



Chemical Biology

# Computational Tools for Chemical Biology

Edited by Sonsoles Martín-Santamaría



## Computational Tools for Chemical Biology

## **Chemical Biology**

### *Editor-in-Chief:*

Tom Brown, *University of Oxford, UK*

### *Series Editors:*

Kira J. Weissman, *Lorraine University, France*

Sabine Flitsch, *University of Manchester, UK*

Nick J. Westwood, *University of St Andrews, UK*

### *Titles in the Series:*

1: High Throughput Screening Methods: Evolution and Refinement

2: Chemical Biology of Glycoproteins

3: Computational Tools for Chemical Biology

### *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: [booksales@rsc.org](mailto:booksales@rsc.org)

Visit our website at [www.rsc.org/books](http://www.rsc.org/books)

# *Computational Tools for Chemical Biology*

Edited by

**Sonsoles Martín-Santamaría**

*Center for Biological Research, CIB-CSIC, Madrid, Spain*

*Email: [smsantamaria@cib.csic.es](mailto:smsantamaria@cib.csic.es)*





Chemical Biology No. 3

Print ISBN: 978-1-78262-700-5

PDF eISBN: 978-1-78801-013-9

EPUB eISBN: 978-1-78801-256-0

ISSN: 2055-1975

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

© The Royal Society of Chemistry 2018

*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.*

*Whilst this material has been produced with all due care, The Royal Society of Chemistry cannot be held responsible or liable for its accuracy and completeness, nor for any consequences arising from any errors or the use of the information contained in this publication. The publication of advertisements does not constitute any endorsement by The Royal Society of Chemistry or Authors of any products advertised. The views and opinions advanced by contributors do not necessarily reflect those of The Royal Society of Chemistry which shall not be liable for any resulting loss or damage arising as a result of reliance upon this material.*

The Royal Society of Chemistry is a charity, registered in England and Wales, Number 207890, and a company incorporated in England by Royal Charter (Registered No. RC000524), registered office: Burlington House, Piccadilly, London W1J 0BA, UK, Telephone: +44 (0) 207 4378 6556.

For further information see our web site at [www.rsc.org](http://www.rsc.org)

Printed in the United Kingdom by CPI Group (UK) Ltd, Croydon, CR0 4YY, UK

# *Preface*

Advances in chemical biology are providing new avenues in science. In particular, within this field, computational chemistry and molecular modelling are tools able to furnish information not (easily) accessible through other experimental methods, and to explain and predict phenomena involving biomolecules. Modern science requires multidisciplinary approaches to address complex problems at the frontier of different disciplines and, in fact, experimental and computational techniques already walk together.

This book aims to offer a broad and complete perspective from chemical biology about the available variety of computational tools, and what they can offer from small molecules to macromolecular systems: docking protocols, molecular dynamics simulations, coarse-grained simulations, quantum mechanical methods and hybrid QM/MM simulations. Chapters discuss and review cutting-edge methods, limitations and future perspectives regarding the development of new methods, and also how to tackle challenges in a wide range of research topics: molecular recognition (nucleic acids, ligand–protein, carbohydrate–protein, and protein–protein interactions), protein design, modelling of nucleic acids, GPCRs, membranes, nanostructures and signalling networks, drug design, prediction of polypharmacology, and lead-discovery, amongst others. Additionally, established and young international leaders in this field coexist together as authors of the different chapters.

It is our hope that the topics covered herein will provide broad information on the possibilities that computational chemistry and molecular modelling can offer to boost research in chemical biology, and also to show the future challenges ahead.

Sonsoles Martín-Santamaría



# Contents

|                  |                                                                                 |          |
|------------------|---------------------------------------------------------------------------------|----------|
| <b>Chapter 1</b> | <b>Computational Chemistry and Molecular Modelling Basics</b>                   | <b>1</b> |
|                  | <i>Samuel Genheden, Anna Reymer, Patricia Saenz-Méndez and Leif A. Eriksson</i> |          |
| 1.1              | Introduction                                                                    | 1        |
| 1.2              | Techniques in Biomolecular Simulations                                          | 2        |
| 1.2.1            | Molecular Mechanics and Force Fields                                            | 2        |
| 1.2.2            | Basic Simulation Techniques                                                     | 4        |
| 1.2.3            | Basic Data Analysis                                                             | 8        |
| 1.2.4            | Software                                                                        | 11       |
| 1.2.5            | Examples                                                                        | 12       |
| 1.3              | Protein Structure Prediction                                                    | 13       |
| 1.3.1            | Sequence Alignment and Secondary Structure Prediction                           | 13       |
| 1.3.2            | Comparative Modelling Approaches                                                | 15       |
| 1.3.3            | Function Prediction                                                             | 18       |
| 1.3.4            | Analysing the Quality of the Modelled Structure                                 | 18       |
| 1.3.5            | Software and Web Based Servers                                                  | 19       |
| 1.4              | Computer-based Drug Design                                                      | 20       |
| 1.4.1            | Pre-requisites for SBDD—Sampling Algorithms and Scoring Functions               | 20       |
| 1.4.2            | Structure Based Drug Design (SBDD)                                              | 24       |
| 1.4.3            | Ligand Based Drug Design (LBDD)                                                 | 26       |
| 1.4.4            | Pharmacophores                                                                  | 26       |

|                  |                                                                                                    |           |
|------------------|----------------------------------------------------------------------------------------------------|-----------|
| 1.4.5            | Compound Optimisation                                                                              | 27        |
| 1.4.6            | Software and Web Based Servers                                                                     | 29        |
|                  | Acknowledgements                                                                                   | 30        |
|                  | References                                                                                         | 30        |
| <b>Chapter 2</b> | <b>Molecular Dynamics Computer Simulations of Biological Systems</b>                               | <b>39</b> |
|                  | <i>James W. Carter, Anna Sofia Tascini, John M. Seddon and Fernando Bresme</i>                     |           |
| 2.1              | Introduction                                                                                       | 39        |
| 2.2              | The Basics of Molecular Dynamics                                                                   | 40        |
| 2.2.1            | Force Fields for Biomolecular Simulations                                                          | 41        |
| 2.2.2            | Multiscale Modelling                                                                               | 44        |
| 2.2.3            | Advanced Force Fields                                                                              | 45        |
| 2.3              | Extracting the Information from MD                                                                 | 46        |
| 2.3.1            | Free Energy Difference Between Two States                                                          | 47        |
| 2.3.2            | Enhanced Configurational Sampling                                                                  | 47        |
| 2.3.3            | Simulating Rare Events                                                                             | 49        |
| 2.3.4            | Computing Elastic Properties in Biomolecular Simulations                                           | 50        |
| 2.4              | MD Simulation vs. Experiment                                                                       | 54        |
| 2.4.1            | NMR and MD: Structure and Dynamics                                                                 | 55        |
| 2.4.2            | Structure of Biomolecules and Diffraction: Solving the Phase Problem with MD                       | 57        |
| 2.5              | Future Directions                                                                                  | 58        |
| 2.6              | Conclusion                                                                                         | 61        |
|                  | Acknowledgements                                                                                   | 63        |
|                  | References                                                                                         | 63        |
| <b>Chapter 3</b> | <b>Designing Chemical Tools with Computational Chemistry</b>                                       | <b>69</b> |
|                  | <i>Silvia Rinaldi and Giorgio Colombo</i>                                                          |           |
| 3.1              | Introduction                                                                                       | 69        |
| 3.2              | Structure Based Approaches for Chemical Biology                                                    | 72        |
| 3.3              | Structural Dynamics as a Source of Novel Chemical Tools                                            | 74        |
| 3.4              | Combining Bioinformatics, Chemoinformatics and Structural Information to Explore Protein Functions | 79        |
| 3.5              | Deep Networks and Big Data in the Discovery of New Drugs and Chemical Tools                        | 81        |
| 3.6              | Conclusions and Perspectives                                                                       | 83        |
|                  | References                                                                                         | 84        |

|                  |                                                                                                      |            |
|------------------|------------------------------------------------------------------------------------------------------|------------|
| <b>Chapter 4</b> | <b>Computational Design of Protein Function</b>                                                      | <b>87</b>  |
|                  | <i>Marc Garcia-Borràs, Kendall N. Houk and Gonzalo Jiménez-Osés</i>                                  |            |
| 4.1              | Introduction                                                                                         | 87         |
| 4.2              | The ‘Inside-out’ Design Protocol                                                                     | 89         |
| 4.2.1            | Description of the Method                                                                            | 89         |
| 4.2.2            | Enzymes Designed Though the ‘ <i>Inside-out</i> ’ Approach: Kemp Eliminases                          | 92         |
| 4.3              | QM/MM Approaches to Enzyme Design                                                                    | 94         |
| 4.3.1            | Description of the Methods                                                                           | 94         |
| 4.3.2            | Engineered Butyrylcholinesterase for Cocaine Detoxification                                          | 96         |
| 4.3.3            | Electron Transfer Reactions Catalysed by Metalloproteins                                             | 99         |
| 4.4              | Summary and Outlook                                                                                  | 101        |
|                  | Acknowledgements                                                                                     | 102        |
|                  | References                                                                                           | 102        |
| <b>Chapter 5</b> | <b>Computational Enzymology: Modelling Biological Catalysts</b>                                      | <b>108</b> |
|                  | <i>Laura Masgrau, Àngels González-Lafont and José M. Lluch</i>                                       |            |
| 5.1              | Introduction                                                                                         | 108        |
| 5.2              | General Framework                                                                                    | 109        |
| 5.2.1            | The Transition State and the Energy Barrier                                                          | 109        |
| 5.2.2            | Quantum Mechanics Molecular Mechanics (QM/MM) Methods                                                | 110        |
| 5.3              | Building the Model(s) of the Enzyme–Substrate Complex(es)                                            | 114        |
| 5.3.1            | Starting Structure and System Setup                                                                  | 114        |
| 5.3.2            | Molecular Dynamics Simulations                                                                       | 114        |
| 5.4              | Potential Energy Methods                                                                             | 115        |
| 5.4.1            | Reaction Path Calculation                                                                            | 115        |
| 5.4.2            | Transition State Localisation                                                                        | 117        |
| 5.4.3            | Analysis                                                                                             | 118        |
| 5.5              | Free Energy Simulations                                                                              | 122        |
| 5.5.1            | Umbrella Sampling Method                                                                             | 123        |
| 5.5.2            | Free Energy Perturbation Theory                                                                      | 127        |
| 5.5.3            | String Method: Minimum Free Energy Paths                                                             | 132        |
| 5.6              | Calculation of the Reaction Rate Constant                                                            | 136        |
| 5.6.1            | Ensemble-averaged Variational Transition State Theory with Multi-dimensional Tunnelling (EA-VTST/MT) | 136        |

|                  |                                                                                                                                                 |            |
|------------------|-------------------------------------------------------------------------------------------------------------------------------------------------|------------|
| 5.7              | Further Considerations about the Relationship between the Activation Free Energy and the Extension of the Sampling of the Configurational Space | 139        |
|                  | References                                                                                                                                      | 141        |
| <b>Chapter 6</b> | <b>Computational Chemistry Tools in Glycobiology: Modelling of Carbohydrate–Protein Interactions</b>                                            | <b>145</b> |
|                  | <i>Alessandra Lacetera, M. Álvaro Berbís, Alessandra Nurisso, Jesús Jiménez-Barbero and Sonsoles Martín-Santamaría</i>                          |            |
| 6.1              | What are the Carbohydrates?                                                                                                                     | 145        |
| 6.2              | From Mono to Polysaccharides: An Overview of the Increasing Complexity                                                                          | 147        |
| 6.2.1            | Monosaccharides                                                                                                                                 | 147        |
| 6.2.2            | Disaccharides: The Glycosidic Linkage and the <i>Exo</i> -anomeric Effect                                                                       | 148        |
| 6.2.3            | Studying the Conformations Around the Glycosidic Linkage                                                                                        | 149        |
| 6.2.4            | Oligosaccharides                                                                                                                                | 149        |
| 6.2.5            | <i>N</i> -glycans                                                                                                                               | 150        |
| 6.2.6            | Polysaccharides                                                                                                                                 | 150        |
| 6.3              | Computational Methodologies for the Study of Carbohydrates                                                                                      | 151        |
| 6.4              | Force Fields for Carbohydrates                                                                                                                  | 153        |
| 6.5              | Modelling Carbohydrate–Protein Interactions                                                                                                     | 155        |
| 6.6              | Conclusions                                                                                                                                     | 159        |
|                  | Acknowledgements                                                                                                                                | 159        |
|                  | References                                                                                                                                      | 159        |
| <b>Chapter 7</b> | <b>Molecular Modelling of Nucleic Acids</b>                                                                                                     | <b>165</b> |
|                  | <i>Hansel Gómez, Jürgen Walther, Leonardo Darré, Ivan Ivani, Pablo D. Dans and Modesto Orozco</i>                                               |            |
| 7.1              | Introduction                                                                                                                                    | 165        |
| 7.2              | QM Methods                                                                                                                                      | 166        |
| 7.2.1            | Basic Methodological Description                                                                                                                | 166        |
| 7.2.2            | Examples of Use                                                                                                                                 | 167        |
| 7.3              | Hybrid QM/MM                                                                                                                                    | 167        |
| 7.3.1            | Basic Methodological Description                                                                                                                | 167        |
| 7.3.2            | Examples of Use                                                                                                                                 | 168        |

|                  |                                                                                                                                       |            |
|------------------|---------------------------------------------------------------------------------------------------------------------------------------|------------|
| 7.4              | Atomistic Force-field Simulations                                                                                                     | 170        |
| 7.4.1            | Basic Methodological Description                                                                                                      | 170        |
| 7.4.2            | Force-field Refinements                                                                                                               | 172        |
| 7.4.3            | Recent Examples of Force-field Studies of Nucleic Acids                                                                               | 175        |
| 7.5              | The Coarse-grain Approach                                                                                                             | 177        |
| 7.5.1            | Basic Methodological Description                                                                                                      | 178        |
| 7.5.2            | Coarse-grained Methods for Predicting RNA Structures                                                                                  | 182        |
| 7.6              | Mesoscopic Models                                                                                                                     | 184        |
| 7.6.1            | Basic Methodological Description                                                                                                      | 185        |
| 7.6.2            | Nucleosome Fibre Simulations                                                                                                          | 186        |
| 7.6.3            | Chromosome Simulations                                                                                                                | 187        |
| 7.7              | Conclusions                                                                                                                           | 188        |
|                  | Acknowledgements                                                                                                                      | 188        |
|                  | References                                                                                                                            | 189        |
| <b>Chapter 8</b> | <b>Uncovering GPCR and G Protein Function by Protein Structure Network Analysis</b>                                                   | <b>198</b> |
|                  | <i>Francesca Fanelli and Angelo Felling</i>                                                                                           |            |
| 8.1              | Introduction                                                                                                                          | 198        |
| 8.2              | Experimental                                                                                                                          | 201        |
| 8.2.1            | Materials                                                                                                                             | 201        |
| 8.2.2            | Methods                                                                                                                               | 201        |
| 8.3              | Results and Discussion                                                                                                                | 205        |
| 8.3.1            | Modelling Allosteric Communication in GPCRs                                                                                           | 205        |
| 8.3.2            | Modelling Allosteric Communication in G Proteins                                                                                      | 213        |
| 8.4              | Conclusions                                                                                                                           | 216        |
|                  | Acknowledgements                                                                                                                      | 216        |
|                  | References                                                                                                                            | 217        |
| <b>Chapter 9</b> | <b>Current Challenges in the Computational Modelling of Molecular Recognition Processes</b>                                           | <b>221</b> |
|                  | <i>Lucía Pérez-Regidor, Joan Guzmán-Caldentey, Carlos F. Rodríguez, Jean-Marc Billod, Juan Nogales and Sonsoles Martín-Santamaría</i> |            |
| 9.1              | Modelling the Dynamics of the Proteins                                                                                                | 221        |
| 9.2              | Three-dimensional Structure Prediction and Homology Modelling                                                                         | 224        |



|       |                                                        |     |
|-------|--------------------------------------------------------|-----|
| 9.3   | Modelling of Protein–Protein Interactions              | 225 |
| 9.4   | Prediction of Protein–Protein Interactions:<br>Docking | 226 |
| 9.5   | Computational Studies of Complex<br>Protein Systems    | 229 |
| 9.6   | Computational Modelling of Nanostructures              | 232 |
| 9.6.1 | Modelling of Gold Nanoparticles                        | 233 |
| 9.6.2 | Modelling of Nanowires                                 | 234 |
| 9.6.3 | Modelling of Nanotubes                                 | 235 |
| 9.6.4 | Modelling of Nanomachines                              | 236 |
| 9.7   | Models of Signalling Networks                          | 237 |
|       | Acknowledgements                                       | 240 |
|       | References                                             | 240 |

**Chapter 10 Novel Insights into Membrane Transport from  
Computational Methodologies** **247**  
*Victoria Oakes and Carmen Domene*

|      |                                            |     |
|------|--------------------------------------------|-----|
| 10.1 | Introduction                               | 247 |
| 10.2 | Computational Methods                      | 248 |
| 10.3 | Unassisted Diffusion Across Lipid Bilayers | 252 |
| 10.4 | Passive Transport by Ion Channels          | 255 |
| 10.5 | Facilitated Diffusion by Transporters      | 259 |
| 10.6 | Signalling <i>via</i> Receptors            | 264 |
| 10.7 | Conclusions                                | 268 |
|      | Acknowledgements                           | 268 |
|      | References                                 | 269 |

**Chapter 11 Application of Molecular Modelling to Speed-up the Lead  
Discovery Process** **281**  
*Iuni M. L. Trist, Maurizio Botta and Anna Lucia Fallacara*

|        |                                                                       |     |
|--------|-----------------------------------------------------------------------|-----|
| 11.1   | Introduction                                                          | 281 |
| 11.1.1 | The ‘Pharmaceutical Crisis’                                           | 281 |
| 11.1.2 | The Drug Discovery Process                                            | 282 |
| 11.1.3 | The Contribution of Molecular Modelling<br>to Improve Drug Discovery  | 284 |
| 11.1.4 | Quantum and Molecular Mechanics in<br>Drug Design                     | 285 |
| 11.1.5 | An Introduction to Structure- and<br>Ligand-based Molecular Modelling | 285 |

|                                                                                                                |                |
|----------------------------------------------------------------------------------------------------------------|----------------|
| <i>Contents</i>                                                                                                | xiii           |
| 11.2 Structure-based Molecular Modelling                                                                       | 286            |
| 11.2.1 Sources of 3D Structures                                                                                | 286            |
| 11.2.2 Docking                                                                                                 | 289            |
| 11.2.3 <i>De Novo</i> Drug Design                                                                              | 291            |
| 11.2.4 Introducing Dynamics                                                                                    | 293            |
| 11.3 Ligand-based Molecular Modelling                                                                          | 296            |
| 11.3.1 Similarity Searching: Same Shape,<br>Same Activity                                                      | 297            |
| 11.3.2 Pharmacophore Modelling                                                                                 | 299            |
| 11.3.3 QSAR                                                                                                    | 300            |
| 11.3.4 Use of In Silico Ligand-based Approaches:<br>A Practical Case Study on Antitubercular<br>Agents         | 304            |
| 11.4 Conclusions                                                                                               | 305            |
| Abbreviations                                                                                                  | 306            |
| Acknowledgements                                                                                               | 307            |
| References                                                                                                     | 307            |
| <br><b>Chapter 12 Molecular Modelling and Simulations Applied to<br/>Challenging Drug Discovery Targets</b>    | <br><b>317</b> |
| <i>Marco De Vivo, Matteo Masetti and Giulia Rossetti</i>                                                       |                |
| 12.1 Introduction                                                                                              | 317            |
| 12.2 Deciphering Metalloenzyme Catalysis <i>via</i><br>Computations                                            | 319            |
| 12.2.1 Ribonuclease H                                                                                          | 319            |
| 12.2.2 Epoxide hydrolase                                                                                       | 321            |
| 12.3 Simulating Membrane Proteins                                                                              | 323            |
| 12.3.1 Membrane Enzymes: The Case of FAAH                                                                      | 323            |
| 12.3.2 Ion Channels: The Case of the Kv11.1<br>Channel                                                         | 324            |
| 12.3.3 GPCR: The Case of the Human Adenosine<br>Receptor A <sub>2A</sub> Embedded in Neuronal-like<br>Membrane | 328            |
| 12.4 Tackling Target Flexibility Through<br>Simulations                                                        | 331            |
| 12.4.1 Lactate Dehydrogenase                                                                                   | 331            |
| 12.4.2 Intrinsically Disordered Proteins                                                                       | 333            |
| 12.4.3 Targeting RNA in Trinucleotide<br>Repeats Diseases                                                      | 335            |
| 12.5 Conclusions                                                                                               | 338            |
| References                                                                                                     | 338            |

|                                                                                         |            |
|-----------------------------------------------------------------------------------------|------------|
| <b>Chapter 13 The Polypharmacology Gap Between Chemical Biology and Drug Discovery</b>  | <b>349</b> |
| <i>Albert A. Antolin and Jordi Mestres</i>                                              |            |
| 13.1 Introduction: Chemical Biology and the Limits of Reductionism                      | 349        |
| 13.1.1 Polypharmacology in Drug Discovery                                               | 349        |
| 13.1.2 Selectivity in Chemical Biology                                                  | 351        |
| 13.2 Systems Pharmacology: Databases and Methods                                        | 353        |
| 13.2.1 Databases of Chemical, Biological and Pharmacological Data                       | 353        |
| 13.2.2 Computational Methods to Predict Polypharmacology                                | 354        |
| 13.3 Case Study 1: The Impact of Chemical Probe Polypharmacology on PARP Drug Discovery | 355        |
| 13.3.1 The History of PARP Biology: From Probes to Drugs                                | 355        |
| 13.3.2 PJ34: A PARP Chemical Tool Binding to PIM Kinases                                | 357        |
| 13.3.3 Differential Off-target Kinase Pharmacology Between Clinical PARP Inhibitors     | 360        |
| 13.4 Case Study 2: Distant Off-target Pharmacology among MLP Chemical Probes            | 363        |
| 13.5 Conclusions and Outlook                                                            | 365        |
| Acknowledgements                                                                        | 366        |
| References                                                                              | 366        |
| <b>Subject Index</b>                                                                    | <b>371</b> |

## CHAPTER 1

# *Computational Chemistry and Molecular Modelling Basics*

SAMUEL GENHEDEN,<sup>a</sup> ANNA REYMER,<sup>a</sup>  
PATRICIA SAENZ-MÉNDEZ<sup>a,b</sup> AND LEIF A. ERIKSSON<sup>\*a</sup>

<sup>a</sup> Department of Chemistry and Molecular Biology, University of Gothenburg, 405 30 Göteborg, Sweden; <sup>b</sup> Computational Chemistry and Biology Group, Facultad de Química, UdelaR, 11800 Montevideo, Uruguay  
\*Email: leif.eriksson@chem.gu.se

## 1.1 Introduction

The use of computers for predicting the structures and properties of biomolecules has closely paralleled computer development since the 1950s, and has been one of the core areas of theoretical or computational chemistry for the past 30 years. Initially, the focus was on force-field based methodologies for studying the structures, dynamics and interactions of biomolecules as such, and the development of accurate models for the main biological solvent, water. With the emergence of accurate quantum chemical techniques suitable for studying (from a quantum chemistry perspective) large systems, density functional theory entered the stage in the 1990s as the key approach for investigating enzymatic mechanisms or properties and reactions of small, but biologically relevant, molecules. The combined use of these tools, so-called QM/MM and lately QM/MM-MD techniques enables precise descriptions of biological phenomena and reactions.

