCTFF copes with a small number of water molecules, e.g. interstitial, cavity, as well as bulk water. Current work is focusing on a rigorous procedure to compute atom types within the Kriging formalism, using feature selection. In other words, it is possible to determine mathematically which geometrical coordinates do not contribute significantly to the prediction of the multipole moment of the atom of interest. The size of the test molecules will be scaled up, and each time intermediate validation will be carefully reported on. Only then, we believe, solid progress can be sustained and a step change be hoped for in the area of biomolecular modelling and drug design in particular. More and more experimental work, e.g. ref. 160 and 161 and references therein, is emerging that disentangles the all-governing Gibbs free energy into an enthalpy term and an entropy term. The refinement of the technique isothermal titration calorimetry (ITC) is responsible for this exciting development. This raises hope that theory and experiment will soon properly meet in the challenging realm of protein–drug affinity.

6.5 Bioisosterism

Drug discovery, and more generally medicinal chemistry, frequently applies the concept of bioisosterism. A bioisostere is a fragment of an active compound with similar physical or chemical properties.^{162,163} In the context of drug design, bioisosteres enhance the desired biological effects such as target selectivity and administration/absorption, distribution, metabolism, excretion/elimination, toxicology (ADMET) profile, without drastically altering the initial chemical structure. A compound with desired biological activity is repeatedly modified during the drug development process, in order to deliver new candidate drugs with similar, but hopefully improved, biological effects.

The main working hypothesis of bioisosterism is that the groups to be substituted are similar in some important physical property. For example, carboxylate and tetrazole are both anions at physiological pH.¹⁶⁴ Groups that are

not geometrically similar nor isovalent may still be bioisosteres if they exhibit similar activity when substituted at the same place in a drug molecule. Figure 6.8 shows examples of some common bioisosteres, identified in a recent study¹⁶⁴ of compounds stored in the MDL Drug Data Report (MDDR), a licensed database compiled from the patent literature.

The prevalence of the use of bioisosteric replacement in drug design can hardly be over-emphasised. Bioisosterism features in many publications and has been reviewed at length.^{162,165–169} The concept of bioisosterism is often considered to be qualitative and intuitive.¹⁷⁰ When a candidate molecule exhibiting activity towards a target is discovered, medicinal chemists will make other, similar, organic compounds in an attempt to improve activity or ADMET properties. To decide which compounds to try, chemists often use experience or intuition and replace active functional groups with others that they believe will behave similarly.

Since bioisosterism is essentially a fragment-based method, QCT is a natural tool to apply in this context. Indeed, the fragments that QCT offer can be viewed as vehicles to express ‘fragment quantum chemical content’ that one can endow bioisosteres with. This idea led to the construction of the ‘Quantum Isostere Database’ (QID),¹⁷¹ a Web-based tool designed to find bioisosteric fragment replacements for lead optimisation using stored *ab initio* data. A wide range of original geometric, electronic and calculated physical properties were stored for each fragment. Physical descriptors with clear meaning were chosen, such as distribution of electrostatic potential energy values across a fragment surface and geometric parameters to describe fragment conformation and shape from *ab initio* structures. Further fundamental physical properties were linked to broader chemical characteristics relevant to biological activity, such as H-bond donor and acceptor strengths. Additional properties with less easily interpretable links to biological activity were also stored to allow future development of QSAR/QSPR models for quantities such as pK_a and solubility.

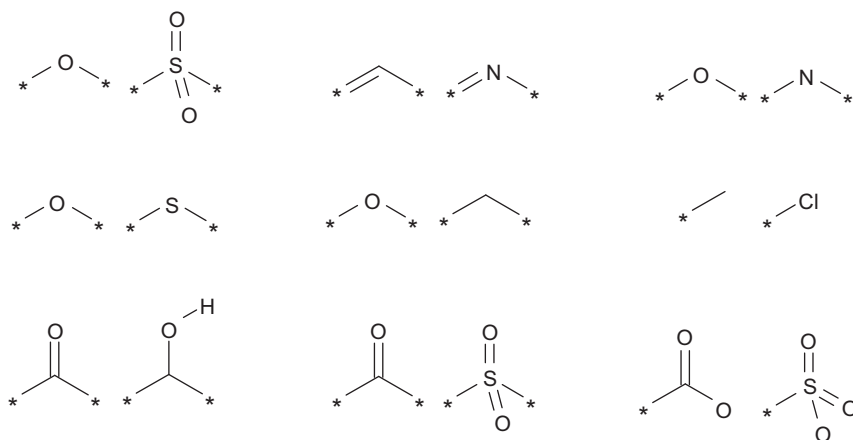


Figure 6.8 Common bioisosteric pairs identified in the MDDR database.

Conformational dependence of the *ab initio* descriptors was explicitly dealt with by storing properties for a variety of low-energy conformers of each fragment. Capping groups were used in *ab initio* calculations to represent different chemical environments, based on background research into transferability of electronic descriptors.⁹ The resulting database has a Web interface allowing medicinal chemists to enter a query fragment, select important chemical features, and retrieve a list of suggested replacements with similar chemical characteristics. This paper¹⁷¹ reported examples of known bioisosteric replacements correctly identified by the QID tool.

This research strand is younger than others discussed and hence receives less emphasis. Research into application of QID is currently on-going.

6.6 Conclusions

Here is discussed, at different lengths, the progress in three different research strands, all based on quantum chemical topology (QCT). One strand aims at constructing quantitative structure–activity relationships from new electronic descriptors extracted from geometry optimised wave functions. A method called quantum topological molecular similarity (QTMS) uses QCT descriptors in combination with partial least squares and possibly a machine learning method. Representative applications are discussed here, in medicinal chemistry, physical organic chemistry and ecochemistry. The independent ability of QTMS to suggest the centre of action in a set of congeneric molecules is an asset to experimental studies. QTMS can confirm experimental hypotheses or suggest preferred pathways. Moving one step away from QCT, the use of simple equilibrium bond lengths as descriptors to capture electronic effects is still waiting to be applied on a routine scale. An important question that keeps nagging QSAR is if a descriptor works for the appropriate quantum chemical reason. If *ab initio* equilibrium bond lengths are a genuine and independent source of information, then they are better than *ad hoc* measures of dubious theoretical background. However, if counting carbons is justified then one should not spend CPU days superimposing high-level electron densities, as in quantum molecular similarity.

The second strand of work aims for a generic method to design a force field *ab ovo*, closer to quantum chemical reality than the popular classical force fields currently used. This is feasible given the abundance of current computing power. Four cornerstones underpin this approach: a definition of an atom in a system, transferability and atom typing, multipole moments and, finally, machine learning to capture the change in these moments in response to a changing environment. When combined, one obtains malleable atoms that adjust their electronic structure to the coordinates of the neighbouring atoms. The force field delivers better accuracy at short range and has been successfully used in MD simulations. In summary, several results of this long-term project are culminating in an attempt to create a step change in the realism of force field design.

Finally, QCT is a natural candidate to make bioisosterism a quantitative concept and simultaneously link it to an underlying quantum chemical reality. An initial web tool called Quantum Isostere Database (QID) has been constructed and initial tests conducted. The QID currently awaits applications in drug design and agrochemistry.

Acknowledgement

Dr Yongna Yuan is thanked for preparing Figure 6.1.

References

1. M. O. Sinnokrot and C. D. Sherrill, *J. Phys. Chem. A*, 2006, **110**, 10656–10668.
2. M. E. Harding, T. Metzroth, J. Gauss and A. A. Auer, *J. Chem. Theory Comput.*, 2008, **4**, 64–74.
3. R. F. W. Bader, *Atoms in Molecules. A Quantum Theory*, Oxford University Press, Oxford, 1990.
4. P. L. A. Popelier, *Atoms in Molecules. An Introduction*, Pearson Education, London, 2000.
5. P. L. A. Popelier and E. A. G. Bremond, *Int. J. Quant. Chem.*, 2009, **109**, 2542–2553.
6. P. L. A. Popelier, in *Structure and Bonding. Intermolecular Forces and Clusters*, ed. D. J. Wales, Springer, Heidelberg, 2005, vol. 115, pp. 1–56.
7. R. F. W. Bader, *Atom in Molecules. A Quantum Theory*, Oxford University Press, 1990.
8. P. L. A. Popelier and F. M. Aicken, *Chem. Phys. Chem.*, 2003, **4**, 824–829.
9. M. Devereux, P. L. A. Popelier and I. M. McLay, *J. Comput. Chem.*, 2009, **30**, 1300–1318.
10. F. Verhulst, *Nonlinear Differential Equations and Dynamical Systems*, Springer-Verlag, Berlin, 1990.
11. A. D. Becke and K. E. Edgecombe, *J. Chem. Phys.*, 1990, **92**, 5397–5403.
12. E. Scrocco and J. Tomasi, *Adv. Quant. Chem.*, 1978, **11**, 115.
13. T. A. Keith, R. F. W. Bader and Y. Aray, *Int. J. Quant. Chem.*, 1996, **57**, 183–198.
14. W. H. Press, S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, *Numerical Recipes in FORTRAN*, 2nd edn., Cambridge University Press, Cambridge, 1994.
15. P. L. A. Popelier, *Comput. Phys. Commun.*, 1996, **93**, 212–240.
16. P. L. A. Popelier, *Theor. Chim. Acta*, 1994, **87**, 465–476.
17. M. Rafat and P. L. A. Popelier, *J. Comput. Chem.*, 2007, **28**, 2602–2617.
18. P. L. A. Popelier, *Comput. Phys. Commun.*, 1998, **108**, 180–190.
19. P. L. A. Popelier, *Mol. Phys.*, 1996, **87**, 1169–1187.
20. N. O. J. Malcolm and P. L. A. Popelier, *J. Comput. Chem.*, 2003, **24**, 437–442.

21. R. F. W. Bader, P. J. MacDougall and C. D. H. Lau, *J. Am. Chem. Soc.*, 1984, **106**, 1594–1605.
22. R. F. W. Bader, R. J. Gillespie and P. J. MacDougall, *J. Am. Chem. Soc.*, 1988, **110**, 7329–7336.
23. R. J. Gillespie, *Coord. Chem. Rev.*, 2000, **197**, 51–69.
24. P. L. A. Popelier, *Coord. Chem. Rev.*, 2000, **197**, 169–189.
25. N. O. J. Malcolm and P. L. A. Popelier, *J. Comput. Chem.*, 2003, **24**, 1276–1282.
26. N. O. J. Malcolm and P. L. A. Popelier, *Faraday Discuss.*, 2003, **124**, 353–363.
27. R. Thom, *Structural Stability and Morphogenesis*, English edn, Benjamin, Reading, MA, 1975.
28. V. I. Arnold, *Usp. Mat. Nauk.*, 1968, **3**, 30.
29. R. F. W. Bader, T. T. Nguyen-Dang and Y. Tal, *Rep. Prog. Phys.*, 1981, **44**, 893–948.
30. X. Krokidis, S. Noury and B. Silvi, *J. Phys. Chem. A*, 1997, **101**, 7277–7282.
31. V. Polo, J. Andres, S. Berski, L. R. Domingo and B. Silvi, *J. Phys. Chem. A*, 2008, **112**, 7128–7136.
32. B. Silvi and A. Savin, *Nature (London)*, 1994, **371**, 683–686.
33. L. Joubert, M. Pavone, V. Barone and C. Adamo, *J. Chem. Theory Comput.*, 2006, **2**, 1220–1227.
34. G. Naray-Szabo and G. G. Ferenczy, *Chem. Rev.*, 1995, **95**, 829–847.
35. P. Balanarayan, R. Kavathekar and S. R. Gadre, *J. Phys. Chem. A*, 2007, **111**, 2733–2738.
36. M. A. Blanco, A. M. Pendas and E. Francisco, *J. Chem. Theory Comput.*, 2005, **1**, 1096–1109.
37. P. L. A. Popelier and D. S. Kosov, *J. Chem. Phys.*, 2001, **114**, 6539–6547.
38. A. M. Pendas, E. Francisco, M. A. Blanco and C. Gatti, *Chem. A, Eur. J.*, 2007, **13**, 9362–9371.
39. M. Rafat and P. L. A. Popelier, *J. Comput. Chem.*, 2007, **28**, 292–301.
40. M. G. Darley and P. L. A. Popelier, *J. Phys. Chem. A*, 2008, **112**, 12954–12965.
41. T. A. Keith and R. F. W. Bader, *J. Chem. Phys.*, 1993, **99**, 3669–3682.
42. J. Cioslowski and G. H. Liu, *J. Chem. Phys.*, 1999, **110**, 1882–1887.
43. B. Webster, *Chemical Bonding Theory*, Blackwell, Oxford, 1990.
44. A. E. Reed, L. A. Curtiss and F. Weinhold, *Chem. Rev.*, 1988, **88**, 899–926.
45. A. J. Stone, *Chem. Phys. Lett.*, 1981, **83**, 233–239.
46. R. S. Mulliken, *J. Chem. Phys.*, 1955, **23**, 1833–1940.
47. R. G. Pearson, *Int. J. Quant. Chem.*, 2007, **108**, 821–826.
48. A. Kovacs, C. Esterhuysen and G. Frenking, *Chem. Eur. J.*, 2005, **11**, 1813–1825.
49. R. McWeeny, *Methods of Molecular Quantum Mechanics*, Academic Press, 2nd edn, San Diego, 1992.

50. P. L. A. Popelier, in *Quantum Biochemistry: Electronic Structure and Biological Activity*, ed. C. F. Matta, Wiley-VCH, Weinheim, 2010, Vol. Part IV, pp. 669–692.
51. R. Carbo, L. Leyda and M. Arnau, *Int. J. Quant. Chem.*, 1980, **17**, 1185–1189.
52. P. Constans, L. Amat and R. Carbo-Dorca, *J. Comput. Chem.*, 1997, **18**, 826–846.
53. L. Amat and R. Carbo-Dorca, *J. Comput. Chem.*, 1997, **18**, 2023–2039.
54. C. Burt, W. G. Richards and P. Huxley, *J. Comput. Chem.*, 1990, **11**, 1139.
55. P. E. Bowen-Jenkins and W. G. Richards, *J. Chem. Soc., Chem. Commun.*, 1986, 133–135.
56. D. L. Cooper, K. A. Mort, N. L. Allan, D. Kinchington and C. McGuigan, *J. Am. Chem. Soc.*, 1993, **115**, 12615–12616.
57. L. Amat, E. Besalu, R. Carbo-Dorca and R. Ponec, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 978–991.
58. P. L. A. Popelier, *J. Phys. Chem. A*, 1999, **103**, 2883–2890.
59. P. L. A. Popelier, in *Molecular Similarity in Drug Design*, ed. P. M. Dean, Chapman and Hall, London, 1995, pp. 215–240.
60. P. L. A. Popelier, *Chem. Phys. Lett.*, 1994, **228**, 160–164.
61. R. F. W. Bader and H. Essen, *J. Chem. Phys.*, 1984, **80**, 1943–1960.
62. P. L. A. Popelier, *Atoms in Molecules; An Introduction*, Pearson Education, London, 2000.
63. C. Lee and S. Smithline, *J. Phys. Chem.*, 1994, **98**, 1135–1138.
64. S. E. O'Brien and P. L. A. Popelier, *Can. J. Chem.*, 1999, **77**, 28–36.
65. S. Wold, *Technometrics*, 1993, **35**, 136–139.
66. P. Geladi and B. R. Kowalski, *Analyt. Chim. Acta*, 1986, **185**, 1–17.
67. D. J. Livingstone, *Data Analysis for Chemists*, Oxford University Press, Oxford, 1995.
68. S. E. O'Brien and P. L. A. Popelier, *J. Chem. Inf. Comp. Sci.*, 2001, **41**, 764–775.
69. UMETRICS, info@umetrics.com, www.umetrics.com, Umeå, Sweden, 1998.
70. R. Ponec, L. Amat and R. Carbo-Dorca, *J. Comput.-Aided Mol. Des.*, 1999, **13**, 259–270.
71. S. Wold, M. Sjostrom and L. Eriksson, in *Encyclopedia of Computational Chemistry*, ed. P. v. R. Schleyer, Wiley, Chichester, 1998, Vol. 3, pp. 2006–2021.
72. SPSS Inc., Chicago, USA, version 10.0.7, 2000.
73. P. L. A. Popelier, U. A. Chaudry and P. J. Smith, *J. Chem. Soc., Perkin II*, 2002, 1231–1237.
74. S. E. O'Brien and P. L. A. Popelier, ECCOMAS Proceedings, Barcelona, Spain, 2000.
75. J. Damborsky and T. W. Schultz, *Chemosphere*, 1997, **34**, 429–446.
76. S. E. O'Brien and P. L. A. Popelier, *J. Chem. Soc., Perkin Trans. 2*, 2002, 478–483.

77. S. Ducki, J. A. Hadfield, L. A. Hepworth, N. J. Lawrence, C. Y. Liu and A. T. McGown, *Bioorg. Med. Chem. Lett.*, 1997, **7**, 3091–3094.
78. D. S. Kosov and P. L. A. Popelier, *J. Chem. Phys.*, 2000, **113**, 3969–3974.
79. L. Bonati, E. Fraschini, M. Lasagni and D. Pitea, *J. Mol. Struct. (Theochem)*, 1994, **303**, 43–54.
80. C. L. Waller and J. D. McKinney, *J. Med. Chem.*, 1992, **35**, 3660.
81. U. A. Chaudry and P. L. A. Popelier, *J. Phys. Chem. A*, 2003, **107**, 4578–4582.
82. T. W. Collette, *Environ. Sci. Technol.*, 1990, **24**, 1671–1676.
83. U. A. Chaudry and P. L. A. Popelier, *J. Org. Chem.*, 2004, **69**, 233–241.
84. R. D. Cramer, III, D. E. Patterson and J. D. Bunce, *J. Am. Chem. Soc.*, 1988, **110**, 5959–5967.
85. M. Wagener, J. Sadowski and J. Gasteiger, *J. Am. Chem. Soc.*, 1995, **117**, 7769–7775.
86. P. J. Smith and P. L. A. Popelier, *J. Comput.-Aided Mol. Des.*, 2004, **18**, 135–143.
87. P. L. A. Popelier, U. A. Chaudry and P. J. Smith, *J. Comput.-Aided Mol. Des.*, 2004, **18**, 709–718.
88. A. J. Shusterman, A. K. Debnath, C. Hansch, W. H. Gregory, R. F. Frank, A. C. Greene and S. F. Watkins, *Mol. Pharm.*, 1989, **36**, 939–944.
89. P. J. Smith and P. L. A. Popelier, *Org. Biomol. Chem.*, 2005, **3**, 3399–3407.
90. R. J. Loader, N. K. Singh, P. J. O'Malley and P. L. A. Popelier, *Bioorg. Med. Chem. Lett.*, 2006, **16**, 1249–1254.
91. C. D. Selassie, R. P. Verma, S. Kapur, A. J. Shusterman and C. Hansch, *J. Chem. Soc., Perkin II*, 2002, 1112–1117.
92. N. K. Singh, R. J. Loader, P. J. O'Malley and P. L. A. Popelier, *J. Phys. Chem. B*, 2006, **110**, 6498–6503.
93. P. L. A. Popelier and P. J. Smith, *Eur. J. Med. Chem.*, 2006, **41**, 862–873.
94. B. Hemmateenejad, A. R. Mehdipour and P. L. A. Popelier, *Chem. Biol. Drug Des.*, 2008, **72**, 551–563.
95. B. Hemmateenejad, A. R. Mehdipour, R. Miri and M. Shamsipur, *J. Comput. Chem.*, 2009, **30**, 2001–2009.
96. U. A. Chaudry, N. K. Singh and P. L. A. Popelier, in *Theoretical Aspects of Chemical Reactivity*, ed. A. Toro-Labbe, Elsevier, Netherlands, 2007, Chap. 15, pp. 301–317.
97. C. X. Xue and P. L. A. Popelier, *J. Phys. Chem. B*, 2008, **112**, 5257–5264.
98. C. X. Xue and P. L. A. Popelier, *J. Phys. Chem. B*, 2009, **113**, 3245–3250.
99. K. Roy and P. L. A. Popelier, *Bioorg. Med. Chem. Lett.*, 2008, **18**, 2604–2609.
100. P. Roy and K. Roy, *QSAR Comb. Sci.*, 2008, **27**, 302–313.
101. A. Golbraikh and A. Tropsha, *J. Mol. Graphics Model.*, 2002, **20**, 269–276.
102. K. Roy and P. L. A. Popelier, *QSAR & Comb. Sci.*, 2008, **27**, 1006–1012.
103. K. Roy and P. L. A. Popelier, *J. Phys. Org. Chem.*, 2009, **22**, 186–196.
104. A. P. Harding, D. C. Wedge and P. L. A. Popelier, *J. Chem. Inf. Mod.*, 2009, **49**, 1914–1924.

105. A. P. Harding and P. L. A. Popelier, *Phys. Chem. Chem. Phys.*, 2011, **13**, 11264–11282.
106. A. P. Harding and P. L. A. Popelier, *Phys. Chem. Chem. Phys.*, 2011, **13**, 11283–11293.
107. G. I. Hawe, I. Alkorta and P. L. A. Popelier, *J. Chem. Inf. Mod.*, 2010, **50**, 87–96.
108. S. Kar, A. P. Harding, K. Roy and P. L. A. Popelier, *SAR QSAR Environ. Res.*, 2010, **21**, 149–168.
109. F. Archambault, C. Chipot, I. Soteras, F. J. Luque, K. Schulten and F. Dehez, *J. Chem. Theor. Comput.*, 2009, **5**, 3022–3031.
110. R. Bukowski, K. Szalewicz, G. C. Groenenboom and A. van der Avoird, *Science*, 2007, **315**, 1249–1252.
111. A. J. Stone and A. J. Misquitta, *Int. Rev. Phys. Chem.*, 2007, **26**, 193–222.
112. J.-P. Piquemal, G. A. Cisneros, P. Reinhardt, N. Gresh and T. A. Darden, *J. Chem. Phys.*, 2006, **124**, 104101.
113. S. L. Price, *Acc. Chem. Res.*, 2009, **42**, 117–126.
114. G. A. Kaminski, H. A. Stern, B. J. Berne and R. A. Friesner, *J. Phys. Chem. A*, 2004, **108**, 621–627.
115. T. Cheeseright, M. Mackey, S. Rose and A. Vinter, *Expert Opin. Drug Discov.*, 2007, **2**, 131–144.
116. D. Elking, T. Darden and R. J. Woods, *J. Comput. Chem.*, 2007, **28**, 1261–1274.
117. T. D. Rasmussen, P. Ren and J. W. Ponder, *Int J. Quant. Chem.*, 2007, **107**, 1390–1395.
118. T. R. Walsh and T. Liang, *J. Comput. Chem.*, 2008, **30**, 893–899.
119. J. W. Ponder and D. A. Case, *Adv. Protein Chem.*, 2003, **66**, 27–85.
120. F. Jensen, *Introduction to Computational Chemistry*, 2nd edn., Wiley, Chichester, 2007.
121. P. L. Freddolino, S. Park, B. Roux and K. Schulten, *Biophys. J.*, 2009, **96**, 3772–3780.
122. A. J. Stone, *Science*, 2008, **321**, 787–789.
123. A. J. Stone and S. L. Price, *J. Phys. Chem.*, 1988, **92**, 3325–3335.
124. S. L. Price, in *Rev. Comput. Chem.*, 2000, **14**, 225–289.
125. C. Sagui, L. G. Pedersen and T. A. Darden, *J. Chem. Phys.*, 2004, **120**, 73–87.
126. S. Brodersen, S. Wilke, F. J. J. Leusen and G. Engel, *Phys. Chem. Chem. Phys.*, 2003, **5**, 4923–4931.
127. W. L. Jorgensen, *J. Chem. Theory Comput.*, 2007, **3**, 1877.
128. S. W. Rick and S. J. Stuart, in *Reviews Comput. Chem.*, 2002, **18**, 89–146.
129. A. J. Stone, *The Theory of Intermolecular Forces*, Clarendon Press, Oxford, 1996.
130. I. C. Hayes and A. J. Stone, *Mol. Phys.*, 1984, **53**, 83–105.
131. B. Jeziorski, R. Moszynski and K. Szalewicz, *Chem. Rev.*, 1994, **94**, 1887–1930.
132. B. Guillot, *J. Mol. Liq.*, 2002, **101**, 219–260.
133. B. Roux and S. Berneche, *Biophys. J.*, 2002, **82**, 1681–1684.

134. T. M. Chang and L. X. Dang, *Chem. Rev.*, 2006, **106**, 1305–1322.
135. P. Jungwirth and D. J. Tobias, *J. Phys. Chem. B*, 2002, **106**, 6361–6373.
136. L. E. Chirlian and M. M. Francel, *J. Comput. Chem.*, 1987, **8**, 894–904.
137. M. M. Francel, C. Carey, L. E. Chirlian and D. M. Gange, *J. Comput. Chem.*, 1996, **17**, 367–383.
138. D. S. Kosov and P. L. A. Popelier, *J. Phys. Chem. A*, 2000, **104**, 7339–7345.
139. M. Rafat, M. Shaik and P. L. A. Popelier, *J. Phys. Chem. A*, 2006, **110**, 13578–13583.
140. P. L. A. Popelier and F. M. Aicken, *Chem. Eur. J.*, 2003, **9**, 1207–1216.
141. P. L. A. Popelier and F. M. Aicken, *J. Am. Chem. Soc.*, 2003, **125**, 1284–1292.
142. P. L. A. Popelier and F. M. Aicken, *Chem. Phys. Chem.*, 2003, **4**, 824–829.
143. A. J. Stone, *The Theory of Intermolecular Forces*, Clarendon Press, Oxford, 1996.
144. P. L. A. Popelier, L. Joubert and D. S. Kosov, *J. Phys. Chem. A*, 2001, **105**, 8254–8261.
145. L. Joubert and P. L. A. Popelier, *Mol. Phys.*, 2002, **100**, 3357–3365.
146. M. Rafat and P. L. A. Popelier, *J. Chem. Phys.*, 2006, **124**, 144102.
147. C. J. F. Solano, A. M. Pendás, E. Francisco, M. A. Blanco and P. L. A. Popelier, *J. Chem. Phys.*, 2010, **132**, 194110.
148. P. L. A. Popelier and M. Rafat, *Chem. Phys. Lett.*, 2003, **376**, 148–153.
149. M. Rafat and P. L. A. Popelier, *J. Chem. Phys.*, 2005, **123**, 204103–204101, 204107.
150. M. Rafat and P. L. A. Popelier, in *Quantum Theory of Atoms in Molecules*, eds. C. F. Matta and R. J. Boyd, Wiley-VCH, Weinheim, 2007, vol. 5, pp. 121–140.
151. M. Rafat and P. L. A. Popelier, *J. Comput. Chem.*, 2007, **28**, 832–838.
152. M. in het Panhuis, P. L. A. Popelier, R. W. Munn and J. G. Angyan, *J. Chem. Phys.*, 2001, **114**, 7951–7961.
153. S. Y. Liem and P. L. A. Popelier, *J. Chem. Theory Comput.*, 2008, **4**, 353–365.
154. C. M. Handley and P. L. A. Popelier, *J. Chem. Theory Comput.*, 2009, **5**, 1474–1489.
155. M. G. Darley, C. M. Handley and P. L. A. Popelier, *J. Chem. Theory Comput.*, 2008, **4**, 1435–1448.
156. N. Cressie, *Statistics for Spatial Data*, Wiley, New York, 1993.
157. C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, The MIT Press, Cambridge, USA, 2006.
158. M. J. L. Mills and P. L. A. Popelier, *Theor. Chem. Acc.*, 2012, in press.
159. M. J. L. Mills and P. L. A. Popelier, *Comput. Theoret. Chem.*, 2011, **975**, 42–51.
160. B. Baum, L. Muley, M. Smolinski, A. Heine, D. Hangauer and G. Klebe, *J. Mol. Biol.*, 2010, **397**, 1042–1054.