With the exponential increase in data to be analysed, obtained through the introduction of automated whole genome and protein sequencing

techniques, the field of bioinformatics rapidly emerged in the early 2000s from the pioneering laborious mapping and comparison of protein and gene sequences in molecular biology, *via* an intense phase, which to a large extent can be viewed as ‘database mining’ and the development of efficient computer based algorithms, into a science of its own, which today has reached a high level of maturity and sophistication. Tools in bioinformatics are nowadays used with great success in structural biology, computational chemistry, genetics, molecular biology, the pharmaceutical industry, pharmacology and more. The aspects of bioinformatics included herein focus on protein structure determination (often referred to as homology modelling), and the tools of database screening and prediction used in drug design.

In this chapter, a brief outline of simulation techniques are given, focusing on the interface between biology and medicinal chemistry; that is molecular mechanics/molecular dynamics to explore the evolution of a system, homology modelling to determine protein structures, and the use of bioinformatics tools such as molecular docking and pharmacophores in drug design. The aim is to provide a brief introduction to a vast and rapidly growing field. In subsequent chapters, more specialised applications are presented, that build upon the foundations given herein. The chapter is in no way intended to be an exhaustive coverage of the entire area of biomolecular simulations, and we have deliberately avoided the inclusion of quantum chemical methods.

The interested reader wishing to dig deeper into the basics of computational modelling is referred to any of the many excellent textbooks available.<sup>1–11</sup>

## 1.2 Techniques in Biomolecular Simulations

### 1.2.1 Molecular Mechanics and Force Fields

The palette of computational chemistry methods has become increasingly versatile. Starting from quantum chemistry, where molecular orbitals and electrons occupying these are described, allows us to calculate any physical or chemical property that directly depends on the electron distribution; reaching all the way to coarse-grained molecular dynamics simulations, where groups of atoms described as beads interacting by laws of Newtonian mechanics, providing valuable insights into the complexity of biological processes on a bigger, cellular level scale. For comparison, a feasible size of a system treated by quantum chemistry calculations, even today, does not exceed a few hundred atoms, whereas the empirical methods, *e.g.* molecular mechanics (MM), can easily handle several hundred thousand atoms, and in case of a coarse-grained approach—several million atoms. Thus, the latter class of methods has become popular among researchers dealing with bio-macromolecular systems, which exist and function in aqueous solutions or lipid environments. The surrounding environment could take up to 90% of

all atoms in a model system, and its presence is crucial for the correct representation of living matter. The giant leap in system size is possible due to reasonable simplicity of the MM potential energy functional. The potential energy is calculated by adding up the energy terms that describe interactions between bonded atoms (bonds, angles and torsions) and terms that describe the non-bonded interactions, such as van der Waals and electrostatic interactions (eqn (1.1)).

$$\begin{aligned}
 V(r^N) = & \sum_{\text{all bonds}} k_l(l - l_0)^2 + \sum_{\text{all angles}} k_\theta(\theta - \theta_0)^2 + \sum_{\text{all torsions}} \frac{1}{2}V_n[1 + \cos(n\omega - \gamma)] \\
 & + \sum_{j=1}^N \sum_{i=j+1}^N \left\{ \varepsilon_{ij} \left[ \left( \frac{r_{0ij}}{r_{ij}} \right)^{12} - 2 \left( \frac{r_{0ij}}{r_{ij}} \right)^6 \right] + \frac{q_j q_i}{4\pi\epsilon_0 r_{ij}} \right\}
 \end{aligned}
 \tag{1.1}$$

The bonded terms represent the stretching of bonds ( $l$ ), bending of valence angles ( $\theta$ ) and rotation of torsional angles ( $\omega$ ); cf. Figure 1.1. Three force constants:  $k_l$ ,  $k_\theta$  and  $V_n$  characterise the energetic cost relative to the equilibrium value, needed to increase the value of a bond length ( $l_0$ ), angle ( $\theta_0$ ) or rotation around a torsional angle. The torsion term represents a periodic rotation of a dihedral angle with periodicity  $n$  and phase  $\gamma$ . The non-bonded energy is the sum of repulsion, attraction and electrostatics between non-bonded atoms. The parameter  $\varepsilon_{ij}$  is related to the well-depth of Lennard-Jones (LJ) potential,  $r_{0ij}$  is the distance at which the LJ potential has its minimum.  $q_i$  is the partial atomic charge,  $\epsilon_0$  is the vacuum permittivity, and

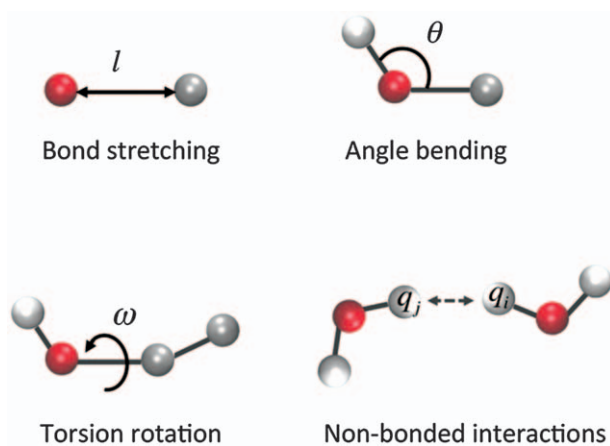


**Figure 1.1** In molecular mechanics, molecular systems are treated by means of classical physics: atoms are represented as charged spheres, which have bonded (bond stretch, angle bend and torsional angle rotation) and non-bonded interactions (van der Waals and electrostatics).

$r_{ij}$  is the distance between atom  $i$  and atom  $j$ . The LJ and Coulomb potentials describe the short-range non-bonded interactions. The evaluations of the long-range electrostatic interactions can be difficult and was often ignored beyond a specific cut-off distance resulting in approximations in a calculation. With the introduction of Ewald summation and particle mesh Ewald (PME) method long-range electrostatic calculations have become significantly more accurate.<sup>12,13</sup>

The simplicity of the potential energy functional form means, on the one hand, fast and easy calculations, and on the other hand that the accuracy of the empirical methods is highly dependent on the set of empirically derived parameters describing atoms and their interactions. These parameters are either derived from *ab initio* or semi-empirical quantum chemistry calculations on small model systems or by fitting to experimental data, *e.g.* X-ray and electron diffraction, NMR and IR spectroscopy. The potential energy functional form and the empirically derived parameters can be both referred to as a force field.

There are a number of empirical force fields families available, having different degrees of complexity, and oriented to treat different kinds of systems. The most popular ones designed for biological macromolecules are AMBER,<sup>14,15</sup> CHARMM,<sup>16</sup> and GROMOS.<sup>17</sup> Other force fields, such as OPLS<sup>18</sup> and COMPASS<sup>19</sup> were originally developed to simulate condensed matter; GAFF<sup>20</sup> a force field developed to simulate organic compounds together with bio-macromolecules; and GLYCAM<sup>21</sup> a force field specifically developed for carbohydrates. Both GAFF and GLYCAM are compatible with AMBER. These force fields vary slightly as to the functional form of the potential energy functional, mainly in the non-bonded terms, as well as values of specific atomic parameters. For more details the reader is referred to a recent review on current advances in empirical force fields for biomolecules.<sup>22</sup> For coarse-grained systems, the most commonly used force field is MARTINI,<sup>23,24</sup> which has been parameterised for lipids, proteins, carbohydrates and nucleic acids. Recently, a tool was developed to parameterise small molecules automatically.<sup>25</sup> The MARTINI model is based on a four-to-one mapping, implying that about four heavy atoms are coarse-grained to a single bead. The beads interact predominantly by Lennard-Jones parameters together with harmonic bonds and angles.<sup>23</sup> Other coarse-grained models commonly used are GROMOS<sup>26</sup> and Elba.<sup>27</sup>

### 1.2.2 Basic Simulation Techniques

To explore the energy landscape described by the molecular mechanics force field, *i.e.* to sample molecular conformations, a simulation is required. This is also the route to relate the microscopic movements and positions of the atoms to the macroscopic or thermodynamic quantities that can be measured experimentally.<sup>28</sup> There are two major simulation methods to sample biomolecular systems: molecular dynamics (MD) and Monte Carlo (MC) (Figure 1.2).

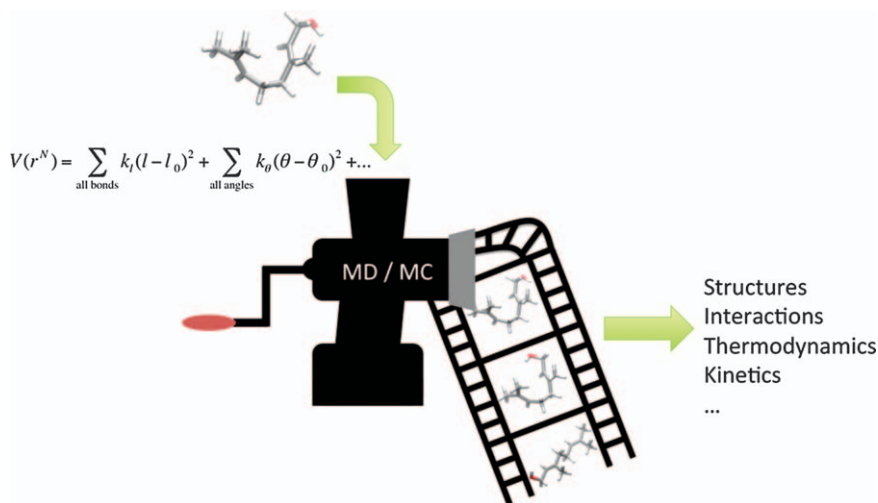


**Figure 1.2** Simulation techniques such as MD and MC use a force field and a starting geometry to generate a set of molecular structures or conformers, which can be analysed to recover *e.g.* structures, interactions and thermodynamic and kinetic quantities.

### 1.2.2.1 Molecular Dynamics

Molecular dynamics is based on Newton's second law of motion, which relates the force,  $\mathbf{F}$ , acted upon an atom to its acceleration,  $\mathbf{a}$ , *i.e.* the second derivative of the position,  $\mathbf{q}$ , with respect to time,  $t$  (eqn (1.2))

$$\mathbf{F} = m\mathbf{a} = m \frac{d^2\mathbf{q}}{dt^2} \quad (1.2)$$

where  $m$  is the mass of the atom. In a molecular simulation, time is discretised and the position after a small, finite time,  $\Delta t$  can be computed using a simple Taylor expansion (eqn (1.3))

$$\mathbf{q}(t + \Delta t) = \mathbf{q}(t) + \frac{d\mathbf{q}(t)}{dt} \Delta t + \frac{d^2\mathbf{q}(t)}{dt^2} \frac{\Delta t^2}{2} + \dots \quad (1.3)$$

and hence it is easy to see that the position  $\mathbf{q}(t)$ , velocity  $d\mathbf{q}(t)/dt$  and acceleration  $d^2\mathbf{q}(t)/dt^2$  are sufficient for propagation of the molecular system. The acceleration can be computed from eqn (1.2) and the force  $\mathbf{F}$  is obtained by differentiating the energy of the system.<sup>29</sup>

An MD simulation is setup by assigning initial velocities and positions to all atoms in the system. The velocities are usually randomly assigned, whereas the positions are typically taken from *e.g.* a crystal structure or idealised geometries. Thereafter, the force acting on each atom is calculated, giving the direction of movement. The atoms are moved in this direction, giving new forces on each atom, and the procedure is then repeated a



number of steps. There are several numerical recipes describing how this integration of motion is done practically, *e.g.* leapfrog, Verlet or velocity-Verlet. They chiefly differ in the numerical stability and whether they in addition to propagating positions also propagate the velocities.<sup>29</sup>

A major limitation to an efficient sampling with MD is the discrete time step,  $\Delta t$ . It is desirable to choose a longer time step, which would give longer simulations with less computational resources. However,  $\Delta t$  is limited by the fastest motion in the simulated system. For an atomistic system, the fastest motion is the bond vibration between a hydrogen and a carbon atom, which limits  $\Delta t$  to about 1 fs. Therefore, these bonds are typically constrained in the simulations, allowing a 2 fs time step. An alternative is to increase the mass of the hydrogen atoms, effectively slowing down the bond vibration.<sup>30</sup> In coarse-grained simulations, a much larger time step is possible, typically between 10 and 40 fs, depending on the model.<sup>23,27</sup>

A simulation obeying Newton's second law of motion can be shown to sample a thermodynamic ensemble with constant number of particles, volume and total energy (kinetic + potential). However, experiments are usually performed at constant temperature and either constant pressure or volume. To sample such an ensemble, the equations of motion have to be modified. In the case of constant temperature, a thermostat is required and there are many such algorithms. Common approaches include (1) modifying the velocities (*e.g.* weak-coupling), (2) introducing fictitious particles in an extended system (*e.g.* Nosé-Hoover), or (3) introducing friction (*e.g.* Langevin dynamics). An extensive discussion of different thermostats is however beyond the scope of this introductory chapter and interested readers are referred to a review on the subject.<sup>31</sup> Similarly to temperature, constant pressure can be introduced by a barostat that modifies the volume of the simulated system. Common approaches include (1) scaling the box dimensions (*e.g.* weak-coupling), (2) introducing fictitious particles (*e.g.* Parinello-Rahman), or (3) introducing a piston. Some thermostats and barostats are better suited for systems far from equilibrium, whereas others are better for production simulations.

### 1.2.2.2 Monte Carlo Simulations

The other major sampling method, Monte Carlo (MC), is a statistical technique where new conformations are generated by a random walk in phase space by assigning random displacements to the internal degrees of freedom, *i.e.* bonds, angles and torsions. Naturally, all conformations are not equally likely and therefore, the sampling is biased such that conformations are generated with a probability prescribed by the thermodynamic ensemble of interest. The overwhelmingly most common way to accomplish this is by performing a Metropolis-Hastings test.<sup>32,33</sup> In a Metropolis MC simulation, a new conformation is accepted with the probability,  $p$  (eqn (1.4))

$$p = \min[1, \exp(-\Delta U/kT)] \quad (1.4)$$

where  $\Delta U$  is the energy difference between the new and old conformation,  $k$  is the Boltzmann constant and  $T$  the absolute temperature. In practice, the energy of the new and old conformation is compared and if the new conformation is lower in energy it is retained for the next step. If it is higher in energy, the Boltzmann factor  $\exp(-\Delta U/kT)$  is compared to a uniform random number between 0 and 1, and if the Boltzmann factor is higher than the random number the new conformation is kept.

An MC simulation consists of a number of moves, which is a recipe on how to sample specific degrees of freedom. This can be a simple translation move where the center of mass of a molecule is displaced, a rotation about a torsion angle or a complicated, concerted move of several protein backbone atoms. A move is selected randomly followed by a Metropolis test of the new conformation and the procedure is repeated for a number of steps.

The Metropolis test illustrated above gives a canonical ensemble, *i.e.* constant number of particles, volume and temperature. However, it can be modified to allow for a volume change such that constant pressure is simulated. Furthermore, entire molecules, *e.g.* water, can be inserted and removed during the simulation, leading to a grand canonical ensemble.<sup>34</sup> Thus, MC simulations are more versatile than MD simulations, but are heavily dependent on the construction of efficient moves. In addition, since MC only depends on the positions of the atoms, dynamic information is lacking, and MC cannot be used to *e.g.* estimate transport properties or diffusion constants.

### 1.2.2.3 Boundary Conditions

An important aspect of both MD and MC is the choice of boundary conditions.<sup>28</sup> Typically, a molecule is solvated by a finite water shell, or inserted in a lipid bilayer, leaving some of the atoms facing vacuum. This is not good physical description of a biological system, and a well-used solution is to extend the system periodically in all three directions to represent a pseudo-infinite system, effectively removing the vacuum. Periodic boundary conditions can be used with various geometries, a cubic box, a rhombic dodecahedron, or a truncated octahedron. The latter scheme is common in simulations of biological macromolecules solvated in water, since it allows the least number of solvent molecules in the system and thus speeds up the computation. Although periodic boundary conditions are the most common choice, there are other solutions in use, *e.g.* spherical boundaries with addition of restraints.<sup>35</sup>

### 1.2.2.4 Enhanced Sampling Techniques

As mentioned earlier, efficient sampling is one of the major limitations of both MD and MC. MD trajectories might not reach all relevant conformations, for example short-lived transient states connected with a biological function. This problem can be addressed by employing enhanced sampling

algorithms, such as metadynamics, steered MD or replica exchange MD. Often, the aim of such enhanced samplings is to build a more complete energy surface and/or to obtain free energy profiles or potential of mean force (PMF) data. Some of the more advanced features are covered in subsequent chapters of this book; we also refer the reader to recent reviews on enhanced sampling techniques.<sup>36,37</sup>

### 1.2.3 Basic Data Analysis

After a simulation has been completed, it needs to be analysed to extract relevant information about the system of interest. This can be quite challenging and depends very much on the type of a simulated system. Here, a few common strategies for analyses will be outlined.

#### 1.2.3.1 Proteins

It is common to estimate the equilibrium of a protein simulation by computing the root mean squared deviation (RMSD; eqn (1.5)) of the backbone atoms compared to the starting conformation,

$$\text{RMSD} = \sqrt{\frac{1}{N} \sum_N (\mathbf{r}_i(t_0) - \mathbf{r}_i(t))^2} \quad (1.5)$$

where  $N$  is the number of atoms,  $\mathbf{r}_i(t)$  the position of atom  $i$ , at time  $t$ . Prior to the analysis the protein need to be fitted onto the starting structure to remove the overall translation and rotation. Although this is a straightforward analysis and gives an indication of local equilibrium, it is a far too simple method to assess the global convergence of the simulation. It is also possible to compute a pair-wise RMSD between each snapshot in a simulation. This could for instance be used in order to evaluate how efficient the sampling has been, or if the simulation has become stuck in a local energy well (Figure 1.3). Whereas the RMSD provides an overall estimate for the entire protein and an approach to assess the degree of movement of individual residues is to compute the root mean squared fluctuation (RMSF; eqn (1.6)), which is simply the variance of the position of an atom:

$$\text{RMSF} = \frac{1}{T} \sum_T (\mathbf{r}(t) - \bar{\mathbf{r}})^2 \quad (1.6)$$

where  $T$  is the total time of the simulation (or number of snapshots) and  $\bar{\mathbf{r}}$  is the average position. The RMSF can be related to the B-factor used in crystallography by multiplying by  $8/3\pi$ . The analysis can be done on a per residue-basis, where all the atoms of a residue is included in the average and can for instance be used to assess the movement of sidechains. Alternatively, one can include only  $\text{C}_\alpha$  atoms in the analysis to assess the backbone movement.