161. E. Freire, *Chem. Biol. Drug Des.*, 2009, **74**, 468–472.
162. L. M. Lima and E. J. Barreiro, *Cur. Med. Chem.*, 2005, **12**, 23–49.
163. J. E. Graham, D. C. Ripley, J. T. Smith, V. H. J. Smith and D. F. Weaver, *J. Mol. Struct. (Theochem)*, 1995, **343**, 105–109.
164. R. P. Sheridan, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 103–108.
165. C. W. Thornber, *Chem. Soc. Rev.*, 1979, **8**, 563–580.
166. C. A. Lipinski, *Annu. Rep. Med. Chem.*, 1986, **21**, 283–291.
167. A. Burger, *Prog. Drug Res.*, 1991, **37**, 287–371.
168. G. A. Patani and E. J. LaVoie, *Chem. Rev.*, 1996, **96**, 3147–3176.
169. C.-G. Wermuth, in *The Practice of Medicinal Chemistry*, ed. C.-G. Wermuth, Academic Press, London, 1996, pp. 202–237.
170. Y. C. Martin, *J. Med. Chem.*, 1981, **24**, 229–237.
171. M. Devereux, P. L. A. Popelier and I. M. McLay, *J. Chem. Inf. Modell.*, 2009, **49**, 1497–1513.

CHAPTER 7

Cheminformatics in Diverse Dimensions

THOMAS ENGEL

Ludwig-Maximilians-Universität München, Department of Chemistry,
Butenandtstr. 5–13, 81377 München, Germany

7.1 Introduction

Scientists may have diverse concepts and perceptions when describing and discussing structural information of chemical compounds. A compound may be characterized by name, by two-dimensional drawing and/or by providing a three-dimensional molecular model. This hierarchy also reflects levels of refinement when representing chemical structures in electronic form – starting from linear notations to 2D chemical graphs and 3D structure representations culminating in molecular property grids. The problem is how to encode these models in electronic format and enable computers to deal with them, since they work with bits of information and perceive neither atoms nor bonds. This interchange must somehow be handled by widely understood input and output formats. The chemists' way of thinking about structures must be translated into computers' internal machine representation through one or more intermediate steps or representations.

The most common method of structure input is to draw a (small) molecular structure with a graphical user interface (structure editor). The drawn structure may next be converted automatically into one of the structural representations needed, or directly into a format understood only by the computer and the software employed. Another approach is to apply existing structural

representations in databases, built up, for example, on experimental data, *e.g.* the PDB database, for further processing with a computer. However, as previously mentioned, the hierarchy or information content of the representation, respectively, is pivotal for further computer processing.

In drug design the most appealing challenge is to discover interactions between a drug and its target. In reality these interactions occur in the 3D world. Therefore, 3D methods provide appropriate tools, such as powerful searching techniques including pharmacophore matching,^{1–8} docking,^{9–12} and similarity/diversity analysis based on molecular fields and surface properties.^{13–16} However, one should ask if the knowledge of 3D structure information really offers advantages for design. Certainly, many tasks and applications such as similarity or structure/sub-structure searching and also a number of structure–activity relationships (SAR) are based on 2D structure information, handled by less complex 2D methods.^{17,18} Finally, whenever needed, added 3D structure information can provide a significant impact on elucidating structure–activity relationships.

The following three sections lend a perspective on the structure representation derived from a computer representation. Thus, code dimensionality in zero dimension (numbers), one dimension (strings) and two dimensions (tables, matrices, etc.) determine this type classification. Within the three classes, all different kinds of dimensionality of structure representation (0D–3D) are included.

Since the number of references to each of the following topics would go far beyond this scope, only the most relevant or updated literature is provided. More detailed information and supporting literature can also be found in two comprehensive reference books: *Handbook of Chemoinformatics*¹⁹ and *Molecular Descriptors for Chemoinformatics*.²⁰

7.2 Zero-dimensional Code Representations

The simplest type of structure representation is the 0D code representation. This form generally reflects indices in databases as data-independent data structure (integer) used to speed up data retrieval operations;²¹ these may include indices calculated by molecular descriptors methods.

Nevertheless, the very well known CAS-RN and hash codes also might be classified into the 2D code representations, both representations are strictly associated with database indexing and therefore arranged within this section.

7.2.1 CAS Registry Numbers

Chemical Abstracts Service (CAS), a division of the American Chemical Society, assigns a CAS-RN to every compound that has been described in the literature. The intention is to make database searches more convenient since compounds may often have multiple names. Almost all molecule databases today allow searching by CAS numbers or other identifiers such as the Beilstein Registry number. The unique CAS registry number is separated by hyphens into three parts with up to 10 digits. The numbers are assigned in increasing order and do not possess inherent meaning.²²

7.2.2 Hash Structural Codes

Hash coding is an established method in computer science, for example in registration procedures.^{23,24} In cheminformatics the structure input occurs as a sequence of characters (names) or numbers that may also be obtained, for example, from a connection table by conversion of a structure drawing (see below). Both names and numbers may be quite long and may not directly be processed by the computer or used for the storage of chemical structures. The ‘hashing’ chops or cuts the input into small pieces to obtain a short, acceptable integer within a certain file length. Subsequent to this process, the integers may be processed, by the computer, with an extremely high transfer speed.

In more detail, the hash procedure (key transformation) computes a number of storage addresses from alphabetic, numeric or alphanumeric keys. The data or the structures are separated into fragments, which are assigned to an identification number (ID). However, the ID of a fragment is not directly accessible with the computer. First, the ID has to be transformed into a pre-set, fixed number of characters (a hash code) by a hash algorithm. This algorithm produces a highly compressed code dependent only on the input information, for example, molecular weight, empirical formula or 2D/3D structure. However, this code includes no data information and is used only as a key to the storage address of the data entry. Thus, a hash code is a unique number that describes and identifies molecular data structures in chemistry such as atoms and bonds or is characteristic of an individual chemical structure.²⁵

The ciphered code is indicated with a defined length, *i.e.* a fixed bit/byte length. A hash code of 32 bits could have 2^{32} (or 4 294 976 296) possible values, whereas one of 64 bits could have 2^{64} values. However, due to the fixed length, several diverse data entries could assign the same hash code (‘address collision’). A higher probability of collision occurs if the number of input data is increased in relation to the range of values (bit length). In fact, the limits of hash coding are reached with about 10 000 compounds with 32 bits and over 100 million with 64 bits, to avoid collisions in databases.²⁶

Accordingly, the hash code is not used as a direct way to access data, rather it serves as an index or key to the filed data entry. Since hash coding receives unique codes by reducing multidimensional data to only one dimension, information can and often will be lost. This loss may prevent reconstruction of the complete data from the hash code.

Intensive data reduction is an efficient method of managing large datasets. Generally, hash codes are used within chemical information processes such as molecule identification and recognition of identical atoms,²⁷ but less in sub-structure and similarity searching.

The hash code has specific characteristics and provides numerous possibilities for use, for example:

- hash codes of molecules that are already pre-computed are suitable for use in full structure searches in database applications
- atomic and bond hash codes are helpful in structure manipulation programs, for example, in reaction prediction or in synthesis design.²⁸

7.2.3 Molecular Descriptors

Molecular descriptors play a fundamental role in the drug discovery process. One accepted basic tenet is that similar molecules have similar activities – this principle leads to structure–activity/property relationships (SAR/SPR). Before the correlation of activities, or other properties of chemical compounds in a SAR/SPR model, the structures of the compounds have to be transferred into numbers – the descriptors. It is a given that the appropriate description of chemical structures for the property to be modelled is an essential task. With a multitude of different types of descriptors available (more than 3000), great care has to be given to the selection of the most appropriate set of descriptors.

A comprehensive treatise on molecular descriptors was published by R. Todeschini and V. Consonni.²⁰ Their terminology states:

The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment.

Descriptors are classified into two main groups depending on their origin:

- non structure-based descriptors: characterized by a number achieved from experimental measurements such as $\log P$, pK_a , molar refractivity, dipole moment, polarizability, boiling point and other physicochemical properties
- structure-based descriptors: theoretical molecular descriptors, calculated from a molecular representation.

To discuss all available, or even the most relevant, descriptors, would go far beyond the scope of this article. Thus, this section focuses on a sample of structure-based descriptors. Herein, the calculation procedure transforms the structural information into a numerical representation of a molecule. The mathematical functions have to be invariant with respect to the molecule's size and number of atoms, allowing some mathematical treatment of the chemical information contained in the molecules, for example model building with statistical methods or artificial neural networks. The information content of a descriptor depends on two major factors; (a) the algorithm employed in the calculation of the descriptor and (b) the molecular representation of the compound. Below, frequently used descriptors are depicted, classified by the dimensionality of their molecular representation. Most of the descriptors are 2D (whole-molecule) that are used in similarity and diversity applications to process a very large number of compounds in high-throughput screening or combinatorial chemistry.^{29–31}

Typical structure descriptors are distinguished in Table 7.1 by their data type (integer or real number, vector, scalar or vector field) and the molecular representation of the compound.

Table 7.1 Examples of 0D code representations of L-tyrosine (*cf.* structure diagram in Table 7.2), the associated dimensionality of the structure representation and the mathematical representation followed by the code. (calculated by Xemistry Web Sketcher^{(*)38} and E-Dragon^{(#)39}).

<i>Type of structure representation</i>	<i>Dimension of the structure</i>	<i>Mathematical representation</i>	<i>Structure code</i>
CAS-RN	1D–3D	index	140-43-2
Hash code	1D–3D	index	56666B48AB05BD2D
Molecular weight	0D	scalar	181.19099
Atom-type counts	0D	scalar	4
Fragment counts (number of total secondary C(sp ³) ^(#))	1D	scalar	2
Wiener path index	2D	scalar	268
Kier-Hall connectivity index (chi-0 ^(#))	2D	scalar	9.845
WHIM descriptor (1st component size directional/unweighted ^(#))	3D	vector	7.244
3D-MoRSE descriptors (signal 03/unweighted ^(#))	3D	vector	−1.936

7.2.3.1 0D Descriptors

This type of molecular representation is independent of the knowledge of the compound structure. It is the simplest type and includes for example atom or bond counts, number of hydrogen, molecular weight, and all other atomic properties that are deducible from the chemical formula. Experimental measurements, i.e. non-structure based data, are also included into this class.

7.2.3.2 1D Descriptors

The concept of a 1D descriptor is to represent a list of structural fragments of a molecule (fragment or functional group counts) without the complete knowledge of the structure. The descriptor is deduced, for example, from condensed formulae or line notations (see Section 7.3.2) and used in substructure searching.

7.2.3.3 2D Descriptors

2D or topological descriptors, sometimes also called indices, represent the constitution of a compound and can be computed from their molecular graph. These types of descriptors are instances of graph invariants, for example molecular connectivity indices and number of nodes. Typical 2D descriptors are the Wiener Index, Randic connectivity index, topological autocorrelation vector, feature trees, BCUT and Kier-Hall connectivity indices.^{30,32}

7.2.3.4 3D Descriptors

Chemical and biological properties are related to the 3D structure of a molecule. In essence, the experimental sources of 3D structure information are

X-ray crystallography, electron diffraction, NMR spectroscopy and neutron scattering. For compounds without experimental data on their 3D structure, automatic methods for the conversion of the connectivity information into a 3D model are required.³³ Thus, in 3D or geometrical descriptors spatial information may be represented in, for example, the 3D-MoRSE, WHIM, GETAWAY, quantum-chemical, size, steric, surface and volume descriptors respectively.

7.2.3.5 4D Descriptors

Hopfinger and co-workers introduced a 4D-QSAR formalism³⁴ as an enhancement of Molecular Shape Analysis (MSA).^{35–37} The 4D descriptors are generated by ensemble sampling of conventional 3D descriptors, i.e. a descriptor of the 3D structural information of a given compound is extended with a set of conformers (the fourth dimension). Examples are descriptors derived from GRID or CoMFA methods.

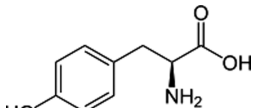
7.3 One Dimensional Code Representations

7.3.1 Nomenclature of Chemical and Biochemical Compounds

Neither a trivial name, nor the systematic nomenclature, both of which represent a chemical structure as an alphanumeric (text) string, is ideal for computer processing. The reason is that various valid compound names can describe one chemical structure. Hence, the basic aim of the IUPAC nomenclature⁴⁰ is to describe particular parts of a structure (fragments, including stereochemistry) in a systematic manner, with special expressions from a vocabulary of terms. This concept allows database systems such as the Chemical Abstracts Service to use IUPAC nomenclature as an index for chemical structures. However, the IUPAC nomenclature does not directly allow the extraction of additional information about the molecule, such as bond orders or molecular weight. As a consequence, the name/structure correlation is unambiguous but not unique. Nowadays, programs can translate names into structures, and structures into names, to make published structures accessible in electronic journals.⁴¹

In biochemistry, essentially amino acid and nucleic acid sequences are represented in one- or three-letter systems. In fact, both types of symbols represent a formula of a compound. Usually, the symbol denotes the *L* configuration of chiral amino acids. As with the trivial names in chemistry, which have no relationship to their chemical structures, the well-known names of the common α -amino acids and nucleobases were usually given to them by their discoverers.^{42,43} The three-letter system for an amino acid is usually derived from the first three letters of its trivial name, for example Arg (but not ARG or arg). Nevertheless, there are difficulties in using the three-letter system in presenting long protein sequences. A one-letter code is much more concise and helpful in summarizing large amounts of data, in aligning and comparing

Table 7.2 Examples of 1D code representations of L-tyrosine and the corresponding structure code (fingerprints are calculated by MOE).⁷⁹

Type of structure representation	Structure code
2D graph	
Trivial name	L-tyrosin
one-/three-letter system	Tyr/Y
IUPAC nomenclature	(2S)-2-amino-3-(4-hydroxyphenyl)propanoic acid
SMILES	Oc1ccc(C[C@H](N)C(=O)O)cc1
SLN	C[1]H:C(:CH:CH:C(:CH:@1)OH)CH2C:s=I]H(NH2)C(=O)OH
Standard InChi	InChI=1/C9H11NO3/c10-8(9(12)13)5-6-1-3-7(11)4-2-6/h1-4,8,11 H,5,10H2,(H,12,13)/t8-m/s1
Structural keys: MACCS	54 84 90 95 104 111 113 123 131 139 143 146 151 152 154 155 156 157 158 159 161 162 163 164 165
TGD (2-point pharmacophore based fingerprint from 2D structure)	600 601 557 556 555 707 706 705 661 558 242 660 241 602 708 559 603 709 560 604 710 561 246 666 247
TAD (2-point pharmacophore based fingerprint from 3D structure)	841 842 780 779 777 991 989 988 926 782 339 924 781 338 840 843 992 783 844 845 994 784 846 785 343 933 345

homologous sequences and in computer techniques for these processes. It may also be used to label residues in 3D representations of protein molecules. Per definition, a sequence starts at the left-hand end with the amino acid residue carrying the free amino group and ends at the right-hand end with the residue carrying the free carboxyl group.

Most of the one-letter codes of the amino acids are derived from the initial letters of the names of the amino acids. Other assignments were made more or less arbitrarily.

IUPAC is responsible both for the rules for naming, as well as for the one- or three-letter systems that were approved in 1968.⁴⁴⁻⁴⁶ Examples of the various 1D code representations are listed in Table 7.2.

7.3.2 Notation of Chemical Compounds

Line notations represent the structure of chemical compounds as a linear sequence of letters and numbers. The first line notations were conceived before the advent of computers. It was realized that the compactness of such a notation was well suited to be handled by computers, since file storage space was expensive at the time of derivation. The heyday of line notations was between 1960 and 1970. Whereas the Wiswesser Line Notation (WLN)⁴⁷⁻⁴⁹ and

the Representation of Organic Structures Description Arranged Linearly (ROSDAL)^{50,51} are now almost obsolete, SMILES^{52,53} and SLN^{54,55} are important representations and are widely used.

7.3.2.1 The SMILES Coding

The SMILES (Simplified Molecular Input Line Entry System) notation has been continually developed by David Weininger since 1986 for chemical data processing. In this line notation the chemical structure information is highly compressed and simplified.^{19,52,53} The SMILES language has found widespread distribution as a universal chemical notation for the representation and exchange of chemical structure information, independent of software or hardware architecture. SMILES notation uses only six basic rules to convert a structure into a character string. The basic SMILES rules are:

1. atoms are represented by their atomic symbols;
2. hydrogen atoms automatically saturate free valences and are omitted (simple hydrogen connection);
3. neighbouring atoms stand next to each other;
4. double and triple bonds are characterized by '=' and '#', respectively; tetrahedral stereochemistry is characterized by '@' and double bonds by '/' oder '\';
5. branches are represented by parentheses;
6. rings are described by allocating digits to the two 'connecting' ring atoms.

SMILES notation has seen many extensions since 1988. The present definition, with examples, can be found in ref. 56 and 57.

Other related coding languages are derived from enhancements of SMILES^{56–58} (XSMILES, SMARTS,⁵⁹ SMIRKS, STRAPS, CHUCKLES,⁶⁰ CHORTLES, CHARTS). Each of them was designed to represent special molecular structures or to allow particular applications (polymers, mixtures, reactions or database handling). A special extension of SMILES is USMILES (sometimes described as Broad SMILES).^{60–62} This 'Unique SMILES' of Daylight is a canonical representation of a structure. This means that the coding is independent of the internal atomic numbering and results always in the same canonical, unambiguous and unique description of the compound, granted by a proprietary algorithm.

The compact textual coding requires no graphical input and additionally permits a fast transmission. These are the important advantages of using SMILES in chemical applications via the Internet and in online services. SMILES notation is also used for the input of structures in the Daylight Toolkit.⁵⁶

7.3.2.2 Sybyl Line Notation

The Sybyl Line Notation (SLN) was also developed to represent molecular structures and is distributed by Tripos Inc. The notation is more or less a modification of SMILES and also uses only six basic rules to specify chemical

structures. SLN is easy to learn and its use is intuitive. Its main distinction from SMILES is that all hydrogen atoms must be specified, as no assumptions are made regarding standard valences or aromaticity. Furthermore, macromolecules, polymers, reactions, structure fragments, substructure queries (Markush structures) and combinatorial libraries can be represented with SLN in a single notation syntax.⁵⁴ All these features make this line notation suitable for database storage as well for data exchange between various programs.

Besides specifications on atoms, bonds, branches and ring closure, SLN additionally provides information on attributes of atoms and bonds, such as charge or stereochemistry. These are indicated in square [] or angle < > brackets behind the entity (for example, *trans*-butane: CH3CH=[s=t]CHCH3). Furthermore, macro atoms allow the shorthand specification of groups of atoms such as amino acids, for example, Ala, Protein2, *etc.* A detailed description of these specifications and also specifications for 2D substructure queries or combinatorial libraries can be found in the literature.⁵⁵

SLN is used in many Internet applications for fast data exchange. A number of commercial software packages, such as ChemDraw,⁶³ CLIFF⁶⁴ and, of course, Tripos products such as CONCORD⁶⁵ operate with this line notation.

7.3.2.3 International Chemical Identifier (InChI)

The IUPAC International Chemical Identifier (InChI) is a non-proprietary unambiguous identifier for chemical compounds. The open-source InChi algorithm,^{66–68} which has been in development since 2000, is implemented nowadays in most structure-drawing software and converts chemical compounds into unique linear alphanumeric strings. The InChI generation has the following workflow:

1. normalization: the chemical compound is transferred into an unambiguous presentation as far as aromaticity, tautomerism, zwitterions, *etc.*;
2. canonization: unique labelling of the atoms based on the Morgan algorithm,⁶⁹ independent of how the structure was drawn;
3. serialization: the tagged atoms are converted into a linear string, the InChI.

The distinct structural information is currently classified and presented in five hierarchically ordered layers and sub-layers of the InChi:

1. main layer (chemical formula, the connection sub-layer/c, and the H atom sub-layer/h);
2. charge layer (charge/q and protons/p sub-layers);
3. stereochemical layer (double bond/b and tetrahedral stereochemistry/t,/m,/s, sub-layers);
4. isotopic layer (/i,/h,/b,/t,/m,/s, sub-layers);
5. fixed-H layer (only in non-standard InChI).

As seen in the list above, each information type of layer is indicated by a lower-case letter after the slash mark, except for the chemical formula sub-layer. Thus, layers and sub-layers are separated by the forward slash '/' delimiter. If any kind of information from the second layer on is missing in the input structure representation, the corresponding layers can be omitted. Furthermore new layers may be added to extend the level of information (for example polymers, Markush structures). To ensure the representation of a compound with a specified level of details/layers, the standard InChi was defined. This guarantees the compatibility between different software, databases and other electronic data sources. All other kind of specific InChi levels are called non-standard InChi. The standard InChI is designated by the prefix 'S' in: 'InChI=1 S/' where 1 is the identifier version number. The 'S' is omitted in the non-standard InChi.

7.3.3 Bit String Representations

In addition to the molecular descriptors (see Section 7.2.3), the concept of a linear descriptor is to represent a single feature of a molecule (fragment counts) at a special position in a bit string or a vector. It is distinguishable from structural keys (fragment code) and fingerprints. Both representations comprise constitutional as well as 3D information about the structure.

7.3.3.1 Structural Keys

The structural key approach describes the chemical composition and structural motifs of molecules represented as a Boolean array. If a certain structural feature is present in a molecule or a sub-structure, a particular bit is set to 1 (true), otherwise to 0 (false). A bit in this array may encode a particular functional group (such as a carboxylic acid or an amide linkage), a structural element (for example, substituted benzene, rings), or, at least, n occurrences of a particular element (for example, a nitrogen atom). Alternatively, the structural key can be defined as an array of integers where the elements of this array contain the frequency of a specific feature in the molecule.

A typical application of structural keys is for sub-structural searches. Here, the key of the query molecule is compared with the stored structural keys of the database entries. This implies that each array element in the structural key has to be defined in a fragment dictionary initially. Therefore, this key can become extremely long and is often inflexible in terms of supplementary fragments. The choice and number of patterns included in the key affect the search across the database. Long keys influence search performance, whereas short keys may screen out many structures that are of no interest. On the other hand, such a structural key can be optimized for the compounds present in the database to be investigated.

One of the most famous search keys is the MDL MACCS key with 166 bits or in an extended version, with 960 bits.^{70–72}

7.3.3.2 Fingerprints

Fingerprints were developed to overcome the shortcomings of structural keys. Several types of molecular fingerprints exist, depending on the structure/sub-structure and their dimensionality (0D to 4D) used to generate fingerprints. A fingerprint is also a binary vector, but in contrast to a structural key there is no need to predefine a fragment dictionary. Initially, all bits of a fingerprint with a fixed size are set to 0. In the second step, a list of patterns is generated for each atom (zero-bond path), for each pair of adjacent atoms and the bond connecting them (one-bond path), and for each group of atoms joined by longer pathways (usually paths of up to seven or eight bonds). Each pattern of the compound is assigned a unique set of bits (typically four or five bits per pattern) along the fingerprint by a hashcoding algorithm.⁷³ The set of bits thus obtained is added to the fingerprint with a logical OR.

Using fingerprints, a database can be screened in the same way as with structural keys by simple Boolean operations. Compared with structural keys, fingerprints have a high information density without loss of specificity.

This string representation of chemical structures facilitates very efficient similarity searches. Following the ‘similar structure – similar property principle’, high-ranked structures in a similarity search are likely to have similar physicochemical and biological properties to those of the target structure.³⁰

7.3.3.3 Pharmacophore-based Fingerprints

‘A pharmacophore is the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response’ according to the IUPAC definition.⁴¹

Pharmacophore-based fingerprints are used to define a pharmacophore with its simple spatial representation of key molecular interactions including charges, ionizable groups, hydrogen bond donor and acceptors, hydrophobic groups and aromatic rings and even virtual points.^{74,75} All possible combinations of these features (two up to more than 10) for all conformers are coded in 3D fingerprints. By analogy, with the use of the 2D fingerprints used in substructure and similarity searching, the pharmacophore key can provide a way to improve search speeds for searches of 3D databases. Furthermore, the pharmacophore concept has found widespread use in medicinal chemistry for hit and lead identification and during the subsequent lead optimization.^{76–78}

7.4 Two-dimensional Code Representations

Computer graphics uses matrices to project 3D space onto 2D such as a computer screen. In a very similar way, additional information about a molecular structure (3D or 2D) may be accumulated in a 2D code representation.

7.4.1 Matrices

The analogy between a structure diagram and a topological graph is the basis for the development of graph theoretical algorithms to process chemical structure information.^{80–82} In mathematical terms, the structure diagrams drawn by a chemist can be considered as ordinary graphs. Graphs consist of nodes (vertices), which are the atoms, and edges, which are the bonds. In organic chemistry, these graphs are often simplified by representing the carbon atoms only as the point where connecting lines meet, the edges (bonds). Such a graph is called a topological graph because it only shows the linkages between atoms and the kind of bonds between them. This type of structure representation contains no data about the 3D structure, the topography, of a molecule.

A graph can be represented as a matrix, the major advantage of which is that the calculation of paths and cycles can be performed easily by well-known matrix operations.

The matrix representation of a structure with n atoms mostly consists of an array of $n \times n$ entries (for example not including the asymmetric incidence matrix). A molecule with its various atoms and bond types can be represented in matrix form in different ways depending on what kind of entries are chosen for the atoms and bonds. Thus, a variety of matrices has been proposed; adjacency, distance, incidence, bond and bond–electron matrices. Generally, drug design resorts to the symmetric adjacency and distance matrices that either represent the constitution (connectivity) or the topological distance information of a molecule.

The n -squared adjacency matrix of a molecule characterizes adjacent nodes of the molecular graph with the value 1 and all not-connected atoms with the value 0 at the appropriate matrix element. Thus, this matrix representation is a Boolean matrix with bits (0 or 1).

The elements of a topological (a) or topographical (b) distance matrix contain values that specify the shortest distance between the atoms involved. Distances can be expressed either (a) as geometric distances (in Å) or (b) as topological distances (number of bonds along the shortest path between the respective atoms).

Hydrogen atoms are sometimes omitted, since their numbers and positions can be calculated from organic structures on the basis of the valence rules of the heavy atoms. Furthermore, the squared symmetric matrices are simplified into a non-redundant matrix representation by depleting all elements that appear twice on one side of the main diagonal.

7.4.2 Connection Table

A major disadvantage of a matrix representation for a molecular graph is that the number of entries increases with the square of the number of atoms in the molecule. A linear function of the number of atoms in the molecule is implemented in the so-called connection table (CT) representation. Herein, the atoms

and the bonds, giving the connections between the atoms, are listed in a tabular form. A connection table has been the predominant form of chemical structure representation in computer systems since the early 1980s. Graph theory methods can equally well be applied to connection table representations of a molecule.

The framework of a basic CT consists of two components; an atom list, where each atom of a molecule is indexed arbitrarily, including the adjacency, and a bond list, where the bond information (bond order) is stored with indices of the atoms that are connected by a bond. Both tables, the atom and the bond lists, are linked through the atom indices.

Similarly to the matrix representations, the connection table can be condensed into a non-redundant representation. Along with duplicates, the hydrogen atoms could also be omitted in 'standard' organic compounds. Programs that have to treat hydrogen atoms may deduce them from the open valences of the corresponding atoms according to the conventions of the valence-bond rules (considering the bond order). In any other case, for example in organometallics, metal complexes, some inorganics or whenever a hydrogen atom plays a major role, the hydrogen atoms have to be included. Nevertheless, the information about the chemical structure is kept intact.

A connection table can be extended by adding other lists, such as lists of free electrons and/or charges on the atoms of the molecule.⁸³ Canonical atom indices are obtained by algorithms, many of which are based on the Morgan algorithm.⁶⁹

Almost all chemical information systems work with their own special type of connection table. They often use various formats distinguishing between internal and external connection tables. Thus, the conversion of one connection table format into another is usually a fairly straightforward task. Although a connection table can be represented in many different ways, the framework remains the same; an atom list and a bond list.

7.4.3 Standard Structure Exchange Formats

Many organizations and software suppliers have developed their own connection table format and quite a few have made provisions for the import or export of other file formats. The SMILES notation has already been mentioned.

7.4.3.1 *The MDL Molfile and SDfile*

The release of the initially proprietary MDL Molfile format to the community at large in 1982 led to its acceptance as a general exchange format for chemical datasets.

A Molfile describes a single molecular structure that can contain disjointed fragments. An extension of the Molfile format, the SDfile (SD stands for structure–data) contains structure and data (properties) for any number of molecules, which makes it especially convenient for handling large sets of

molecules – for data transfer between databases or from databases to data analysis tools, for example. MDL also designed three other, but in some ways internally similar, file formats for handling queries (RGfiles), single reactions (RXNfiles) and reactions as well as molecules together with their associated data (RDfiles). All the MDL file formats are referred to together as CTfiles (chemical table files) and described extensively in references 84 and 85. The default filename extension for Molfiles is ‘.mol’, and for SDfiles, ‘.sdf’ or ‘.sd’ are used.

Each Molfile consists of two parts; the so-called *header block* specific to Molfiles (lines 1–3) and a *connection table* – *Ctab* (from line 4), that is fundamental to all MDL’s CTfile formats. The connection table (*Ctab*) contains the description of the collection of atoms constituting the given compound, which can be wholly or partially connected by bonds. Such a collection can represent molecules, molecular fragments, sub-structures, substituent groups and so on. In case of a Molfile, the *Ctab* block describes a single molecule.

Within an SDfile, each molecule is represented by its Molfile with additional data items describing its non-structural properties (molecular weight, heat of formation, molecular descriptors, biological activity, *etc.*). The information on a molecule is terminated by a delimiter line (containing only ‘\$\$\$\$’). Each data item starts with a *data header* line, which reflects a molecular property name. Next, one or more rows contain the actual data and they are terminated by an empty line.

The historical development led to the situation that the Molfile is not a rigidly standardized format. Molfiles from different sources may differ in some details, depending on the software that generated them. Thus, the extension of the file name is always *.mol but the structure of the file format can be different from the MDL Molfile (for example, Sybyl-, MSI-Molfile, Molfile V3000, *etc.*).

7.4.3.2 PDB File Format

Whereas the Molfile typically is applied to ‘small’ chemical compounds, the PDB file format has become the standard representation for exchanging macromolecular information derived from X-ray diffraction and NMR studies, primarily for large molecules such as proteins and nucleic acids including the 3D structure information. The exchange format has been developed since 1971 at Brookhaven National Laboratories. There, the Protein Data Bank⁸⁶ was established serving also as an eponym for the format. In 1998 the database was moved to the Research Collaboratory for Structural Bioinformatics (RCSB).

PDB files were designed for storage of crystal structures and related experimental information on biological macromolecules. Over the years, the PDB file format was extended to handle results from other experimental (NMR, cryoelectron microscopy) and theoretical methods of 3D structure determination. PDB files contain, first of all, the Cartesian coordinates of the atoms forming the molecule, primary and secondary structures (in the case of proteins), bibliographic citations and crystallographic structure factors as well as X-ray diffraction or NMR experimental data. PDB files can store data for a single molecule, one molecule with associated water molecules, multiple

molecules or an ensemble of alternative models for a single molecule (from NMR experiments). The current PDB file format version is 3.2 and its specification can be found in ref. 87.

Each line of a PDB file must contain exactly 80 printable ASCII characters (selected from the lower and upper case Latin letters, the digits 0–9, and the symbols `-=[]\; ', . / ~ ! @ # \$ % ^ & * () _ + { } | : ' < > ? in addition to the space and end-of-line indicator. The end-of-line indicator is a system-specific line-feed character in Unix, a sequence of carriage return and line-feed characters in DOS and Windows and a carriage return in MacOS. If the actual data occupy less than 80 characters or columns, using the punched card analogy, then spaces are appended up to the 80th column (inclusive). Each line is self-identifying – its first six characters contain a left-justified and blank-filled record name (keyword), which must be one of the predefined names. Columns 7–70 contain data (built from fields), and columns 71–80 are usually empty in older PDB files, but can contain other information in files conforming to newer PDB specifications. The format and contents of the fields within the record are dependent on the keyword. Furthermore, all records in a PDB file must appear in a predefined order. This record/field terminology allows the treatment of a PDB file as an ordered collection of record types.

It has been recognized, however, that the PDB format was unable to adequately express many different aspects of experimental data for macromolecules. Moreover, the PDB representation has a fixed format with many non-coherent rules and the order of data in the file is strongly defined. This makes PDB files complicated for parsing and manual handling (the latter is nearly impossible). Finally, changes in the PDB file format specification enforced the reformatting of existing files to make them conform to the new format specification.

7.4.3.3 Further Structure Exchange Formats

An alternative and much more flexible approach is represented by the STAR file format,^{88,89} which can be used for building self-describing data files. Additionally, special dictionaries can be constructed that specify, more precisely, the contents of the corresponding data files. The two most widely used such dictionaries (and file formats) are the CIF (Crystallographic Information File) format^{90–92} – the International Union of Crystallography's standard for representation of small molecules – and mmCIF,⁹³ which is intended as a replacement for the PDB format for the representation of macromolecular structures. CIF is also a 3D structure information file format with more than three incompatible file versions and is used in crystallography. CIF should not be confused with the Chiron Interchange Format, which is also extended with *.cif.

CML,^{94–97} which is an extension of XML (www.xml.org), tries to unify all the chemical information available, not only for Internet processing.

Most of the standard file formats for 3D chemical structure information contain both the 3D Cartesian atom coordinates and the connection table. The *Z-matrix* is another method for describing a molecule in 3D space, wherein

internal coordinates are adopted such as bond lengths, bond angles and torsion angles. *Z*-matrices are commonly used as input to quantum mechanical (*ab initio* and semi-empirical) calculations as they correctly describe the spatial arrangement of the atoms of a molecule. Note that there is no explicit information on the connectivity present in the *Z*-matrix, as there is, for example in a connection table. Nevertheless, quantum mechanics derives the bonding and non-bonding intramolecular interactions from the molecular electronic wavefunction, starting from atomic wavefunctions and a crude 3D structure. In contrast, most molecular mechanics packages require the initial molecular geometry as 3D Cartesian coordinates plus the connection table, as they have to assign appropriate force constants and potentials to each atom and each bond in order to relax and optimize the molecular structure. Furthermore, Cartesian coordinates are preferable to internal coordinates if the spatial situations of ensembles of different molecules have to be compared and, of course, both representations are interconvertible.

The various internal and external file formats make it necessary to have programs that convert one format into another. One of the first conversion programs for chemical structure information was Babel (around 1992).⁹⁸ It supports almost 50 data formats for input and output of chemical structure information. CLIFF is another file format converter based on the CACTVS technology and which supports nearly the same number of file formats.⁶⁴

During the process of conversion, a program may drop some information produced by other software because the format conventions cannot handle this additional information.

7.5 Conclusion

Chemical structures are the universal and the most natural language for chemists, but not for computers. The input and output of chemical structures in computer programs and databases requires specific methods for their representation in computer-readable form. Computers work with bits and they perceive neither atoms nor bonds. In contrast, chemists try to build models of the world with molecules. The models are often conceptually quite simple; 2D plots of molecular structures, or projections of 3D structures onto a plane. However, all the various models of structure representations mentioned in this chapter and applied in drug design are approximations. Therefore, there is no right or wrong molecular representation used for drug discovery, but it should always be context-dependent and well deliberated. For successful SAR modelling, the appropriate choice of molecular descriptors is especially essential.

References

1. J. H. Van Drie, D. Weininger and Y. C. Martin, *J. Comput.-Aided Mol. Des.*, 1989, **3**, 225–251.
2. O. F. Guner, D. R. Henry and R. S. Pearlman, *J. Chem. Inf. Comput. Sci.*, 1992, **32**, 101–109.

3. D. E. Clark, G. Jones and P. Willett, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 197–206.
4. J. Greene, S. Kahn, H. Savoj, P. Sprague and S. Teig, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 1297–1308.
5. T. Langer and R. D. Hoffmann, *Pharmacophores and Pharmacophore Searches*, Wiley-VCH, Weinheim, 2006, 375.
6. S.-Y. Yang, *Drug Discovery Today*, 2010, **15**(11/12), 444–450.
7. Q. Gao, L. Yang and Y. Zhu, *Curr. Comput.-Aided Drug Des.*, 2010, **6**(1), 37–49.
8. K.-H. Kim, N. D. Kim and B.-L. Seong, *Expert Opin., Drug Discovery*, 2010, **5**(3), 205–222.
9. Y. Sun, T. J. A. Ewing, A. G. Skillman and I. D. Kuntz, *J. Comput.-Aided Mol. Des.*, 1998, **12**, 597–604.
10. S. Makino, T. J. A. Ewing and I. D. Kuntz, *J. Comput.-Aided Mol. Des.*, 1999, **13**, 513–532.
11. F. N. Novikov and G. G. Chilov, *Mendeleev Commun.*, 2009, **19**(5), 237–242.
12. P. Kolb, R. S. Ferreira, J. J. Irwin and B. K. Shoichet, *Curr. Opin. Biotechnol.*, 2009, **20**(4), 429–436.
13. R. Carbo, M. Arnau and L. Leyda, *Int. J. Quant. Chem*, 1980, **17**(6), 1185–1189.
14. A. Y. Meyer and W. G. Richards, *J. Comput.-Aided Mol. Des.*, 1991, **5**, 427–439.
15. G. Klebe, U. Abraham and T. Mietzner, *J. Med. Chem.*, 1994, **37**, 4130–4146.
16. J. Auer and J. Bajorath, *Methods Mol. Biol. (Totowa, NJ, U.S.)*, 2008, **453** (Bioinformatics, Volume 2), 327–347.
17. P. Willet, J. M. Barnard and G. M. Downs, *J. Chem. Inf. Comput. Sci.*, 1998, **38**(6), 983–996.
18. N. Nikolova and J. Jaworska, *QSAR Comb. Sci.*, 2004, **22**(9–10), 1006–1026.
19. J. Gasteiger, *Handbook of Chemoinformatics*, Wiley-VCH, Weinheim, 2003. Vol. 4 , 1870.
20. R. Todeschini and V. Consonni, *Molecular Descriptors for Chemoinformatics*, Wiley-VCH, Weinheim, 2009, Vol. 2 , 1257 pp.
21. Y. S. Prabhakar and M. K. Gupta, *Sci. Pharm*, 2008, **76**(2), 101–132.
22. CAS Homepage, Updated 2010, Accessed 9 Nov 2010, <http://www.cas.org/>
23. P. Willett *Chemical Information Systems*, Research Studies Press, Letchworth, UK, 1986.
24. W. T. Wipke, S. K. Krishnan and G. I. Ouchi, *J. Chem. Inf. Comput. Sci.*, 1978, **18**, 32–37.
25. J. Zupan, *Algorithms for Chemists*, John Wiley & Sons, Chichester, 1989.
26. R. G. Freeland, S. A. Funk, L. J. O’Korn and G. A. Wilson, *J. Chem. Inf. Comput. Sci.*, 1979, **19**, 94–97.
27. W.-D. Ihlenfeldt and J. Gasteiger, *J. Comput. Chem.*, 1994, **15**(8), 793–813.
28. H. E. Helson, *Rev. Comput. Chem.*, 1999, **13**, 313–399.

29. P. Willett, in *Handbook of Chemoinformatics*, ed. J. Gasteiger, Wiley-VCH, Weinheim, 2003, pp. 904–915.
30. J. D. Holliday, C.-Y. Hu and P. Willet, *Combinatorial Chemistry and High Throughput Screening*, 2002, **5**, 155–166.
31. W. A. Warr, in *Handbook of Chemoinformatics*, ed. J. Gasteiger, Wiley-VCH, Weinheim, 2003, pp. 1604–1639.
32. B. Lucic, S. Nikolic and N. Trinajstić, *Internet Electronic J. Mol. Des.*, 2008, **7**(9), 195–206.
33. J. Sadowski, in *Handbook of Chemoinformatics*, ed. J. Gasteiger, Wiley-VCH, Weinheim, 2003, pp. 231–261.
34. A. J. Hopfinger, S. Wang, J. S. Tokarski, B. Jin, M. Albuquerque, P. J. Madhav and C. Duraiswami, *J. Am. Chem. Soc.*, 1997, **119**, 10509–10524.
35. A. J. Hopfinger, *J. Am. Chem. Soc.*, 1980, **102**, 7196–7206.
36. A. J. Hopfinger, *J. Med. Chem.*, 1981, **24**, 818–822.
37. C. H. Andrade, K. F. M. Pasqualoto, E. I. Ferreira and A. J. Hopfinger, *Molecules*, 2010, **15**, 3281–3294.
38. Chemical Computing Group, MOEv2008.10, developed and distributed by Chemical Computing Group; Updated 2010, Accessed 3 Nov 2010, <http://www.chemcomp.com>
39. Virtual Computing Chemistry Lab, E-Dragon, Updated 2009, Accessed 3 Nov 2010, <http://www.vcclab.org>
40. International Union of Pure and Applied Chemistry, Commission on the Nomenclature of Organic Chemistry, The Nomenclature of Organic Chemistry. Sections A, B, C, D, E, F and H, Pergamon Press, Oxford, 1979.
41. [IUPAC] IUPAC Publications, Updated 2 Nov 2010, Accessed 3 Nov 2010, <http://www.iupac.org/Publications>
42. H. B. Vickery and C. L. A. Schmidt, *Chem. Rev.*, 1931, **9**, 169–318.
43. H. B. Vickery, *Adv. Protein Chem.*, 1972, **26**, 81–71.
44. IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN), Recommendations 1983, *Eur. J. Biochem.*, 1993, **213**(1), 2.
45. Prepared G. P. Moss, *IUPAC Nomenclature Homepage*, Updated 18 Jan 2010, Accessed 3 Nov 2010, <http://www.chem.qmul.ac.uk/iupac/>
46. IUPAC Commission on the Nomenclature of Organic Chemistry, UK, *Pure Appl. Chem.*, 1984, **56**(5), 595–624. Located at <http://www.iupac.org/publications/pac/1984/pdf/5605x0595.pdf>: Accessed 3 Nov 2010.
47. W. J. Wiswesser, *J. Chem. Inf. Comput. Sci.*, 1982, **22**, 88–93.
48. E. G. Smith, *The Wiswesser Line-Formula Chemical Notation*, McGraw-Hill, New York, 1968.
49. J. J. Vollmer, *J. Chem. Educ.*, 1983, **60**(3), 192–6.
50. J. M. Barnard, C. J. Jochum and S. M. Welford, in *Chemical Structure Information Systems: Interfaces Communication and Standards*, ed. W. A. Warr, ACS Symposium Series No. 400, American Chemical Society, Washington, DC, 1989, pp. 76–81.
51. H.-G. Rohbeck, in *Software Development in Chemistry 5*, ed. J. Gmehling, Springer-Verlag, Berlin, 1991, pp. 49–58.

52. D. Weininger, *J. Chem. Inf. Comput. Sci.*, 1990, **30**(3), 237–243.
53. D. Weininger, *J. Chem. Inf. Comput. Sci.*, 1988, **28**(1), 31–36.
54. S. Ash, M. A. Cline, R. W. Homer, T. Hurst and G. B. Smith, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 71–79.
55. R. W. Homer, J. Swanson, R. J. Jilek, T. Hurst and R. D. Clark, *J. Chem. Inf. Model.*, 2008, **48**(12), 2294–2307.
56. Daylight, Chemical Information Systems Inc., Accessed 3 Nov 2010, <http://www.daylight.com/>
57. J. Gasteiger and T. Engel, *Chemoinformatics – A Textbook*, Wiley-VCH, Weinheim, 2003, 600 pp.
58. J. Hinze and U. Welz, in *Software Development in Chemistry 10*, ed. J. Gasteiger, Springer-Verlag, Berlin, 1991, pp. 59–65.
59. K. Schomburg, H-C. Ehrlich, K. Stierand and M. Rarey, *J. Chem. Inf. Model.*, 2010, **50**(9), 1529–1535.
60. M. A. Siani, D. Weininger and J. M. Blaney, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 588–593.
61. R. G. A. Bone, M. A. Firth and R. A. Sykes, *J. Chem. Inf. Comput. Sci.*, 1999, **39**(5), 846–860.
62. D. Weininger, A. Weininger and J. L. Weininger, *J. Chem. Inf. Comput. Sci.*, 1989, **29**(2), 97–101.
63. CambridgeSoft, Updated 2010, Accessed 4 Nov 2010, <http://www.cambridgesoft.com/>
64. Molecular Networks/Inspiring Chemical Discovery, Updated 2010, Accessed 4 Nov 2010, <http://www.mol-net.de/>
65. Tripos – A Certara Company, Updated 2009, Accessed 4 Nov 2010, <http://www.tripos.com/>
66. S. R. Heller and A. D. McNaught, *Chem. Int.*, 2009, **31**(1), 7–9.
67. P. Murray-Rust, H. S. Rzepa, S. M. Tyrrell and Y. Zhang, *Org. Biomol. Chem.*, 2004, **2**, 3192–3203.
68. International Pure Applied Chemistry, Updated 23 July 2010, Accessed 4 Nov 2010, <http://www.iupac.org/inchi/>
69. H. L. Morgan, *J. Chem. Doc.*, 1965, **5**, 107–113.
70. The definitions of MDL's 166 MACCS search keys can be found in the ISIS/Base help file, section 49.2.4: Specifying searchable keys as a query. <http://www.symyx.com/>
71. R. D. Brown and Y. C. Martin, *J. Chem. Inf. Comput. Sci.*, 1997, **37**(1), 1–9.
72. J. L. Durant, B. A. Leland, D. R. Henry and J. G. Nourse, *J. Chem. Inf. Comput. Sci.*, 2002, **42**(6), 1273–1280.
73. C. A. James and D. Weininger, *Daylight Theory Manual*, Ch. 6., 2008, Daylight Chemical Information Systems, Aliso Viejo, CA, USA, Updated Nov 2010, Accessed 4 Nov 2010, <http://www.daylight.com/>
74. C. G. Wermuth, C. R. Ganellin, P. Lindberg and L. A. Mitscher, *Pure Appl. Chem.*, 1998, **70**(5), 1129–1143.
75. A. R. Leach, V. J. Gillet, R. A. Lewis and R. Taylor, *J. Med. Chem.*, 2010, **53**(2), 539–558.

76. D. Horvath, B. Mao, R. Gozalbes, F. Barbosa and S. L. Rogalski, *Methods and Principles in Medicinal Chemistry*, Vol. 23 (Chemoinformatics in Drug Discovery), ed. T. Oprea, Wiley-VCH, 2005. pp. 117–140.
77. X. Chen, A. R. Rusinko III, A. Tropsha and S. S. Young, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 887–896.
78. S. L. Dixon, A. M. Smondyrev and S. N. Rao, *Chemical Biology and Drug Design*, 2006, **67**(5), 370–372.
79. Xemistry Web Sketcher V2, Updated 2010, Accessed 5 Nov 2010, <http://www.xemistry.com/>
80. A. T. Balaban, *J. Chem. Inf. Comput. Sci.*, 1995, **35**(3), 339–350.
81. H. P. Schultz, *J. Chem. Inf. Comput. Sci.*, 1989, **29**(3), 227–228.
82. O. Ivanciuc, in *Handbook of Chemoinformatics*, ed. J. Gasteiger, Wiley-VCH, Weinheim, 2003, Chapter II, Section 4, pp. 103–138.
83. J. Gasteiger, M. G. Hutchings, B. Christoph, L. Gann, C. Hiller, P. Löw, M. Marsili, H. Saller and K. Yuki, *Topics Curr. Chem*, 1987, **137**, 19–73.
84. Ctfile Formats, MDL Information Systems, CA, 2010, Updated 2008, Accessed 8 Nov 2010, <http://www.symyx.com/downloads/public/ctfile/ctfile.pdf/>
85. A. Dalby, J. G. Nourse, W. D. Hounshell, A. K. I. Gushurst, D. L. Grier, B. A. Leland and J. Laufer, *J. Chem. Inf. Comput. Sci.*, 1992, **32**, 244–255.
86. F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. E. Meyer, M. D. Brice, J. R. Rogers, O. Kennard, T. Shimanouchi and M. Tasumi, *J. Mol. Biol.*, 1977, **112**, 535–542.
87. PDB format description:, Updated 2008, Accessed 8 Nov 2010, [ftp://ftp.wwpdb.org/pub/pdb/doc/format_descriptions/Format_v32_A4.pdf/](ftp://ftp.wwpdb.org/pub/pdb/doc/format_descriptions/Format_v32_A4.pdf)
88. S. R. Hall, *J. Chem. Inf. Comput. Sci.*, 1991, **31**, 326–333.
89. S. R. Hall and N. Spadaccini, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 505–508.
90. S. R. Hall, F. H. Allen and I. D. Brown, *Acta Crystallogr*, 1991, **A47**, 655–685.
91. I. D. Brown and B. McMahon, *Acta Crystallogr*, 2002, **B58**(3, No. 1), 317–324.
92. P. E. Bourne, H. M. Berman, B. McMahon, K. D. Watenpaugh and J. W. Westbrook, P. M. and D. Fitzgerald, *Methods Enzymol.*, 1997, **277**, 571–590.
93. CIF 1.1 Specification, Macromolecular Crystallographic Information File: Accessed 8 Nov 2010, <http://www.iucr.org/resources/cif/spec/>
94. P. Murray-Rust and H. S. Rzepa, *J. Chem. Inf. Comp. Sci.*, 1999, **39**, 928.
95. P. Murray-Rust and H. S. Rzepa, *J. Chem. Inf. Comp. Sci.*, 2001, **41**, 1113–1123.
96. G. Gkoutos, P. Murray-Rust, H. S. Rzepa and M. Wright, *J. Chem. Inf. Comp. Sci.*, 2001, **41**, 1124–1130.
97. Chemical Markup Language, Updated 2010, Accessed 9 Nov 2010, <http://www.xml-cml.org/>
98. Main Page – OpenBabel, Updated 21 Oct 2010, Accessed 9 Nov 2010, <http://openbabel.sourceforge.net/>

CHAPTER 8

Analysing Molecular Surface Properties

DAVID C. WHITLEY

Centre for Molecular Design, School of Pharmacy and Biomedical Sciences,
Institute of Biomedical and Biomolecular Sciences, University of Portsmouth,
Mercantile House, Hampshire Terrace, Portsmouth PO1 2EG, UK

8.1 Introduction

Molecular surfaces have long played an important role in many aspects of molecular modelling, from simple visualization of molecules to questions of molecular recognition in ligand–receptor binding. Efforts have focussed both on characterizing and comparing the three-dimensional shape of molecular surfaces¹ and on studying the distributions of surface properties to describe features such as electrostatics and lipophilicity. Many techniques arising from this area are employed in computer-aided drug design, where the emphasis on the molecular surface embedded in three dimensions complements one- or two-dimensional structure-based approaches. Typical applications of these methods include the assessment of surface and property similarity in virtual screening of candidate drug molecules, and the inclusion of surface-based descriptors during the development of quantitative structure activity relationships (QSARs) to explore links with biological activity.

The aim of this chapter is to provide an overview of the methods used to define, calculate and analyse molecular surfaces and molecular surface properties. The first section reviews the different ways of representing molecular surfaces, ranging from geometric, atom-based models to iso-contours of the electron density approximated by sets of orthogonal functions. Some of the

commonly studied molecular surface properties are summarized (see Section 8.2). Section 8.3 considers the various approaches used in analysing the similarity of molecular surface properties, the majority of which fall into three types. Projection-based methods map the surface property onto a two-dimensional sphere and assess similarity between molecules by rotating the corresponding spheres relative to each other to find optimal matches. Distribution-based approaches generally rely on binning the property values at a set of representative surface points to obtain fingerprints that are then compared. Feature-based methods identify distinguished points on the surface and compare molecules by matching corresponding points from each surface. Most of the feature-based methods are concerned, essentially, with the extrema of functions on surfaces and lie naturally within the scope of Morse theory. The final section outlines the analysis of surface properties in terms of the Morse-Smale complex derived from the gradient vector field of the property. In particular, this defines a surface graph that acts as an invariant, capturing the major features of the property. The bifurcations of the gradient field are also seen to be relevant to several aspects of the analysis of surface properties, including their conformational dependence.

8.2 Molecular Surface Models

Perhaps the simplest molecular models are the space-filling Corey-Pauling-Koltun (CPK) models² in which a molecule is depicted as the union of a set of three-dimensional balls representing the atoms, with each ball assigned the van der Waals radius of the corresponding atom. The van der Waals surface (VDWS) is the topological boundary of the CPK model, formed by the parts of the atomic surfaces not contained in the interior of any other atom. This is a compact, orientable, piecewise-smooth surface consisting of a union of spherical polygons, each of which is the visible part of an atomic surface. While sufficient for simple visualization of small molecules, for macromolecules the VDWS may contain large clefts that are inaccessible by ligands or solvent molecules. It also lacks smoothness at the intersections of the spherical polygons, where the normal vector changes discontinuously. To address these deficiencies two related surfaces were introduced based on the interaction between a CPK model and a probe sphere representing a solvent molecule (invariably water).

The solvent accessible surface³ (SAS) is formed by the trajectory of the centre of the probe sphere as it rolls over the VDWS. The SAS is the topological boundary of the CPK model obtained by adding the probe sphere radius to each of the atomic van der Waals radii. This provides a more realistic representation of the molecule in interaction with a solvent but exhibits the same lack of smoothness as the VDWS. The solvent-excluded surface^{4,5} (SES), also known as the molecular surface (MS) or the Connolly surface, is the surface swept out by the probe sphere as it contacts the molecule. The MS is formed from two distinct parts. The *contact surface* is the subset of the VDWS touched by the probe sphere. It consists of a set of convex spherical polygons, each of which is a subset

of an atomic van der Waals sphere. The *re-entrant surface* is traced by the molecule-facing boundary of the probe as it contacts two or more atoms. This includes toroidal patches, formed as the probe rolls in contact with two atoms and concave spherical triangles, created when the probe rests on three atoms simultaneously. Figure 8.1 illustrates the definitions of the VDWS, SAS and SES.

In simple cases, such as the drug molecules illustrated (see Figure 8.2), the SES consists entirely of convex spherical caps, toroidal belts and concave spherical pits, which fit together to form a surface with a continuous normal

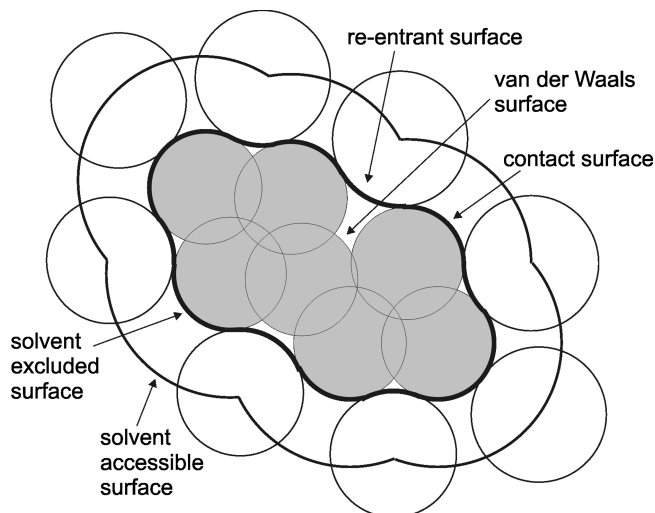


Figure 8.1 Atom-based molecular surfaces: the VDWS is the boundary of the atomic spheres (shaded circles); the SAS is the outer solid curve, traced by the centre of the probe sphere (open circles); the SES is the heavy solid curve traced by the molecule-facing boundary of the probe sphere.

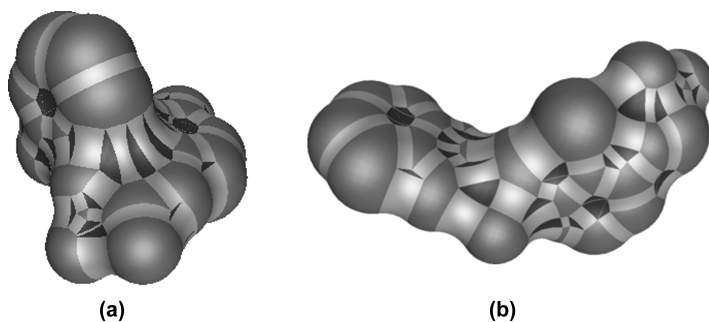


Figure 8.2 Solvent excluded surfaces: (a) Lorazepam, (b) Lansoprazole. The contact surface consists of convex spherical polygons (mid grey); the re-entrant surface consists of toroidal patches (light grey) and concave spherical triangles (dark grey).

vector. (Throughout, all molecular surfaces are drawn with Geomview.)⁶ However, in many cases the re-entrant surface contains singularities formed by self-intersections of the probe sphere as it traverses the VDWS. An additional complication arises if the molecule contains cavities, in which case the SES is disconnected, with one component for the outer, visible surface and one component for each cavity.⁷

Connolly⁸ derived analytic expressions for the spherical and toroidal faces of the MS and produced some of the earliest software to calculate these surfaces.⁹ Sanner *et al.*¹⁰ applied the method of alpha-shapes¹¹ to obtain a more efficient algorithm for the MS and extended Connolly's analysis of the singularities on the surface. Subsequently, various aspects of the calculation of these surfaces were addressed, including treatment of the MS singularities,^{12,13} constructions using parallel processing,¹⁴ NURBS¹⁵ and marching tetrahedra,¹⁶ and several algorithms were developed to calculate the surface area and volume of the SAS.^{17,18,19,20} Some applications require only a representative set of points on the surface, which can be found by triangulating the faces of the MS. The resulting set of vertices, the 'dot surface', is often distributed unevenly over the surface, and the GEPOL²¹ method provides a more uniform distribution. Gerstein *et al.*²² provide a definitive reference for the early work in this area and Ivanov *et al.*²³ compared the performance of several of the first-generation programs. More recent work includes a level-set method²⁴ to calculate the MS of very large proteins and the construction of the VDWS, SAS and MS as rational Bézier surfaces.²⁵ In a slightly different vein, minimal molecular surfaces, obtained by minimizing the mean curvature of a certain 'hypersurface function', have also been proposed.²⁶

Edelsbrunner's molecular skin surface²⁷ (MSS) offers a theoretically appealing alternative to the rolling probe surface. Its definition, as an envelope of 'shrunk' spheres derived from the atomic spheres, is considerably more complicated than the MS, but it provides a surface with a continuous normal vector, no self-intersections or singularities, and is formed from spherical and hyperboloid patches (rather than the spherical and toroidal patches of the SES).