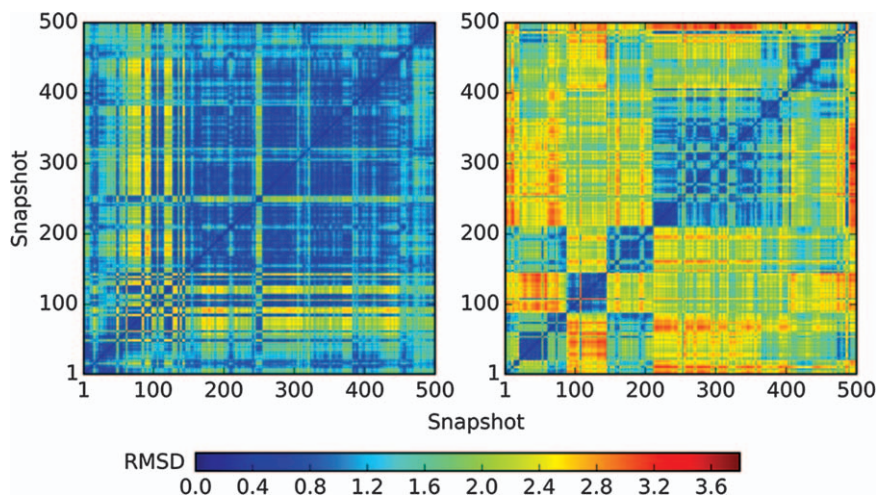


**Figure 1.3** The pairwise RMSD between each snapshot in two MC simulations of a small drug-like molecule. The simulation to the left is stuck in a local energy-well as shown by the small RMSD between the snapshots, whereas the simulation to the right displays larger RMSD, indicating more diverse sampling.

To assess the overall compactness of a protein, the radius of gyration can be computed from eqn (1.7):

$$R_g^2 = \frac{1}{M} \sum_i m_i (r_i - R_c)^2 \quad (1.7)$$

where  $M$  is the total mass of the protein,  $m_i$  the mass of atom  $i$  and  $R_c$  the mass center. This is a simple analysis to determine if the protein is compact or extended. In order to obtain more specific analysis on the protein structure, one can analyse the secondary structure. It is possible to classify each amino acid to determine if it is part of a helix, a beta sheet or a loop. This is useful to monitor during the simulation in order to detect large conformational changes, *i.e.* loss of secondary structure. More complicated motions can be investigated with a principal component analysis (PCA). This is a statistical technique that reduces the dimensionality of problem; instead of looking at all  $3N$  coordinates in a simulation, PCA reduces this to a few principal components that describes the major movements. The principal components are computed from the eigenvalues of a covariance matrix, describing the covariance of the positions of selected atoms. For a protein, typically the  $C_\alpha$  atoms are analysed. The principal component can be projected onto the simulated system and visualised, enabling straightforward inspection of the major motions. This could for instance be a breathing movement of two protein domains or the outward movement of a loop area upon binding of a small molecule.

### 1.2.3.2 Nucleic Acids

The conformational space of DNA or RNA is quite diverse and dynamic, reflecting their ability to change depending on the physicochemical properties of the surrounding environment, local sequence motif, and interactions with other molecules. Thus, apart from RMSD and analysis of the inter and intramolecular networks of contacts, assessment of nucleic acids simulations is directed towards capturing such conformational interplay. The geometry of DNA, and to some extent RNA, can be described in terms of helical parameters (pitch and diameter of the helix), groove parameters (depth and width), furanose ring conformation, six torsional angles of the backbone, rotational (tip and inclination) and translational base pair parameters, six intra-base parameters (buckle, propeller, opening, shear, stretch, and stagger), and six inter-base parameters (tilt, roll and twist, shift, slide, and rise). The most popular programs that analyse nucleic acids simulations in terms of the mentioned degrees of freedom include Curves+ and Canal,<sup>38</sup> and 3DNA.<sup>39</sup> As nucleic acids exist and function as salts, the behaviour of the surrounding counterions is an integral part of the analysis, and could now be done with, *e.g.*, the program Canion.<sup>40</sup>

### 1.2.3.3 Membranes

To assess the equilibration of a membrane simulation, several simple geometric properties are typically calculated such as the area and volume per lipid as well as the thickness of the membrane.<sup>41</sup> The area and volume can be calculated straightforwardly from the box dimensions and by assuming the water density. There are several kinds of thicknesses that can be computed, but a simple one is to measure the distance between the peaks of the density of the phosphate atoms (or equivalent). It is also common to calculate an order parameter of the fatty acyl chain (eqn (1.8))

$$S = \frac{1}{2} \langle 3\cos^2\theta - 1 \rangle \quad (1.8)$$

where  $\theta$  is the angle between the bilayer normal and a carbon–deuterium bond in the acyl chain, and the brackets indicate an average over the simulation. For coarse-grained simulations, the bond between two neighboring atoms replaces the carbon–deuterium bond. It is therefore not correct to compare order parameters from atomistic and coarse-grained simulations. In both cases, the order parameter gives information on the phase of the membrane, *i.e.* if it is in the fluid or liquid-ordered phase.

### 1.2.3.4 Small Molecules

When performing simulations on small molecules (either as solvated entities, in a larger ‘bulk’ system, or as part of *e.g.* a biomolecular complex), an interesting analysis to perform is clustering. This will provide information

on different kinds of conformations the molecule attains during a simulation and can be used to both assess if the simulation is stuck in a local conformation and to investigate the probability of different conformations. There is a plethora of different clustering methods available and it is outside the scope of this text to discuss them at any length; the interested reader is referred to the literature.<sup>42</sup>

### 1.2.4 Software

Over the last two decades with the development of new simulation algorithms and new technologies in hardware platform design, molecular simulations have dramatically increased in size, length and system complexity. The appearance of a variety of standardised molecular modelling software packages, including GROMACS,<sup>43</sup> Amber,<sup>44</sup> CHARMM,<sup>45</sup> GROMOS,<sup>46</sup> and NAMD,<sup>47</sup> has transformed the field of computational chemistry by commoditising molecular simulations and making it accessible to a broader group of researchers. All these packages have complementary strengths and profiles, with GROMACS and NAMD being two of the most popular. Considering GROMACS and NAMD as only MD engines there is no dramatic difference as to their performance, both work with a variety of force fields, and have GPU acceleration implemented. However, small differences should be mentioned, such as the possibility to perform QM/MM simulation in GROMACS, or NAMD's extensibility to user-written scripts. Both packages are distributed free of charge with source code. Moreover, for NAMD, there are downloadable binaries for a variety of platforms. This can be useful for a beginner in computational chemistry, as compilation of MD software might not always be straightforward. Both GROMACS and NAMD are parallel molecular dynamics engines, designed for high-performance simulations of large biomolecular systems, with GROMACS being better for simulations of smaller systems on medium-size supercomputers. To achieve the best performance for a particular system on a particular supercomputer we recommend initial benchmarking. For both GROMACS and NAMD a wide variety of tutorials are available. External software packages, like PLUMED,<sup>48</sup> can provide additional functionality, such as enhanced molecular dynamics techniques mentioned above.

Various pieces of software are used for visualisation and analysis of molecular dynamics trajectories. Among the most popular and freely accessible tools are molecular modelling programs VMD<sup>49</sup> and USCF Chimera.<sup>50</sup> VMD (visual molecular dynamics) is a specialised molecular visualisation program for displaying, animating, and analysing molecular dynamics trajectories, extensively used with any MD software. USCF Chimera, on the other hand, is a highly extensible program for interactive visualisation and analysis of molecular structures and related data, including density maps, supramolecular assemblies, sequence alignments, docking results, trajectories, and conformational ensembles. Both programs could be used to create professional illustrations. VMD and USCF Chimera can also perform basic structural analysis, but for more extensive assessment of trajectories, such as



clustering or modifications of system topologies, AmberTools and, in particular, cpptraj<sup>51</sup> is recommended.

### 1.2.5 Examples

Molecular simulations have been used in numerous applications, to obtain the structure and dynamics of many biomolecules in order to elucidate biochemical functions, processes and pathways. In principle, any molecular system can be simulated, and two typical examples are shown in Figure 1.4. Although many of the simulations have studied individual macromolecules solvated in water or a model membrane, there has also been some effort in looking at larger assemblies. Already a decade ago, Schulten and co-workers used a massively parallel supercomputer to study the complete satellite tobacco mosaic virus, described with an all-atom force field.<sup>52</sup> They were able to accumulate 50 ns of simulation time of a system containing roughly 1 million atoms and could conclude that the virus capsid became unstable upon removal of the core RNA molecules. Using a coarse-grained model, Sansom and co-workers simulated the influenza A virion, a significantly larger system.<sup>53</sup> By using a CG model, they could simulate at the micro-second timescale and in addition investigate alterations to the membrane envelope and sensitivity to temperature.

Long-time scales, such as micro- or milliseconds, are generally not accessible when using an all-atom force field. The exception is if special-purpose hardware is used as in the work from Shaw and co-workers. They reported the first continuous millisecond simulation of a protein described with an all-atom force field when they studied the folded structure of BPTI.<sup>54</sup> Several distinct conformational states were found, separated by large kinetic barriers. Shaw and co-workers has also used simulation techniques to find flaws in current protein force fields, which could not previously be detected due to the typically short simulations. An alternative to running long,

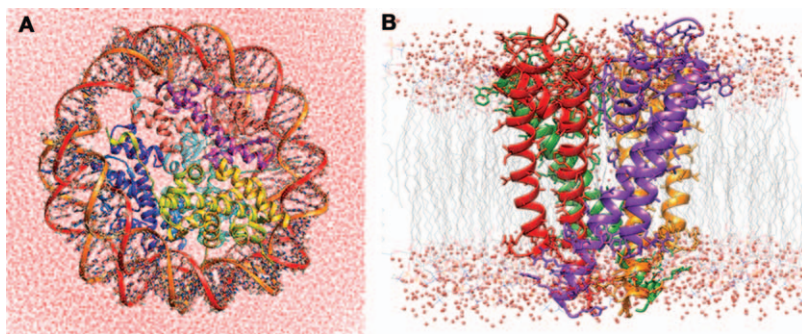


**Figure 1.4** Examples of complex systems that can be simulated by all atom MD: (A) nucleosome particle (PDB id: 5B2I) in 20 Å octahedral box of explicit water, total 205 084 atoms; (B) potassium channel (PDB id: 1K4C) in 100×100 Å POPC membrane patch, total 54 000 atoms.

continuous trajectories that has become popular lately is to assemble many short simulations using Markov state models.<sup>55</sup> They have for instance been used to study protein folding and how this immensely complicated process is affected by solvent effect and electrostatics.<sup>56</sup>

The above-mentioned examples represent extremes, both in terms of system size and length of the simulations. Molecular simulations of much more modest dimensions are routinely used to gain insight into biological processes and to complement wet-lab experiments. We find a very fruitful application of simulations in the field of enzymology.<sup>57</sup> Here the substrate, a few active site residues and coordinating waters and ions are described with quantum mechanics, whereas the rest of the system is described with a molecular mechanics force field. Such simulations have been used in numerous applications to for instance elucidate enzymatic mechanism, understanding the nature of the catalytic power and enzyme design.<sup>57,58</sup> Another common application of molecular simulation is the estimation of binding free energies of small molecules, *e.g.* drugs, to their targets. Jorgensen and co-workers routinely use such simulations to aid in their drug design pipeline.<sup>59</sup> This is done by systematically introducing small chemical groups such as a hydroxyl or methyl group in a lead compound and evaluating its contribution with alchemical free energy simulation.

The above illustrations provide a few examples of what is possible with molecular simulations, with many more provided in the subsequent chapters of this book.

## 1.3 Protein Structure Prediction

Protein structure prediction is often listed among activities within the bioinformatics area, and essentially covers approaches enabling us to go from primary sequence (be it nucleic or amino acids), *via* secondary and tertiary structure, to quaternary structure and possibly also function of the resulting protein. This follows the central assumption that a protein's primary sequence and the inherent properties of the amino acid side chains dictate the final folded three-dimensional structure. Besides the above predictions, which are generally obtained through knowledge-based potentials or algorithms, or by comparing to already existing structures of systems with similar amino acid sequence, analysis of the quality of the resulting model is an essential part of protein structure prediction. We will in this section go through the different steps involved in structure prediction, including tools for analysis and some of the available software and web based solutions.

### 1.3.1 Sequence Alignment and Secondary Structure Prediction

Assuming we know the primary structure; that is, the amino acid sequence of our protein (or, if we have the DNA sequence, have translated this to the



corresponding amino acids, and assuming there are no post-translational modifications that will alter the sequence), the next step is normally to search a database of protein structures for homologous sequences to which we can compare our query through sequence alignment. Normally, the co-ordinate repository of X-ray, NMR and EM structures of the RCSB protein databank from Brookhaven National Laboratory ([www.rcsb.org](http://www.rcsb.org)), or any of its sister sites (PDBe, PDBj, BMRB), or a related database including refined protein structures, is used. To date the protein databank contains over 110 000 solved protein structures, with an annual growth of around 8500–9000 new entries, as well as structures of DNA, RNA and protein–nucleic acid complexes.

The search against databases to identify homologous sequences is normally performed using BLAST (Basic Logical Alignment Search Tool)<sup>60</sup> or FASTA.<sup>61</sup> Nowadays BLAST, which comes in several variants depending on type of algorithm, sequence and database, is the more common. A BLAST search systematically compares three-letter segments of the query sequence, referred to as words, to the database of templates step by step in a heuristic approach. For example, a sequence AHKRV is searched as the words AHK, HKR, KRV; this initial search is referred to as ‘seeding’. Comparison of words from the query sequence with words the database of known sequences is made both based on identity (each residue is matched perfectly), and similarity (similar function/property/size, but not identical), and a total score is calculated using a scoring matrix such as BLOSUM62 (BLOck Substitution Matrix).<sup>62</sup> For example, according to BLOSUM62, an arginine matched by another arginine is given the value +5, arginine *vs.* lysine is +2, and arginine *vs.* cysteine is –2. Also, other scoring matrices and approaches exist, such as identical scoring matrix, minimal mutation distance matrix and point-accepted mutation (PAM).

After the 3-letter word search is done, the word length is extended to nearest and next-nearest neighbours, and possible alternative alignments assessed and scored. An example of the latter is shown in Figure 1.5A assuming a query sequence AHRKCCVGA to be matched against the template sequence AGRKKCVGGA, where different parts given as gaps or insertions provide different scores and result in shorter segments or additional loops of the modelled protein.

If we have several templates to compare against, we must, in addition to the possible alignments as above, also consider which of those alternative yet slightly mismatching sequences that fits the best; *e.g.* assuming we again have our query sequence AHRKCCVGA, is AHRKSVCVGGA or AHRACKVCVGA a better template (*cf.* Figure 1.5B)? In the case of multiple sequences to which we compare our query, this is referred to as multiple sequence alignment, of which the most common methods are the iterative PSI-BLAST<sup>63,64</sup> and CLUSTALW.<sup>65</sup> Sequence alignment approaches are also commonly used to explore similarity of a certain protein between different species, to identify conserved residues and motifs, and similar.

|                   |                             |                                                                                          |                                  |
|-------------------|-----------------------------|------------------------------------------------------------------------------------------|----------------------------------|
| <b>A</b>          | AHRKCCVGA<br><i>Query</i>   | AGRKKCCVGGGA<br><i>Template</i>                                                          |                                  |
| <i>Alignments</i> | AHRK-CCVG-A<br>AGRKKCCVGGGA | AHR-KCCV-GA<br>AGRKKCCVGGGA                                                              | AHRK-CCV-GA<br>AGRKKCCVGGGA      |
| <b>B</b>          | AHRKCCVGA<br><i>Query</i>   | AHRKSVCVGGGA<br><i>Template 1</i>                                                        | AHRACKVCVGA<br><i>Template 2</i> |
| <i>Alignments</i> |                             | AHR--K-C-CVG-A<br>AHR--K-SVCVGGGA<br>AHR--KSV-CVGGGA<br>AHRACK-V-CVG-A<br>AHRACKVC--VG-A |                                  |

**Figure 1.5** Examples of different alignments to same sequence (A), and multiple sequence alignment (B).

Once the best-scoring template(s) have been determined, we next compute the most likely secondary structure elements of the query sequence, again by stepwise comparing 3- and 5-letter segments and their likelihood of forming any of the common motifs. Each amino acid is scored (often using a scale from 0 to 9) based on probability to attain a certain structural motif in its local environment, and the prediction is compared to the template structure(s) that have been selected from the (PSI-)BLAST search. Secondary structure predictions are commonly displayed or mapped graphically in some way, *e.g.* red bar for helix, yellow arrow for  $\beta$ -sheet (*cf.* Figure 1.6A as an example). The aim is to determine which parts of our query sequence that locally are more likely to form  $\alpha$ -helices, which sections that have a propensity for formation of  $\beta$ -sheets, where loops or coil-regions will be. This, again, is used to match the similarity against the obtained templates, and assist in the folding predictions. In addition, predictions are also frequently made based on the properties of amino acid sidechains, of which segments or parts of the sequence that are more hydrophobic or more hydrophilic. The rationale for this in accurate structure determination becomes obvious if we *e.g.* compare a globular protein present in the cytosol (hydrophilic surface and hydrophobic interior) with a membrane spanning ion channel (hydrophobic residues on the outside, interacting with the membrane lipids, and hydrophobic residues in the interior, lining the pore).

### 1.3.2 Comparative Modelling Approaches

Having determined suitable templates for our query sequence, and most likely secondary structure elements, *i.e.* the primary fold of the sequence, we next organise or pack our query structure according to the templates in order to generate a tertiary structure model. This is referred to as comparative

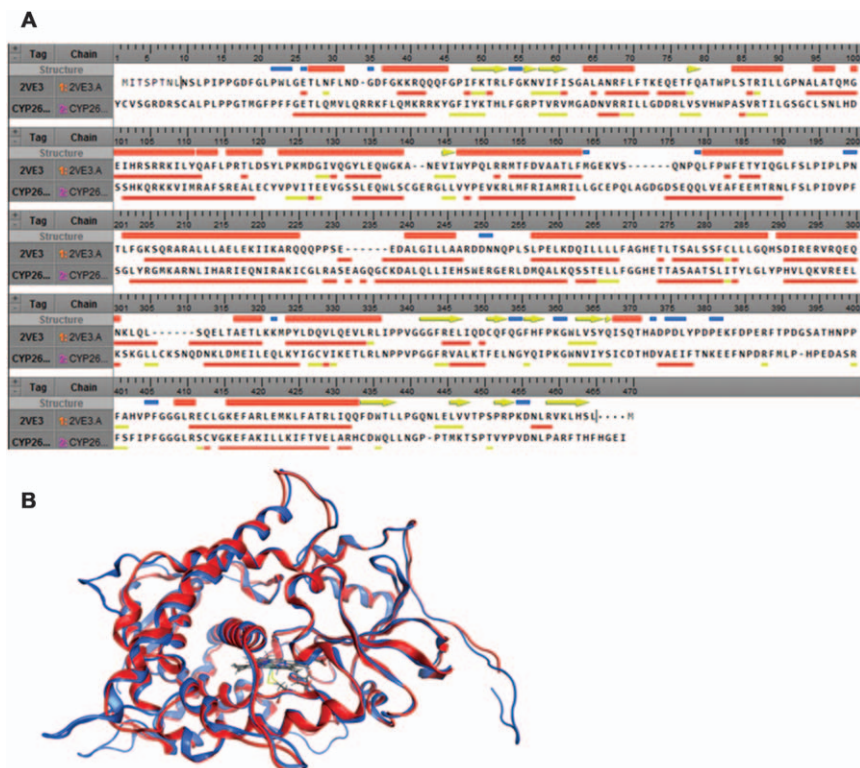


**Figure 1.6** (A) Sequence alignment of template CYP120A1 (PDB ID 2VE3) and amino acid sequence of CYP26A1 based on this template. Top row in each segment displays consensus secondary structure (red for  $\alpha$ -helix, yellow for  $\beta$ -sheet and blue for loops, and below each sequence the analysed or predicted secondary structure is displayed. (B) Superposed homology model of CYP26A1 (blue) and its template CYP120A1 (red).

protein modelling, and relies on the fact (as far as we can conclude from the currently available determined protein structures) that, although the number of possible proteins is essentially infinite, the number of folds is limited to approximately 2000 different ‘types’. That is, provided the sequence similarity (or  $\sim$ identity) between two proteins is sufficient, the two structures will in all likelihood have essentially the same backbone topology in the aligned regions.

Normally one separates the field of comparative protein modelling into homology modelling and threading/fold recognition. In homology modelling, we assume that if two sequences are so closely related that they can be satisfactorily aligned, they will also attain the same three-dimensional structure. This approach views the problem from the standpoint that folds are more evolutionarily conserved than the actual sequences. Clearly, the more identical the sequences, the better the model—if two sequences share 70% identity—the accuracy of the modelled structure is claimed to be

similar to that of a crystal structure with resolution ( $C\alpha$  RMSD) in the range of 1–2 Å; at 25% identity, the structure corresponds to a resolution of 2–4 Å.

If there are no apparent homologous proteins identified, secondary structure prediction is necessary and compared to the template database, where after the sequence is ‘threaded’ according to the fold of the best-matching recognised template(s). This is also referred to as 3D–1D fold recognition, as it links a primary sequence to a three-dimensional structure. 3D–1D alignment is sometimes also included as an intermediate step in the above homology modelling.

There are a number of variants to the above, such as fragment assembly and segment matching; which in essence means that smaller parts of the sequence are modelled separately and then combined into a full protein structure.

In the unfortunate event that no reasonable templates can be identified, the ‘last resort’ is referred to as *ab initio* structure prediction. In this case the secondary structure elements are assembled stochastically, normally using a Monte Carlo type of algorithm, combined with refinement and (possibly) shorter simulations in order to generate a large number of potential three-dimensional models which are assessed and either discarded or improved further in successive iterations. One must, however, be very cautious when it comes to the interpretation of protein structures generated entirely without prior knowledge or templates, as the uncertainty in predicting the appropriate spatial arrangement between secondary structure elements is very high.