Several algorithms to triangulate the MSS have been developed^{28,29,30} but fast calculation of a high quality triangulation, especially for macromolecules, appears to remain a challenge. Figure 8.3 shows the MSS triangulation obtained by the procedure of Kruithof and Vegter²⁹ implemented in the Computational Geometry Algorithms Library (CGAL).³¹ The initial triangulation on the left contains many unnecessary narrow triangles, but these can be removed by applying the CGAL surface mesh simplification routine to produce the improved triangulation shown on the right. More recently a method to visualize the MSS has been developed using a ray-casting technique with calculations carried out on the graphics processing unit.³²

These geometric, atom-based models of the molecular surface are, of course, somewhat artificial (the problems associated with self-intersection and singularities of the MS are testament to this) and they lack a solid theoretical basis: the values adopted for the atomic radii by different authors show surprisingly large variations,²² while use of the value 1.4 Å for the probe sphere radius is commonplace. The difficulty is that at a fundamental level the 'shape' of a

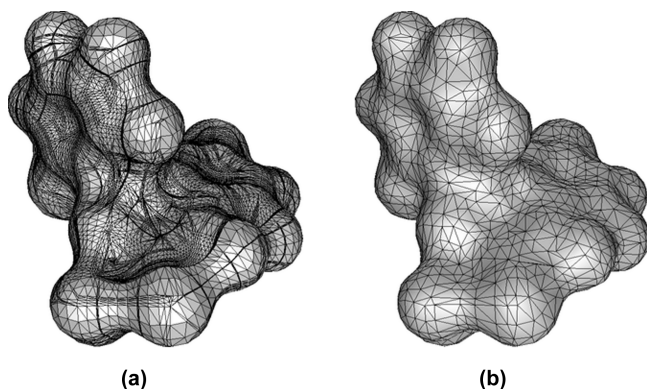


Figure 8.3 Triangulated molecular skin surfaces for Lorazepam. (a) CGAL MSS; (b) CGAL simplified MSS.

molecule is described by its electron density distribution. From this point of view a molecule has no ‘surface’ as such, and the geometric models are perhaps best viewed as approximations to iso-contours of the electron density. It is therefore natural to consider alternative, more direct methods for obtaining these iso-surfaces.

Considered as an iso-contour, a molecular surface S is simply a level set

$$S = G^{-1}(c) = \{x \in \mathbb{R}^3 | G(x) = c\} \quad (8.1)$$

for some suitably chosen real-valued function $G: \mathbb{R} \rightarrow \mathbb{R}$. This has the great advantage that provided the iso-contour value c is a regular value of G , that is $\text{grad } G(x) \neq 0$ for each $x \in G^{-1}(c)$, then S has the structure of a two-dimensional submanifold of $\mathbb{R}^3$. This means that S has a local two-dimensional coordinate system near each point in S which is as smooth as G : if G has continuous k -th order derivatives, then so have the local coordinates. The challenge is to find a tractable but realistic form for G to represent the electron density.

One way to achieve this is to represent atoms by Gaussian functions rather than hard spheres and use the sum of the atomic Gaussians as a surrogate for the electron density of the molecule. Grant and Pickup³³ showed that Gaussians of the form $\lambda \exp(-\alpha r^2)$, where r is a radial coordinate, with suitable values of the parameters λ and α , could be used successfully to calculate molecular surface areas. This Gaussian representation subsequently evolved into ROCS (Rapid Overlay of Chemical Structures),³⁴ one of the most popular current software packages for molecular shape comparison. Arguing that they provide a more correct asymptotic description of the electron density, Gabdoulline and Wade³⁵ used atomic Gaussians of the form $\lambda \exp(-\alpha r^2)$ to define the molecular surface implicitly as $S = G^{-1}(1)$ where

$$G(x) = \sum_i \exp[-(|x - c_i| - r_i)/d] \quad (8.2)$$

Here the i -th atom has centre c_i and radius r_i , and the smoothing factor d can be adjusted in the range $0 < d \leq 1$ to vary the amount of detail visible in the resulting surface. While this provides a simple analytic formula for G , which is useful in some circumstances, it remains an atom-based approximation to the true electron density.

In principle, quantum mechanical methods can be used to calculate the electron density and triangulations of the iso-density surfaces can then be obtained by the marching cubes^{36,37} or marching tetrahedra¹⁶ algorithms. These triangulation procedures, however, are known to produce meshes that often have relatively poor quality, with many superfluous, thin triangles with small internal angles. A secondary refinement procedure may then be necessary to produce a satisfactory triangulation. To illustrate this, Figure 8.4a shows the iso-surface defined by (X.2) for Lorazepam, triangulated by the marching cubes algorithm implemented in the GNU Triangulated Surfaces Library (GTS).³⁸ Application of the GTS coarsening routine produces the improved triangulation shown in Figure 8.4b. Good quality triangulations can also be found directly using more expensive algorithms. The method of Boissonnat and Oudot,³⁹ for example, uses a restricted Delauney triangulation to generate provably good surface meshes. Its implementation in the CGAL surface mesh routine produces the result shown in Figure 8.4c.

In order to perform analysis on the surface, for instance to calculate curvatures or locate the local extrema of surface properties such as electrostatic potential or hydrophobicity, the local coordinates on the surface must be calculated. For the atom-based surfaces analytic formulae are available for the various surface patches. For iso-contour surfaces, local coordinates can be found using continuation methods,⁴⁰ if a reasonably simple expression such as Equation (8.2) for the contour function G is available, or via an approximating triangulation. However, for large molecules such as proteins, a great number of coordinate patches is required. A contrasting approach, which provides a single, global coordinate system for the whole surface, uses truncated spherical harmonic expansions to approximate the surface.^{41,42,43}

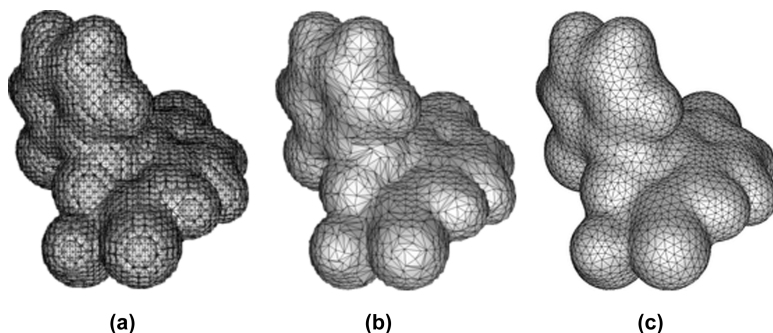


Figure 8.4 Triangulated iso-contour surfaces for Lorazepam, based on Equation (8.2) with $d = 0.1$. (a) GTS marching tetrahedra surface; (b) GTS coarsened marching tetrahedra surface; (c) CGAL surface mesh.

If the surface is *star-shaped*, so that it can be represented in spherical polar coordinates by a single-valued radial function $r(\theta, \phi)$, then the surface can be expanded as an infinite series of real spherical harmonic functions $y_{lm}(\theta, \phi)$,

$$r(\theta, \phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^l a_{lm} y_{lm}(\theta, \phi) \quad (8.3)$$

Truncating the series at order $l=L$ gives an approximation $r_L(\theta, \phi)$ to the surface and Figure 8.5 shows the effect of varying the value of L to obtain approximating surfaces of increasing resolution. The surfaces (see Figure 8.5) are calculated by ParaSurf.⁴⁴ In practice, relatively low values of $L < 20$ are used in applications and orders as low as $L=6$ are often sufficient for visualization and molecular shape comparison.⁴⁵

By the orthogonality of the spherical harmonics, the coefficients a_{lm} in Equation (8.3) are defined by

$$a_{lm} = \int_{S^2} r(\theta, \phi) y_{lm}(\theta, \phi) d\Omega \quad (8.4)$$

Evaluation of these integrals, over the two-dimensional unit sphere S^2 , is the main hurdle in practical applications of this approach. The original numerical integration schemes have been improved upon, first by the use of a tessellation of S^2 obtained by repeated subdivision of an icosahedron⁴⁶ and later by the

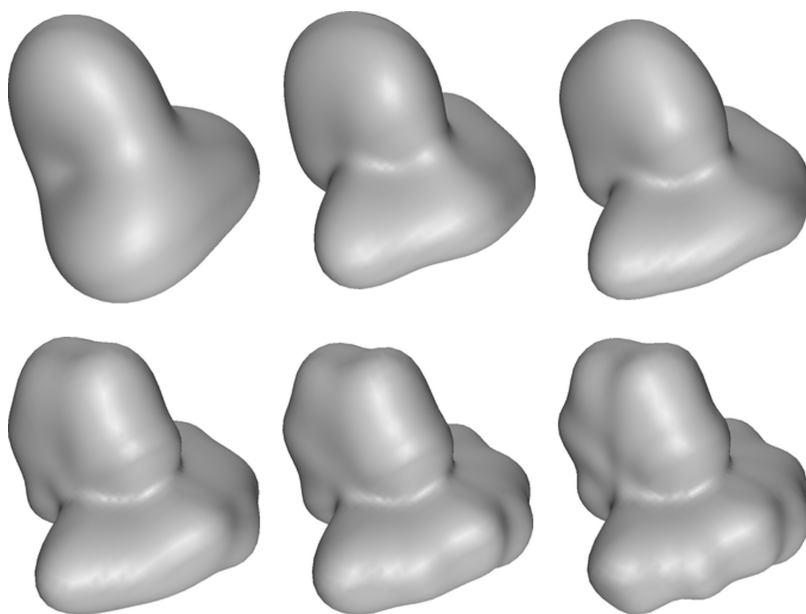


Figure 8.5 Spherical harmonic approximations to the iso-density surface of Lorazepam, for $L = 3, 6, 8, 10, 12$ and 15 .

application of spherical t -designs.⁴⁷ The real spherical harmonic functions are defined by

$$y_{lm}(\theta, \phi) = \begin{cases} \sqrt{\frac{(2l+1)}{2\pi} \frac{(l-m)!}{(l+m)!}} P_l^m(\cos \theta) \cos(m\phi) & m > 0 \\ \sqrt{\frac{2l+1}{4\pi}} P_l^0(\cos \theta) & m = 0 \\ \sqrt{\frac{(2l+1)}{2\pi} \frac{(l+m)!}{(l-m)!}} P_l^{-m}(\cos \theta) \sin(-m\phi) & m < 0 \end{cases} \quad (8.5)$$

Their major computational cost lies in the evaluation of the associated Legendre polynomials $P_l^m(\cos \theta)$. The Hierarchical Equal Area isoLatitude Pixelization (HEALPix),^{48,49} developed originally to address problems in cosmology, has vertices lying on a fixed number of parallels of latitude, $\theta = \text{constant}$. Thus spherical harmonic calculations based on a HEALPix grid require the associated Legendre polynomials to be evaluated only at a fixed (and relatively small) number of θ values, offering a potential speed improvement over earlier schemes that essentially have no control over the latitudinal position of vertices and the consequent number of Legendre polynomial evaluations required.

The principal drawback to the spherical harmonic approach is its restriction to *star-shaped* surfaces: the molecule must have a centre of coordinates (CoC) so that lines radiating from the CoC cut the molecular surface only once. For the majority of small molecules it suffices to place the CoC at the centre of gravity (CoG) of the atom centres and then shrink-wrap⁵⁰ the molecule around the CoC to obtain a star-shaped surface that is close enough to the original surface for most applications. In some cases, such as U-shaped or V-shaped structures with long arms, the CoG may lie outside the surface, in which case heuristics must be applied to shift the CoC towards the bend or hinge region. An absolute restriction, however, is that surfaces described in this way are topological spheres, with genus 0. Surfaces with higher genus, such as a torus produced by a molecule with a large ring, cannot be accommodated by this method since shrink-wrapping the surface changes its topological type.

Several research groups have employed spherical harmonics in applications using a variety of underlying surfaces, integration schemes and expansion orders. Cai *et al.*⁵¹ have used a shrink-wrapped SES, triangulated by repeated subdivision of an icosahedron⁵² and $L = 10$ expansions to study protein–ligand recognition. In their study of shape variations of binding pockets and their ligands, Kahraman *et al.*⁵³ used surfaces triangulated by a spherical 21-design and $L = 14$ expansions. Lin and Clark⁴⁵ investigated the expansion orders needed to accurately reproduce shrink-wrapped iso-contours of the electron density calculated by semi-empirical methods. In addition, spherical harmonic techniques have been used extensively in studies of protein–protein docking, a topic that is dealt with in Chapter 3 of this volume.

The restriction to star-shaped surfaces can be removed by extending the spherical harmonic approach to a full three-dimensional description of the

molecule. Just as spherical harmonics can be used to express functions defined on the two-dimensional sphere, other sets of orthogonal polynomials can be used to express functions defined on regions of three-dimensional space. Mak *et al.*⁵⁴ used series of three-dimensional Zernike polynomials

$$f(r, \theta, \phi) = \sum_{nlm} c_{nlm} R_{nl}(r) y_{lm}(\theta, \phi) \quad (8.6)$$

to expand functions on the three-dimensional unit ball, where the radial function is defined by

$$R_{nl}(r) = \sum_{k=0}^{(n-l)/2} N_{nlk} r^{n-2k} \quad (8.7)$$

if $n - l$ is even and zero otherwise, and N_{nlk} is a normalization factor. By calculating the coefficients c_{nlm} for a voxelized image of a molecule, scaled to fit inside the unit ball and using a ‘thresholding’ method on the truncated reconstruction function Equation (8.6), Mak *et al.* were able to produce approximating surfaces for molecules that not only violate the star-shaped condition but have positive genus, so are not topological spheres. The application of spherical harmonics and 3D Zernike polynomials to protein shape analysis has been reviewed recently by Venkatraman *et al.*⁵⁵

8.3 Molecular Surface Properties

One of the original motivations for determining molecular surfaces was to enable the calculation of molecular surface areas and volumes and they have subsequently been employed to calculate refinements such as the polar surface area and myriad variants catalogued in the *Handbook of Molecular Descriptors*.⁵⁶ Here, our interest lies not in such global molecular properties but in local properties that can be measured at individual points on the surface S and expressed as real-valued functions $f: S \rightarrow \mathbb{R}$.

The most important geometrical local surface properties describe the curvature of the surface in terms of the principal curvatures κ_1 and κ_2 at each point p on the surface.⁵⁷ The classical measures combining the principal curvatures into a single-valued property are the *mean curvature*

$$H = (\kappa_1 + \kappa_2)/2 \quad (8.8)$$

and the Gaussian curvature

$$K = \kappa_1 \kappa_2 \quad (8.9)$$

These are calculated for iso-contour surfaces from the function G in Equation (8.1) and its first two derivatives,⁵⁸ and their integrals over a triangulated

surface can be found within the framework of Minkowski functionals.⁵⁹ The curvature of molecular surfaces is often described by the *shape index*⁶⁰

$$SI = -\frac{2}{\pi} \arctan\left(\frac{\kappa_1 + \kappa_2}{\kappa_1 - \kappa_2}\right) \quad (8.10)$$

where $\kappa_1 > \kappa_2$. The shape index indicates the convex or concave nature of the surface, ranging from -1 at concave points with $\kappa_1 = \kappa_2 > 0$, through 0 at saddle points with $\kappa_1 = -\kappa_2$, to $+1$ at concave points with $\kappa_1 = \kappa_2 < 0$. The *surface topography index*⁶¹ has similar properties to the shape index. The magnitude of the curvature is captured by the *curvedness*⁶⁰

$$R = \sqrt{(\kappa_1^2 + \kappa_2^2)}/2 \quad (8.11)$$

Most physical local surface properties are the restrictions to S of quantum-chemical properties defined on a three-dimensional neighbourhood of the molecule. By far the most commonly studied of these is the molecular electrostatic potential (MEP), the interaction energy between a molecule and a unit positive charge at $x \in \mathbb{R}^3$, defined by

$$MEP(x) = \sum_{i=1}^n \frac{Z_i}{|c_i - x|} - \int \frac{\rho(x')}{|x' - x|} dx' \quad (8.12)$$

Here the summation is over the atoms; the i -th atom has centre c_i and nuclear charge Z_i and ρ is the electron density. The integration of the electron density by *ab initio* or density functional theory (DFT) methods remains prohibitively expensive when dealing with the number of molecules involved in typical chemoinformatics applications and a simplified model based on partial atomic charges is often used:

$$MEP(x) = \sum_{i=1}^n \frac{q_i}{|c_i - x|} \quad (8.13)$$

where q_i is the partial charge on atom i . There are many sources for partial atomic charges, including the partial equalization of orbital electronegativities (PEOE) method,⁶² molecular mechanics force fields such as MMFF⁶³ or AMBER⁶⁴ and semi-empirical programs such as MOPAC.⁶⁵ Distributed multipole models^{66,67} offer an alternative method for calculating the MEP that avoids the spherical symmetry of the charge density around each atom implicit in Equation (8.11). For macromolecules and solvated systems the MEP is usually calculated by numerical solution of the Poisson–Boltzmann equation, by finite difference, finite element or boundary element methods.^{68,69}

Other surface properties that have been studied include the local ionization energy (IEL),⁷⁰ local electron affinity (EAL)^{71,72} and local polarizability (POL).^{71,73}

A further class of molecular surface properties have been proposed to describe lipophilicity, a feature of central importance in predicting biological activity. Measured by the logarithm of the octanol–water partition coefficient, $\log P$, it plays a significant role in many QSAR models. In contrast to the quantum-chemical descriptors, however, there is no underlying physical model for lipophilicity. Based on the existence of several additive models of the form $\log P = \sum_{i=1}^n L_i$, where L_i is the lipophilic contribution of the i -th atom and by analogy with the partial atomic charge models of the MEP, Audry *et al.*⁷⁴ introduced the molecular lipophilicity potential (MLP) as

$$MLP(x) = \sum_{i=1}^n \frac{L_i}{1 + d_i(x)} \quad (8.14)$$

Here $d_i(x) = |c_i - x|$. Subsequently, other functional forms were employed, including the exponential^{75,76}

$$MLP(x) = \sum_{i=0}^n L_i \exp[-\alpha d_i(x)] \quad (8.15)$$

and a normalized sum of Fermi-type functions⁷⁷

$$MLP(x) = \frac{\sum_{i=1}^n L_i g(d_i(x))}{\sum_{i=1}^n g(d_i(x))} \quad (8.16)$$

where

$$g(d) = \frac{1}{1 + \exp[\beta(d - \delta)]} \quad (8.17)$$

Systems of atomic lipophilic contributions L_i were determined by Broto *et al.*⁷⁸ and later by Ghose and co-workers.^{79,80,81}

The MLP provides a qualitative indication of the distribution of lipophilicity over the molecular surface but appears not to have a direct quantitative link with $\log P$. To establish such a connection, the molecular free energy surface density (MolFESD)^{82,83} was defined so that its integral over the molecular surface equals the free energy of transfer between solvents, ΔG , (and hence determines $\log P$, via the relation $\Delta G = -2.303 RT \log P$). Surface integral models for free energies of solvation have since been derived using the MEP, IEL, EAL and POL.⁸⁴

8.4 Molecular Surface Similarity

The similarity principle, that compounds with similar structures tend to have similar physicochemical properties and biological activities, is a central tenet of chemoinformatics and underpins many *in silico* approaches to drug design.^{85,86,87}

The typical aim of similarity-based methods is to identify potentially bio-active molecules by measuring their similarity to a known active compound or their complementarity to a receptor binding site. Assessing similarity at the level of the molecular surface, rather than the molecular structure graph, allows the identification of molecules with similar shapes and properties but very different chemical structures, facilitating what has become known as scaffold hopping. Methods of assessing surface similarity can be divided broadly into three types: projection, distribution or feature-based.

Some of the first work on surface similarity used gnomonic projections,⁸⁸ mapping properties from the molecular surface onto a circumscribing sphere by radial projection outward from a central origin. Optimal similarity matches between molecules were found by searching the space of relative rotations of the corresponding spheres.⁸⁹ Electrostatic complementarity between proteins and ligands was studied by comparing the MEP of the ligand with the MEP induced by the protein on the ligand surface, examining the correlation between these values at points on the ligand surface generated by inverse gnomonic projection of an icosahedral tessellation of the sphere.⁹⁰

Gnomonic projection is the inverse of shrink-wrapping the molecule to obtain a star-shaped surface that can be described by a spherical harmonic expansion. This representation is particularly amenable to similarity calculations, since the distance between two spherical harmonic surfaces can be expressed in terms of the expansion coefficients α_{lm} in Equation (8.3) and the coefficients are transformed linearly when the molecule is rotated. Thus the relatively expensive calculation of the coefficients is required only once for each molecule, with only inexpensive updates needed during searches for optimal relative rotations of query and target molecules. This allows rapid searching and clustering of small-molecule databases,⁹¹ virtual high throughput screening and protein–ligand docking,⁹² with accurate results obtained using low-order expansions. The expansion coefficients can also be used to define rotationally invariant fingerprints, which provide a faster though less sensitive method of comparison. Moreover, the spherical harmonic approach is not restricted to shape alone, as other surface properties can also be expressed as expansions of the form Equation (8.3) and superior overlays are often found using a combination of properties.^{93,94}

The distribution-based methods consider property values at a set of points spread over the molecular surface. Spatial auto-correlation descriptors⁹⁵ for electrostatic and hydrophobic potentials were among the first of this type of method to be introduced. They were used in combination with neural networks to model binding affinities and later applied in a 3D-QSAR study.⁹⁶

The MaP⁹⁷ descriptor is derived from a set of points on a 3D grid that lie close to the SES, found using a modified version of GEPOL.²¹ The points are classified as H-bond donor, H-bond acceptor, hydrophobic or hydrophilic, and for each pair-wise combination of these properties, the possible inter-point distances are divided into bins. For each pair of surface points, the bins for the corresponding properties and distance are incremented, using a fuzzy counting procedure for the distances to adjust for the discrete spacing of the grid points. The MaP descriptor is the resulting vector of bin counts. This is related to

earlier atom-based methods that have been applied to similarity calculations,^{98,99} but to date it appears only to have been used to derive 3D-QSAR models, where a feature of the method is that the variables (bin counts) selected in the QSAR equations can be back-projected to the molecule to obtain a chemical interpretation of the results.

MOLPRINT 3D¹⁰⁰ encodes the local environments of surface points to provide a tool for virtual screening. Interaction energies with various probes may be calculated by the GRID program at points on the Connolly dot surface (generated, unusually, with twice the default atomic radii). The energy ranges for each probe are divided into bins and the neighbours of each surface point are considered in concentric layers. A bit string is created for each point, with bits corresponding to energy bins in each layer. A bit is set if an interaction energy in the corresponding range is present in that layer. In this way a fingerprint for each probe is attached to each surface point. By restricting attention to layers 0–4 around each point, local descriptors with a degree of conformational independence are obtained. In virtual screening experiments, compounds were ranked by their similarity to a query structure, measured either by the Tanimoto coefficient, or by a naïve Bayes classifier trained on features selected from the fingerprints using an information gain criterion. Retrieval rates comparable to those of commonly used 2D and 3D fingerprint methods were obtained and the feature selection procedure was shown to identify features that could be associated with experimental binding patterns.

Several studies have addressed questions of similarity and complementarity of molecular surfaces by seeking to identify regions where some surface property is roughly conserved, selecting a representative point from each region and matching these points to overlay ligands or dock ligands into protein-binding sites. Mostly, the property matched is based on the surface curvature, specifically the degree of convexity or concavity, with physical properties considered in more recent work. To find optimal matches between sets of representative points from two surfaces, various combinatorial methods have been used, with the majority opting for either geometric hashing¹⁰¹ or clique detection¹⁰² in a specially constructed correspondence graph. In either case a rigid body transformation mapping one surface to the other is obtained.

Zachmann *et al.*¹⁰³ derived a procedure for partitioning molecular surfaces into regions of similar curvature by fitting quadratic forms to groups of neighbouring vertices in a triangulation of the surface. Lin *et al.*¹⁰⁴ located a ‘sparse critical point’ (SCP) on each face of the SES by projecting the centre of gravity of the face onto the surface. These points were then used in a geometric hashing scheme to perform molecular docking.¹⁰⁵ Goldman and Wipke¹⁰⁶ took the SCPs that lie on spherical faces of the SES (ignoring the toroidal, saddle faces) and derived a set of quadratic shape descriptors (QSD) by fitting a quadratic form to the Connolly dot surface in a neighbourhood of each SCP. Molecular overlays were obtained by an exhaustive search for matches between the QSDs, using the shape index as a similarity measure. The QSD method has also been applied to protein–ligand docking.¹⁰⁷ The surface patch alignment (SPAt) method of Cosgrove *et al.*¹⁰⁸ follows a similar procedure. SPAt partitions an atom-based

Gaussian iso-density surface into patches of comparable curvature (classified as convex, concave, saddle, cylinder or flat) by calculating the principal curvatures at a set of reference points selected from the vertices of a marching tetrahedra triangulation. Surface matching is then performed by clique detection in a correspondence graph formed by pairing patches of the same type from two surfaces, scoring the cliques by how well paired reference points are aligned by the rigid body transformation induced by the clique. Pickering *et al.*¹⁰⁹ combine the shape index of regions on the Connolly dot surface with clique detection to compare ligand binding sites on protein surfaces, while Baum and Hege¹¹⁰ generate alignments using partial surface similarities.

Exner *et al.*¹¹¹ introduced ideas from fuzzy set theory to segment molecular surfaces into patches of similar property values (surface topography index, MEP, MLP and a hydrogen bond density) and examined protein–ligand surface complementarity, using geometric hashing to carry out docking.¹¹² SURFCOMP¹¹³ uses the same fuzzy set theory approach to compare the MEP and MLP at a set of ‘critical points’ on the surfaces to be matched, adding a local comparison of the surfaces in neighbourhoods of the critical points based on harmonic maps.¹¹⁴ The critical points, classified as ‘peaks’ and ‘valleys’, are selected by examining the principal curvatures of quadratic patches fitted to a dot surface. Surface matching is done by clique detection, with an elaborate system of filters employed to restrict the number of admissible pairings between critical points on the two surfaces, thereby reducing the size of the correspondence graph and improving the speed of the combinatorial search.

In most cases the strategy adopted to assess the performance of these methods is to take a set of structurally dissimilar ligands known to bind to the same or a similar protein active site, align the complexes by a rigid body RMS fit of the backbone atoms on corresponding residues and use the induced alignment of the ligands as the experimental result to be reproduced. The ligands may then be aligned by the method in question and RMSD values reported between the experimental and calculated ligand atom positions. Proschak *et al.*¹¹⁵ compare the performance of their Shapelets method (a minor variant of SPAt, differing mainly in the way the surface is partitioned) to SURFCOMP on a set of four dihydrofolate reductase (DHFR) ligands and to both SURFCOMP and SPAt on a set of eight thermolysin inhibitors. Apart from one particularly poor alignment in the DHFR dataset that skews the overall result, SURFCOMP with MEP generally outperforms the other methods, though it is more computationally expensive than Shapelets.