Regions of high flexibility, such as loops or the C- and N-termini, are normally poorly resolved or missing in X-ray crystal structures, and loop modelling is hence one of the approaches by which homology modelling can be used in order to improve a protein structure. Care should, however, be taken if the modelled loop is longer than ten amino acids.

Once a model is available, side chains need be optimised (packed) properly. This is done by successively evaluating the energetics for different rotameric states of the sidechains, either by actual energy calculations or using rotamer libraries, and determining the lowest possible (*i.e.* most stable) overall configuration. Finally, the model is generally subjected to some form of energy relaxation or minimisation. In Figure 1.6B, an example of a homology model is shown, and superposed to its template. The overall  $C\alpha$  RMSD of 1.75 Å over 433 residues, with 31.5% sequence identity between query and template. As seen, the agreement for the ordered secondary structure regions is very high, whereas the main deviations are noted in the loops and termini.

Several programs also have the capability to generate models from different templates, and merge the best-matching local segments thereof to construct a hybrid multiple template model. In general, one main template is in those cases used for the core structure, and replacing smaller fragments that are less accurately determined, such as loops or stretches where the sequence similarity to the ‘core template’ is particularly low.

### 1.3.3 Function Prediction

Predicting the explicit role or function of a protein, or different parts thereof, is more difficult, and is compared to the above protein structure determination still at a very early stage of development. In essence, protein function prediction relies heavily on identifying homologous regions either by sequence motifs or by 3D structure alignment, to identify possible domains and, by analogy with the identified templates, their specific roles. However, herein lies also the aspect of paralogs—proteins that have evolved from a common ancestor into structurally very conserved entities but where the functional role is entirely different—which makes the task even more difficult. It lies beyond the scope of the current introductory chapter to also cover such aspects, although we do mention some servers and other tools in Section 1.3.5 below.

### 1.3.4 Analysing the Quality of the Modelled Structure

Once a structure has been modelled, it is crucial to also assess its quality. Some quality checks are already embedded in the routines employed in the model development (such as BLAST *E*-value, BLOSUM score, side chain dihedrals and packing score), but it is recommended that a thorough assessment is made using some of the many servers and programs available. Some of the key tools are included herein.

The first assessment to be made is to produce a Ramachandran plot of the obtained structure, which displays the values of the  $\phi$  and  $\psi$  angles of the protein backbone, and which provides a picture of the stereochemical quality of the amino acids. The RAMPAGE server (<http://mordred.bioc.cam.ac.uk/~rapper/rampage.php>) is one of the main tools for this. A large number of outliers (*i.e.* values of the backbone torsional angles that do not fall into the allowed or generously allowed regions) indicates that the model has significant problems.

Secondly, the folding reliability can be evaluated using the Verify3D server ([http://services.mbi.ucla.edu/Verify\\_3D](http://services.mbi.ucla.edu/Verify_3D)). This assessment evaluates the likelihood that a particular residue in a particular sequence context part-takes in the predicted 3D fold, and provides an estimate of correct *vs.* incorrect folding for each amino acid. Scores below zero indicate serious folding problems, and 80% of the residues of a protein should attain values  $\geq 0.2$  in the 3D–1D profile for the model to have acceptable folding reliability.

The absolute quality of a model can be obtained using the QMEAN Z-score server (<http://swissmodel.expasy.org/qmean/cgi/index.cgi>), which calculates the quality of the model by combining six different structural descriptors such as secondary structure elements, solvent accessibility and torsional angles. The normalised mean value is then compared against the corresponding values of a non-redundant set of high-resolution experimental structures of similar size ( $\pm 10\%$ ) solved through X-ray crystallography.

Also, many other quality and property checks can be performed besides the basic tools listed above. Examples are determination of druggable sites and their relative scores, relating to docking simulations; see Chapter 1.4,

using *e.g.* ProteinsPlus, <http://proteinsplus.zbh.uni-hamburg.de> (recently changed from DoGSiteScorer), and metal binding site interactions, if present, through the sever CheckMyMetal (CMM), [http://csgid.org/csgid/metal\\_sites](http://csgid.org/csgid/metal_sites).

### 1.3.5 Software and Web Based Servers

There are several programs available for protein structure determination, both as standalone codes to be installed on the users own computer or cluster, and as web based servers. The current compilation is not intended to be exhaustive, but meant to provide a sample of different options available; most (but not all) of the programs listed are free for academic users. Each program has its pros and cons, and the reader is advised to read up on the different codes and approaches first, and to preferably test more than one code in order to build up experience in what functions the best for his or her particular needs. Lastly, a slight word of caution: Although there is a plethora of web based programs available, one must always remember that submitting your computation to someone else's computer (server) means you have no control over the results, including aspects pertaining to safety/security.

I-TASSER—server and downloadable (<http://zhanglab.ccmb.med.umich.edu/I-TASSER>).<sup>66</sup>

LOMETS—meta-server combining nine different programs (<http://zhanglab.ccmb.med.umich.edu/LOMETS>).<sup>67</sup>

MODELLER—a standalone program and server; several graphical interface programs are also available that use MODELLER (<http://salilab.org/modeller>).<sup>68</sup>

MOE—a standalone program, with license fee (<http://www.chemcomp.com>).<sup>69</sup>

PHYRE2—server (<http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index>).<sup>70</sup>

PRIME—part of the Schrödinger package; standalone, with license fee (<http://www.schrodinger.com/Prime>).

RAPTORX—server and downloadable (<http://raptorx.uchicago.edu>).<sup>71</sup>

ROBETTA—server, and as part of the downloadable Rosetta3 package (<https://web.archive.org/web/20150819163428/http://www.robetta.org>).<sup>72</sup>

SWISS-MODEL—server (<http://swissmodel.expasy.org>).<sup>73</sup>

YASARA—standalone, minor license fee (<http://www.yasara.org>).<sup>74</sup>

For protein function or domain prediction, fewer programs are available as yet. However, we refer the interested reader to any of the following servers:

PFAM—Protein Families Database (<http://pfam.xfam.org>).<sup>75</sup>

dcGO—(<http://supfam.org/SUPERFAMILY/dcGO>).<sup>76</sup>

PROSITE—Database of protein domains, families and functional sites (<http://prosite.expasy.org>).<sup>77</sup>



## 1.4 Computer-based Drug Design

Computer-based drug design (CBDD) or computer-aided drug design (CADD) refer to the application of different computational methodologies and algorithms for developing bioactive compounds. It is currently an independent discipline within computational chemistry, mainly because it focuses on predicting/designing the next potential bioactive molecule to synthesise and test. It is well known that drug research and development is not only time-consuming, but also very expensive. It has been estimated that developing a new drug from idea to market would take *ca.* 14 years with an associated cost from 800 million to 2 billion dollars.<sup>78,79</sup> In fact, the overall cost is increasing every year, mainly for specialised drugs for smaller patients populations.<sup>80</sup> This emphasises the benefits of applying computational tools in the early stages of drug discovery, thereby reducing the cost, the required time and the inherent risks such as late-stage failure.<sup>81</sup> The well-known ‘fail fast, fail early’ pharmaceutical mantra is the goal.<sup>82,83</sup> While high-throughput screening (HTS) of large compound libraries is still the major source for discovering new hits in drug design, CBDD is currently playing a key role in the search for novel bioactive compounds, both in pharmacy and academy.<sup>84</sup> A comparison between the two techniques has been reported in the screening for novel inhibitors of Protein Tyrosine Phosphatase-1B (PTP1B), an enzyme implicated in diabetes. The HTS of 400 000 compounds resulted in 85 hits actually inhibiting the enzyme (0.021%). On the other hand, 365 high-scoring molecules were obtained from the virtual screening, 127 of which inhibited PTP1B (34.8%). These results clearly showed that CBDD increased the hit rate over random (HTS) screening.<sup>85</sup> Thus, the application of computational tools allows for covering a larger part of chemical space and at the time the number of compounds that must be synthesised, is drastically reduced.

CBDD can be classified into two main classes: structure-based drug design (SBDD) and ligand-based drug design (LBDD).<sup>86,87</sup> SBDD is based on the knowledge of the 3D structure of the target protein, using virtual screening techniques to search for molecules having complementarities toward the selected target. For SBDD, molecular docking, virtual screening and molecular dynamics are the most important underlying methodologies.<sup>88</sup> LBDD does not require knowledge of a protein, instead using the information provided by known active and inactive compounds to find potential hits by similarity searches or quantitative structure–activity relationship studies (QSAR).<sup>87</sup> The latter is usually the selected methodology when there is no structural information available of the target system.

### 1.4.1 Pre-requisites for SBDD—Sampling Algorithms and Scoring Functions

Molecular docking is a methodology that attempts to predict the conformation of ligands within the receptor binding site.<sup>89</sup> The identification of

the ‘best pose’, *i.e.* the ligand internal conformation and orientation towards the receptor, involves searching the ligand conformational space (sampling or posing) and ranking of the predicted binding conformations (scoring).

#### 1.4.1.1 Sampling Algorithms

Docking most frequently deals with ligand flexibility and in some cases with protein flexibility. Virtual screening employing a large ligand library (virtual high-throughput screening, VHTS) is a very time- and resource-consuming procedure. Therefore, usually the ligand and receptor are both treated as rigid bodies (rigid ligand and rigid receptor docking) in the initial screening. Even though the search space is restricted, large libraries can be rapidly explored and filtered. More often, molecular docking employing smaller libraries treats the ligand as flexible, while the receptor is kept fixed (flexible ligand and rigid receptor docking). Finally, incorporating receptor flexibility is the most accurate and costly methodology, and it is usually employed for refining previous docking rounds (flexible ligand and flexible receptor docking).<sup>90,91</sup> Treatment of ligand flexibility includes systematic, stochastic, and simulation methods.<sup>92–94</sup>

**1.4.1.1.1 Systematic Methods.** These methods account for ligand flexibility by exploring the conformational space of the molecule. After search of a ligand’s degrees of freedom, the method converges to the most likely binding mode. As with all ‘down-hill’ methods, a systematic search can converge to a local minimum rather than the global one, a problem that can be overcome by performing several searches starting from different initial ligand conformations.<sup>95</sup> When exploring all possible degrees of freedom in a ligand (exhaustive search), the number of possible combinations is usually prohibitive, facing the so-called problem of combinatorial explosion. An alternative to exhaustive search is to employ incremental construction algorithms.<sup>93</sup> Docking programs such as DOCK,<sup>96</sup> FlexX<sup>97</sup> and Glide<sup>98</sup> apply an incremental construction search method, in which the ligand is first divided into fragments. One fragment is selected as anchor (usually the larger fragment) and docked in the binding site. The remaining fragments are incrementally added until the entire ligand is built.<sup>90</sup>

**1.4.1.1.2 Stochastic Methods.** Stochastic methods perform a random search of the conformational space by making random changes to the ligand or a population of ligands. Such changes include translational, rotational and internal modification of the ligand’s coordinates. This strategy allows for finding the global minimum, covering also a larger conformational space. MC, genetic algorithms and tabu search are typical algorithms belonging to this class.<sup>92,94</sup>

MC methods make random modifications to the ligand structure and the resulting conformation is tested according to the metropolis criterion, which accepts conformations with a lower energy, and higher energy states when



the Boltzmann factor is greater than a random number between 0 and 1 (see also Section 1.2.2). Programs such as Autodock<sup>99</sup> and MOE<sup>69</sup> employ MC methods for sampling.

Genetic algorithms (GA) have their roots in the Darwin's theory of evolution and natural selection. All structural parameters are encoded in genes, and a particular pose is referred to as a 'chromosome'. The random search algorithm then generates several chromosomal mutations (*i.e.* several poses), which are in turn evaluated in terms of energy. The best adapted chromosome will be the one with the lowest energy and thus selected to be used in the next generation. The next generation is populated with poses having increased favourable structural characteristics. After several generations (several conformational search cycles), the energy minimum conformation is reached.<sup>100</sup> Programs implementing genetic algorithms are Autodock<sup>101</sup> and Gold.<sup>102</sup>

Tabu search (TS) is a heuristic method originally proposed by Glover in 1986.<sup>103</sup> The algorithm proceeds stepwise from a conformation, generating a number of moves to the current solution. The moves are scored, ranked using the energy function and examined. The method keeps a list of the previously visited solutions, and a move is considered 'tabu' if it generates a solution that is not sufficiently different from the previous ones. The algorithm calculates the root mean square deviation (RMSD) between the current move and the all previously recorded solutions. Only those movements having a RMSD smaller than a cut-off are accepted. The tabu search continues for a user-defined number of iterations.<sup>104</sup> Examples of programs implementing tabu search are MOE<sup>69</sup> and PRO\_LEADS.<sup>105</sup>

**1.4.1.1.3 Simulation Methods (MD).** MD simulations (*cf.* Section 1.2.2) are also used in the context of molecular docking, allowing for representing the flexibility of both the ligand and the receptor. However, MD is not the best method for simulation of ligand–target interactions, mostly for its intrinsic difficulty to cross high-energy barriers, leading to a poor sampling. Considering that MD simulations are efficient at exploring the local hyper surface, the best approach is to use a systematic or random search in order to find the most likely conformation for the ligand, followed by MD simulations.<sup>90</sup> For techniques dealing with receptor flexibility in particular, see Section 1.4.2.

### 1.4.1.2 Scoring Functions

Molecular docking programs predict binding conformations employing sampling algorithms, and their evaluation to estimate the energy of the ligand–target interaction is crucial. To this end, scoring functions are employed aiming to rank the complexes and discriminate correct poses from incorrect ones. Therefore, the design and proper use of scoring functions is of utmost importance in SBDD. Scoring functions can be classified into four

types: force field-based, empirical, knowledge-based and consensus scoring functions.<sup>90,92–94,106–109</sup>

**1.4.1.2.1 Force Field-based Scoring Functions.** Force field-based scoring functions estimates the binding energy using classic molecular mechanics formulations, calculating the sum of the non-bonded interactions (*i.e.* electrostatic and van der Waals terms). The electrostatic terms are calculated by a Coulombic formulation using a distance-dependent dielectric function to account for charge–charge interactions, whereas the van der Waals terms are usually described by a Lennard-Jones potential function (see Section 1.2.1). The parameters of the Lennard-Jones term can modify the ‘hardness’ of the potential, which in turns changes the distance between the receptor and ligand atoms. The most important limitations of force field-based scoring functions include the introduction of cut-off distances for the treatment of non-bonded interactions, which reduce the accuracy in calculating long-range effects involved in binding. In addition, force field scoring functions do not estimate entropic contributions and solvation energies.<sup>110</sup> The results of scoring with force field-based functions can be refined through calculation of linear interaction energy (LIE),<sup>111,112</sup> inclusion of generalised solvation through the Born model (MM/GBSA)<sup>113</sup> and free-energy perturbation methods (FEP).<sup>114–117</sup> Programs such as Gold,<sup>102</sup> Dock<sup>96</sup> and AutoDock<sup>101</sup> employ force field-based scoring functions.

**1.4.1.2.2 Empirical Scoring Functions.** Empirical scoring functions fit parameters to reproduce experimental data, such as binding energy, originally proposed by Böhm.<sup>118,119</sup> The binding energy is decomposed into several weighted terms, such as hydrogen bond, hydrophobic contact terms, desolvation energy, ionic interactions and binding entropy. The coefficients of each term are calculated from a regression analysis using experimental information from a training set of ligand–protein complexes with known binding affinities. Although the empirical scoring functions are simple to evaluate, the major drawback is their dependence on the training set employed and thus the transferability of the weighted parameters.<sup>120–124</sup> ChemScore<sup>120</sup> and FlexX<sup>97</sup> are examples of programs using empirical scoring functions.

**1.4.1.2.3 Knowledge-based Scoring Functions.** These functions are designed to reproduce experimentally determined complex structures. They are based on the assumption that the more favourable interatomic distances occur with higher frequency and the algorithms model those frequency distributions as pairwise atom-type potentials. The score is calculated as a sum of the individual interactions. The functions are computationally simple allowing for screening large compound databases, but as they rely on the training set employed for deriving the parameters an

extensive use is limited.<sup>125–128</sup> Knowledge-based potentials have been implemented in potential of mean force (PMF)<sup>126,129–131</sup> and DrugScore.<sup>132</sup>

**1.4.1.2.4 Consensus Scoring Functions.** A more recent strategy is the introduction of consensus scoring functions. This scoring scheme combines different scoring functions aiming to improve single scores and increase the ligand enrichment (*i.e.* the percentage of occurrence of strong binders among high-scoring ligands). However, when the employed scoring functions are significantly correlated, the method could magnify the calculation errors, rather than attenuate them.<sup>133–135</sup> CScore implements consensus scoring by combining Gold, DOCK, PMF, ChemScore and FlexX scoring functions.<sup>136</sup>

## 1.4.2 Structure Based Drug Design (SBDD)

SBDD makes use of high-throughput virtual screening (HTVS) techniques to search for bioactive compounds, identifying hits out of thousands of molecules by detecting complementarities between the ligand and the biological target. Selected compounds are ranked employing different scoring functions. Eventually the selected hits are experimentally evaluated to assay the biological activity on the selected target.<sup>137,138</sup> SBDD consists on the following key steps: (1) preparation of the target receptor, (2) compound database selection and preparation, and (3) molecular docking (*i.e.* determination of a favourable binding pose for each compound and ranking of the docked structures).

The first step involves the preparation of the target receptor, which indeed is of uttermost importance because the representation of the active site affects the quality of ligand posing and scoring. Experimentally determined structures of many receptors are available, mostly through X-ray crystallography and NMR spectroscopy. When the receptor structure is not experimentally available, it is possible to create a model starting from the sequence and applying homology modelling (see Section 1.3). Once the receptor model has been selected it has to be prepared for molecular docking studies, usually by adding hydrogen atoms, removing water molecules unless they bear important interactions, calculating partial charges and assigning tautomerisation states.<sup>139</sup> The initial selection of the structure is critical in the sense that small conformational changes arising from ligand binding highly influence the results, *e.g.* when using holo-, apo-proteins or homology models as targets.<sup>140</sup>

The second step is the selection and preparation of the small-molecule database. Several public databases containing millions of compounds and chemical information are freely accessible.<sup>141,142</sup> The most common chemical databases used in VHTS are ZINC,<sup>143,144</sup> PubChem,<sup>145,146</sup> DrugBank,<sup>147,148</sup> ChemSpider,<sup>149,150</sup> ChemBank,<sup>151,152</sup> eMolecules,<sup>153</sup> ChEMBL,<sup>154,155</sup> ChemDB,<sup>156,157</sup> and Binding DB.<sup>158,159</sup>

Preparing the original libraries requires different aspects depending on the software employed to perform the screening, but commonly involves the correct assignment of stereochemistry, partial charges and ionisation state according to the selected pH. In addition, several filters may be applied in order to enrich according to expected physicochemical properties of the potential ligands. Usually, filtering the database according to Lipinski's rule of five is performed to ensure drug-likeness.<sup>160–163</sup>

The third step involves docking experiments of the prepared small-molecule database into the prepared receptor binding site, and the analysis of the resulting docked conformations (Figure 1.7).

Exploration of ligand flexibility was described in previous sections of this chapter. The biological molecules are intrinsically mobile and consequently the representation of molecular flexibility of receptors is an important aspect of SBDD.<sup>165</sup> Incorporating receptor flexibility is a challenge in molecular docking due to the evident computational cost of modelling multiple degrees of freedom. Methods for accounting for receptor flexibility include the use of soft potentials (soft-docking), use of rotamer libraries, inclusion of side chain flexibility, and to perform ensemble docking.<sup>166</sup> Soft-docking decreases the van der Waals repulsion term energy in the scoring function allowing for partial overlapping between the receptor and the ligand. This method is simple but does not include suitable flexibility.<sup>167</sup> Employing rotamer libraries involve searching within the library to obtain possible conformations of the residue side chains. Even though it is efficient in terms of computation, it is highly dependent on the database used and ignores backbone flexibility.<sup>168</sup> Including side chain flexibility consist in sampling several side chains of the receptor simultaneously with the ligand sampling using for instance genetic algorithms. The main drawback is that only selected side chains are accounted for (the others are treated as rigid) and the backbone is not considered flexible. Finally, using an ensemble of protein conformations as obtained from *e.g.* NMR experiments allows for docking

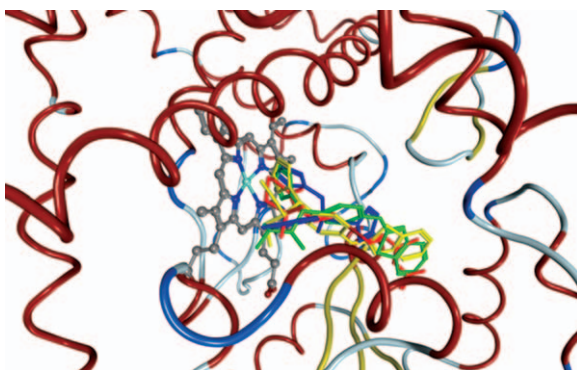


**Figure 1.7** Structures of atRA (all-*trans* retinoic acid) in red, fluconazole in blue, R115866 in yellow and R116010 in green, docked in the active site of CYP26A1.<sup>164</sup>

the ligand into several receptor structures (*i.e.* different conformations), considering in each round a flexible ligand and the receptor as a rigid body. This method completely accounts for flexibility, but only for those protein conformations included in the sampling.<sup>169,170</sup> Even though several pitfalls of SBDD are well-known,<sup>171,172</sup> it has been successfully employed in identifying potent hits in many drug discovery studies.<sup>173–181</sup>

### 1.4.3 Ligand Based Drug Design (LBDD)

LBDD involves the analysis of ligands known to interact with the selected target. Several molecular descriptors are calculated for a set of reference compounds, *i.e.* compounds known to be active, which in turn are applied as molecular filters. Thus, the filtering is employed to select compounds sharing characteristics with the reference set. These methods do not require any information of the structure of the receptor. A distinct approach is the construction of a QSAR model predicting the biological activity from chemical structure.<sup>182,183</sup>

Different molecular descriptors can be calculated and the selected set depends on the biological function to be predicted. Molecular descriptors can be 2D (depending only on the topological connectivity) or 3D (depending on the geometry). 2D descriptors include physical properties such as atomic charges, polarisability,  $\log P$  (logarithm of partition coefficient between *n*-octanol and water), solubility in water, volume, number of hydrogen bond donor/acceptor atoms, and molecular weight. 3D descriptors include properties such as total energy (and its components), ionisation potential, and HOMO/LUMO energy.<sup>94,184–186</sup>

After accounting for molecular descriptors, fingerprint techniques (similarity search) may be used to search databases for compounds similar in structure to a query (usually a lead compound).

Quantitative Structure–Activity Relationship (QSAR) models describe the mathematical relation between structural features and target response of a set of compounds.<sup>187,188</sup> The method involves the inclusion of active and inactive ligands, thus creating a set of mathematical descriptors. The subsequent step consists of the generation of a model establishing the relationship between those descriptors and the experimental biological activity of the compounds. Finally, the model is applied to predict the activity of compounds of interest.<sup>94</sup>

### 1.4.4 Pharmacophores

Pharmacophore models are of fundamental importance in drug design when no structural data is available. IUPAC have defined a pharmacophore as ‘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. The pharmacophore can be considered as the largest common denominator shared by a

set of active molecules'.<sup>189</sup> Thus, a pharmacophore is a collection of structural properties relevant for biological activity, a purely abstract concept, rather than a real molecule. Pharmacophoric descriptors include hydrogen-bond donors and acceptors, hydrophobic, aromatic, acidic, basic and ionisable groups.

A pharmacophore model can be ligand-generated, *i.e.* by superposing a set of bioactive molecules and extracting common chemical features responsible for the biological activity, or structure-based, *i.e.* by determining the main interactions between the target and the active ligands.<sup>190</sup> The latter can be obtained by analysing the structure of ligand–receptor complexes (either from crystal structures or from docking experiments), particularly the chemical features of the active site and the interactions with an active compound (Figure 1.8). The pharmacophore model must then fit the selected features (Figure 1.9).<sup>191</sup>

A more challenging problem is to generate a ligand-based pharmacophore model, and involves the following steps: (1) identifying the relevant properties, (2) superposing the molecules according to those properties, and (3) generating the pharmacophore model. The most demanding issue to address is the development of algorithms for effective molecular superposition, ensuring that a maximum number of chemical features overlap geometrically.<sup>192,193</sup> This in turn incorporates the problem of conformational flexibility, that can be addressed by the pre-enumerating method (multiple conformations of each molecule are included into a database), or by performing a conformational analysis during the pharmacophore modelling process as requested by the alignment algorithm (the so called on-the-fly method).<sup>190,193</sup> Once the ligands have been aligned, a pharmacophore feature map is extracted. A more general property definition increases the population of compounds matching the pharmacophore. This allows for identifying new compounds but also increasing the rate of false positives.<sup>94,194</sup>

Ligand-based pharmacophore modelling has become an essential computational strategy for drug discovery in the absence of structural information about the target. Several programs incorporate pharmacophore construction, such as Catalyst (part of Biovia Discovery Studio),<sup>195,196</sup> Phase<sup>197,198</sup> accessible by Schrödinger's graphical interface Maestro,<sup>199</sup> and MOE.<sup>69,200</sup>

### 1.4.5 Compound Optimisation

The last step of computational drug discovery involves the modification of the hits in order to improve the biological activity by changing the chemical structure, the hit-to-lead process. This optimisation involves increasing the drug potency (two- or three-fold), selectivity and pharmacokinetics, including absorption, distribution, metabolism, excretion and potential for toxicity (ADMET).<sup>201,202</sup>

In order to increase the biological potency of detected hits, similarity search employing pharmacophoric models is a valuable tool.<sup>203</sup>

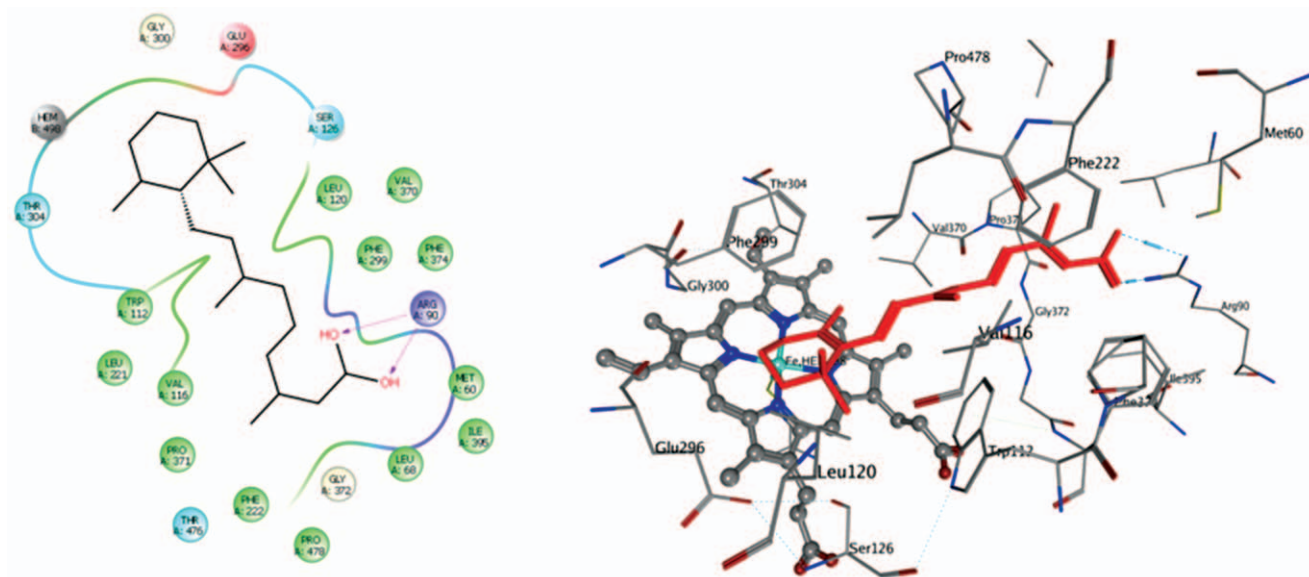


**Figure 1.8** Main ligand interactions of atRA docked into the active site of CYP26A1, 2D (left) and 3D (right) diagrams.



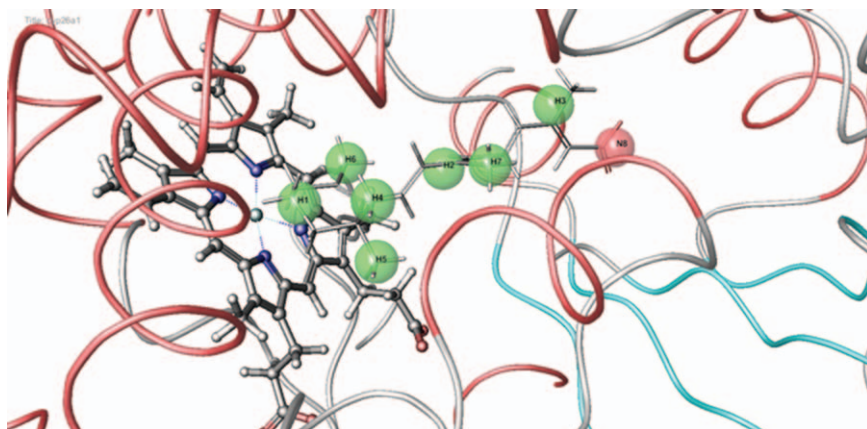


**Figure 1.9** Pharmacophore model for atRA docked into the active site of CYP26A1. Green spheres indicate hydrophobic sites interactions, while the red sphere points out a required hydrogen bond acceptor in that position.

Focusing on ADMET properties, several filters can be employed to sort-out compound libraries, such as Lipinski's rule of five as mentioned in Section 1.4.2. This set of rules is related to those properties thought as necessary for good oral bioavailability and mainly targeting eukaryotic receptors.<sup>204</sup> However, when focusing for instance on antibiotics, the target might be located in the peptidoglycan matrix or the outer surface on the inner membrane. Then, permeation through the inner lipid membrane is not required to kill the pathogens and Lipinski's rules are simply not followed.<sup>205</sup> Therefore, different filtering rules may be needed depending on the particular biological target of interest. Moreover, several biologically active compounds violate more than one of the Lipinski's rules, such as atorvastatin (Lipitor<sup>®</sup>) and montelukast (Singulair<sup>®</sup>),<sup>206</sup> evidencing that automatic filtering might artificially remove potential leads.

Metabolic stability of drugs is a desirable property in the sense that when it is lowered, the drug diminishes its efficacy and increases the risk of generating toxic metabolites. Cytochrome P450 enzymes (CYPs) are major drug-metabolizing enzymes and the prediction of compounds that would be metabolised by, or inhibit CYPs must be assessed.<sup>207,208</sup> In terms of SBDD, the off-target prediction (focusing on available structures of CYPs) must be performed, aiming to determine the affinity of potential hits towards different receptors other than the main biological target. Besides accounting for metabolism, the off-target prediction may include any relevant human protein where inhibition would lead to toxic side effects.

#### 1.4.6 Software and Web Based Servers

In previous sections of this chapter several programs were mentioned, describing the sampling algorithms, scoring functions and type of drug design



scheme included (SBDD or LBDD). Apart from those software packages, several web based servers for molecular docking and virtual screening are available. DOCK Blaster is a web server version of UCSF DOCK allowing for screening ZINC databases subsets.<sup>209,210</sup> SwissDock allow for ligand selection (ZINC ID, URL specification, an internal curated database, or an uploaded file).<sup>211,212</sup> DockThor is an online receptor-ligand docking facility allowing for uploading receptor, ligands and cofactor structures.<sup>213,214</sup> PharmMapper server is an integrated pharmacophore matching platform for potential target identification (off-target binding), very useful when predicting potential toxicity of developed hits.<sup>215,216</sup> These are some examples, many more web-based resources can be found.

## Acknowledgements

Funding from the Swedish research council (LAE, AR), the Faculty of science at University of Gothenburg (LAE), the Wenner-Gren foundations (SG) and the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007–2013) under REA grant agreement no 608743 (PSM) are gratefully acknowledged.

## References

1. A. Hinchliffe, *Molecular Modelling for Beginners*, Wiley & Sons, 2nd edn, 2008.
2. F. Jensen, *Introduction to Computational Chemistry*, Wiley & Sons, 2nd edn, 2007.
3. C. J. Cramer, *Essentials of Computational Chemistry: Theories and Models*, Wiley & Sons, 2nd edn, 2004.
4. A. M. Lesk, *Introduction to Bioinformatics*, Oxford University Press, 2002.
5. D. W. Mount, *Bioinformatics; Sequence and Genome Analysis*, Cold Spring Harbor Laboratory Press, 2nd edn, 2004.
6. J. Xiong, *Essential Bioinformatics*, Cambridge University Press, 2006.
7. D. Higgins and W. Taylor, *Bioinformatics: Sequence, Structure and Databanks*, Oxford University Press, 2000.
8. A. R. Leach, *Molecular Modelling: Principles and Applications*, Pearson, 2nd edn, 2001.
9. M. P. Allen, D. J. Tildesley, *Molecular Simulations of Liquids*, Oxford University Press, 1987.
10. *Drug Design: Structure- and Ligand-based Approaches*, ed. K. M. Merz Jr., D. Rigne and C. H. Reynolds, Cambridge University Press, 2010.
11. D. C. Young, *Computational Drug Design: A Guide for Computational and Medicinal Chemists*, Wiley & Sons, 2009.
12. T. A. Darden, D. York and L. Pedersen, *J. Chem. Phys.*, 1993, **98**, 10089–10092.
13. C. Sagui and T. A. Darden, *Annu. Rev. Biophys. Biomol. Struct.*, 1999, **28**, 155–179.

14. A. Pérez, I. Marchán, D. Svozil, J. Sponer, T. E. Cheatham, C. A. Loughton and M. Orozco, *Biophys. J.*, 2007, **92**, 3817–3829.
15. Y. Duan, *et al.*, *J. Comput. Chem.*, 2003, **24**, 1999–2012.
16. N. Foloppe and A. D. MacKerell, Jr., *J. Comput. Chem.*, 2000, **21**, 86–104.
17. C. Oostenbrink, A. Villa, A. E. Mark and W. F. van Gunsteren, *J. Comput. Chem.*, 2004, **25**, 1656–1676.
18. W. L. Jorgensen, D. S. Maxwell and J. Tirado-Rives, *J. Am. Chem. Soc.*, 1996, **118**, 11225–11236.
19. H. Sun, *J. Phys. Chem. B*, 1998, **102**, 7338–7364.
20. J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, *J. Comput. Chem.*, 2004, **25**, 1157–1174.
21. K. N. Kirschner, *et al.*, *J. Comput. Chem.*, 2008, **29**, 622–655.
22. L. Monticelli, D. P. Tieleman, *Methods in Molecular Biology*, Clifton, N.J., 2013, pp. 197–213.
23. S. J. Marrinck, H. J. Risselada, S. Yefimov, D. P. Tieleman and A. H. de Vries, *J. Phys. Chem. B*, 2007, **111**, 7812–7824.
24. S. J. Marrinck and D. P. Tieleman, *Chem. Soc. Rev.*, 2013, **42**, 6801.
25. T. Bereau and K. Kremer, *J. Chem. Theory Comput.*, 2015, **11**, 2783–2791.
26. S. Riniker and W. F. van Gunsteren, *J. Chem. Phys.*, 2011, **134**, 084110.
27. M. Orsi and J. W. Essex, *PLoS One*, 2011, **6**, e28637.
28. C. Mura and C. E. McAnany, *Mol. Simul.*, 2014, **40**, 732–764.
29. M. E. Tuckerman and G. J. Martyna, *J. Phys. Chem. B*, 2000, **104**, 159–178.
30. K. A. Feenstra, B. Hess and H. J. C. Berendsen, *J. Comput. Chem.*, 1999, **20**, 786–798.
31. P. H. Hünenberger, *Adv. Polym. Sci.*, 2005, **173**, 105–149.
32. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, *J. Chem. Phys.*, 1953, **21**, 1087.
33. W. K. Hastings, *Biometrika*, 1970, **57**, 97–109.
34. G. A. Ross, M. S. Bodnarchuk and J. W. Essex, *J. Am. Chem. Soc.*, 2015, **137**, 14930–14943.
35. G. King and A. Warshel, *J. Chem. Phys.*, 1989, **91**, 3647.
36. R. C. Bernardi, M. C. R. Melo and K. Schulten, *Biochim. Biophys. Acta, Gen. Subj.*, 2015, **1850**, 872–877.
37. V. Spiwok, Z. Sucer and P. Hosek, *Biotechnol. Adv.*, 2015, **33**, 1130–1140.
38. R. Lavery, M. Moakher, J. H. Maddocks, D. Petkeviciute and K. Zakrzewska, *Nucleic Acids Res.*, 2009, **37**, 5917–5927.
39. X.-J. Lu and W. K. Olson, *Nucleic Acids Res.*, 2003, **31**, 5108–5121.
40. M. Pasi, J. H. Maddocks and R. Lavery, *Nucleic Acids Res.*, 2015, **43**, 2412–2433.
41. D. Poger and A. E. Mark, *J. Chem. Theory Comput.*, 2010, **6**, 325–336.
42. J. Shao, S. W. Tanner, N. Thompson and T. E. Cheatham, *J. Chem. Theory Comput.*, 2007, **3**, 2312–2334.
43. D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark and H. J. C. Berendsen, *J. Comput. Chem.*, 2005, **26**, 1701–1718.
44. D. A. Case, *et al.*, *J. Comput. Chem.*, 2005, **26**, 1668–1689.

45. B. A. Brooks, *et al.*, *J. Comput. Chem.*, 2009, **30**, 1545–1614.
46. M. Christen, *et al.*, *J. Comput. Chem.*, 2005, **26**, 1719–1751.
47. J. C. Phillips, *et al.*, *J. Comput. Chem.*, 2005, **26**, 1781–1802.
48. M. Bonomi, *et al.*, *Comput. Phys. Commun.*, 2009, **180**, 1961–1972.
49. W. Humphrey, A. Dalke and K. Schulten, *J. Mol. Graphics*, 1996, **14**, 33–38.
50. E. F. Pettersen, *et al.*, *J. Comput. Chem.*, 2004, **25**, 1605–1612.
51. D. R. Roe and T. E. Cheatham, *J. Chem. Theory Comput.*, 2013, **9**, 3084–3095.
52. P. L. Freddolino, A. S. Arkipov, S. B. Larson, A. McPherson and K. Schulten, *Structure*, 2006, **14**, 437–449.
53. T. Reddy, *et al.*, *Structure*, 2015, **23**, 584–597.
54. D. E. Shaw, *et al.*, *Science*, 2010, **330**, 341–346.
55. V. S. Pande, K. Beuchamp and G. R. Bowman, *Methods*, 2010, **52**, 99–105.
56. G. Jayachandran, V. Vishal and V. S. Pande, *J. Chem. Phys.*, 2006, **124**, 164902.
57. M. W. van der Kamp and A. J. Mulholland, *Biochemistry*, 2013, **52**, 2708–2728.
58. H. M. Senn and W. Thiel, *Curr. Opin. Chem. Biol.*, 2007, **11**, 182–187.
59. W. L. Jorgensen, *Acc. Chem. Res.*, 2009, **42**, 724–733.
60. S. F. Altschul, W. Gish, W. Miller, E. Myers and D. J. Lipman, *J. Mol. Biol.*, 1990, **215**, 403–410.
61. D. J. Lipman and W. R. Pearson, *Science*, 1985, **227**, 1435–1441.
62. S. Henikoff and J. G. Henikoff, *Proc. Natl. Acad. Sci. U. S. A.*, 1992, **89**, 10915–10919.
63. S. F. Altschul, *et al.*, *Nucleic Acids Res.*, 1997, **25**, 3389–3402.
64. A. A. Schaffer, *et al.*, *Nucleic Acids Res.*, 2001, **29**, 2994–3005.
65. J. D. Thompson, D. G. Higgins and T. J. Gibson, *Nucleic Acids Res.*, 1994, **22**, 4673–4680.
66. J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson and Y. Zhang, *Nat. Methods*, 2015, **12**, 7–8.
67. S. Wu and Y. Zhang, *Nucleic Acids Res.*, 2007, **35**, 3375–3382.
68. A. Fiser and A. Sali, *Methods Enzymol.*, 2003, **374**, 461–491.
69. Montreal, Quebec, MOE 2015:10 edn, 2016.
70. L. A. Kelley, *et al.*, *Nat. Protoc.*, 2015, **10**, 845–858.
71. M. Källberg, H. Wang, S. Wang, J. Peng, Z. Wang, H. Lu and J. Xu, *Nat. Protoc.*, 2012, **7**, 1511–1522.
72. A. Leaver-Fay, *et al.*, *Methods Enzymol.*, 2011, **487**, 545–574.
73. M. Biasini, *et al.*, *Nucleic Acids Res.*, 2014, **42**(W1), W252–W258.
74. E. Krieger and G. Vriend, *Bioinformatics*, 2014, **30**, 2981–2982.
75. R. D. Finn, *et al.*, *Nucleic Acids Res.*, 2016, **D44**, D279–D285.
76. H. Fang and J. Gough, *Nucleic Acids Res.*, 2013, **D41**, D536–D544.
77. C. J. A. Sigrist, *et al.*, *Nucleic Acids Res.*, 2013, **D41**, D344–D347.
78. Y. Tang, W. Zhu, K. Chen and H. Jiang, *Drug Discovery Today Technol.*, 2006, **3**, 307–313.