Using semi-empirical quantum-chemical descriptors, Jakobi *et al.*¹¹⁶ identify a set of ‘unique surface extremes’ of the MEP, IEL, EAL and POL on molecular fragments and use them in the manner of pharmacophores to assess fragment similarity and compare fragments suitable for scaffold hopping.

ParaMatch¹¹⁷ aligns molecules by matching both positions and property values at the vertices of a triangulated surface, using a method employed previously by Robinson *et al.*¹¹⁸ The rotational and translational components of the alignment are considered separately. A fixed set of possible rotations is chosen, then for each rotation the set of all possible translations between pairs

of vertices from the two surfaces that have property values within some tolerance is calculated. An optimal translation is found by clustering this set of inter-vertex translations, treating the translations as 3D points under a gravitational type potential. Using the MEP and a curvature descriptor as properties, alignments of DHFR ligands and HIV-1 reverse transcriptase inhibitors were studied and compared with a spherical harmonic canonical alignment. The results were only partially successful, suggesting that alignments based on combinations of properties may be necessary in some cases. Although considerably slower than the spherical harmonic method, ParaMatch may be better able to match molecules of dissimilar size, where the spherical harmonic alignment is hampered by the restriction that it overlays the centres of gravity.

A detailed discussion of the relevance of the local maxima and minima of the MEP and IEL for 11 small molecules is provided by Bulat *et al.*¹¹⁹ In this case density functional theory is used to calculate the properties on the electron isodensity surface, with extrema located by examining values in 2-neighbourhoods of vertices in a marching tetrahedra triangulation.

8.5 Molecular Surface Property Graphs

Many of the feature-based approaches to the analysis of molecular surface properties involve questions of surface topology and geometry and the structure of functions on surfaces, which are areas of long-standing mathematical study. We now outline some aspects of this general framework and show how to use it to extract an invariant, in the form of a surface graph, which captures the essential features of the surface property in a natural and systematic way. Much technical detail is omitted here; see Zomorodian^{120,121} for an accessible introduction to this material from the viewpoint of computational topology and Palis and de Melo¹²² for a rigorous treatment in the context of dynamical systems theory.

The most important local features of a scalar molecular surface property $f:S \rightarrow \mathbb{R}$ are its critical points, where $\text{grad } f(p)=0$, and the local behaviour of f near a critical point is determined by the Hessian matrix of partial second derivatives

$$Hf(p) = \begin{pmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{pmatrix} \quad (8.18)$$

Critical points where $\det Hf(p) \neq 0$ are called *non-degenerate* or *hyperbolic* and are classified according to their *index*, which is the number of negative eigenvalues of $Hf(p)$. If $\text{index}(p)=2$ then p is a local maximum; if $\text{index}(p)$ then p is a saddle point; and if $\text{index}(p)=0$ then p is a local minimum. The function f is called a *Morse function* if all its critical points are non-degenerate.

A natural way to capture the overall behaviour of f on S arises from consideration of the gradient vector field

$$\dot{x} = \text{grad } f(x) \quad (8.19)$$

The solutions of Equation (8.19) determine a flow $\phi: \mathbb{R} \times S \rightarrow S$ on the surface, where $\phi(t, x)$ satisfies the initial value problem

$$\frac{\partial \phi(t, x)}{\partial t} = \text{grad } f(\phi(t, x)), \quad \phi(0, x) = x \quad (8.20)$$

The solution curve $\{\phi(t, x) \mid t \in \mathbb{R}\}$ passing through a point $x \in S$ is called the *trajectory* or *orbit* of x . The trajectories of the gradient field cross the level sets of f orthogonally, in the direction of increasing values of f . The critical points of f are fixed points of the field, and non-degenerate maxima, saddles and minima of f correspond to sinks, saddles and sources of the field.

Associated with each hyperbolic fixed point p of the gradient field are the stable and unstable manifolds

$$W^s(p) = \left\{ x \in S \mid \lim_{t \rightarrow +\infty} \phi(t, x) = p \right\} \quad (8.21)$$

$$W^u(p) = \left\{ x \in S \mid \lim_{t \rightarrow -\infty} \phi(t, x) = p \right\} \quad (8.22)$$

These consist of all the points whose trajectories are asymptotic to p under the forward and reverse flow, respectively. The stable manifold has dimension $\text{index}(p)$ and the unstable manifold has dimension $2 - \text{index}(p)$. The stable manifold of a sink and the unstable manifold of a source are two-dimensional topological discs containing the fixed point, while the stable and unstable manifolds of a saddle are curves passing through the saddle.

As gradient fields have no periodic orbits, the stable and unstable manifolds of saddles always lead to other fixed points of the field. However, cases where they lead to another saddle point, forming a *saddle connection*, are atypical, since saddle connections can be removed by arbitrarily small perturbations of the field, a point we will return to later. A gradient field of a Morse function that has no saddle connections is called a *Morse-Smale* gradient field. Thus, for a Morse-Smale gradient field the stable manifolds of saddles originate from sources and the unstable manifolds of saddles terminate at sinks.

For a sink, $W^s(p)$ is also called the basin of attraction of p , while for a source, $W^u(p)$ is the basin of repulsion. The boundaries of the basins of attraction are formed by the stable manifolds of the saddles and the boundaries of the basins of repulsion are formed by the unstable manifolds of the saddles. Due to their role in delineating the basins, the stable and unstable manifolds of the saddles are also known as *separatrices*. The decomposition of the surface by a Morse-Smale gradient field into fixed points, separatrices and the remaining open two-dimensional discs forms the *Morse-Smale complex* of the field.¹²³ The one-dimensional skeleton of this complex is a directed graph that links the critical points, encoding their relative positions on the surface. In the present context we call this graph the *molecular surface property graph* (MSPG). It has vertices at the critical points of f and directed edges formed by the separatrices. First introduced for properties on the Connolly surface,¹²⁴ MSPGs have

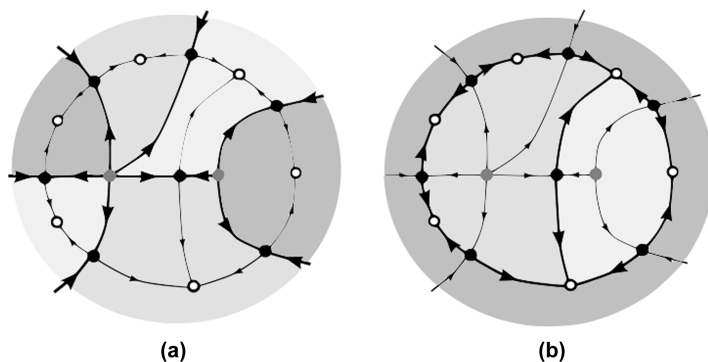


Figure 8.6 A Morse-Smale complex, showing the separatrices (arcs), sinks (open circles), saddles (black circles) and sources (grey circles). (a) basins of attraction, (b) basins of repulsion.

recently been studied in detail on implicit surfaces of the form of Equation (8.2) for partial atomic charge models of the MEP and the formulation of the MLP in Equation (8.14).¹²⁵ A Morse-Smale complex, showing the MSPG and the basins of attraction and repulsion, is sketched in Figure 8.6. This example shows the stereographic projection of a gradient field on a sphere, with a source at the north pole.

An algorithm to calculate the Morse-Smale complex for a triangulated surface, using techniques from piecewise-linear Morse theory, is presented by Edelsbrunner *et al.*¹²³ A method for segmenting a molecular surface along lines that divide the two-dimensional faces of the Morse-Smale complex has also been described.¹²⁶ Here, we outline a procedure for calculating the MSPG for properties on iso-density surfaces that first locates the critical points and then identifies the separatrices by numerically integrating the field from specific points near the saddles.

For an iso-surface (Equation 8.1) where the iso-value c is a regular value of G , the surface S can be covered by a collection of charts providing local coordinates (u, v) on S . Suitable charts can be found by a continuation method⁴⁰ or by taking a sufficiently accurate triangulation of S and projecting the triangles orthogonally onto the surface. The critical points of f then satisfy

$$\text{grad } f(u, v) = \left(\frac{\partial f}{\partial u}, \frac{\partial f}{\partial v} \right) = (0, 0) \quad (8.23)$$

This is a pair of simultaneous, non-linear equations that can be solved using Newton's method, given a good enough starting point for the iteration. A straightforward strategy for locating the critical points is therefore to take a grid of points in each coordinate chart and test if the Newton-Raphson steps for Equation (8.23) from each point converge to a critical point within some pre-set number of iterations. The derivatives of f with respect to the local

coordinates required for Newton's method (and also to classify the critical points as maxima, minima and saddles) are calculated by Rheinboldt.¹²⁷

Now it is impossible to know whether such a brute force search has found *all* the critical points of f , but there are two invariants that can be used to check that at least a consistent set of critical points has been found. The first of these is a local invariant called the *winding number*.¹²⁸ It can be calculated for any closed curve C on S that does not pass through a critical point. During a traversal of C , the direction vector $\mathbf{d} = \text{grad } f(x)/|\text{grad } f(x)|$ of the gradient field makes an integer number of rotations. Counting clockwise rotations of $\mathbf{d}$ as $+1$ and counter-clockwise rotations as -1 , the winding number $\omega(C)$ is the number of rotations made by $\mathbf{d}$ during a clockwise trip around C . The winding number is related to the number of maxima ($N_{\max}$), saddles (N_{sad}) and minima ($N_{\min}$) lying inside C by the alternating sum

$$\omega(C) = N_{\max} - N_{\text{sad}} + N_{\min} \quad (8.24)$$

Applying this to the boundary of each coordinate chart identifies those charts where an incomplete set of critical points has been found and the grid of starting points for the critical point search must be refined.

In addition to these local checks, there is a global constraint on the number of critical points of each type on the whole surface, governed by the Euler characteristic $\chi(S)$. This is a topological invariant of S , satisfying $\chi(S) = 2 - 2\gamma(S)$, where $\gamma(S)$ is the genus (the number of handles) of S . For example, a spherical molecular surface has $\gamma(S) = 0$ and $\chi(S) = 2$. The Poincaré-Hopf index formula¹²⁸ relates $\chi(S)$ to the number of critical points on S :

$$\chi(S) = N_{\max} - N_{\text{sad}} + N_{\min} \quad (8.25)$$

Having found a consistent set of critical points, the separatrices can be traced by integrating the gradient field from carefully chosen starting points near the saddles. For a hyperbolic saddle p , the eigenvectors of the Hessian $HF(p)$ are linear approximations to the separatrices in the tangent space to the surface at p . Taking two points close to p on the unstable eigenvector, one on either side of p , projecting these points to the surface and integrating the field forwards in time provides an approximation to the unstable manifold $W^u(p)$. Similarly, integrating the field backwards in time from the projections of points close to p on the stable eigenvector gives an approximation to the stable manifold $W^s(p)$. In each coordinate chart the gradient field is simply a two-dimensional system of ordinary differential equations and a standard Runge-Kutta method suffices for the numerical integration. The integration is continued until the orbit approaches a sink or a source.

This algorithm has been implemented for the Gabdoulline-Wade surface (Equation 8.2).¹²⁵ It can also be applied to spherical harmonic surfaces, where the existence of a global coordinate system avoids the book-keeping required to track the local coordinate charts of a general implicit surface. The derivatives of the spherical harmonic expansions needed in this case are calculated by

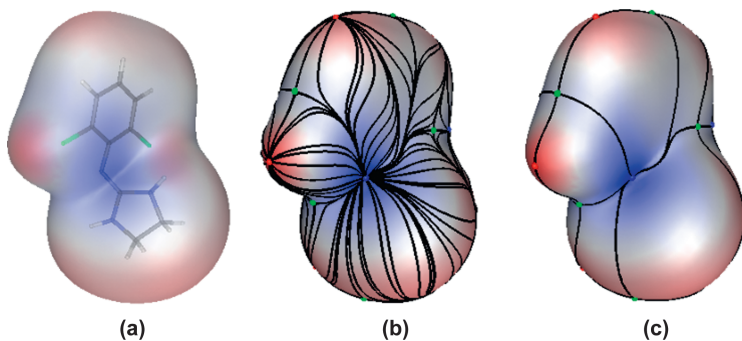


Figure 8.7 The $L = 6$ spherical harmonic expansion of the MEP for clonidine. (a) MEP, (b) gradient vector field, (c) MSPG. Critical points shown as small spheres: maxima (red), saddles (green), minima (blue).

differentiating well-known recurrence formulae for the Legendre polynomials,¹²⁹ and the critical point consistency checks are applied to the winding numbers of the field around the faces in the triangulation used to calculate the spherical harmonic coefficients. Figure 8.7 shows the gradient vector field and the Morse-Smale complex for the order 6 expansion of the MEP for clonidine. The sketches of the Morse-Smale complex and the basins in Figure 8.6 correspond to this example.

Two MSPGs G_1 and G_2 can be compared by finding their maximum common edge subgraph¹³⁰ G_{12} and calculating their similarity as¹³¹

$$s(G_1, G_2) = \frac{(V_{12} + E_{12})^2}{(V_1 + E_1)(V_2 + E_2)} \quad (8.26)$$

where G_i has V_i vertices and E_i edges ($i = 1, 2$) and G_{12} has V_{12} vertices and E_{12} edges.

Morse-Smale gradient flows on surfaces have two particularly important properties. First, they are *typical*, in that they form an open and dense subset (in the appropriate topology) of all the gradient fields on the surface.¹²² In this sense a ‘randomly’ chosen gradient field will be a Morse-Smale field. Second, their qualitative behaviour does not change under small perturbations, a notion that is formalized in the concept of structural stability. Two vector fields X and Y on a surface S are *topologically equivalent* if there is a homeomorphism mapping orbits of X to orbits of Y , preserving their orientation. A vector field is *structurally stable* if all nearby fields (again, in the appropriate technical sense) are topologically equivalent. Among gradient vector fields on surfaces the structurally stable fields are precisely the Morse-Smale fields. Structural stability of a gradient field on a surface is equivalent to the following two conditions:

1. each fixed point is hyperbolic and
2. there are no orbits joining saddle points.

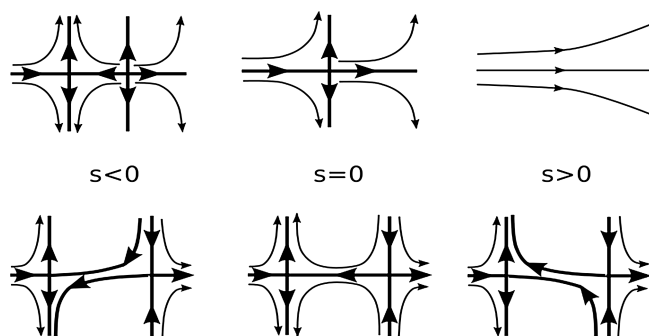


Figure 8.8 Bifurcations of one-parameter families of surface gradient fields: saddle-node (top), saddle-connection (bottom). The bifurcations occur as the parameter s passes through the value $s = 0$; the separatrices are shown as bold lines.

Although Morse-Smale gradient fields are stable under small perturbations, larger variations can produce situations where condition 1 or 2 is violated and a *bifurcation* occurs in which the topological type of the field changes. In one-parameter families of gradient fields on surfaces only two bifurcations are typically seen.¹³² The first is the *saddle-node* bifurcation (Figure 8.8(top)) in which two hyperbolic critical points move towards each other until they coincide in a non-hyperbolic critical point and then vanish. This is a local bifurcation that can be detected by the fact that $Hf(p)$ has a zero eigenvalue at the bifurcation point. The second is the *saddle-connection* bifurcation (Figure 8.8(bottom)), where the family of vector fields passes through a situation where the stable manifold of one saddle coincides with the unstable manifold of another, re-arranging the relative positions of the separatrices in the process. This is a global bifurcation that cannot be detected by local analysis of the critical points.

In studying molecular surface properties these bifurcations arise in three different contexts. First, the surfaces and properties depend on various model parameters (atomic radii, iso-levels, *etc.*) and bifurcation theory may be used to identify parameter ranges in which the models are structurally stable. Second, conformational change provides a natural parametric variation of the molecular surface. For a specific property, the conformations of a molecule may be separated into classes of topologically equivalent fields to produce a property-based partition of conformational space. Finally, MSPGs often contain critical points corresponding to rather small variations in the surface properties that are of little practical interest. These relatively minor features can be removed by forcing ‘virtual’ saddle-node bifurcations between adjacent critical points whose property values lie within some limit, reconnecting the edges of the MSPG as if a real saddle-node bifurcation had occurred. In this way a simplified MSPG is obtained, retaining only salient critical points¹³³ in a manner that is related to the idea of persistent homology.^{120,121}

8.6 Concluding Remarks

The volume of literature on molecular shape and similarity is daunting but, as a recent perspective¹³⁴ shows, the importance of these topics to drug design and medicinal chemistry ensures that developments in this area are still required and will doubtless continue apace. Despite restricting attention to specifically surface-based methods, it has only been possible here to trace a relatively narrow track through the wide range of methods and applications that have been pursued, and several major topics (the use of surfaces in docking and the characterization of binding site cavities, for instance) have gone unmentioned. The improvements in computational power that have transformed so many disciplines have enabled attention in this area to shift from the most basic visualization of molecular surfaces to the calculation and analysis of increasingly realistic surface properties. Despite their inherent artificiality, the early geometrical models have proved immensely useful (and enduring – the MSMS¹⁰ calculation of the Connolly surface remains in widespread use 15 years on). Future work, however, will most likely be based on improved triangulations of iso-surfaces.

Feature-centred approaches to analysing surface properties lead naturally to consideration of their gradient fields, whose behaviour in terms of dynamics and bifurcation theory is well established. A notable aspect is that the saddle points, neglected by most authors in favour of the local maxima and minima, are seen from this point of view as central through the way their separatrices organize the overall structure of the field. Isolating the MSPG as the major invariant of interest allows graph matching methods, familiar to chemists from work on comparing molecular structure graphs, to be applied to obtain property-based alignments and similarity measures. The combinatorial nature of the graph matching, shared with other feature-based methods, means it cannot compete in performance terms with alignment programs such as ROCS³⁴ or USR.¹³⁵ However, the ability to locate surface features with direct relevance to molecular interactions¹¹⁹ offers potentially greater insights. Early indications are that the situation can be more complicated than might at first be expected (properties often have many critical points and conformational change around, say, a single rotatable bond can involve many bifurcations) but focussing on the Morse-Smale complex provides a systematic framework in which to analyse molecular surface properties.

References

1. P. Mezey, *Shape in Chemistry: An Introduction to Molecular Shape and Topology*, VCH, New York, 1994.
2. R. B. Corey and L. Pauling, *Rev. Sci. Instrum.*, 1953, **24**, 621.
3. B. Lee and F. M. Richards, *J. Mol. Biol.*, 1971, **55**, 379.
4. F. M. Richards, *Ann. Rev. Biophys. Bioeng.*, 1977, **6**, 151.
5. J. Greer and B. L. Bush, *Proc. Natl. Acad. Sci. USA*, 1978, **75**, 303.

6. M. Phillips, S. Levy and T. Munzner, *Not. Am. Math. Soc.*, 1993, **8**, 77.
Geomview is available at <http://www.geomview.org/>
7. P. Alard and S. J. Wodak, *J. Comput. Chem.*, 1991, **12**, 918.
8. M. L. Connolly, *J. Appl. Crystallogr.*, 1983, **16**, 548.
9. M. L. Connolly, *J. Mol. Graph.*, 1993, **11**, 139.
10. M. F. Sanner, A. J. Olson and J.-C. Spehner, *Biopolymers*, 1996, **38**, 305.
11. H. Edelsbrunner and E. P. Mucke, *ACM Trans. Graph.*, 1994, **13**, 43.
12. Y. N. Vorobjev and J. Hermans, *Biophys. J.*, 1997, **73**, 722.
13. F. Eisenhaber and P. Argos, *J. Comput. Chem.*, 1993, **14**, 1272.
14. A. Varshney and F. P. Brooks, *IEEE Comput. Graph. Appl.*, 1994, **14**, 19.
15. C. Bajaj, H. Y. Lee, R. Merkert and V. Pascucci, in *Proceedings of the Fourth ACM Symposium on Solid Modeling and Applications*, ACM Press, New York, 1997, p. 217.
16. S. L. Chan and B. O. Purisima, *J. Comput. Chem.*, 1998, **19**, 1268.
17. T. J. Richmond, *J. Mol. Biol.*, 1984, **178**, 63.
18. G. Perrot, B. Cheng, K. D. Gibson, J. Vila, K. A. Palmer, A. Nayeem, B. Maigret and H. A. Scheraga, *J. Comput. Chem.*, 1992, **13**, 1.
19. V. Gogonea and E. Osawa, *J. Comput. Chem.*, 1995, **16**, 817.
20. R. Fraczekiewicz and W. Braun, *J. Comput. Chem.*, 1998, **19**, 319.
21. J. L. Pascual-Ahuir and E. Silla, *J. Comput. Chem.*, 1990, **11**, 1047.
22. M. Gerstein, F. M. Richards, M. S. Chapman and M. L. Connolly, *Intl. Tables Crystallogr.*, 2006, **F**, 539.
23. P. S. Ivanov, P. Jain, D. Osei-Kuffuor, V. S. Stanev, R. Tripathi, P. N. Zhivkov and H. J. Bernstein, in *Proceedings of The National Conference on Undergraduate Research (NCUR)*, Salt Lake City, 2003.
24. T. Can, C.-I. Chen and Y.-F. Wang, *J. Mol. Graph. Model.*, 2006, **25**, 442.
25. W. Chen, J. Zheng and Y. Cai, *Comput.-Aid. Des.*, 2010, **42**, 267.
26. P. W. Bates, G. W. Wei and S. Zhao, *J. Comput. Chem.*, 2007, **29**, 380.
27. H. Edelsbrunner, *Discrete Comput. Geom.*, 1999, **21**, 87.
28. H.-L. Cheng, T. K. Dey, H. Edelsbrunner and J. Sullivan, *Discrete Comput. Geom.*, 2001, **25**, 525.
29. N. Kruithof and G. Vegter, *Comput.-Aid. Des.*, 2004, **36**, 1075.
30. H.-L. Cheng and X. Shi, *Comput. Geom.*, 2009, **42**, 196.
31. CGAL, Computational Geometry Algorithms Library, <http://www.cgal.org/>
32. M. Chavent, B. Levy and B. Maigret, *J. Mol. Graph. Model.*, 2008, **27**, 209.
33. J. A. Grant and B. T. Pickup, *J. Phys. Chem.*, 1995, **99**, 3503.
34. J. A. Grant, M. A. Gallardo and B. J. Pickup, *J. Comput. Chem.*, 1996, **17**, 1653.
35. R. R. Gabdouliline and R. C. Wade, *J. Mol. Graph.*, 1996, **14**, 341.
36. W. E. Lorensen and H. E. Cline, *SIGGRAPH Comput. Graph.*, 1987, **21**, 163.
37. T. Heiden, T. Goetze and J. Brickmann, *J. Comput. Chem.*, 1999, **14**, 246.
38. GNU Triangulated Surface Library, <http://gts.sourceforge.net/>
39. J.-D. Boissonnat and S. Oudot, *Graph. Model.*, 2005, **67**, 405.
40. M. Henderson, *Int. J. Bifur. Chaos Appl. Sci. Eng.*, 2002, **12**, 451.

41. N. Max and E. D. Getzhoff, *IEEE Comput. Graph. Appl.*, 1988, **8**, 42.
42. B. S. Duncan and A. J. Olson, *Biopolymers*, 1993, **33**, 219.
43. S. E. Leicester, J. L. Finney and R. P. Bywater, *J. Mol. Graph.*, 1988, **6**, 104.
44. J.-H. Lin and T. Clark, *J. Chem. Inf. Model.*, 2005, **45**, 1010.
45. R. J. Morris, R. J. Najmanovich, A. Kahraman and J. M. Thornton, *Bioinformatics*, 2005, **21**, 2347.
46. D. W. Ritchie and G. J. L. Kemp, *J. Comput. Chem.*, 1999, **20**, 383.
47. R. J. Morris, *J. Mol. Graph. Model.*, 2005, **24**, 356.
48. K. M. Gorski, E. Hivon, A. J. Banday, B. D. Wandelt, F. K. Hansen, M. Reinecke and M. Bartelmann, *ApJ.*, 2005, **622**, 759.
49. M. R. Calabretta and B. F. Roukema, *Mon. Not. R. Astron. Soc.*, 2007, **381**, 865.
50. J. H. van Drie, *J. Chem. Inf. Comp. Sci.*, 1997, **37**, 38.
51. W. Cai, X. Shao and B. Maigret, *J. Mol. Graph. Model.*, 2002, **20**, 313.
52. W. Cai, M. Zhang and B. Maigret, *J. Comput. Chem.*, 1998, **19**, 1805.
53. A. Kahraman, R. J. Morris, R. A. Laskowski and J. M. Thornton, *J. Mol. Biol.*, 2007, **368**, 283.
54. L. Mak, S. Grandison and R. J. Morris, *J. Mol. Graph. Model.*, 2008, **26**, 1035.
55. V. Venkatraman, L. Sael and D. Kihara, *Cell Biochem. Biophys.*, 2009, **54**, 23.
56. R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, 2000.
57. M. P. Do Carmo, *Differential Geometry of Curves and Surfaces*, Prentice-Hall, 1976.
58. R. Goldman, *Comput. Aid. Geom. D.*, 2005, **22**, 632.
59. G. E. Schröder-Turk, W. Mickel, S. C. Kapfer, F. M. Schaller, B. Breidenbach, D. Hug and K. Mecke, ArXiv:1009.2340 [cond-mat.soft], 2010.
60. B. S. Duncan and A. J. Olson, *Biopolymers*, 1993, **33**, 231.
61. W. Heiden and J. Brickmann, *J. Mol. Graph.*, 1994, **12**, 106.
62. J. Gasteiger and M. Marsili, *Tetrahedron*, 1980, **36**, 3219.
63. T. A. Halgren, *J. Comput. Chem.*, 1996, **17**, 490.
64. J. W. Ponder and D. A. Case, *Adv. Prot. Chem.*, 2003, **66**, 27.
65. J. P. P. Stewart, *J. Comput.-Aid. Mol. Des.*, 1990, **4**, 1.
66. C. E. Whitehead, C. M. Brenneman, N. Sukumar and M. D. Ryan, *J. Comput. Chem.*, 2003, **24**, 512.
67. A. H. C. Horn, J.-H. Lin and T. Clark, *Theor. Chem. Acc.*, 2005, **114**, 159. Erratum: *Theor. Chem. Acc.*, 2007, **117**, 461.
68. N. A. Baker, *Rev. Comput. Chem.*, 2005, **21**, 349.
69. B. Z. Lu, Y. C. Zhou, M. J. Holst and J. A. McCammon, *Commun. Comput. Phys.*, 2008, **3**, 973.
70. P. Sjöberg, J. S. Murray and P. Politzer, *Can. J. Chem.*, 1990, **68**, 1440.
71. B. Ehresman, B. Martin, A. H. C. Horn and T. Clark, *J. Mol. Model.*, 2003, **9**, 342.
72. T. Clark, *J. Mol. Model.*, 2010, **16**, 1231.
73. B. Martin, P. Gedeck and T. Clark, *Int. J. Quant. Chem.*, 2000, **77**, 473.