79. S. C. Basak, *Curr. Comput.-Aided Drug Des.*, 2012, **8**, 1–2.
80. Forbes, *Crisis In Pharma R&D: It Costs \$2.6 Billion To Develop A New Medicine; 2.5 Times More Than in 2003*, <http://www.forbes.com/sites/theapothecary/2014/11/26/crisis-in-pharma-rd-it-costs-2-6-billion-to-develop-a-new-medicine-2-5-times-more-than-in-2003/#239cbbce1641>.
81. C. M. Song, S. J. Lim and J. C. Tong, *Briefings Bioinf.*, 2009, **10**, 579–591.
82. I. M. Kapetanovic, *Chem.-Biol. Interact.*, 2008, **171**, 165–176.
83. B. C. Duffy, L. Zhu, H. Decornez and D. B. Kitchen, *Bioorg. Med. Chem.*, 2012, **20**, 5324–5342.
84. J. H. Van Drie, *J. Comput.-Aided Mol. Des.*, 2007, **21**, 591–601.
85. T. N. Doman, S. L. McGovern, B. J. Witherbee, T. P. Kasten, R. Kurumbail, W. C. Stallings, D. T. Connolly and B. K. Shoichet, *J. Med. Chem.*, 2002, **45**, 2213–2221.
86. D. M. Kruger and A. Evers, *ChemMedChem*, 2010, **5**, 148–158.
87. Y. Westermaier, X. Barril and L. Scapozza, *Methods*, 2015, **71**, 44–57.
88. S. Kalyaanamoorthy and Y. P. Chen, *Drug Discovery Today*, 2011, **16**, 831–839.
89. I. Halperin, B. Ma, H. Wolfson and R. Nussinov, *Proteins*, 2002, **47**, 409–443.
90. X.-Y. Meng, H.-X. Zhang, M. Mezei and M. Cui, *Curr. Comput.-Aided Drug Des.*, 2011, **7**, 146–157.
91. J. P. Changeux and S. Edelstein, *F1000 Biol. Rep.*, 2011, **3**, 19.
92. D. B. Kitchen, H. Decornez, J. R. Furr and J. Bajorath, *Nat. Rev. Drug Discovery*, 2004, **3**, 935–949.
93. L. G. Ferreira, R. N. Dos Santos, G. Oliva and A. D. Andricopulo, *Molecules*, 2015, **20**, 13384–13421.
94. G. Sliwoski, S. Kothiwale, J. Meiler and E. W. Lowe, Jr., *Pharmacol. Rev.*, 2014, **66**, 334–395.
95. Z. Zsoldos, D. Reid, A. Simon, S. B. Sadjad and A. P. Johnson, *J. Mol. Graphics Modell.*, 2007, **26**, 198–212.
96. T. J. A. Ewing, S. Makino, A. G. Skillman and I. D. Kuntz, *J. Comput.-Aided Mol. Des.*, 2001, **15**, 411–428.
97. M. Rarey, B. Kramer, T. Lengauer and G. Klebe, *J. Mol. Biol.*, 1996, **261**, 470–489.
98. 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.
99. D. S. Goodsell and A. J. Olson, *Proteins*, 1990, **8**, 195–202.
100. E. M. Krovat, T. Steindl and T. Langer, *Curr. Comput.-Aided Drug Des.*, 2005, **1**, 93–102.
101. 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.
102. M. L. Vendonk, J. C. Cole, M. J. Hartshorn, C. W. Murray and R. D. Taylor, *Proteins*, 2003, **52**, 609–623.
103. F. Glover, *Comput. Oper. Res.*, 1986, **13**, 533–549.

104. C. A. Baxter, C. W. Murray, D. E. Clark, D. R. Westhead and M. D. Eldridge, *Proteins*, 1998, **33**, 367–382.
105. D. R. Westhead, D. E. Clark and C. W. Murray, *J. Comput.-Aided Mol. Des.*, 1997, **11**, 209–228.
106. A. R. Leach, B. K. Shoichet and C. E. Peishof, *J. Med. Chem.*, 2006, **49**, 5851–5855.
107. P. L. Kastritis and A. M. J. J. Bonvin, *J. Proteome Res.*, 2010, **9**, 2216–2225.
108. R. T. Kroemer, *Curr. Protein Pept. Sci.*, 2007, **8**, 312–328.
109. S.-Y. Huang, S. Z. Grinter and X. Zou, *Phys. Chem. Chem. Phys.*, 2010, **12**, 12899–12908.
110. P. Englebienne and N. Moitessier, *J. Chem. Inf. Model.*, 2009, **49**, 2564–2571.
111. J. Michel, M. L. Verdonk and J. W. Essex, *J. Med. Chem.*, 2006, **49**, 7424–7439.
112. I. D. Wall, A. R. Leach, D. W. Salt, M. G. Ford and J. W. Essex, *J. Med. Chem.*, 1999, **42**, 5142–5152.
113. V. Tsui and D. Case, *Biopolymers*, 2000, **56**, 275–291.
114. P. Kollman, *Chem. Rev.*, 1993, **93**, 2395–2417.
115. J. M. Briggs, T. J. Marrone and J. A. McCammon, *Trends Cardiovasc. Med.*, 1996, **6**, 198–206.
116. H. A. Carlson and W. L. Jorgensen, *J. Phys. Chem.*, 1995, **99**, 10667–10673.
117. J. Aqvist, V. B. Luzhkov and B. O. Brandsdal, *Acc. Chem. Res.*, 2002, **35**, 358–365.
118. H. J. Böhm, *J. Comput.-Aided Mol. Des.*, 1992, **6**, 593–606.
119. H. J. Böhm, *J. Comput.-Aided Mol. Des.*, 1998, **12**, 309–323.
120. M. D. Eldridge, C. W. Murray, T. R. Auton, G. V. Paolini and R. P. Mee, *J. Comput.-Aided Mol. Des.*, 1997, **11**, 425–445.
121. G. M. Verkhivker, D. Bouzida, D. K. Gehlhaar, P. A. Rejto, S. Arthurs, A. B. Colson, S. T. Freer, V. Larson, B. A. Luty, T. Marrone and P. W. Rose, *J. Comput.-Aided Mol. Des.*, 2000, **14**, 731–751.
122. A. N. Jain, *J. Comput.-Aided Mol. Des.*, 1996, **10**, 427–440.
123. R. D. Head, M. L. Smythe, T. I. Oprea, C. L. Waller, S. M. Green and G. R. Marshall, *J. Am. Chem. Soc.*, 1996, **118**, 3959–3969.
124. C. W. Murray, T. R. Auton and M. D. Eldridge, *J. Comput.-Aided Mol. Des.*, 1998, **12**, 503–519.
125. G. Verkhivker, K. Appelt, S. T. Freer and J. E. Villafranca, *Protein Eng.*, 1995, **8**, 677–691.
126. I. Muegge and Y. C. Martin, *J. Med. Chem.*, 1999, **72**, 791–804.
127. H. Gohlke, M. Hendlich and G. Klebe, *J. Mol. Biol.*, 2000, **295**, 337–356.
128. S.-Y. Huang and X. Zou, *J. Comput. Chem.*, 2006, **27**, 1866–1875.
129. J. Shimada, A. V. Ishchenko and E. I. Shakhnovich, *Protein Sci.*, 2000, **9**, 765–775.
130. I. Muegge, *Perspect. Drug Discovery Des.*, 2000, **20**, 99–114.
131. I. Muegge, *J. Comput. Chem.*, 2001, **22**, 418–425.

132. H. F. Velec, H. Gohlke and G. Klebe, *J. Med. Chem.*, 2005, **48**, 6296–6303.
133. P. S. Charifson, J. J. Corkery, M. A. Murcko and W. P. Walters, *J. Med. Chem.*, 1999, **42**, 5100–5109.
134. M. Feher, *Drug Discovery Today*, 2006, **11**, 421–428.
135. N. M. O'Boyle, J. W. Liebeschuetz and J. C. Cole, *J. Chem. Inf. Model.*, 2009, **49**, 1871–1878.
136. R. D. Clark, A. Strizhev, J. M. Leonard, J. F. Blake and J. B. Matthew, *J. Mol. Graphics Modell.*, 2002, **20**, 281–295.
137. E. Lionta, G. Spyrou, D. K. Vassilatis and Z. Cournia, *Curr. Top. Med. Chem.*, 2014, **14**, 1923–1938.
138. R. P. Gangwal, M. V. Damre, N. R. Das, G. V. Dhoke, A. Bhadauriya, R. A. Varikoti, S. S. Sharma and A. T. Sangamwar, *J. Mol. Graphics Modell.*, 2015, **57**, 89–98.
139. A. N. Jain and A. Nicholls, *J. Comput.-Aided Mol. Des.*, 2008, **22**, 133–139.
140. S. L. McGovern and B. K. Shoichet, *J. Med. Chem.*, 2003, **46**, 2895–2907.
141. A. J. M. Barbosa and A. Del Rio, *Curr. Top. Med. Chem.*, 2012, **12**, 866–877.
142. S. O. Jonsdottir, F. S. Jorgensen and S. Brunak, *Bioinformatics*, 2005, **21**, 2145–2160.
143. ZINC, <http://zinc.docking.org/>.
144. J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad and R. G. Coleman, *J. Chem. Inf. Model.*, 2012, **52**, 1757–1768.
145. PubChem, <https://pubchem.ncbi.nlm.nih.gov/>.
146. Q. Li, T. Cheng, Y. Wang and S. H. Bryant, *Drug Discovery Today*, 2010, **15**, 1052–1057.
147. DrugBank, <http://www.drugbank.ca/>.
148. D. S. Wishart, *Nucleic Acids Res.*, 2006, **34**, D668–D672.
149. ChemSpider, <http://www.chemspider.com/>.
150. H. E. Pence and A. Williams, *J. Chem. Educ.*, 2010, **87**, 1123–1124.
151. ChemBank, <http://chembank.broadinstitute.org/>.
152. K. P. Seiler, G. A. George, M. P. Happ, N. E. Bodycombe, H. A. Carrinski, S. Norton, S. Brudz, J. P. Sullivan, J. Muhlich, M. Serrano, P. Ferraiolo, N. J. Tolliday, S. L. Schreiber and P. A. Clemons, *Nucleic Acids Res.*, 2007, **36**, D351–D359.
153. eMolecules, <https://www.emolecules.com/>.
154. ChEMBL, <https://www.ebi.ac.uk/chembl/>.
155. A. P. Bento, A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Krüger, Y. Light, L. Mak, S. McGlinchey, M. Nowotka, G. Papadatos, R. Santos and J. P. Overington, *Nucleic Acids Res.*, 2014, **42**, D1083–D1090.
156. ChemDB, <http://chemdb.ics.uci.edu/>.
157. J. H. Chen, E. Linstead, S. J. Swamidass, D. Wang and P. Baldi, *Bioinformatics*, 2007, **23**, 2348–2351.
158. BindingDB, <https://www.bindingdb.org/bind/index.jsp>.
159. T. Liu, Y. Lin, X. Wen, R. N. Jorissen and M. K. Gilson, *Nucleic Acids Res.*, 2007, **35**, D198–D201.



160. C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Adv. Drug Delivery Rev.*, 1997, **23**, 3–25.
161. C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Adv. Drug Delivery Rev.*, 2001, **46**, 3–26.
162. C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Adv. Drug Delivery Rev.*, 2012, **64**, 4–17.
163. C. A. Lipinski, *Adv. Drug Delivery Rev.*, 2016, 33–41.
164. M. K. A. Awadallah, T. Alshammari, L. A. Eriksson and P. Saenz-Mendez, *Molecules*, 2016, **21**, 351.
165. S. J. Teague, *Nat. Rev. Drug Discovery*, 2003, **2**, 527–541.
166. C. B-Rao, J. Subramanian and S. D. Sharma, *Drug Discovery Today*, 2009, **14**, 394–400.
167. F. Jiang and S.-H. Kim, *J. Mol. Biol.*, 1991, **219**, 79–102.
168. A. R. Leach, *J. Mol. Biol.*, 1994, **235**, 345–356.
169. R. M. A. Knegtel, I. D. Kuntz and C. M. Oshiro, *J. Mol. Biol.*, 1997, **266**, 424–440.
170. C. N. Cavasotto and R. A. Abagyan, *J. Mol. Biol.*, 2004, **337**, 209–225.
171. T. Scior, A. Bender, G. Tresadern, J. L. Medina-Franco, K. Martínez-Mayorga, T. Langer, K. Cuanalo-Contreras and D. K. Agrafiotis, *J. Chem. Inf. Model.*, 2012, **52**, 867–881.
172. J. Kirchmair, P. Markt, S. Distinto, G. Wolber and T. Langer, *J. Comput.-Aided Mol. Des.*, 2008, **22**, 213–228.
173. O. M. Becker, D. S. Dhanoa, Y. Marantz, D. Chen, S. Shachman, S. Cheruku, A. Heifetz, P. Mohanty, M. Fichman, A. Sharadendu, R. Nudelman, M. Kauffman and S. Noiman, *J. Med. Chem.*, 2006, **49**, 3116–3135.
174. F. M. Ruiz, R. Gil-Redondo, A. Morreale, A. R. Ortiz, C. Fábrega and J. Bravo, *J. Chem. Inf. Model.*, 2008, **48**, 844–854.
175. N. Li, F. Wang, S. Niu, J. Cao, K. Wu, Y. Li, N. Yin, X. Zhang, W. Zhu and Y. Yin, *BMC Microbiol.*, 2009, **9**, 129.
176. K. J. Simmons, I. Chopra and C. W. G. Fishwick, *Nat. Rev. Microbiol.*, 2010, **8**, 501–510.
177. I. Pauli, R. N. dos Santos, D. C. Rostirolla, L. K. Martinelli, R. G. Ducati, L. F. S. M. Timmers, L. A. Basso, D. S. Santos, R. V. C. Guido, A. D. Andricopulo and O. Norberto de Souza, *J. Chem. Inf. Model.*, 2013, **53**, 2390–2401.
178. Z. Miller, K.-S. Kim, D.-M. Lee, V. Kasam, S. E. Baek, K. H. Lee, Y.-Y. Zhang, L. Ao, K. Carmony, N.-R. Lee, S. Zhou, Q. Zhao, Y. Jang, H.-Y. Jeong, C.-G. Zhan, W. Lee, D.-E. Kim and K. B. Kim, *J. Med. Chem.*, 2015, **58**, 2036–2041.
179. K. Matsuno, Y. Masuda, Y. Uehara, H. Sato, A. Muroya, O. Takahashi, T. Yokotagawa, T. Furuya, T. Okawara, M. Otsuka, N. Ogo, T. Ashizawa, C. Oshita, S. Tai, H. Ishii, Y. Akiyama and A. Asai, *ACS Med. Chem. Lett.*, 2010, **1**, 371–375.
180. L. Wang, Q. Gu, X. Zheng, J. Ye, Z. Liu, J. Li, X. Hu, A. Hagler and J. Xu, *J. Chem. Inf. Model.*, 2013, **53**, 2409–2422.

181. S. Dadashpour, T. Tuylu Kucukkilinc, O. Unsal Tan, K. Ozadali, H. Irannejad and S. Emami, *Arch. Pharm.*, 2015, **348**, 179–187.
182. H. Geppert, M. Vogt and J. Bajorath, *J. Chem. Inf. Model.*, 2010, **50**, 205–216.
183. P. Ripphausen, B. Nisius and J. Bajorath, *Drug Discovery Today*, 2011, **16**, 372–376.
184. H. Eckert and J. Bajorath, *Drug Discovery Today*, 2007, **12**, 225–233.
185. P. Willet, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 983–996.
186. A. Bender and R. C. Glen, *Org. Biomol. Chem.*, 2004, **2**, 3204.
187. A. R. Katritzky and E. V. Gordeeva, *J. Chem. Inf. Comput. Sci.*, 1993, **33**, 835–857.
188. S. Zhang, *Drug Des. Discovery*, 2011, **716**, 23–38.
189. IUPAC, 1998, <http://www.chem.qmul.ac.uk/iupac/medchem/ix.html#p7>.
190. S.-Y. Yang, *Drug Discovery Today*, 2010, **15**, 444–450.
191. J. Zou, H.-Z. Xie, S.-Y. Yang, J.-J. Chen, J.-X. Ren and Y.-Q. Wei, *J. Mol. Graphics Modell.*, 2008, **27**, 430–438.
192. G. M. Spitzer, M. Heiss, M. Mangold, P. Markt, J. Kirchmair, G. Wolber and K. R. Liedl, *J. Chem. Inf. Model.*, 2010, **50**, 1241–1247.
193. G. Wolber, *Drug Discovery Today*, 2008, **13**, 23–29.
194. C. Acharya, A. Coop, J. E. Polli and A. D. MacKerell, *Curr. Comput.-Aided Drug Des.*, 2011, **7**, 10–22.
195. Catalyst, <http://accelrys.com/products/collaborative-science/biovia-discovery-studio/pharmacophore-and-ligand-based-design.html>.
196. Dassault-Systèmes, *BIOVIA, Discovery Studio Modeling Environment*, San Diego, 2016.
197. Phase, 4.4, *Schrödinger LLC*, New York, 2016.
198. S. L. Dixon, A. M. Smondyrev, E. H. Knoll, S. N. Rao, D. E. Shaw and R. A. Friesner, *J. Comput.-Aided Mol. Des.*, 2006, **20**, 647–671.
199. Maestro, 10.3, *Schrödinger LLC*, New York, 2016.
200. P. Labute, C. Williams, M. Feher, E. Sourial and J. M. Schmidt, *J. Med. Chem.*, 2001, **44**, 1483–1490.
201. I. Poggesi, P. S. Burton, J. T. Goodwin and M. Germani, in *Optimizing the “Drug-Like” Properties of Leads in Drug Discovery*, ed. R. T. Borchardt, E. H. Kerns, M. J. Hageman, D. R. Thakker and J. L. Stevens, Springer, New York, 2006, pp. 195–220.
202. C. Manly, J. Chandrasekhar, J. Ochterski, J. Hammer and B. Warfield, *Drug Discovery Today*, 2008, **13**, 99–109.
203. K. H. Bleicher, H.-J. Böhm, K. Müller and A. I. Alanine, *Nat. Rev. Drug Discovery*, 2003, **2**, 369–378.
204. M. S. Lajiness, M. Vieth and J. Erickson, *Curr. Opin. Drug Discovery Dev.*, 2004, **7**, 470–477.
205. R. O’Shea and H. E. Moser, *J. Med. Chem.*, 2008, **51**, 2871–2878.
206. R. G. Bickerton, G. V. Paolini, J. Besnard, S. Muresan and A. L. Hopkins, *Nat. Chem.*, 2012, **4**, 90–98.
207. C. J. Omicinski, R. P. Remmel and V. P. Hosagrahara, *Toxicol. Sci.*, 1999, **48**, 151–156.



208. H. van de Waterbeemd and E. Gifford, *Nat. Rev. Drug Discovery*, 2003, **2**, 192–204.
209. DockBlaster, <http://blaster.docking.org/>.
210. J. J. Irwin, B. K. Shoichet, M. M. Mysinger, N. Huang, F. Colizzi, P. Wassam and Y. Cao, *J. Med. Chem.*, 2009, **52**, 5712–5720.
211. SwissDock, <http://www.swissdock.ch/>.
212. A. Grosdidier, V. Zoete and O. Michielin, *Nucleic Acids Res.*, 2011, **39**, W270–W277.
213. DockThor, <http://dockthor.lncc.br/index.php?pg=home>.
214. C. S. de Magalhães, D. M. Almeida, H. J. C. Barbosa and L. E. Dardenne, *Inf. Sci.*, 2014, **289**, 206–224.
215. PharmMapper, <http://lilab.ecust.edu.cn/pharmmapper/>.
216. X. Liu, S. Ouyang, B. Yu, Y. Liu, K. Huang, J. Gong, S. Zheng, Z. Li, H. Li and H. Jiang, *Nucleic Acids Res.*, 2010, **38**, W609–W614.

## CHAPTER 2

# *Molecular Dynamics Computer Simulations of Biological Systems*

JAMES W. CARTER,<sup>†</sup> ANNA SOFIA TASCINI,<sup>†</sup> JOHN M. SEDDON  
AND FERNANDO BRESME\*

Imperial College London, Department of Chemistry, SW7 2AZ London, UK

\*Email: f.bresme@imperial.ac.uk

## 2.1 Introduction

The pioneering work of Alder and Wainwright on computer simulations of hard spheres laid the foundations of the molecular dynamics (MD) simulation method. This method has evolved significantly in the last fifty years, reaching a high level of sophistication that now permits the investigation of complex biological molecules and processes. Molecular dynamics is now a discipline in its own right, that both complements experimental measurement and acts as a unique predictive tool, providing much needed insight at molecular scales, where experiments are not always possible or are difficult. The development of multiscale models, which was recognised with the award of the 2013 Nobel Prize in Chemistry, has contributed to the increasing use of MD simulations in biology.

The aim of this chapter is to introduce the reader to classical MD simulations and the theory and practicalities associated with them. The chapter

---

<sup>†</sup>These authors contributed equally to this work.

also demonstrates the breadth of available, well established MD techniques that can be applied to biomolecular systems, with an explanation of how each method works and why it is useful.

The first part of the chapter focuses on force fields, with emphasis on those used in biomolecular simulations. We further discuss multiscale approaches and more advanced topics such as polarisation and implicit solvent models. The focus then shifts to the calculation of properties of biomolecules and the use of rare-event and enhanced sampling methods, such as replica exchange, which are often required in biomolecular simulations.

MD is particularly helpful in determining the mechanical and bending properties of lipid membranes. We describe the relevant simulation approaches and discuss selected case studies that reveal some of the issues associated with computations of the area per lipid and compressibility of biological membranes. This section also demonstrates how MD can interface with experimental and theoretical methods to form the cornerstone of a multidisciplinary approach to current problems in biophysics.

Finally, the importance of MD to analyse experimental data is highlighted, in particular NMR and scattering techniques where MD forms a vital part of the data analysis procedure. The chapter concludes with a discussion of current cutting-edge research in biomolecular MD and the likely future direction of this fast-moving field. The concepts and techniques introduced in this chapter are illustrated throughout with examples taken from a wide range of biomolecular simulations from the literature, showing how these ideas have been implemented to address current research questions in biophysics.

## 2.2 The Basics of Molecular Dynamics

Molecular dynamics is a computational technique in which the behaviour of atoms and molecules is described explicitly either using classical or, more recently, quantum mechanics. We will focus in this chapter on the classical implementation. Newton's equations of motion for a system consisting of  $N$  particles are given by:

$$m_i \frac{\partial^2 \mathbf{r}_i}{\partial t^2} = \mathbf{F}_i(\mathbf{r}^N), \quad (2.1)$$

where  $t$  is the time,  $\mathbf{r}_i$ ,  $m_i$  and  $\mathbf{F}_i$  are the position, mass and force on particle  $i$ , respectively, and  $\mathbf{r}^N$  represents the positions of all  $N$  particles.

The evolution of the system over time can be obtained by numerically integrating eqn (2.1) in discrete time steps ( $\Delta t$ ). Many algorithms have been developed for this purpose, which build on the basic Verlet algorithm.<sup>1</sup> In addition, periodic boundary conditions are used to eliminate the edge effects resulting from the finite size of the system ( $\sim 10^3$ – $10^6$  particles).<sup>2</sup> It follows from eqn (2.1) and the Verlet algorithm

that generating an MD trajectory requires the initial positions, velocities and forces of all particles.

Newton's equations of motion conserve the total energy ( $E$ ), and the resulting configurations are part of the microcanonical (constant  $NVE$ ) ensemble. In many problems of interest in biology it is more convenient to work at constant temperature instead. MD thermostats exploit the relationship between temperature and average kinetic energy to control temperature by scaling particle velocities.<sup>3-7</sup> While generally all of these algorithms predict structural and thermodynamic properties within the error of the computations, not all of them reproduce the correct canonical ( $NVT$ ) ensemble. This point must be noted if fluctuation properties are required, as these algorithms cannot be used to estimate thermodynamic properties, *e.g.*, compressibility, from the fluctuation route. A range of algorithms have also been developed to hold pressure constant and allow the volume to fluctuate.<sup>3,8,9</sup> These algorithms are widely used to simulate the tensionless state of biological membranes.

### 2.2.1 Force Fields for Biomolecular Simulations

A force field is a mathematical function of the positions of all particles that defines the potential energy of a configuration ( $U(\mathbf{r}^N)$ ). Once the force field is defined, the instantaneous forces acting on each particle can be computed from the derivative of the potential:  $\mathbf{F}_i = -\nabla_i U(\mathbf{r}^N) = -\partial U(\mathbf{r}^N)/\partial \mathbf{r}_i$ .

In practice, the majority of force fields share the same basic functional form (eqn (2.2)). Each term in the force field function represents a different type of interaction. Intramolecular interactions include 'bonds', 'angles' and 'dihedrals' to account for covalent bonds and molecular flexibility. Intermolecular interactions are included often using the Lennard-Jones (LJ) potential, which accounts for Pauli repulsion and van der Waals attraction. Further, electrostatic interactions between charges,  $Q$ , are handled using the Coulomb potential. Figure 2.1 shows an example of how these interactions operate.

$$\begin{aligned}
 U(\mathbf{r}^N) = & \sum_{ij \in \{\text{Bonds}\}} k_{ij}(r_{ij} - r_{ij,0})^2 + \sum_{ijk \in \{\text{Angles}\}} k_{ijk}(\theta_{ijk} - \theta_{ijk,0})^2 \\
 & + \sum_{ijkl \in \{\text{Dihedrals}\}} k_{ijkl}(1 + \cos(n_{ijkl}\varphi_{ijkl} - \varphi_{ijkl,0})) \\
 & + \sum_{ij \in \{\text{LJ}\}} 4\epsilon_{ij} \left( \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 \right) + \sum_{ij \in \{\text{Coulomb}\}} \frac{Q_i Q_j}{4\pi\epsilon_0 r_{ij}}.
 \end{aligned} \tag{2.2}$$

At relevant biological temperatures ( $\sim 310$  K), molecular vibrations are in the ground state and the harmonic bond term in eqn (2.2) is replaced by a rigid constraint,<sup>10</sup> which restores the bond length to its equilibrium value each time step, removing fast bond vibrational motion from the simulation. It should

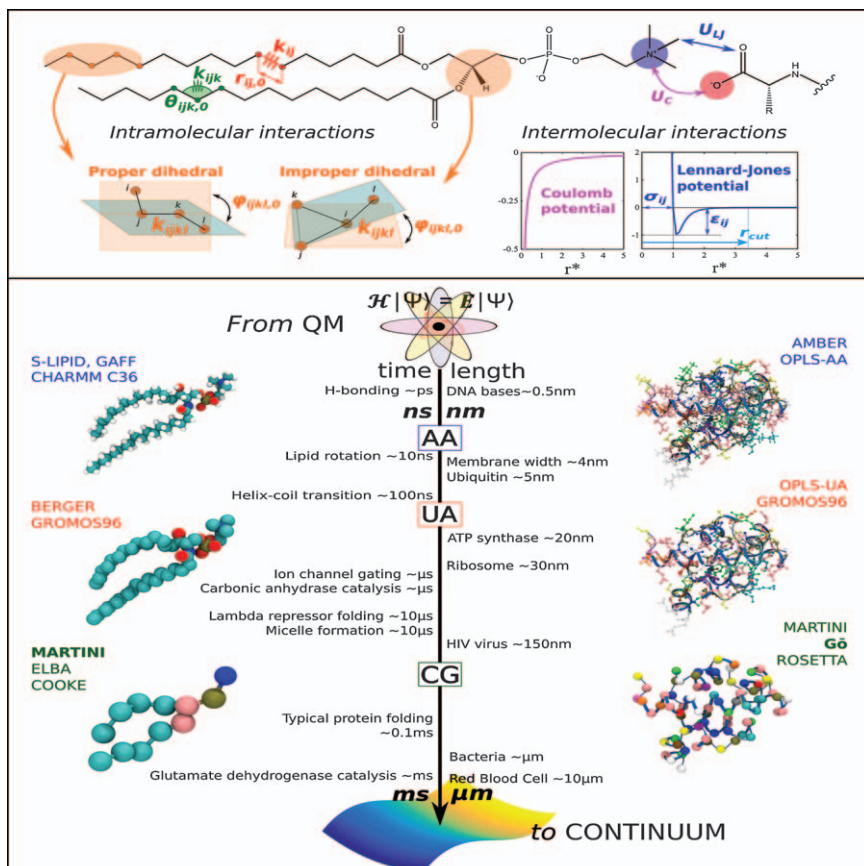


**Figure 2.1** (Top) Illustration of the different contributions to the force field (eqn (2.2)) applied to a lipid (DPPC) and a general amino acid.  $k$  are force constants and  $r_{ij,0}$ ,  $\theta_{ijk,0}$ ,  $\varphi_{ijkl,0}$  are equilibrium bond lengths, angles and dihedral angles, respectively. Dihedral potentials (periodicity  $n_{ijkl}$ ) are applied to linear chains of atoms and improper dihedrals are used for tetrahedral groups. The shape of the Lennard-Jones and Coulomb potentials are also shown. (Bottom) Representative structures for a lipid (DPPC (from LipidBook)) and protein ( $\lambda$ -repressor (taken from PDB structure 3KZ3)) using different resolution models alongside commonly used force fields and the time and length scales typically accessible.

also be noted that there are a number of different functional forms commonly used for the dihedral potential, including a harmonic potential often used for improper dihedrals to give rigidity to rings and tetrahedral sites.

The Lennard-Jones function is a pairwise additive potential in which the effects of interatomic interactions between three or more particles are included in the ‘effective’ two-body parameters,  $\epsilon_{ij}$  and  $\sigma_{ij}$  (see eqn (2.2)). These parameters define the interaction strength and effective atomic diameter

respectively (see Figure 2.1) and are commonly given for single atom types ( $\sigma_i$ ,  $\epsilon_i$ ). Interactions between different types are then derived from combination rules, such as the Lorentz and Berthelot rules,<sup>11</sup> in which  $\sigma_{ij}$  and  $\epsilon_{ij}$  are obtained from arithmetic and geometric averages of the single atom values, respectively:  $\sigma_{ij} = (\sigma_i + \sigma_j)/2$ ,  $\epsilon_{ij} = \sqrt{\epsilon_i \epsilon_j}$ .

The calculation of non-bonded forces is the most computationally time-consuming part of an MD simulation. For short range interactions such as the LJ potential, the interaction approaches zero sufficiently fast that it can be neglected beyond a short cut off distance,  $r_{\text{cut}}$  (usually around  $2.5\text{--}5\sigma$ ). The potential can be modified to ensure the force is continuous at  $r_{\text{cut}}$ , in order to avoid artefacts associated with force discontinuities. Dispersion corrections to the energy and pressure to account for the cut off can also be included<sup>2</sup> and are of particular importance in constant pressure simulations if properties corresponding to the full potential are required, *e.g.* density or surface tension.

The electrostatic term in eqn (2.2) decays very slowly ( $\propto r^{-1}$ ) with distance, hence these interactions are often better treated using the Ewald summation method.<sup>2</sup> This method is considered the state of the art for computation of electrostatic interactions, however it imposes a periodicity in the system that must be considered carefully in computations of solvation energies.<sup>12</sup>

Atomistic force fields can be broadly divided into two types: all atom (AA) force fields, in which every atom in a molecule is represented explicitly as a distinct interaction site, and united atom (UA) force fields, in which non-polar, alkyl hydrogens are merged with the adjacent carbon atom interaction site reducing the number of constituent atoms in each molecule (see Figure 2.1). Some commonly used atomistic force fields include AMBER, CHARMM, GROMOS (UA only) and OPLS, with the first three developed alongside MD software packages of the same name. A large number of different parameter sets have been published, with some of these available within MD software packages or, increasingly, from online databases such as LipidBook<sup>13</sup> or the database of antimicrobial compounds.<sup>14</sup>

The parameterisation of force fields can be based on quantum calculations of small fragments of the target molecule or by fitting to experimental properties, both to structural data such as bond lengths from crystal structures, and thermodynamic data, as in the case of organic molecules fitted to reproduce liquid densities and vaporisation enthalpies. When choosing a force field, it is vital to understand the philosophy behind its parameterisation to appreciate the various limitations of the model. Parameters from different force fields should not be mixed blindly. When setting up a simulation it is also important to choose simulation parameters (*e.g.* potential cut off) bearing in mind those used in the original parameterisation.

A significant increase in attainable length and time scales can be achieved by reducing atomistic detail in favour of a coarse-grained (CG)

representation. By combining atoms within functional groups or natural molecular subunits, such as amino acids, into a smaller number of CG beads, molecules can be represented by fewer interaction sites while maintaining some chemical specificity. The reduction in the number of degrees of freedom in the CG model results in a smoother energy landscape and permits the use of a longer time step (from  $\sim 1$  fs in AA and UA simulations to  $\sim 10$  fs in CG). CG models feature faster dynamics<sup>15</sup> making the study of larger systems and slower processes possible. Many CG force fields are obtained from atomistic simulation data using systematic coarse-graining approaches such as Boltzmann inversion.<sup>16</sup> A rigorous theoretical framework now exists for coarse-graining, culminating in the MS-CG (multiscale coarse-graining) method in which a variational principle is used to obtain a CG force field from an atomistic simulation.<sup>17</sup>

One of the most widely used CG models is the MARTINI model, originally developed for lipids,<sup>18</sup> but now extended to proteins, nucleic acids and carbohydrates. It has been employed to model a wide range of biological phenomena of ever increasing scale and complexity, from single lipid phase behaviour,<sup>19</sup> to huge multicomponent mixtures,<sup>20</sup> mitochondrial processes<sup>21</sup> and entire viral capsids.<sup>22</sup> The model is based on a 4 : 1 mapping of heavy (non-hydrogen) atoms to CG beads (see Figure 2.1) and is parameterised to reproduce experimental thermodynamic data. The ELBA CG model uses a similar level of resolution, but provides a more detailed description of electrostatic interactions and has experimented with non-spherically symmetric potentials.<sup>23</sup> At a lower resolution, the Cooke model<sup>24</sup> is a three-bead lipid model used primarily to study bending properties.

An even larger diversity of CG models exists for proteins. Structure-based models such as Gō models are constructed specifically to reproduce the native structure of proteins, neglecting all non-native interactions.<sup>25</sup> Despite their simplicity these models can be used to study protein folding<sup>26</sup> and the impact of mutations.<sup>27</sup> More complex, unbiased models which attempt to predict protein structure have also been developed and are regularly tested during CASP experiments (see ref. 28 for review). In general, there are two different approaches to deriving CG protein models. In ‘physics-based’ models, fundamental principles of chemical physics are used to build the model, often using a potential function similar to eqn (2.2). Alternatively, the development of ‘knowledge-based’ models is driven by experimentally derived statistical data that is used to highlight common protein characteristics.

### 2.2.2 Multiscale Modelling

Many biomolecular processes occur over a wide range of length and/or timescales. Examples include membrane-protein interactions and the bending or phase behaviour of large multicomponent membranes. This

multiscale behaviour cannot be captured by a single atomistic or CG simulation and hence a number of methods to convert between simulations at different resolutions have been developed.

‘Backmapping’ involves converting a configuration from one level of detail to another, usually from CG to UA/AA. It is especially useful for generating good initial configurations for atomistic simulations, for example it has been found that backmapping can predict, in some cases, the orientation of membrane proteins better than purely AA approaches.<sup>29</sup> Mapping between different resolution models has also been combined with replica exchange (see Section 2.3.2) to improve conformational sampling.<sup>30</sup> Simulations at different resolutions are run in parallel and a mapping procedure is used to swap configurations across resolution levels according to a Monte Carlo (MC) criterion.

UA/AA and CG representations can also be combined into a single ‘hybrid’ simulation, where the CG representation is often used to model the solvent. In adaptive resolution methods, conversion between higher and lower resolution models can be done dynamically during the simulation based on the position of the particle relative to the solute,<sup>31</sup> thus ensuring that the solute is never in direct contact with the CG solvent.

Finally, MD simulations have also been combined with non-MD continuum simulations. Microscopic properties such as transport coefficients calculated *via* MD can be transferred to the continuum simulation and macroscopic properties like density can be transferred back to the atomistic representation, hence creating a feedback loop between length scales.<sup>32</sup> This approach offers the possibility of truly multiscale simulations spanning microscopic to macroscopic scales.

### 2.2.3 Advanced Force Fields

Polarisability, *i.e.* the change in the distribution of charge within a molecule in response to the surrounding environment, can be significant in aqueous solvent and at biomolecular interfaces. Incorporating this effect into MD simulations should represent a significant improvement in accuracy. There are a variety of classical approaches to include polarisation in MD simulations, although polarisable force fields have not currently been parameterised as extensively as non-polarisable ones.

One approach is to treat the partial charges as dynamic variables, with an associated fictitious mass. In the fluctuating charge (FQ) or charge equilibrium (CHEQ) model, charges are allowed to fluctuate during each MD step in response to changes in the surrounding electrostatic field, to maintain equalisation of electronegativity within each molecule.<sup>33</sup> The CHEQ model has been implemented in the CHARMM force field and applied to lipid and protein systems, where it revealed a dependence of amino acid charge on position within a protein sequence.<sup>34</sup>

The Drude oscillator or shell model links each atom to an additional ‘Drude’ particle *via* a harmonic potential. Splitting the atomic charge over



these two sites forms a dipole that can change in magnitude and direction, and hence can respond to variation in the surrounding electrostatic field. Variants of this model have been implemented in both GROMACS and CHARMM and have even been used in CG models.<sup>35</sup> The Drude approach has been successfully used to simulate lipids, proteins and nucleic acids.<sup>36</sup>

An alternative approach to polarisation is based on the explicit inclusion of induced dipoles in the potential energy function. The electrostatic term in the force field (eqn (2.2)) is expanded by including dipole–dipole and dipole–point charge interactions. Since the value of each dipole depends on the local electrostatic field, which is itself dependent on the values of the other dipoles, an iterative procedure is used to obtain a self-consistent solution for the dipole values. This method has been implemented in AMBER and in the AMOEBA force field,<sup>37</sup> which includes a more thorough treatment of electrostatic interactions involving a multipole expansion up to quadrupoles.

Simulating the aqueous environment that surrounds biomolecules involves calculating water–water interactions far away from the solute, which have negligible impact on the behaviour of the biomolecule. This has motivated the development of methods to treat solvation effects implicitly. Removing explicit water molecules reduces the number of particles, increasing simulation speed, and leads to faster dynamics, due to the elimination of frictional forces involving the solvent. Implicit solvent models aim to account for the solvation free energy within the solute potential by considering three separate contributions: the formation of a cavity within the solvent and resulting disruption of solvent hydrogen bonding, solute–solvent dispersion interactions, and charge effects.<sup>38</sup> The first two components are often assumed to be proportional to the solute solvent accessible surface area, however, improvements on this basic assumption have been suggested including a separate treatment of solvent–solute van der Waals interactions to introduce a dependence on solute atom type.<sup>39</sup> Charge effects are described by the Poisson–Boltzmann equation, with a relative dielectric constant used for water. However, solving this equation at every MD time step is very computationally intensive and so it is usually simplified and solved within the assumptions of the Generalised Born model. An interesting extension of implicit solvent models is to modelling membrane proteins, where the membrane surrounding the protein is also treated implicitly.<sup>40</sup>

## 2.3 Extracting the Information from MD

MD trajectories contain the positions, velocities and forces of each particle at each time step in the simulation, providing structural and dynamic information. The properties of interest, *e.g.*, thermodynamic properties ( $X$ ) are often given by ensemble averages ( $\langle X \rangle$ ) and can be calculated by taking

advantage of the ergodic hypothesis, which states that ensemble averages are equal to time averages:

$$\langle X \rangle = \iint X(\mathbf{r}^N, \mathbf{p}^N) P(\mathbf{r}^N, \mathbf{p}^N) d\mathbf{r}^N d\mathbf{p}^N = \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t X(\mathbf{r}^N, \mathbf{p}^N; t') dt', \quad (2.3)$$

where  $P(\mathbf{r}^N, \mathbf{p}^N) = e^{-E(\mathbf{r}^N, \mathbf{p}^N)/k_B T} / \int e^{-E(\mathbf{r}^N, \mathbf{p}^N)/k_B T} d\mathbf{r}^N d\mathbf{p}^N$  is the probability distribution defining the canonical ensemble. However, when applying this equation, care must be taken to assess whether the system is really ergodic over the simulation timescale.

### 2.3.1 Free Energy Difference Between Two States

Not all properties can be expressed as ensemble averages, most significantly the free energy ( $A$ ), which is given by:  $A = -k_B T \ln Q$  (where  $Q = \int e^{-E(\mathbf{r}^N, \mathbf{p}^N)/k_B T} d\mathbf{r}^N d\mathbf{p}^N$  is the canonical partition function).  $A$  is proportional to the volume of phase space (the set of all possible system configurations with positions,  $\mathbf{r}^N$ , and momenta,  $\mathbf{p}^N$ ) and so cannot be simply approximated by a time average using the ergodic principle. However, differences in free energy can be calculated as ensemble averages using thermodynamic integration where the derivative of the free energy is integrated between two different states:

$$\Delta A = A(\lambda_1) - A(\lambda_0) = \int_{\lambda_0}^{\lambda_1} \frac{\partial A}{\partial \lambda} d\lambda = \int_{\lambda_0}^{\lambda_1} \left\langle \frac{\partial U}{\partial \lambda} \right\rangle_{\lambda} d\lambda. \quad (2.4)$$

In simulations, unlike in experiments, the free energy derivatives can correspond to unmeasurable or non-physical quantities. For example, by introducing a parameter,  $\lambda$ , which switches between two different potential energy functions (*e.g.* setting  $U(\lambda) = \lambda U_1 + (1 - \lambda) U_0$ ), ‘alchemical transitions’ can be used to measure the strength of protein binding of different ligands.<sup>41</sup>

### 2.3.2 Enhanced Configurational Sampling

For many systems with rugged energy landscapes, like those in complex biomolecules, the simulation timescale may be too short to properly explore the available phase space. For example proteins may become trapped in a local minimum for the entire duration of a simulation, hence producing a non-ergodic trajectory. Advanced simulation techniques have been developed to enhance sampling and avoid the non-ergodic problem. These approaches are discussed below.

In simulated annealing, simulations are initiated at high temperature, where a more extensive exploration of phase space is possible, and then gradually cooled during the simulation. This causes energy barriers to become more and more significant relative to the thermal energy ( $\sim k_B T$ ) and

encourages the system to relax into a minimum free energy state (see Figure 2.2). The method has been widely used in structure determination<sup>42,43</sup> both by NMR and X-ray diffraction as it provides a means to generate a range of configurations representing distinct metastable low energy states to compare and fit to structural data.