74. E. Audry, J. P. Dubost, J. C. Colleter and P. Dallet, *Eur. J. Med. Chem.*, 1986, **21**, 71.
75. J. L. Fauchere, P. Quarendon and L. Katterer, *J. Mol. Graph.*, 1988, **6**, 203.
76. P. Gaillard, P.-A. Carrupt, B. Testa and A. Boudon, *J. Comput.-Aid. Mol. Des.*, 1994, **8**, 83.
77. W. Heiden, G. Moeckel and J. Brickmann, *J. Comput.-Aid. Mol. Des.*, 1993, **7**, 503.
78. P. Broto, G. Moreau and C. Vandycke, *Eur. J. Med. Chem.*, 1984, **19**, 71.
79. A. K. Ghose and G. M. Crippen, *J. Comput. Chem.*, 1986, **7**, 565.
80. A. K. Ghose, A. Pritchett and G. M. Crippen, *J. Comput. Chem.*, 1988, **9**, 80.
81. V. N. Viswanadhan, A. K. Ghose, G. R. Revankar and R. K. Robins, *J. Chem. Inf. Comput. Sci.*, 1989, **29**, 163.
82. P. Pixner, W. Heiden, H. Merx, G. Moeckel, A. Moller and J. Brickmann, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 1309.
83. R. Jager, S. M. Cast and J. Brickmann, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 237.
84. B. Ehresmann, M. J. de Groot and T. Clark, *J. Chem. Inf. Model.*, 2005, **45**, 1053.
85. M. A. Johnson and G. M. Maggiora (eds.), *Concepts and Applications of Molecular Similarity*, Wiley, New York, 1990.
86. P. M. Dean (ed.), *Molecular Similarity in Drug Design*, Blackie, Glasgow, 1995.
87. A. Bender and R. C. Glen, *Org. Biomol. Chem.*, 2004, **2**, 3204.
88. P.-L. Chau and P. M. Dean, *J. Mol. Graph.*, 1987, **5**, 97.
89. P. M. Dean and P.-L. Chau, *J. Mol. Graph.*, 1987, **5**, 152.
90. P.-L. Chau and P. M. Dean, *J. Comput.-Aid. Mol. Des.*, 1994, **8**, 513.
91. L. Mavridis, B. D. Hudson and D. W. Ritchie, *J. Chem. Inf. Model.*, 2007, **47**, 1787.
92. W. Cai, J. Xu, X. Shao, V. Leroux, A. Beaudrait and B. Maigret, *J. Mol. Model.*, 2008, **14**, 393.
93. D. T. Manallack, *J. Mol. Model.*, 2008, **14**, 797.
94. V. I. Perez-Nueno, V. Venkatraman, L. Mavridis, T. Clark and D. W. Ritchie, *Mol. Info.*, 2010, **30**, 151.
95. M. Wagener, J. Sadowski and J. Gasteiger, *J. Am. Chem. Soc.*, 1995, **117**, 7769.
96. S. Moro, M. Bacilieri, C. Ferrari and G. Spalluto, *Curr. Drug Disc. Tech.*, 2005, **2**, 13.
97. N. Stiefl and K. Baumann, *J. Med. Chem.*, 2003, **46**, 1390.
98. R. P. Sheridan, M. D. Miller, D. J. Underwood and S. K. Kearsley, *Chem. Inf. Comput. Sci.*, 1996, **36**, 128.
99. R. D. Brown and Y. C. Martin, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 572.
100. A. Bender, H. Y. Mussa, G. S. Gill and R. C. Glen, *J. Med. Chem.*, 2004, **47**, 6569.
101. H. J. Wolfson and I. Rigoutsos, *IEEE Comput. Sci. Eng.*, 1997, **4**, 10.
102. J. Kerbosch and C. Bron, *Commun. ACM.*, 1973, **16**, 575.

103. C.-D. Zachmann, W. Heiden, M. Schlenkrich and J. Brickmann, *J. Comput. Chem.*, 1992, **13**, 76.
104. S. L. Lin, R. Nussinov, D. Fischer and H. J. Wolfson, *Proteins: Struct., Funct. Bioinform.*, 1994, **18**, 94.
105. D. Fischer, S. L. Lin, H. L. Wolfson and R. Nussinov, *J. Mol. Biol.*, 1995, **248**, 459.
106. B. B. Goldman and W. T. Wipke, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 644.
107. B. B. Goldman and W. T. Wipke, *Proteins: Struct., Funct., Genet.*, 2000, **38**, 79.
108. D. A. Cosgrove, D. M. Bayada and A. P. Johnson, *J. Comput.-Aid. Mol. Des.*, 2000, **14**, 573.
109. S. J. Pickering, A. J. Bulpitt, N. Efford, N. D. Gold and D. R. Westhead, *Comput. Chem.*, 2001, **26**, 79.
110. D. Baum and H.-C. Hege, in *Computational Life Sciences II, Lecture Notes in Computer Science, Vol. 4216*, ed. M. R. Berthold, R. Glen and I. Fischer, Springer, Berlin/Heidelberg, 2006, p. 183.
111. T. E. Exner, M. Keil and J. Brickmann, *J. Comput. Chem.*, 2002, **23**, 1176.
112. T. E. Exner, M. Keil and J. Brickmann, *J. Comput. Chem.*, 2002, **23**, 1188.
113. C. Hofbauer, H. Lohninger and A. Aszódi, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 837.
114. D. Zhang and M. Herbert, *IEEE Conference on Computer Vision and Pattern Recognition*, 1999, **2**, 2524.
115. E. Proschak, M. Rupp, S. Derksen and G. Schneider, *J. Comput. Chem.*, 2008, **29**, 108.
116. A.-J. Jakobi, H. Mauser and T. Clark, *J. Mol. Model.*, 2008, **14**, 547.
117. B. D. Hudson, D. C. Whitley, M. G. Ford, M. Swain and J. W. Essex, *J. Mol. Model.*, 2008, **14**, 49.
118. D. D. Robinson, P. D. Lyne and W. G. Richards, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 503.
119. F. A. Bulat, A. Toro-Labbe, T. Brink, J. S. Murray and P. Politzer, *J. Mol. Model.*, 2010, **16**, 1679.
120. A. J. Zomorodian, *Topology for Computing*, Cambridge University Press, 2009.
121. A. J. Zomorodian, in *Algorithms and Theory of Computation Handbook, Second Edition*, ed. M. J. Atallah and M. Blanton, CRC Press, 2009, Volume 2, Chapter 3.
122. J. Palis and W. de Melo, *Geometric Theory of Dynamical Systems. An Introduction*, Springer, New York, 1982.
123. H. Edelsbrunner, J. Harer and A. Zomorodian, *Discrete Comput. Geom.*, 2003, **30**, 87.
124. D. C. Whitley, in *Proceedings of the 9th European Symposium on Quantitative Structure-Activity Relationships*, ed. M. Ford, D. Livingstone, J. Dearden and H. Van der Waterbeemd, Blackwell, Oxford, 2003, p. 428.
125. V. Venkatraman, *Graphical Representations of Molecular Surface Properties*, PhD Thesis, University of Portsmouth, 2007.

126. V. Natarajan, Y. Wang, P.-T. Bremer, V. Pascucci and B. Hamann, *Comput. Aided Geom. D.*, 2006, **23**, 495.
127. W. Rheinboldt, *Comp. Math. Applic.*, 1996, **32**, 15.
128. V. Guillemin and A. Pollack, *Differential Topology*, Prentice-Hall, New York, 1974.
129. M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover Publications, New York, 1965.
130. E. Krissinel and K. Henrick, *Softw. Pract. Exper.*, 2004, **34**, 591.
131. J. Raymond, E. Gardiner and P. Willett, *Comput. J.*, 2002, **45**, 631.
132. J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*, Springer, New York, 1983.
133. Y.-S. Liu, M. Liu, D. Kihara and K. Ramani, in *Proceedings of the 2007 ACM Symposium on Solid and Physical Modelling*, ACM, New York, 2007, p. 277.
134. A. Nicholls, G. B. McGaughey, R. P. Sheridan, A. C. Good, G. Warren, M. Mathieu, S. W. Muchmore, S. P. Brown, J. A. Grant, J. A. Haigh, N. Nevins, A. N. Jain and B. Kelley, *J. Med. Chem.*, 2010, **53**, 3862.
135. P. J. Ballester and W. G. Richards, *Proc. R. Soc. A*, 2007, **463**, 1307.

CHAPTER 9

Atomistic Modelling of Drug Delivery Systems

PAUL COX

Centre for Molecular Design, School of Pharmacy and Biomedical Sciences,
Institute of Biomedical and Biomolecular Sciences, University of Portsmouth,
Mercantile House, Hampshire Terrace, Portsmouth PO1 2EG, UK

9.1 Introduction

The development of new and effective drug delivery systems is of fundamental importance to improve the pharmacological profiles of many classes of therapeutic molecules. However, in comparison with drug design work, very few molecular modelling studies have focused on drug delivery systems. Mathematical modelling methods have been used to investigate the properties of bulk systems.¹ The aim of this chapter is to describe the application of modelling methods to probe drug delivery systems at the atomic level and therefore these mathematical methods are beyond the scope of this chapter. Readers interested in mathematical methods are referred to the excellent review paper by Siepmann and Siepmann.¹

There are several very good technical reasons why atomistic modelling of drug delivery systems has lagged behind work on drug design.

1. Large numbers of atoms are often required to produce a good model. In addition, solvent molecules often play a crucial role in the drug delivery process and these may have to be included in large numbers. Computational times are often, therefore, in practice relatively high.

2. Some systems can involve ‘unusual’ chemistry and therefore the necessary parameters are less readily available and some development may be required.
3. In order to investigate dynamic properties, such as diffusion rates, considerable computational resource is often required in order to investigate how a system evolves over time.

These factors, coupled with economic factors that drive the pharmaceutical industry to focus on the discovery of new drugs, combine to ensure that atomistic modelling of drug delivery systems have been relatively neglected. However, the continuing improvement in computer power and storage space, means that systems with larger numbers of atoms and at higher levels of theory are now much more accessible and, as a result, this is an area that is starting to grow significantly. The key approaches that have been used to investigate other chemical and biological systems at the atomic level are all applicable to the investigation of drug delivery systems. Molecular mechanics, molecular dynamics, docking methods and quantum mechanical approaches have all been used in this area in order to aid our interpretation and understanding of drug delivery systems. The aim of this chapter is to highlight some examples of the application of these methods and to highlight the type of information that can be obtained.

9.2 Molecular Mechanics – Energy Minimisation

Molecular mechanics (MM) methods allow the prediction of both the detailed geometry and the energetics of these systems. This approach is often used in conjunction with other experimental techniques, to yield detailed information about a system under investigation. Once this has been achieved, the methods can often be exploited to help further tailor a system for a specific application. The molecular mechanics approach in itself is relatively simple. It treats atoms as hard spheres and bonds as simple springs. The classical laws of physics are then applied in order to calculate the energy of the system. It has the advantage that the equations involved are relatively simple, allowing large numbers of atoms to be included in any model. The disadvantages of the approach are that reliable parameters need to be available for the system under investigation and that studies requiring a detailed investigation of the electronic structure within a system need to use higher levels of theory (see Quantum Mechanical Methods, Section 9.4).

9.2.1 Encapsulation of Multinuclear Platinum Complexes

An excellent illustration of the use of energy minimisation to help characterise a drug delivery system is provided by the work carried out by Wheate *et al.*^{2,3} and Bali *et al.*³ on multinuclear platinum complexes. The platinum complex cisplatin has been widely used as a treatment for cancer. The use of

cisplatin is, however, limited because of two factors. Firstly, natural and acquired resistance to the drug can limit its effectiveness. Secondly, its side-effects limit the dose that may be given. The novel multinuclear complexes investigated in these studies have novel DNA binding modes which allow them to overcome the problems of resistance. They are also active at lower concentrations, which helps to minimise side-effects. However, novel drug delivery methods are needed for these compounds because not only are they more toxic than cisplatin, they are also particularly susceptible to degradation before they reach their intracellular targets.

In order to overcome these problems, cucurbiturils, which are barrel shaped molecules, were investigated as hosts for the pharmacological delivery of the novel multinuclear platinum compounds. By encapsulating them inside the cucurbiturils, the toxicity may be reduced and also degradation within the body prevented. The modelling work carried out in this investigation used a molecular mechanics approach to optimise the geometry of the cucurbituril–drug complexes. Point charges were determined *via* QM calculations, reflecting the complex chemical nature of the multinuclear drugs. These charges were then incorporated into the molecular mechanics calculations prior to energy minimisation.

In combination with experimental work, the results highlighted the importance of the geometry of the platinum complex within the cucurbiturils. Within a larger cucurbituril structure (see Figure 9.1, centre) the drug molecule is more folded than other analogues, but occupies the space more efficiently and the drug is shielded better. Steric hindrance in a smaller cucurbituril structure (see Figure 9.1, right) limits the folding of the drug inside the cucurbituril structure and therefore provides less effective shielding. These results are consistent with the observed experimental observations that confirm the greater protection of

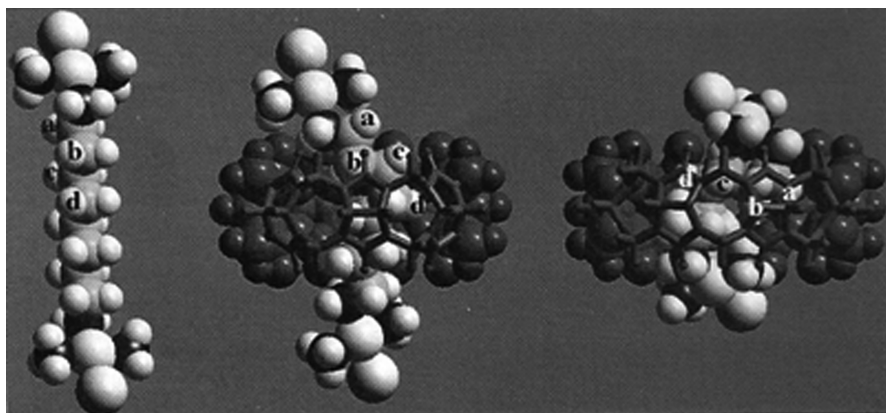


Figure 9.1 *Trans*-[[PtCl(NH₃)₂]₂](μ-NH₂(CH₂)₈NH₂)²⁺ drug optimised (a) outside a cucurbituril structure (b) inside a large cucurbituril structure and (c) inside a smaller cucurbituril structure. Reproduced from ref. 3.

the drug by the larger cucurbituril structure. By providing a clear insight into the relationship between the cucurbituril structures, the drug molecules and the rate of degradation, modelling provides a route for the design and optimisation of a drug delivery system tailored for these promising compounds.

9.2.2 Carbon Nanotubes

An additional illustrative example of the use of molecular mechanics to help develop a potential drug delivery system is provided by work carried out on carbon nanotubes by Roldo *et al.*^{4,5} There is considerable and rising interest in the use of nanotubes in drug delivery because their morphology gives them the ability to pierce cells and thus the potential to deliver their drug cargo directly to the site of action. However, inherent insolubility means that they have a tendency to aggregate in solution, which is an undesirable feature for an efficient drug delivery system. Several methods of improving the solubility of carbon nanotubes have been investigated, one of which is to attempt binding suitable polymer molecules to their external surface.

In these studies, a combination of molecular mechanics and molecular dynamics simulations were carried out in order to investigate the interaction of single walled nanotubes (SWNTs) and biocompatible chitosan polymer derivatives. Here, the molecular dynamics calculations are used to try to help ensure that these calculations, which involve large numbers of atoms and polymers with highly flexible side-chains, reach the global minimum energy state rather than a local minimum. (Later in the chapter we will consider the use of the MD approach to probe the diffusion of drug molecules within delivery agents.) The polymeric compounds investigated in this study were *N*-butyl-*O*-sulfate chitosan (NBSC), *N*-octyl-*O*-sulfate chitosan (NOSC) and *N*-palmitoyl-*O*-sulfate chitosan (NPSC). The key difference between these structures is the length of the alkyl chains which radiate out from the polymer backbone and contain 3, 7 or 14 CH₂ groups, respectively.

The modelling investigation involved the construction of a (8,8) single walled nanotube with an overall length of 150 Å. In this case, the nanotube was terminated with hydrogen atoms attached to the appropriate carbon atoms in order to ensure that no carbon atoms had incomplete coordination. Polymer chains consisting of 70 repeat units were initially energy minimised in the absence of the nanotube. Polymer molecules were then docked within 5 Å from the surface of the nanotube and a combination of molecular dynamics and molecular mechanics was used to determine the lowest energies for the polymers relative to the nanotube. This minimisation was carried out by performing molecular dynamics at high temperature (750 K) for several thousand time-steps (typically 200 000 steps with a time-step of 1×10^{-15} s) followed by energy minimisation using a conjugate gradient algorithm. This sequence was repeated until no further decrease in the energy was obtained. Calculations were performed first using one polymer chain and then several chains were used in order to try to improve the overall binding energy.

The energy minimised structures of the three polymer molecules showed that they all adopt a helical backbone with the alkyl chains radiating outwards. The energy minimised structures for the three polymer molecules are shown in Figure 9.2.

Wherever possible, modelling results should be compared with experimental results in order to establish the validity of the approach. In this case, it was possible to compare the thickness of the polymer-coated nanotubes with measurements obtained from atomic force microscopy studies. Good agreement between theory and modelling was obtained.^{4,5}

The results of the modelling study showed that NPSC has a significantly higher binding energy and a greater coverage of the surface of the nanotube (see Figure 9.3 and Table 9.1). Experimental work confirmed the modelling results, showing that NPSC has the highest affinity for the nanotubes, leading to a significantly

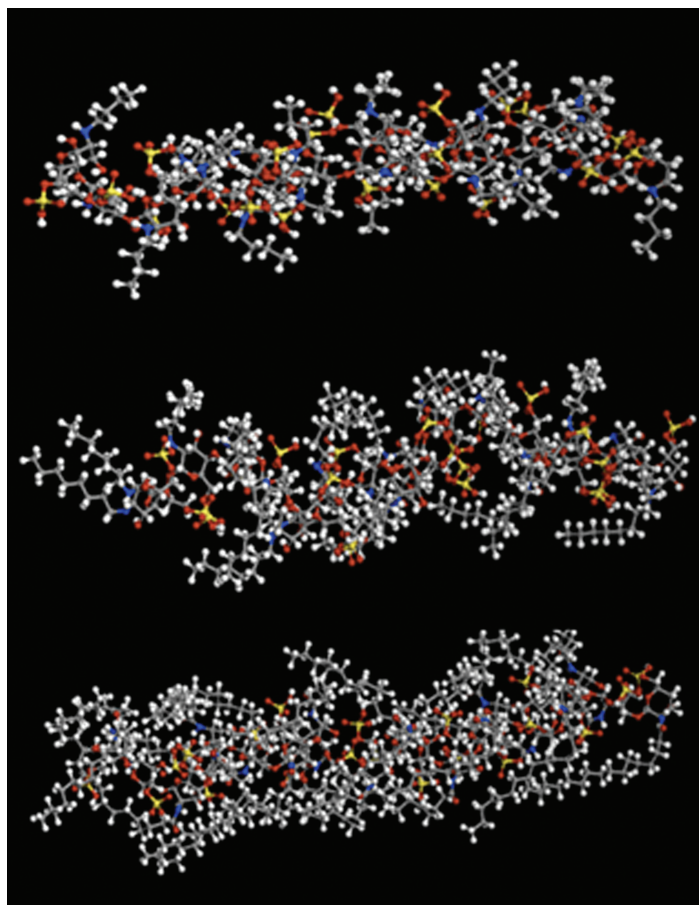


Figure 9.2 The energy minimised conformation of chains of NBSC (top), NOSC (middle) and NPSC (bottom).

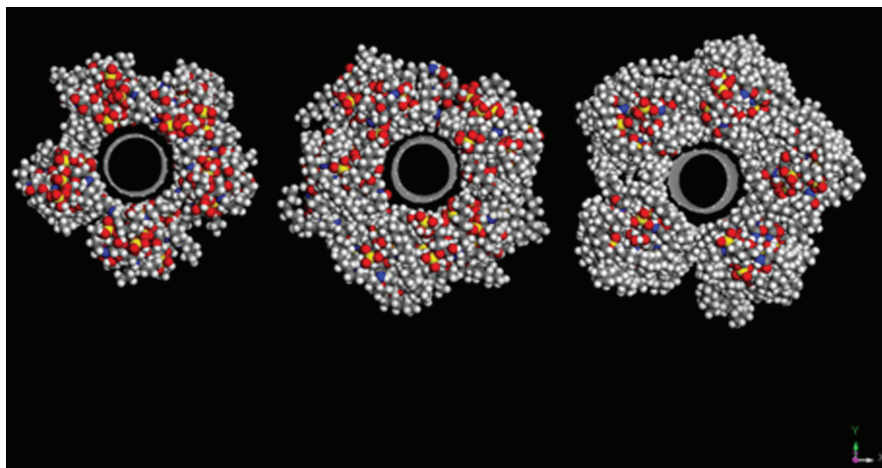


Figure 9.3 The optimised location of five polymer chains around a SWNT for NBSC, NOSC and NPSC, left to right, respectively.

Table 9.1 Binding energies obtained for the three polymer molecules with 1 and 5 chains.

<i>Polymer</i>	<i>Number of polymer chains bound to SWNT</i>	<i>Binding energy per chain/kcal mol⁻¹</i>
NBSC	1	-123.55
NOSC	1	-123.06
NPSC	1	-272.21
NBSC	5	-208.08
NOSC	5	-286.47
NPSC	5	-300.93

better suspension of the nanotubes in aqueous solution, demonstrating the way in which modelling can successfully drive experimental work in this area.

In this particular study, the effect of solvent was included by adjusting the relative permittivity of the solution to an appropriate value. This is one way that the effect of the solvent can be incorporated into calculations. Alternatively this can be achieved by explicitly including water molecules. Both have strengths and weaknesses. The advent of computers with increased processing power is making explicit inclusion of water a more attractive route.

9.2.3 Fullerenes

An additional important area for work in the area of drug delivery systems is concerned with the investigation of the potential use of fullerenes. Their promise lies in the fact that these compounds offer the possibility of being able to deliver drugs directly through cell membranes and into target cells. Owing to their small size (~ 1 nm), fullerenes have high potential for bioavailability.

The fullerene core is very hydrophobic and may be modified chemically by the addition of functional groups. However, like carbon nanotubes discussed in the previous section, their poor water solubility has severely limited their application. One potential way that this problem can be overcome is by encapsulation within liposomes. Liposomes comprise an aqueous core with an outer phospholipid bilayer. This may be used to transport compounds with low aqueous solubility to their target site.

Liposomes have been widely used in drug delivery systems. They have several key advantages:

1. their versatile structure means that they can easily be tailored for specific applications;
2. lipophilic drugs can be accommodated in their bilayers (or hydrophilic ones in their aqueous core);
3. they are without apparent toxicity, are non-immunogenic and are fully biodegradable.

One of the features of calculations for this type of system is the use of periodic boundary conditions to create an infinite system to represent the liposome bilayer. This can be achieved by extending a simulation box in the direction perpendicular to the bilayer itself. Here, it is important to ensure that the extension to the simulation box is sufficiently large that there is not any interaction between the infinite bilayer replicates that are created when the periodic boundary conditions are applied. In this case, water molecules are used to fill the space between the bilayers. However, even with a relatively small simulation box this can mean that several thousand independent water molecules need to be included in the calculation, thus making the calculations relatively time consuming, and also increasing the need for the application of molecular dynamics techniques to help minimise problems associated with obtaining local rather than global minima when energy minimisation is applied to a system with a large numbers of particles.

In this example,⁶ a combination of molecular dynamics and energy minimisation was used to assess the effect of the concentration of fullerene molecules on the integrity of the liposome.

Simulations were carried out using a simulation box constructed using two 4×4 arrays of liposome (egg PC) molecules. Initially, the simulation box was extended to 50 Å in the direction parallel to the bilayer and the space above and below was saturated with water molecules. In total, 5382 water molecules were incorporated into the simulation box. Molecular dynamics calculations were carried out at temperatures of 750, 500 and 300 K for a simulation time of 10 ps prior to energy minimisation using a combination of the steepest descents and conjugate gradient algorithms. The size of the cell was allowed to vary during the optimisation. Simulations were performed with 0, 2, 4, 6 and 8 fullerene molecules per simulation box.

The optimised lipid bilayer in the absence of C_{60} is shown in Figure 9.4. A view of the surface viewed from above (with water molecules removed for clarity) is also shown (Figure 9.5).

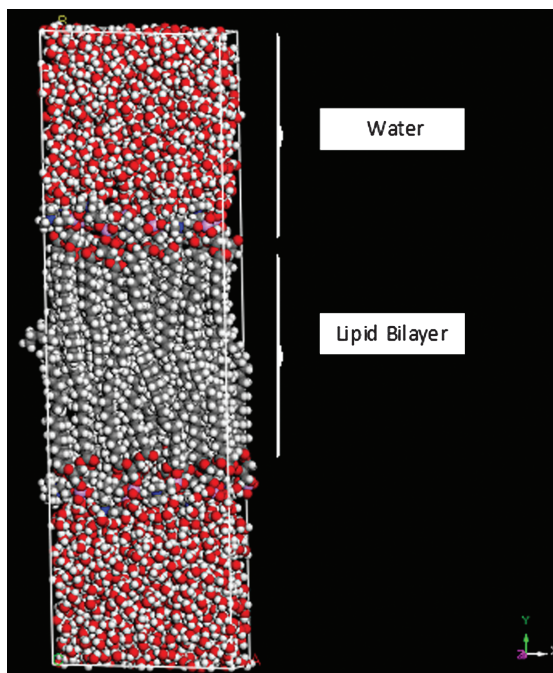


Figure 9.4 The optimised lipid bilayer.

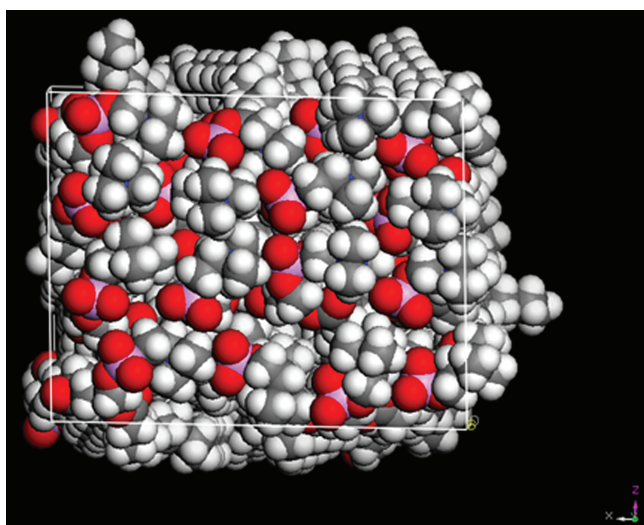


Figure 9.5 The optimised bilayer viewed from above (water molecules omitted from view for clarity).

At low concentrations of C_{60} (2 molecules per simulation box), the location of the carbonaceous material is optimised towards the centre of one of the layers. It can be seen that the hydrophobic tails of adjacent PC layers ‘wrap’ themselves around the spherical C_{60} molecule in order to maximise the van der Waals interactions. At this low concentration of C_{60} , molecules are easily absorbed into the liposome and little disruption of the bilayer is observed. The surface viewed from above appears unchanged, with the C_{60} molecule shielded from view (Figure 9.6).

At the highest concentrations of C_{60} (8 molecules per simulation box) the hydrophobic tails of the phospholipid molecules again ‘mould’ themselves around the fullerene molecules but at these concentrations this moulding effect means that adjacent fullerenes are displaced higher up the bilayer in order to optimise van der Waals interactions (Figure 9.7). This results in significant disruption to the surface of the bilayer and means that C_{60} molecules are now clearly visible when the surface is viewed from above (Figure 9.8). This disruption of the surface is also reflected in the calculated surface area of the bilayer at different concentrations of C_{60} . At low concentration of fullerene (2 molecules per simulation box) and with no fullerene present, the surface area remains largely unchanged, with values of 839.7 and $835.5 \text{ \AA}^2$, respectively. However, at higher concentrations of fullerene (8 molecules per simulation box) the surface area increases significantly to $1003.5 \text{ \AA}^2$, reinforcing the suggestion that low concentrations of fullerene have little effect on the surface of the

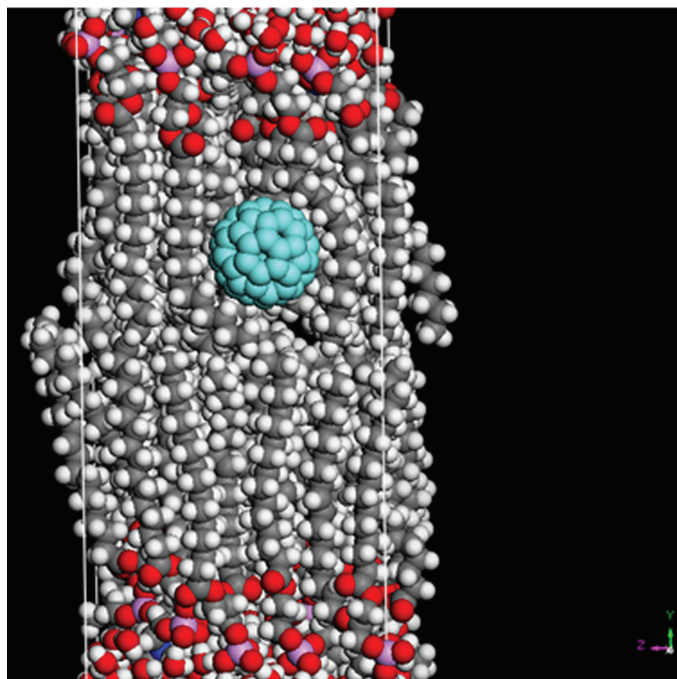


Figure 9.6 The optimised location C_{60} molecule in the lipid bilayer when present at low concentration (water molecules omitted from view for clarity).