Sometimes it is advantageous to allow drastic changes in configuration *via* large jumps between distant regions of phase space that can then be explored locally by MD. This can be achieved by introducing a MC step into the simulation. Possible MC moves include overall molecular rotations, translations and intermolecular rearrangements, but also unphysical

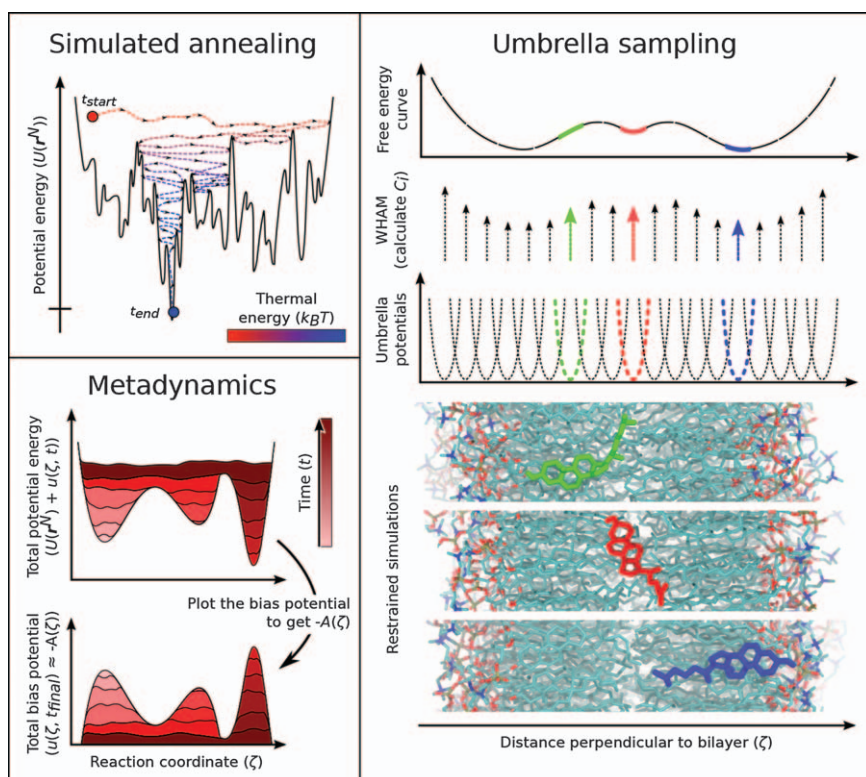


**Figure 2.2** (Top-left) Schematic figure showing the principle behind simulated annealing, the simulation trajectory is represented as a dashed line coloured according to temperature. (Right) Illustration of the umbrella sampling procedure used to determine the free energy barrier of cholesterol flip-flop and a sketch of the WHAM derived free energy curve. Three representative snapshots from restrained simulations using the umbrella potentials are also included. (Bottom-left) Metadynamics method, the top graph shows the total potential energy over time as the bias potential increases, filling up the potential energy landscape; the bottom plot shows the bias potential which is used to determine the free energy,  $-A(\zeta)$ .

moves such as insertion and deletion of particles and changes to particle identities. Some biomolecular systems are well suited to this approach, for example, the structural similarity between lipid molecules permits MC moves that exchange lipid identity allowing a more extensive probing of phase behaviour.<sup>44</sup>

One of the most successful ways of combining MC in MD simulations is replica exchange (REMD). Multiple simulations at different temperatures are run in parallel and an MC step is used to frequently exchange configurations between them. The higher temperatures are used to accelerate the crossing of local energy barriers. The success of this technique depends on the chosen temperature range, the simulation length and the number of replicas. A critical parameter is the average replica exchange probability between each replica,  $p_{(1 \rightarrow 2)}$ , which should be kept constant across replicas. This can be achieved by choosing an exponential temperature distribution.<sup>45</sup> REMD is widely used to study protein folding and kinetics.<sup>46</sup>

The large amounts of data produced by MD simulations using techniques such as REMD, can be analysed within a Markov State Model (MSM) framework.<sup>47</sup> A MSM is a kinetic model that describes the transitions between a number of discrete ‘microstates’ visited during a simulation, without being restricted to a specific reaction coordinate. By constructing a transition matrix, information on the characteristic dynamics of the system can be obtained from its eigenvalues and eigenvectors. This method has been used to identify metastable protein conformations with potential as drug targets.<sup>48</sup> In addition, it is also an indispensable part of the analysis of the huge amounts of data obtained from distributed computing projects such as Folding@home.<sup>49</sup>

Finally, the conformational dynamics of biomolecules, *e.g.*, peptides, can be accelerated by changing the relative masses of the solvent and solute to decouple their relative motion, exploiting the fact that classical configurational ensemble averages are independent of mass. This idea has been used to sample the conformational space of a nine-residue peptide in which protein side-chain and backbone motion was also decoupled.<sup>50</sup>

### 2.3.3 Simulating Rare Events

The free energy barriers associated with biological processes are often prohibitively large (10's of  $\text{kJ mol}^{-1}$ ) to be sampled during the timescale of a standard MD simulation and so require methods that go beyond the ones described above to promote the sampling of specific events. Many techniques rely on the definition of a reaction coordinate,  $\zeta(\mathbf{r}^N)$ , which should change smoothly and monotonically over the course of the event and should be chosen carefully to ensure it fully captures the process of interest, *e.g.* a conformational change. Examples of possible order parameters to study protein folding include root mean squared deviation from the native state structure or fraction of native contacts formed.

One of the simplest ways to preferentially sample states in a particular region of phase space is to add a restraining potential. In the umbrella sampling (US) method,<sup>51</sup> the transition between two states is divided into a number of ‘windows’, each characterised by a different value of  $\zeta(\mathbf{r}^N)$ . A separate simulation is run for each window with a, usually harmonic, restraining potential that biases the simulation to sampling states within that window. The effect of the bias potential,  $u_i(\zeta)$ , can then be removed to recover the unbiased free energy curve ( $a_i$ ) for window  $i$ , by using  $a_i(\zeta) = -k_B T \ln(P_{bias,i}(\zeta)) - u_i(\zeta) + C_i$ , where  $P_{bias,i}(\zeta)$  represents the biased probability of visiting states along  $\zeta$ , obtained from histograms computed from simulation  $i$ .  $C_i$  is a window dependent constant that shifts  $a_i$  so that it lines up with the free energy curves from other windows. The weighted histogram analysis method (WHAM)<sup>52</sup> provides an iterative approach to obtain self-consistent values for  $C_i$ , which minimises the error between windows. Figure 2.2 illustrates the application of this technique to studying cholesterol flip-flop.

Biasing potentials can also be used to push a simulation *away* from a particular region of phase space. In metadynamics<sup>53</sup> the biasing potential,  $u(\zeta, t)$ , is used to disfavour states already visited by the simulation and accumulates over time as the system explores phase space.  $u(\zeta, t)$  can in principle take different forms, but is commonly a Gaussian. As the simulation progresses, the bias potential forces the simulation into unexplored regions of phase space by reducing the depth of potential energy wells and the relative heights of barriers between minima (see Figure 2.2) until eventually the height is small enough ( $\sim k_B T$ ) that the simulation can surmount the barrier and reach adjacent minima. Over the course of the simulation, the biasing potential will grow, filling the energy surface until the landscape is flat and at this point the free energy can be related directly to the total bias potential:  $\lim_{t \rightarrow \infty} u(\zeta, t) = -A(\zeta)$ . Metadynamics has been combined with replica exchange to investigate the large conformational space inhabited by proteins in their denatured states.<sup>54</sup>

Steered MD relies on a relation from non-equilibrium statistical mechanics summarised by the Jarzynski identity:<sup>55</sup>  $e^{-\Delta A/k_B T} = \langle e^{-W/k_B T} \rangle$ , where  $W$  represents the work done in converting between two states. The right-hand side needs careful consideration since  $W$  is a function of the path taken between the two states and the ensemble average is over all paths which connect these states. This can make the technique inefficient, given the level of sampling required to achieve good statistics. Steered MD mimics experimental setups used in pulling experiments (*e.g.* optical tweezers or AFM)<sup>56</sup> and it has been applied to study protein–ligand unbinding, by physically pulling a bound inhibitor away from the protein.<sup>57</sup>

### 2.3.4 Computing Elastic Properties in Biomolecular Simulations

The elastic properties of membranes play an essential role in determining the function of transmembrane proteins, by influencing protein anchoring

and insertion,<sup>58</sup> and the fusion of membrane structures.<sup>59,60</sup> The fluctuations of biological membranes, which are defined by the elastic properties of the membrane, can be described theoretically using the Helfrich Hamiltonian (HH).<sup>61</sup> This model neglects the molecular nature of the membrane and defines the membrane free energy<sup>61,62</sup>

$$\mathcal{F} = 2\kappa(H - c_0)^2 + \bar{\kappa}H_G, \quad (2.5)$$

in terms of the spontaneous curvature,  $c_0$ , which is zero for membranes where both leaflets have the same composition and non-zero for heterogeneous compositions, such as in true biological membranes.  $H$  and  $H_G$  define the mean and Gaussian curvatures,  $H = \frac{1}{2} \left( \frac{1}{R_1} + \frac{1}{R_2} \right)$  and  $H_G = \frac{1}{R_1 R_2}$  respectively, in terms of the radii of curvature  $R_i^{-1}$ . The bending modulus,  $\kappa$  in eqn (2.5), quantifies the energy cost associated with deviations from  $c_0$ , and the saddle splay modulus,  $\bar{\kappa}$ , scales the contribution due to  $H_G$ . The HH has been used to quantify the membrane entropy,<sup>61,62</sup> and bending moduli.<sup>59,63,64</sup> However, this approach has limitations as it cannot describe small wavelength fluctuations involving motion of lipids in and out of the membrane plane (protrusions) or fluctuations of the membrane thickness. Computer simulations still represent the most powerful route to quantify these fluctuation modes.

The computation of mechanical properties and the area per lipid requires the definition of the instantaneous shape (IS) of the membrane (see Figure 2.3). The IS can be obtained from the local height,  $h(x, y)$ , of the membrane above a reference plane. Grid methods,<sup>65,66</sup> and intrinsic surface based approaches<sup>67-69</sup> can be used to calculate  $h$ .

In the intrinsic approach the instantaneous surface,  $\xi(\vec{R})$  where  $\vec{R} = (x, y)$ , is constructed by selecting a series of pivots, which in membranes can be readily defined by the phospholipid head groups. The IS surface can then be represented as a Fourier series,

$$\xi(\vec{R}, q_u) = \sum_{|\vec{q}| \leq q_m} \hat{\xi}_{\vec{q}} e^{i\vec{q} \cdot \vec{R}}, \quad (2.6)$$

where  $\vec{q} = 2\pi \left( \frac{n_x}{L_x}, \frac{n_y}{L_y} \right)$ ,  $n_x, n_y = 0, \pm 1, \pm 2, \dots$ , and  $L_x$  and  $L_y$  represent the lengths of the simulation cell. The sum in eqn (2.6) involves an upper limit,  $q_m = 2\pi/\lambda_c$ , defined by  $\lambda_c \approx \sqrt{A_l}$ , where  $A_l$  is the area per lipid,  $(L_x \times L_y)/N_l$ , and  $N_l$  is the number of lipids per leaflet. It is accepted that membrane undulations with wavevectors longer than  $q_m$  are not physically meaningful.

The analysis of membrane thermal undulations is the most popular simulation approach to quantify the bending modulus,<sup>63,66,69-71</sup> although methods based on the active deformation of the membrane have also been proposed.<sup>72,73</sup> The fluctuation approach relies on the analysis of the



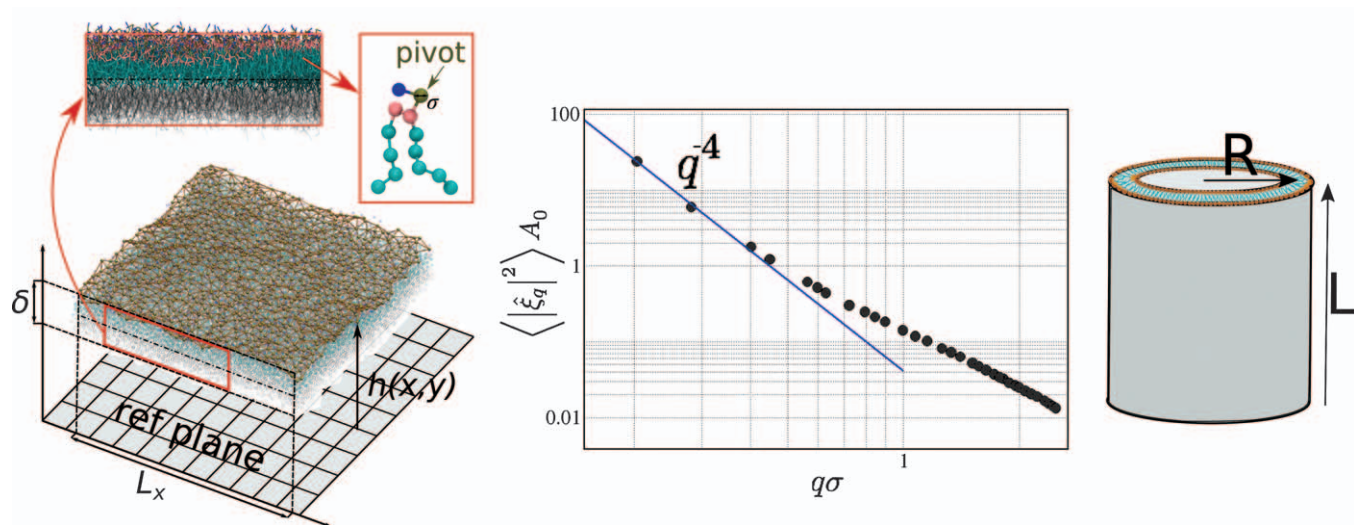


**Figure 2.3** (Left) Snapshot of a POPC bilayer modelled using the MARTINI force field. Lipids in the lower leaflet are coloured in grey and the dot-dashed line indicates the middle of the bilayer. The solvent has been omitted for clarity. The network of interconnected phospholipid head groups (pivots) represents the instantaneous shape of the upper leaflet of the membrane. (Middle) Fluctuation spectrum of a tensionless POPC bilayer at 320 K. The blue straight line highlights the  $q^{-4}$  dependence which is observed at low wavevectors and that defines the bending modulus of the membrane. Data taken from ref. 69. (Right) Sketch of a cylindrical membrane of radius  $R$  and length  $L$ , illustrating the set up employed in the computation of the bending modulus.

membrane local height. The coefficients in the Fourier series defined in eqn (2.6), give the fluctuation spectrum:

$$\langle |\hat{\zeta}_q|^2 \rangle = \frac{k_B T}{(\kappa q^4 + \gamma q^2) A_0}, \quad (2.7)$$

where  $\gamma$  is the the membrane surface tension and  $A_0 = L_x \times L_y$  is the bilayer projected area. The tensionless state,  $\gamma = 0$ , is the most relevant in biophysics. In this state the fluctuations are determined by the bending modulus only, which follows directly from eqn (2.7). Figure 2.3 shows the fluctuation spectrum of a POPC (MARTINI) bilayer in the tensionless state. The spectrum shows the expected power law behaviour,  $q^{-4}$ , at low wave-vectors ( $q$ ), corresponding to the regime dominated by the bending modes. At wavevector  $q \approx 2\pi/\delta$ , where  $\delta$  is the membrane thickness, a change in power law can be seen, which is associated with discrete lipid fluctuations (protrusions).<sup>69,74</sup>

The fluctuation spectrum method has been applied to a number of membranes. For DPPC (MARTINI) several values have been reported for the bending modulus (in  $k_B T$ ):  $\beta\kappa = 9.6$ ,<sup>75</sup>  $13.7$ ,<sup>76</sup>  $19.3$ ,<sup>77</sup>  $36$ ,<sup>78</sup>  $23$ .<sup>69</sup> Values of the order of  $20k_B T$  are in the range expected from experimental studies.<sup>79,80</sup> Simulations of coarse-grained models of POPC bilayers also give values for the bending modulus of  $21k_B T$ . The results reported in the literature show a large variability.<sup>81</sup> This fact justifies the continuous development of methods as well as force fields to compute these mechanical properties of biomolecules.

The bending modulus has also been computed from the tensile force needed to maintain the shape of a cylindrical membrane.<sup>64</sup> The bending modulus is defined by the axial force,  $F$ , acting on the cylinder (see Figure 2.3), and the cylinder radius,  $R$ . For CG membranes, values of  $\beta\kappa = 11.7 \pm 0.2$ , were reported, showing good agreement with the results of the fluctuation spectrum route,  $\beta\kappa = 12.5 \pm 1$ . The bending modulus can also be estimated from the membrane thickness, using simple approaches based on classical elasticity theory<sup>70</sup> where the bending rigidity is determined by the area compressibility modulus,  $K_A$ , and the bilayer thickness,  $\delta$ , as  $\kappa = K_A \delta^2/n$ ,<sup>82</sup> where  $n = 24$  in the polymer brush model.<sup>83</sup>

The elastic properties of membranes are also defined by  $K_A$ , which measures the tensile strength of the bilayer and is directly proportional to the bilayer area fluctuations,  $\langle A^2 \rangle - \langle A \rangle^2$ .  $K_A$  has been extensively investigated in computer simulations over the last two decades. The major problem in the interpretation of different simulation results is the lack of a unique definition of the membrane area. The cross-sectional area,  $\langle L_x \times L_y \rangle$ , is commonly employed to estimate the area per lipid. However this approach is not thermodynamically consistent, since the cross-sectional area is not an extensive thermodynamic variable. Hence the area per lipid obtained in this way, and the corresponding compressibility, depend on the membrane area.<sup>63,72</sup> This problem has prompted the development of new definitions of



the membrane area, which fulfill thermodynamic consistency, and allow the unambiguous calculation of the compressibility.<sup>84</sup>

The analysis of different computational approaches for calculating the area per lipid shows that the standard approach, based on the cross-sectional area, despite being thermodynamically inconsistent, predicts areas that are very close to the real one. The area per lipid is often used to test the accuracy of force fields. The area per lipid for POPC bilayers varies between 0.638–0.67 nm<sup>2</sup> depending on the type of force field, CG or AA/UA, and the temperature, 303–310 K.<sup>85–87</sup> These values are within the range of those estimated in experiments, 0.63–0.68 nm<sup>2</sup>. The variability across different experiments is connected to the experimental approach, either X-ray or NMR.<sup>79</sup>

In addition to the computation of elastic properties, MD also enables the calculation of the local pressure tensor, hence making it possible to resolve local stresses,  $\sigma(z)$ , in biological membranes. This approach relies on the microscopic definition of the stress tensor by Irving and Kirkwood,<sup>88</sup> later developed further by Noll,<sup>89,90</sup> resulting in the Irving–Kirkwood–Noll theory. The local stress is determined by a kinetic,  $\sigma_{\alpha\beta}^K(z)$ , and potential,  $\sigma_{\alpha\beta}^V(z)$ , term, such that  $\sigma_{\alpha\beta}(z) = \sigma_{\alpha\beta}^K(z) + \sigma_{\alpha\beta}^V(z)$ . Vanegas *et al.*<sup>91</sup> developed a method to calculate the local stress, which is available as a GROMACS tool. The current implementation does not, however, include full long-range electrostatic interactions. Once the local stresses are known, the local pressure can be obtained from the contributions of the local stress tangential,  $P_T(z)$ , and normal,  $P_N(z)$ , to the membrane plane:  $\Pi(z) = P_T(z) - P_N(z)$ . This relationship is very useful, since it allows the computation of the surface tension ( $\gamma$ ), the product of the bending modulus ( $\kappa_b$ ) and spontaneous curvature ( $c_0$ ), and the Gaussian modulus ( $\kappa_G$ ) (eqn (2.8)):

$$\int_0^\delta \Pi(z) dz = -\gamma; \quad \int_0^\delta \Pi(z) z dz = \kappa_b c_0; \quad \int_0^\delta \Pi(z) z^2 dz = -\kappa_G. \quad (2.8)$$

The integration is performed over the membrane thickness,  $\delta$ , where  $\Pi(z) \neq 0$ . This approach has been used to obtain elastic properties of skin lipids, such as ceramide 2 (see Figure 2.4), revealing an interplay between membrane composition and microscopic stress.<sup>92</sup>

## 2.4 MD Simulation vs. Experiment

The improvements in computer hardware and algorithms, and the development of coarse-grained models has substantially increased the length and timescales that can be tackled using MD simulations, allowing a closer connection with experiments. Following early simulations of protein flexibility in the 1970s,<sup>93</sup> comparisons with experimental data have expanded significantly and are crucial for testing the accuracy of force fields. Simulations have also been employed to resolve seemingly contradictory experimental data, obtained using different techniques. An example of this idea is

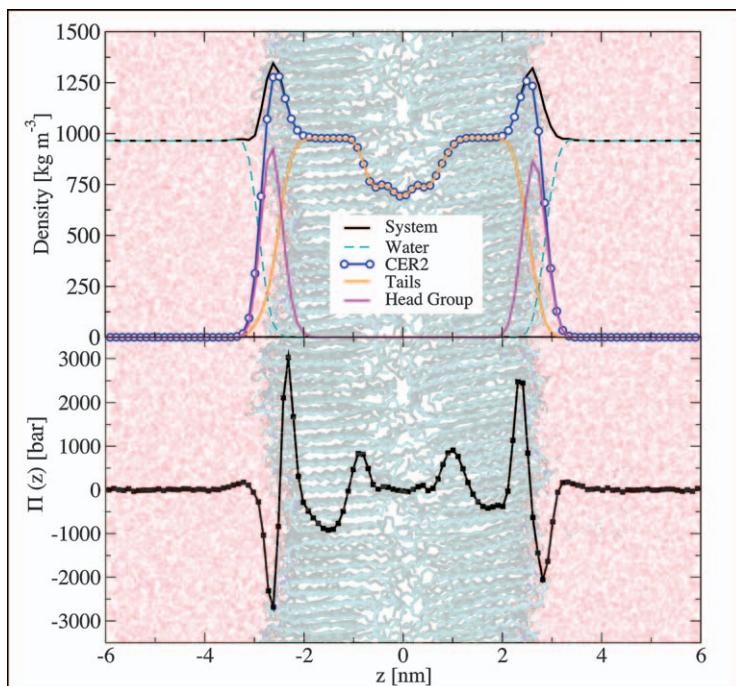


**Figure 2.4** Lateral density profile (top) and pressure profile (bottom) of a ceramide 2(CER2) lipid bilayer, overlaid on a simulation snapshot. The peaks and troughs of the pressure profile reflect structural features of the bilayer.

the application of MD to reconcile NMR and X-ray data and produce a unified model for the catalytic mechanism of