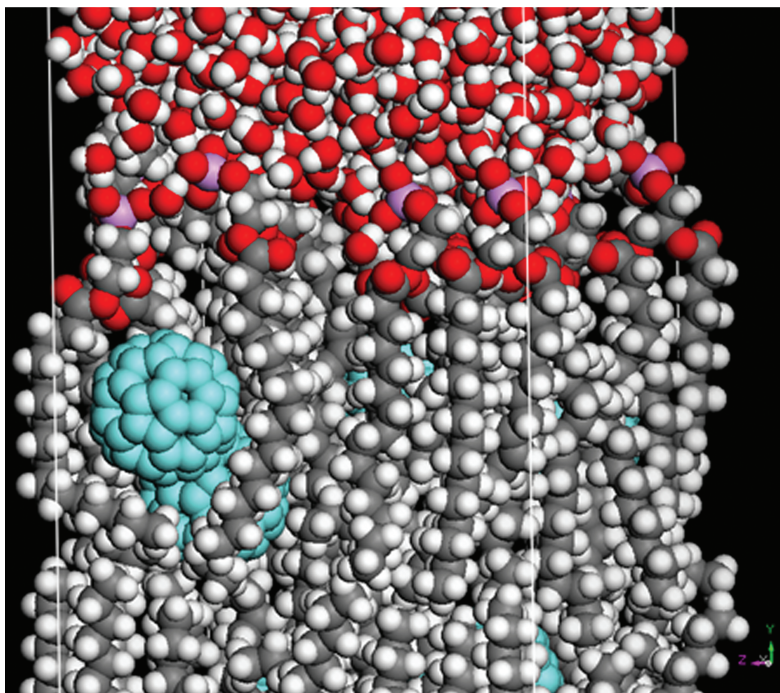


Figure 9.7 The optimized location of C_{60} molecules in the lipid bilayer when present at high concentrations (water molecules omitted from view for clarity).

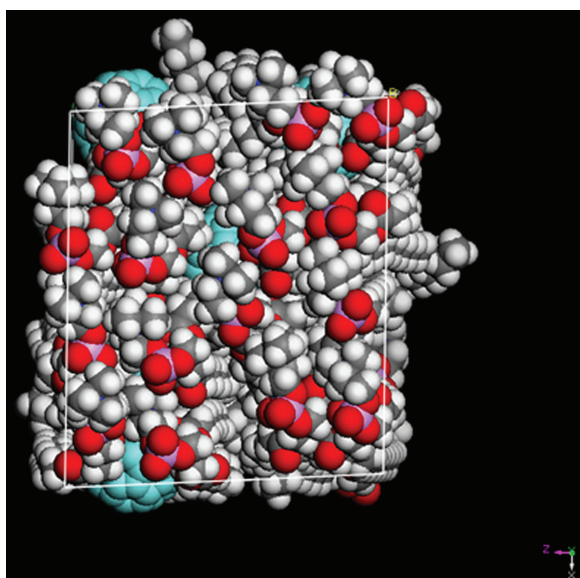


Figure 9.8 The surface of the lipid bilayer viewed from above when C_{60} is present at high concentration (water molecules omitted from view for clarity).

bilayer but that high concentrations can disrupt it substantially. At both concentrations, incorporation of the fullerene into the bilayer results in a favourable binding energy for the C₆₀ molecules (−261.28 and −288.52 kcal mol^{−1} at low and high concentrations respectively) although we note that the binding energy per C₆₀ molecule is higher at low concentration (−130.64) than at high concentration (−36.07 kcal mol^{−1}).

The theoretical observations in this work are supported by experimental measurements that show that these lipid bilayers become ‘leaky’ at the same level of loading at which the modelling suggest that the inclusion of the C₆₀ molecules begins to significantly distort the size of the simulation box. Membrane integrity studies reveal that incorporation of C₆₀ molecules destabilise the liposomal formulations based on the retention values of the fluorescent hydrophilic marker calcein. The % latency of vesicle-encapsulated calcein is used as a measure of membrane integrity. The results show that PC-liposomes loaded with low levels of C₆₀ are stable since after 24 h of incubation high levels of the initially encapsulated calcein are retained in the vesicles. In contrast, when C₆₀ is incorporated in the lipid bilayers at approximately the concentrations of the higher loading in the simulations there is a significant decrease of the retention of calcein after the initial measurement. This supports the theoretical prediction that at high loading levels, C₆₀ molecules disrupt the surface of the bilayer and thus retention of the calcein decreases.

9.3 Molecular Dynamics

The previous examples have relied upon molecular mechanics approaches to optimise the energy of the system under investigation in order to predict the energetics and geometry of the systems under investigation. However, in terms of modelling drug delivery processes a key technique is that of molecular dynamics, which allows us to investigate the time evolution of a system at a given temperature. A molecular mechanics force field is usually at the heart of this type of calculation but particles are initially assigned velocities and directions in order to give the system appropriate levels of kinetic energy. A time step is specified (usually of the order of 1×10^{-15} s) and then essentially Newton’s second law of motion is used to determine the velocities and positions of the atoms as the simulation progresses through time.

9.3.1 Zeolites as Drug Delivery Agents

An interesting example of molecular dynamics applied to a drug delivery is provided by the search for novel inorganic delivery agents. Many inorganic compounds, such as zeolites, are being explored as potential drug carriers for oral delivery. Zeolites are formed by linking SiO₄ and AlO₄ tetrahedra to form crystalline solids.⁷ Currently over 150 different zeolite frameworks have had their structure determined and many of these possess channels that range from ~0.6 to 1.4 nm – potentially large enough to accommodate a wide range of different drug molecules. One of the difficulties that experimentalists face when

investigating zeolites as drug delivery agents is selecting the appropriate structure to choose for the delivery of a particular drug molecule. Modelling methods appear to have an important role to play in this area.

In this study,⁸ molecular dynamics was used to simulate the diffusion of the drug molecule inside its zeolite host at the appropriate temperature (body temperature). Crucially, this allows the flexibility of both the drug and the framework to be taken into account – key factors in determining whether a molecule can diffuse in and out of the zeolite's internal channel system.

An effective example of the type of information that can be determined from this approach is provided by a study into the diffusion of two different drug molecules, theophylline and salbutamol, inside the framework of zeolite BEA – an important zeolite with a multi-connected channel system (Figure 9.9). At first sight, the theophylline and salbutamol molecules might be expected to show similar diffusion behaviour inside the zeolite framework because the two molecules have very similar dimensions in terms of their widths i.e. the second biggest dimension (if you think about posting a parcel through a circular mailbox then it isn't the longest dimension that is likely to be the barrier to successfully putting the parcel through, it will be the next largest dimension).

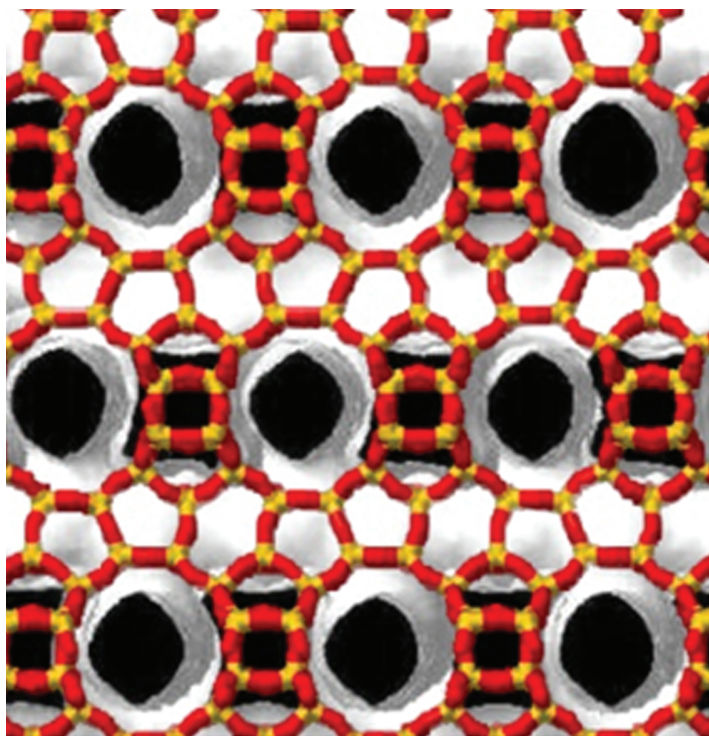


Figure 9.9 Zeolite BEA. The framework atoms are shown in red (cations) and yellow (anions). The internal pore channel system is shown in white. Effectively a drug molecule inside this structure can move in the x , y and z directions.

The structures of theophylline and salbutamol are shown in Figure 9.10. In this study, molecular dynamics (MD) simulations for both salbutamol and theophylline were performed for a total simulation time of 1×10^{-9} s with a time-step of 1×10^{-15} s. The simulations were performed at 310 K. A $3 \times 3 \times 2$ cell of zeolite beta with one drug molecule inside was used as the simulation box. Periodic boundary conditions were used to create an infinite system and full flexibility of both the zeolite framework and the guest molecule was allowed. Flexibility of the framework is important here because the channel dimensions vary slightly because of the vibration of the atoms at this temperature.

Interestingly, the simulations for the two different drug molecules showed significantly different behaviour. Analysis of the MD results showed that salbutamol was able to diffuse through the framework whereas theophylline was not. This is shown in Figure 9.11 by the calculated root-mean squared displacement for the drug molecules as a function of time. The upward slope for salbutamol shows the molecule was diffusing during the simulation while the flat line for theophylline indicates that there was no net movement through the zeolite host. Observation of the trajectories of the two molecules (Figure 9.12) shows that it is the flexibility of the salbutamol molecule that allows it to move through the channel system.



Figure 9.10 The chemical structures of (a) theophylline and (b) salbutamol.

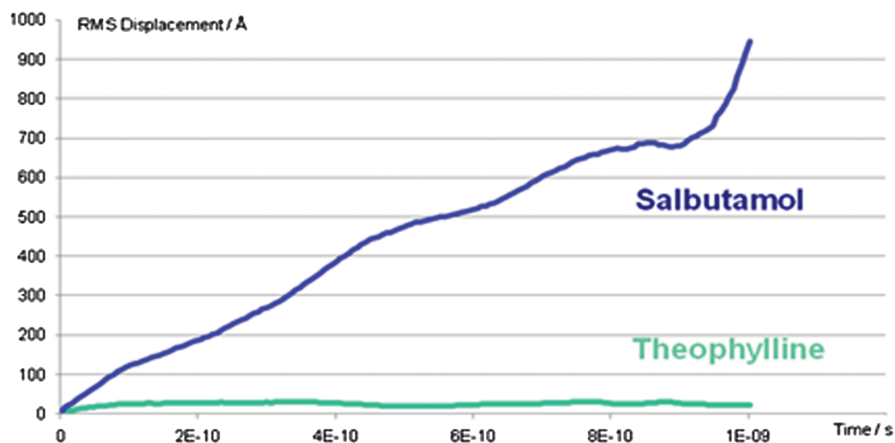


Figure 9.11 Root mean squared displacement graphs for salbutamol and theophylline inside the zeolite.

The relatively rigid structure of the theophylline molecule means that it cannot progress through the zeolite's channels – the van der Waals' interactions between the drug and zeolite are too high and the theophylline molecule is too inflexible to overcome this barrier. However, the salbutamol molecule is much more flexible and when the alignment of the functional groups is favourable, the effective width of the molecule is reduced and it is able to move along a channel (Figure 9.12a). It is important to note that the diffusion of salbutamol within the zeolite is still significantly hindered by the need for it to adopt certain conformations in order to move and hence the channel system exerts a

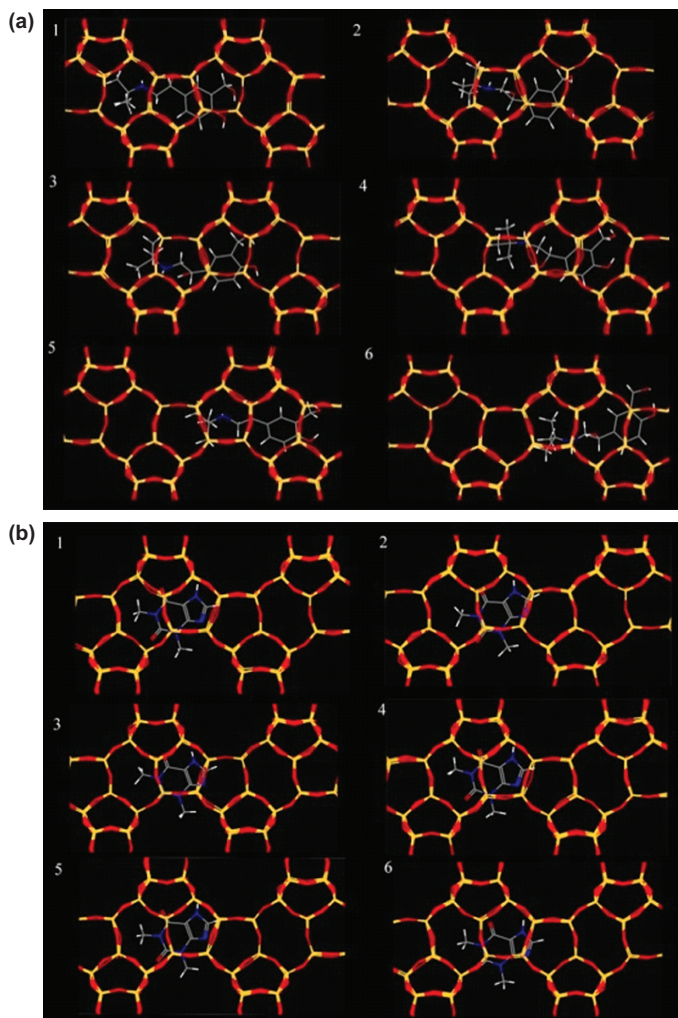


Figure 9.12 (a) The flexibility of the salbutamol molecule allows it to move through the zeolite's channel system. (b) The rigidity of the theophylline molecule means that it gets 'stuck' at channel intersection points and therefore it is not likely to penetrate into the zeolite's internal channel system in significant quantities.

significant constraint on the diffusion of the molecule. These results suggest any experimentally observed release for theophylline is likely to be controlled only by the surface of the zeolite nanocrystals.

Experimental studies confirm the validity and usefulness of the results of these simulations. Drug loading studies via supercritical fluid technology (SCF) reveal significant differences in the amount of drug that is encapsulated inside the zeolite host. For salbutamol and theophylline the %w/w located inside the host framework is 22.5 and 5.1, respectively. This supports the modelling work that suggests that theophylline will not penetrate through the pores, resulting in a low loading level. A value as low as 5.1% for theophylline may simply be attributable to molecules on the surface of the zeolite particles, or trapped at the opening of the channels, rather than penetration of channel system. In contrast, a significantly higher proportion of the salbutamol molecules penetrate inside the internal channel system of the zeolite, as shown by the increase of 17.4% in the amount of drug uptake.

Release rates from zeolite samples loaded with the two drugs are shown in Figure 9.13. For the first 30 min, a steep release is observed for both drugs. However, at prolonged times, the release of salbutamol is significantly delayed. This delayed release can be attributed to the fact that diffusion of salbutamol molecules from within the zeolite's channels are hindered by the types of drug–zeolite interactions observed in the simulations. As a result, complete release of theophylline is reached after 180 min whilst release of salbutamol was prolonged up to 240 min.

These results clearly highlight the potential value of modelling to interpret and guide this type of study in the future. Molecular dynamics has the potential

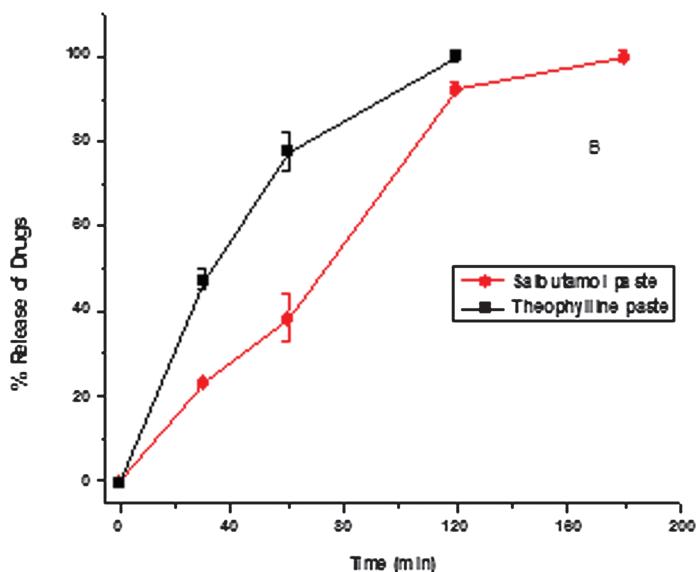


Figure 9.13 The release profiles for salbutamol and theophylline from zeolite BEA.

to screen drug–zeolite combinations prior to further experimental investigation. Moreover, it may be possible to use this type of simulation to ‘fine tune’ the loading and release of appropriate drug molecules by probing the effect of subtle alterations to the zeolite lattice, such as changing the Si/Al ratio in the framework, thereby fully optimising the drug delivery system.

9.3.2 Other Examples of MD Simulations

Molecular dynamics (MD) has proved to be a versatile approach for investigating the time-dependent properties of drug delivery systems as well as for simulating the diffusion of drug molecules from their host. Some of the key areas investigated include the following systems.

1. Carbon nanotubes (see Section 9.2.2): MD has been used to investigate the encapsulation and release of a variety of different drug molecules from these important materials. An excellent recent review has been carried out by Hilder and Hill,⁹ that highlights the use of theoretical investigations into the loading and unloading of drugs into nanotubes. Key MD studies in this area include work on the interaction between the peptide drug zadaxin and nanotubes,^{10,11} studies of the competitive replacement process for peptides attached to the inner or outer surfaces of the tubes,¹² determination of the preferred location of the anticancer drug gemcitabine¹³ and the permeation of water through carbon¹⁴ and TiO₂ nanotubes.¹⁵
2. Liposomes (see Section 9.2.3): MD is used to predict permeability in bilayer membranes. A good review of this area is the paper by Xiang and Anderson.¹⁶ Examples of the range and scope of work in area can be found in the studies of diffusion coefficients for small molecules in DMPC bilayers,¹⁷ the calculation of partition coefficients for a range of polar and non-polar molecules¹⁸ and the binding of anti-cancer drugs for potential liposomal delivery.¹⁹
3. Polymeric delivery systems: diffusion of drugs encapsulated within polymeric systems is highly amenable to MD simulation. Examples include the uptake of the drugs doxorubicin, silymarin and aliclazide by different polymeric systems,²⁰ the study of polymer–drug interactions in copolymer nanospheres,²¹ simulation of ‘smart’ systems that respond to environmental conditions²² and the investigation of small molecule diffusion in membrane materials.²³
4. Dendrimers: these are roughly spherical or disc-shaped compounds that share the common feature of branches emerging from a central core in a radial fashion. Dendrimers have been widely investigated as vehicles for the encapsulation of hydrophobic drugs and have been shown to be effective drug delivery agents for a range of organic compounds. MD has been used to provide information about the detailed structure of dendrimers as well as providing insights into their rational design for drug delivery purposes. Examples of studies in this area include the investigation of the association of ibuprofen with a poly(amidoamine)

dendrimer²⁴ and the bioactive agent shikonin with hyperbranched polyesters,²⁵ examination of the conformational behaviour of peptide dendrimers²⁶ and studies of the use of poly(amidoamine) dendrimers as drug delivery agents for platinum and ruthenium-based drugs.

9.4 Quantum Mechanical Methods

Quantum mechanical methods have been much less widely used in simulating drug delivery systems. This is essentially as a result of the high computational requirements for this type of simulation. One important property that can be obtained from QM simulations that cannot be obtained from MM-based methods is a detailed picture of the distribution and energies for the electrons within molecules.

9.4.1 Delivery of Vancomycin using Mesoporous Materials

Zeolites are classified as microporous materials, since their pore sizes are less than 2 nm. This means that their channels are too small to accommodate larger molecules such as the antibiotic vancomycin. A closely related family of materials are known as mesoporous silicas. These compounds have much larger channels than zeolites and their dimensions can be carefully controlled by the synthesis, with channel sizes up to around 30 nm. The channel walls of these materials are amorphous and the presence of silanol groups (Si–OH) on the surface which makes these compounds challenging to simulate but makes them more amenable to be tailored for their potential application. One way this can be achieved is by functionalising the surface silanols with hydrophobic groups.

The study by Doadrio *et al.*²⁷ made use of an interesting combination of techniques in order to yield a significant insight into the drug release properties of a mesoporous material, SBA-15, before and after functionalisation. Experimental work showed that release rates of vancomycin from SBA-15 are much higher for 'pure' SBA-15 compared with the same material functionalised with octyltrimethoxysilane (C₈) alkyl chains. Semi-empirical QM methods using the AM1 Hamiltonian were used to obtain detailed information on the distribution of charge and the electrostatic potential for both the mesoporous material and the drug (Figure 9.14). A grid-based docking approach has then been used to identify the most favourable binding site for the drug in the modified and unmodified forms of SBA-15. This type of grid-based approach has normally been used for locating the binding sites for drugs in protein structures. They have the advantage of being a fast method of locating a binding site and also help to overcome the problems of locating local rather than global energy minima – particularly in this type of situation when there is not an obvious place to dock the drug molecule prior to performing an energy minimisation. (Another example of the grid-based approach applied to investigate drug delivery systems is provided by the work of Gullick *et al.* in the characterisation and design of captopril pro-drugs for enhanced percutaneous absorption.²⁸)

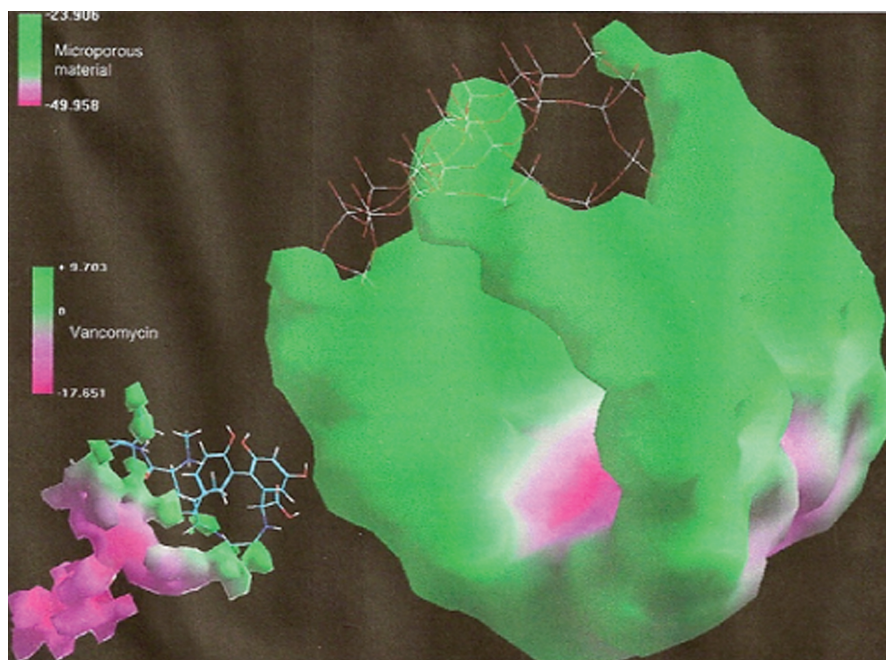


Figure 9.14 Electrostatic maps for vancomycin and the mesoporous framework. Reprinted from *Microporous Mesoporous Mater.*, Doadrio *et al.*, 539–566, Copyright (2010), with permission from Elsevier.

The results show that addition of the C_8 chains reduce the negative charge associated with the surface of SBA-15, which therefore reduces the electrostatic interaction with the positive head group associated with vancomycin. The authors are then able to relate this reduced interaction to the faster release rate obtained for the functionalised SBA-15. Moreover, the results show that vancomycin does not penetrate deeply into the channel system and that only the small positively charged region of the molecule successfully enters the mesopores, even though there appears plenty of room for further penetration. This is an interesting result and demonstrates that not only can important information be learned about release rates but that there is also the potential to understand and predict loading levels too.

9.4.2 Design of Conducting Polymers for Drug Delivery

Another area where molecular modelling techniques have been successfully applied is for studying electronically conducting polymers. This type of polymer may have significant applications in drug delivery where higher conductivities give greater control over the delivery process.²⁹ The first step involved in the synthesis of this type of polymer electrochemically is to remove an electron from the monomer. The resulting radical cations are unstable and readily couple with other cations, with the elimination of two protons, to form dimers. Successive monomer units continue to be added until a polymer chain

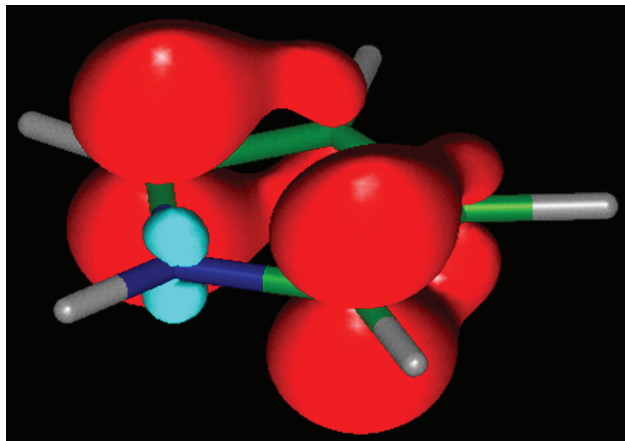


Figure 9.15 The spin density contours for pyrrole.

forms. Quantum mechanical calculations are used to map the location of the unpaired electron within the monomer cation and the growing polymer chains. The result of this type of calculation for the monomer pyrrole is shown (Figure 9.15). The red contour lobes that map the averaged location of the unpaired electron are largest at the carbon positions adjacent to the nitrogen atom (the alpha positions). These results suggest that two pyrrole molecules are likely to couple together at the alpha positions, because this is where the unpaired electrons spend most of their time. Indeed, pyrrole is known to favour coupling at these positions and the calculated magnitude of the spin density at each of the atomic positions in the pyrrole cation ring agrees well with the values obtained from electron spin resonance experiments.^{30,31}

Having established the accuracy of this type of calculation by referring to the pyrrole system, the approach can be extended to predict the coupling positions of other monomer units without having to synthesise the compounds. The position at which molecules couple as the polymer chain grows is critical, because it affects the polymer's conductivity. Ideally, coupling should occur at positions that maintain an alternating single–double bond sequence throughout the molecule (conjugation). Linear growth without significant branching or cross-linking is also desirable.

The conductivities of polymers made from the three monomers have been determined in the laboratory. The results show that conductivity increases from polymer (a) to polymer (c). This observation can be rationalised via modelling by calculating the spin densities associated with the cations of the three molecules. The results of these calculations are shown in Figure 9.16.

For maximum conductivity, these molecules should couple in the alpha positions next to the hetero (nitrogen, sulfur or oxygen) atoms in the five-membered rings at either end of the molecule. The results of the calculated spin density for molecule (a) show two significant regions of spin density centred at the desired alpha carbon positions next to the nitrogen atoms in the five-membered rings. However, there

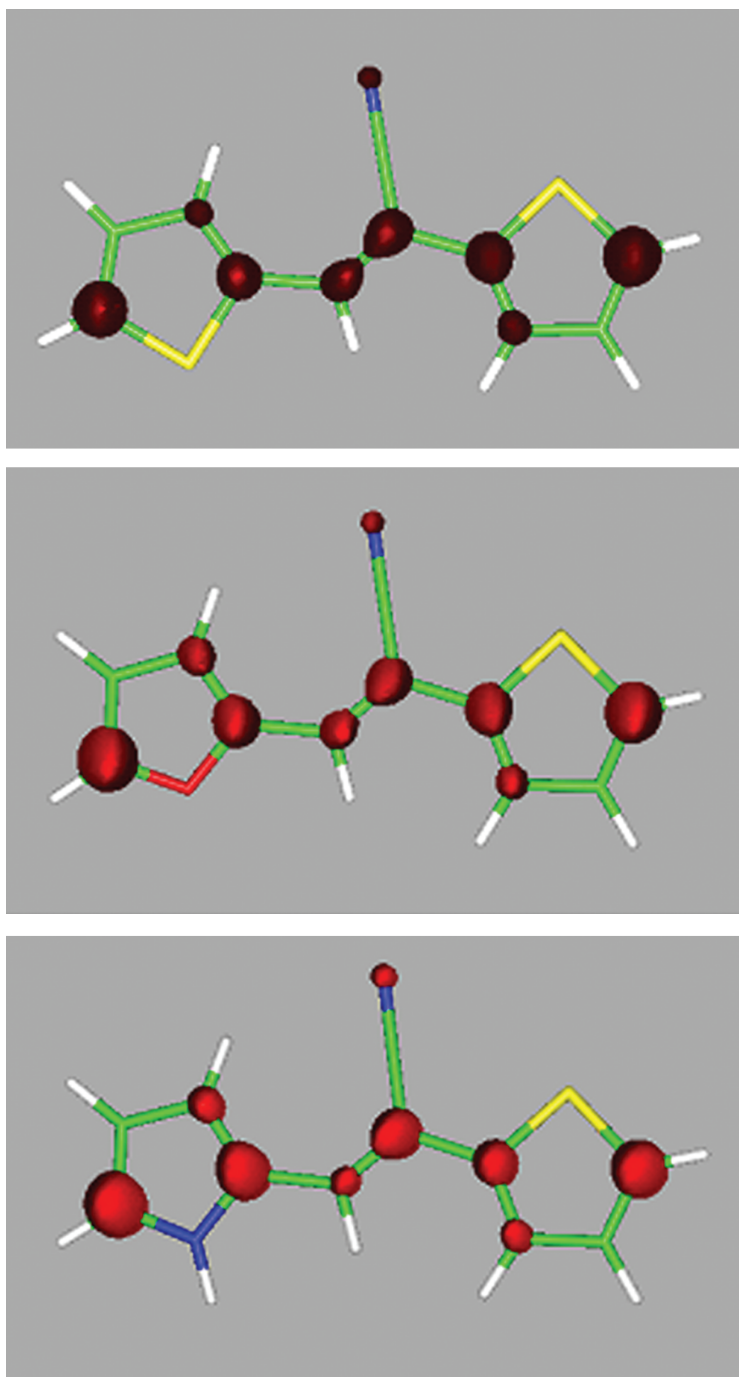


Figure 9.16 Spin density contours for the three monomer molecules.

is a comparable region of high spin density at the double bond in the middle of the molecule. Coupling at these sites would cause considerable branching and, as a result, polymer (a) should have low conductivity. By altering the substituents on molecule (a) and the heteroatoms on the rings we can direct the spin density away from the double bond in the monomer unit. Experimental investigations confirm the success of the approach – both molecules (b) and (c) conduct significantly better than polymer (a) when synthesised under equivalent conditions.

In this way modelling can be used in a predictive mode to screen a series of molecules. Researchers can observe the effects of substitutions and fine tune spin density distributions to favour coupling at desirable sites before any experimental work is carried out. Although this is not the only criterion against which high conductivity in these molecules needs to be screened, it should help to eliminate experiments that fall at the first hurdle because they produce non-conducting polymers by coupling at unfavourable sites.

9.5 Summary

For many years, modelling studies of drug delivery systems have lagged behind those concerned with the identification of potential new drug molecules. However, simulations of drug delivery systems at the atomic level are now widely performed on a variety of different systems. The improvements in computer speed means that these simulations now make increasing use of molecular dynamics and higher levels of theory in order to improve the reliability of the results obtained. The examples highlighted in this chapter show that not only can modelling studies help us to interpret experimental observations, they are also beginning to help drive experimental programmes of work by predicting modifications to systems such that they are optimised for their particular purpose. A number of different types of simulation method are available to probe these systems at the atomistic level and often a range of techniques are used together in order to obtain the necessary insight into the system under investigation. These trends are set to continue in the future, with modelling playing a key role in the design and optimisation of drug delivery systems and not just the design of novel drug compounds.

References

1. J. Siepmann and F. Siepmann, *Int. J. Pharm.*, 2008, **364**(2), 328–343.
2. N. J. Wheate, D. P. Buck, A. I. Day and J. G. Collins, *Dalton Trans.*, 2006, **3**, 451–458.
3. M. S. Bali, D. P. Buck, A. J. Coe, A. I. Day and J. G. Collins, *Dalton Trans.*, 2006, **45**, 5337.
4. S. Piovesan, P. A. Cox, J. R. Smith, D. G. Fatouros and M. Roldo, *Phys. Chem. Chem. Phys.*, 2010, **12**(48), 15636.
5. D. G. Fatouros, K. Power, O. Kadir, I. Dekany, S. N. Yanopoulos, N. Bouropoulos, A. Bakandristos, M. D. Antonijevic, G. D. Zouganelis and M. Roldo, *Nanoscale*, 2009, **3**(3), 1218.

6. N. Bouropoulos, O. L. Katsamenis, S. Norman, P. Kallinteri, M. E. Favretto, S. N. Yannopoulos, A. Bakandritsos and D. G. Fatouros, 2011, manuscript in preparation.
7. L. E. Smart and E. A. Moore, in *Solid State Chemistry: An Introduction*, 3rd edn, CRC Press, Boca Raton, USA.
8. D. G. Fatouros, D. Douroumis, V. Nikolakis, S. Ntais, A. M. Moschovi, V. Trivedi, B. Khima, M. Roldo, H. Nazar and P. A. Cox, *J. Mater. Chem.*, 2011, **21**(21), 7789.
9. T. A. Hilder and J. M. Hill, *Small*, 2009, **5**(3), 300.
10. Y.-C. Liu and Q. Wang, *J. Chem. Phys.*, 2007, **126**, 124901.
11. Q. Chen, Q. Wang, Y. Liu, T. Wu, Y. Kany, J. D. Moore and K. E. Gubbins, *J. Chem. Phys.*, 2009, **131**, 15101.
12. Y. Cheng, Q. X. Pei and H. J. Gao, *Nanotechnology*, 2009, **20**(14), 145101.
13. U. Arsawang, O. Saengsawang, T. Rungrotmongkoi, P. Sornmee, K. Wittayanarakul, T. Remsungne and S. Hannongbua, *J. Mol. Graph. Model.*, 2011, **29**(5), 591.
14. A. Striolo, *Nanotechnology*, 2007, **18**, 475704.
15. W. H. Yang, J. J. Fang, P. Liu, J. F. Li and K. Y. Cai, *Mol. Phys.*, 2011, **109**(6), 969.
16. T. X. Xiang and B. D. Anderson, *Adv. Drug Deliver. Rev.*, 2006, **58**(12–13), 1357.
17. B. Jin and J. A. Hopfinger, *Pharm. Res.*, 1996, **13**, 1786.
18. A. Pohorille and M. A. Wilson, *J. Chem. Phys.*, 1996, **204**, 337.
19. T. X. Xiang, Z. Q. Jiang, L. Song and D. Anderson, *Mol. Pharm.*, 2006, **3**(5), 589.
20. M. Subashini, P. V. Devarajan, G. S. Sonavane and M. Doble, *J. Mol. Model.*, 2011, **17**(5), 1141.
21. A. D. Costache, L. Sheihet, K. Zaveri, D. D. Knight and J. Kohn, *Mol. Pharm.*, 2009, **6**(5), 1620.
22. S. P. Adiga and D. W. Brenner, *Nano Letters*, 2002, **2**(6), 567.
23. D. Hofmann, L. Fritz, J. Ulbrich, C. Schepers and M. Bohning, *Macromol. Theor Simul.*, 2000, **9**(6), 293.
24. I. Tanis and K. Karatasos, *J. Phys. Chem. B*, 2009, **113**(31), 10984.
25. I. Tanis, K. Karatasos, A. N. Assimopoulou and V. P. Papageorgiou, *Phys. Chem. Chem. Phys.*, 2011, **13**(22), 10808.
26. L. C. S. Filipe, M. Machuqueiro and A. M. Baptista, *J. Am. Chem. Soc.*, 2011, **133**(13), 5042.
27. A. L. Doadrio, J. C. Doadrio, J. M. Sanchez-Montero, A. J. Salinas and M. Vallet-Regi, *Microporous Mesoporous Mater.*, 2010, **132**(3), 559–566.
28. G. P. Moss, D. R. Gullick, P. A. Cox, C. Alexander, M. J. Ingram, J. D. Smart and W. J. Pugh, *J. Pharm. Pharmacol.*, 2006, **58**(2), 167.
29. D. Svirskis, J. Travas Sejdic, A. Roberts and S. Garg, *J. Control. Release*, 2010, **146**(1), 6.
30. J. R. Smith, P. A. Cox, N. M. Ratcliffe and S. A. Campbell, *Trans. Instit. Metal Finishing*, 2002, **80**, 52–55.
31. J. R. Smith, P. A. Cox, S. A. Campbell and N. M. Ratcliffe, *J. Chem. Soc. Faraday Trans.*, 1995, **91**(15), 2331–2338.

Subject Index

References to figures are given in *italic type*. References to tables are given in **bold type**.

- ab initio* methods, 3, 101, 108, 112
 - bond length calculations, 131, 142
- actin, 9–16
- additive QM/MM methods, 5–6
 - extended solvation, 6–7
- B3LYP, 32–3, 51, 110, 132, 137
- Babel, 179
- BCP space, 124, 127–30
- benzenes, 121, 129
- benzimidazoles, 138–9
- Bessell moments, 149
- bifurcation theory, 125
- binding energies, 14–15
- binning, 100
- biodegradability, 135
- bioisosterism, 154–6
- biomolecule residues, 16
- bit strings, 173–4
- bond critical point (BCP) space,
 - 124, 127–30
- bond dissociation enthalpy, 138
- bond lengths, 131, 142
- bovine pancreatic trypsin inhibitor (BPTI), 59
- cancer chemotherapy, 27, 211–13
- Carbó molecular similarity
 - index, 129
- carbon nanotubes (CNT), 213–15
 - molecular dynamics, 225
 - see also* fullerenes
- carboxylate, 154
- carboxylic acids, 133, 141
- Cartesian grids
 - local surface properties, 100–2
 - protein docking, 61–4
- CAS registry numbers, 165
- Case, David, 144
- CASSCF, 114
- catastrophe theory, 125
- CCSD(T), 121
- characterization techniques, 56–7
- CHelp procedure, 146
- chemical nomenclature,
 - 169, 170–3
- Chemicals Abstracts Service (CAS),
 - 165, 169
- Chemistry at Harvard Molecular Mechanics (CHARMM), 11
- CHEMSHELL software, 11
- Chinese medicines, 135–6
- chitosans, 213
- chlorpromazine, 91
- cisplatin, 27, 28, 40–2, 211–13
- CLIFF, 179
- clonidine, 202
- clozapine, 96, 100, 102
- CML format, 178–9
- Common Unified Device Architecture (CUDA), 71, 71–2
- comparative molecular field analysis (CoMFA), 2, 90, 100
- computation time
 - ab initio* methods, 143
 - protein docking, 59–60

- quantum mechanical
 - calculations, 88–9
 - semi-empirical methods (SM), 115–16
- Computational Geometry Algorithms Library (CGAL), 187
- conducting polymers, 227–30
- connection tables, 175–6
- Connolly surface, 185
- contact surface, 185–6
- copper complexes, 41–2
 - DNA intercalation, 45–7, 46
 - hexacoordinate, 36–7
- core-rim interface model, 58
- Corey-Pauling-Koltun (CPK) model, 185–7
- correlation coefficient (PLS regression), 132
- Coulomb interaction, 91, 149
- coupled-cluster with single,
 - double and perturbative triple (CCSD(T)), 121
- crambin, 150
- critical points, 122–3, 198, 198–9
- cryo-electron microscopy, 57
- cucurbiturils, 212–13
- CUDA SIMT model, 71–2
- cytochrome P450, 28, 32
 - alkene oxidation, 50–3
- cytosine, 139
- d electron orbital, 36–7
 - transition metal complexes, 35–9
- databases, 58–9
- dendrimers, 225–6
- density functional theory (DFT), 28–9, 30–3, 120–1
 - Kohn-Sham formulation, 30–1
 - local density approximation (LDA), 31
 - spin state energies, 32–3
 - time-dependent, 113–14
- density functional tight binding (DFTB), 111
- cis*-diaminodichloroplatinum(II), 40–1, 47–8, 47
- dispersion interactions, 13, 15–16, 112
 - density functional theory (DFT), 32
 - transition metal complexes, 32
- DNA, 139
 - copper complex
 - intercalation, 45–7
 - platinum complex binding, 47–8
- donor-acceptor interactions, 91–2
- DOT, 63
- Drude oscillators, 145
- drug delivery systems, 210–11
 - carbon nanotubes, 213–15
 - fullerenes, 215–19
 - liposomes, 216, 225
 - molecular mechanics, 211–13
 - quantum mechanics, 226–30
 - zeolites, 221–6
- electron affinity, 92, 193
- electron density, 122, 124, 144
 - as basis for similarity
 - measures, 126, 127, 130
 - expressed as multipole
 - moments, 144–6, 148
 - isocontours, 188–92
 - mean curvature, 192–3
 - Laplacian, 128
- electron donor-acceptor interactions, 91–2
- electronegativity, 92
- electrostatic embedding
 - approach, 5–6
 - extended solvation, 6–7
- electrostatic interactions, 5–6, 144
 - long-range modelling, 6–7
 - protein docking, 62–3
- electrostatic potential, molecular *see* molecular electrostatic potential
- empirical valence bond (EVB) model, 52
- enzyme-catalysed reactions, 15–16
- excitation energy transfer (EET), 114
- excited states, 113–14
- extended solvation, 6–7
- extracted solvation methods, 6–7

- fast Fourier transform (FFT), 63
 calculation on GPUs, 73–4
 see also protein docking, fast
 Fourier transform-based
Federal Drug Administration
 (FDA), 40–2
file formats, 176–8
fingerprints, 174
Fischer, Emil, 2
FlexX, 17–18, 50, 51
Fock matrix, 115
force fields, 3, 143–56
 CHARMM, 11
 FlexX, 17–18
 MM3*, 42–3
 molecular mechanics, 33
 multipole moments, 144, 148
 partitioning, 146
 polarisability and, 144–6
 Q2MM, 52–3
 QM-derived point charges, 90
 quantum chemical
 topology, 143–4
 transferability, 147–8
 validation, 143–4
Fourier transform *see* fast Fourier
 transform
frontier molecular orbital (FMO)
 theory, 91–2
FTDOCK, 63
fullerenes, 215–19
 carbon nanotubes, 213–15

G-actin, 9–16
Gauss-Laguerre function, 65
Gaussian-type orbitals (GTO), 65–6
generalised gradient approximation
 (GGA), 31
generalized solvent boundary
 potential (GSBP), 7
genetic algorithms, 137
geometric hashing, 61
Glide, 17
GMP math library, 74–5
gnomonic projection, 195
gold, 29

gradient vector fields, 122, 199
 Morse-Smale, 200–3
graphics processor units (GPU), 70–5
GRID, 100
ground state energy, 30
guanine, 139

haloperidol, 127–8
Hammett acidity constants, 138
Handbook of Chemoinformatics, 165
hardness, 93
hash coding, 166
hen egg lysozyme, 66
Hex algorithm, 61, 74, 75–9
 docking results, 77–9
Hierarchical Equal Area isoLatitude
 Pixelisation (HEALPix), 191
highest occupied molecular orbital
 (HOMO), 91–2
hydrogen bonding, 100–2
 latrunculin/actin binding
 study, 11–12
 semi-empirical methods
 (SM), 112–13
hydrophobicity, 12–13, 63, 216
hydroxamate inhibitors, 17
hydroxide anion, 101
(*E*)-1-(4-hydroxyphenyl)but-
 1-en-3-one, 135–6
HyHel-5 antibody, 66
hyperbolic critical points, 198

identification numbers, 166
imidazoles, 123–4
INDO/S, 114
induced fit model, 2
influenza inhibition, 138
integrated molecular orbital/
 molecular mechanics
 (IMOMM), 7–8
interacting quantum atoms (IQA)
 approach, 125
interaction potential, 143
intermolecular interactions, 91–3
International Chemical Identifier
 (InChI), 172–3

- International Union of
 Crystallography (IUC), 178
isothermal titration calorimetry
 (ITC), 154
- Jacobi polynomials, 68
Jahn-Teller effect, 29, 37, 38
JUMMUZ, 43, 44
- Kepler, Johannes, 122–3
Kohn-Sham (KS) formulation, 30–1
KP1019, 28
Kriging, 151–3
- lansoprazole, 186
latrunculins, 9–16
 actin binding study
 docking, 10–11
 force field minimization, 11
 hydrogen bonding, 11–12
 methodology, 10
 relative binding energies, 14–15
 structure preparation, 10
- Legendre polynomials, 191
leucine, 57
LFSE, 35–6
ligand field stabilisation energy
 (LFSE), 34
 angular overlap model, 37–8
 d orbital configuration and, 36–7
ligand field theory (LFT), 37
ligand-receptor interactions,
 latrunculin analogues-actin, 9–16
linear scaling techniques, 88–9,
 114–15
lipid bilayer, 216–18, 217
lipophilicity, 194
 copper complexes, 42–4
 platinum complexes, 41–2
liposomes, 216, 225
local density approximation (LDA), 31
local properties, 92–3, 193–4
 electron affinity, 92, 193
 on grids, 100–2
 hardness, 93
 ionisation energy, 92, 193
 surface autocorrelations, 98
 surface projection, 93–7
- LocalSCF, 115
lock-and-key model, 2
log *P* *see* octanol/water partition
 coefficient
lorazepam, 186, 188, 189
lowest unoccupied molecular orbital
 (LUMO), 92
- machine learning, 150, 150–2
MaP descriptor, 195–6
matrices (structure
 representation), 175
matrix calculations, 71, 72, 73
matrix metalloproteinase-9
 (MM9), 17
matrix metalloproteinases, 28
MDL Drug Data Report
 (MDDR), 155
MDL MACCS key, 173
MDL Molfile, 176–7
mesoporous materials, 226–7
metal complexes *see* transition metal
 complexes
metalloproteins, 16
 see also cytochrome P450
metclopamide, 49–50
methionine, 57
2 α -methyl-9 α -fluorocortisol, 137
MM3* force field, 42–3
MNDO, 108–9
MNDO/d, 109–11
MNDO/H, 11, 14–15
MNDO99, 11
Modelling of Protein Interactions in
 Genomes, 59
modified neglect of differential
 overlap *see* MNDO
molecular descriptors, 167–9
 see also structural representations
Molecular Descriptors for
 Chemoinformatics, 165
molecular dynamics, 39, 60, 112,
 213, 216
 carbon nanotubes, 225

- molecular dynamics (*continued*)
 conducting polymers, 227–30
 zeolites, 220–5
- molecular electrostatic potential (MEP),
 42, 63, 89–90, 125, 193, 197, 198
 surface projection, 96–7
 see also electrostatic interactions
- molecular lipophilicity potential
 (MLP), 194
- molecular maps, 139
- molecular mechanics (MM)
 drug delivery systems, 211–13
 carbon nanotube, 212–15
 fullerenes in liposomes, 215–20
 force fields, 33
 ligand field effects, 35–9
 transition metal complexes, 33–5
 ligand effects, 35–9
 zinc, 50
 unique labelling problem, 33–5
- molecular skin surface (MSS), 187
- molecular surface models *see* surface
 models
- molecular surface property graph
 (MSPG), 199–202
- MolFit, 63
- MOLPRINT 3D, 196
- Morse function, 198
- Morse-Smale complex, 199–203
- MOZYME, 115
- MP2, 32
- MS-CASPT2, 114
- MSINDO, 111–12
- multipole moments, 144, 148
 Bessell moments, 149–50
- MYOVIEW, 28
- NAMI-A, 28
- NCI database, 88
- neural networks, 151–2
- nitroaromatics, 141
- NPSC (*N*-palmitoyl-*O*-sulfate
 chitosan), 213
- nuclear magnetic resonance (NMR)
 spectroscopy, 57
- Nvidia GPUs, 71–2
- octanol/water partition
 coefficient, 41, 194
- octyltrimethoxysilane, 226–7
- OMx, 111
- own *n*-layered integrated molecular
 orbital and molecular mechanics
 (ONIOM), 8
- N*-palmitoyl-*O*-sulfate chitosan
 (NPSC), 213
- ParaMatch, 197
- partial equalization of orbital
 electronegatives (PEOE), 193
- partial least squares (PLS)
 regression, 131–2, 142
- PDB file format, 177–8
- perturbation theory, 145
- pharmacokinetics, 41, 154–6
- pharmacophore fingerprints, 174
- phenanthrolenes, 45–7, 46
- phenol, 127
- phenols, 135, 138, 140
- p*-phenylacetic acids, 133
- π -stacking, 15, 112
- pK_a values, 137, 141–2
- platinum complexes, 27, 29,
 40–2
 DNA binding, 47–8
 encapsulation, 211–12
- PM3, 108, 114
- Poincaré-Hopf index formula, 201
- polar surface area, 41
- polarisability, 92, 93, 136
- polarisation effects, 144–5
 quantum chemical topology force
 field models, 150–2
- polarised continuum models, 91
- polychlorinate dibenzo-*p*-dioxins
 (PCDD), 136
- polymers, 227–30
- polypeptides, 111
- POSIX, 70
- principal component analysis
 (PCA), 131
- protein crystallography, 39–40
- Protein Data Bank, 57

- protein docking, 59–60
 - by geometric hashing, 61
 - Cartesian grid-based, 61–4
 - electrostatic interactions, 62–3
 - fast Fourier transform-based
 - Cartesian grid, 61–4
 - spherical polar basis, 64–70
 - graphics processor units, 70–5
 - Hex* algorithm, 75–9
 - multidimensional spherical polar FFTs, 70
 - rigid body docking, 60–1
 - spherical polar Fourier (SPF) functions, 64–7
- protein-protein interactions (PPI), 56–7
 - complexation, 60
 - databases, 58–9
 - interaction surfaces, 57–8
 - see also* protein docking
- pyridines, 142
- pyrrole, 228
- Q2MM force field, 52–3
- QCTFF force field, 151–2
- QTAIM approach, 125
- quantitative structure-activity relationship (QSAR), 2, 39–40
 - copper complexes, 43–4
 - quantum topological molecular similarity and, 135–6
 - validation sets, 139–40
- quantum chemical topology (QCT), 121–5
 - bioisosterism, 154–6
 - critical points, 122–3
 - force fields, 144–54
 - advantages, 153–4
 - multipole moments, 148–50
 - state of art, 144–6
 - transferability, 147–8
 - gradient vector fields, 122
 - partitioning method, 146
 - see also* quantum topological chemical similarity
- Quantum Isostere Database (QID), 155
- quantum mechanics (QM), 87–8
 - computational demands, 88–9
 - drug delivery systems, 226–30
 - conducting polymers, 227–30
 - intermolecular interactions, 91–3
 - linear scaling models, 88–9
 - local properties, 91–3
 - on grids, 100–2
 - surface projection, 93–7, 93–8
 - molecular electrostatic potential (MEP), 89–90
 - surface autocorrelations, 98
 - surface integral models, 99–100
 - vancomycin delivery using mesoporous materials, 226–7
 - see also* quantum chemical topology
- quantum mechanical/molecular mechanical (QM/MM) methods, 1–2, 112–13
 - basic concepts, 4–5
 - additive, 5–6
 - extended solvation, 6–7
 - cytochrome P450 alkene oxidation, 51–3
 - latrunculin-actin binding, 9–16
 - methodological overview, 3–4
 - structure-based drug design (SBDD), 16–17
 - subtractive, 7–8
 - see also* semi-empirical methods (SM)
- quantum molecular similarity (QMS), 126–42
- quantum topological molecular similarity (QTMS), 126–42
 - acid dissociation constant prediction, 137, 141–2
 - BCP space, 127–30
 - bond length, 130–1
 - mutagenicity studies, 137–8
 - phenol toxicity, 140
 - ranking and regression, 131–5, 142
 - use in determining QSARs, 136–7
 - VIP values, 132–5

- Rapid Overlay of Chemical Structures (ROCS), 188
RDFiles, 177
re-entrant surface, 186
Research Collaboratory for Structural Bioinformatics (RCSB), 177
RGFile, 177
ring critical point (RCP), 124
rotationally invariant fingerprints (RIF), 97, 97–8
- saddle connections, 199
salbutamol, 221–4, 222, 223, 224
SBA-15, 226–7
scaffold hopping, 195
scoring refinement, QM/MM, 16–18
SDFFile, 176–7
search keys, 173
semi-empirical methods (SM), 3–4, 107–8, 197–8
 computation time, 115–16
 DFTB, 111
 electron and excitation energy transfer, 114
 excited states, 113–14
 hydrogen bonds, 112–13
 linear scaling methods, 114–15
 MNDO, 108–9
 MNDO/d, 109–11
 MNDO/H, 11, 14–15
 MSINDO, 111–12
 OM x , 111
 parametrization, 108–9
 see also quantum mechanics/
 molecular mechanics (QM/MM)
- separatrix, 123
serine, 57
similarity measures
 molecular surface, 194–8
 see also quantum topological
 molecular similarity
simultaneous instructions multiple threads (SIMT), 71
single-walled nanotubes (SWNT), 213
SMILES coding, 171
 solvent accessible surface (SAS), 185
 solvent effects, 112
 see also polarisation effects
 sparse critical point (SCP), 196
 spherical harmonic expansion, 93–8
 electron density isocontour representations, 189–91
 spherical polar Fourier (SPF) functions, 64–7
 multi-dimensional, 70
 spin state, 29–30
 spin state energies, transition metal complexes, 32–3
 STAR file format, 178–9
 steroids, 137
 structural keys, 173
 structure exchange formats, 176–8, 178–9
 format conversion, 179
 SMILES, 171
 structure preparation, 10
 structure representations
 zero-dimensional, 165–7
 bit strings, 173–4
 CAS registry numbers, 165
 hash coding, 166
 input, 164–5
 International Chemical Identifier (InChi), 172–3
 molecular descriptors, 167–9
 one-dimensional, 169–73
 SMILES, 171
 Sybyl line notation, 171–2
 two-dimensional, 174–9
 see also structure exchange formats
 structure-based drug design (SBDD), 2–3
 application of QM/MM methods, 8–9
 rescoring, 16–17
 subtractive QM/MM methods, 7–8
 surface autocorrelations, 98
 surface models
 Corey-Pauling-Koltun (CPK), 185–7

- electron density
 - isocontours, 187–91
 - graphs, 198–203
 - mean curvature, 192
 - molecular skin surface (MSS), 187
 - similarity measures, 194–8
 - surface integral, 99–100
 - see also* local properties
- surface patch alignment (SPAt), 196–7
- surface topography index, 193
- Sybyl line notation, 171–2
- T. pyriformis*, 142
- tandem-affinity-purification mass spectrometry, 57
- technetium, 27
- tetrazole, 154–5
- theophylline, 221–4, 222, 223, 224
- Thom, Ren, 125
- TIP3P, 148–9
- total surface area, platinum complexes, 41–2
- toxicity, 140, 142–3
- trajectory, 199
- transition metal complexes,
 - platinum, 27, 29, 40–2, 47–8, 211–12
- transition metal complexes
 - applications, 27–8
 - bond lability, 30
 - coordination geometries, 34–5
 - copper, 42–3
 - density functional theory (DFT), 30–3
 - DNA intercalation, 45–7
 - encapsulation, 211–13
 - explicitly considered
 - receptors, 44–53
 - ligand effects, 35–9
 - modelling methods, 29
 - molecular mechanics, 33–5
 - protein crystallography, 39–40
 - spin state, 29–30
 - zinc, 50–3
 - transition states, 50
 - trimethoprim, 98, 99
 - tryptophan, 57
 - tyrosine, **168**
- unique labelling problem, 33–5
- VALVUZ, 44
- van der Waals interactions, 5, 123–4, 149–50
- van der Waals surface (VDWS), 185
- vancomycin, 226–7
- variable importance in projection (VIP) plots, 132–5
- virial field, 125
- virtual high-throughput screening, 29
- voxels, 64
- water, 101, 101, 150–1
- Wiener index, 168
- winding number, 201
- yeast two-hybrid mass spectroscopy, 57
- Z-matrix, 178–9
- ZDOCK, 63
- ZEBJIZ, 43, 44
- zeolites, 220–5
- zinc complexes, 50–3
- zwitterions, 88

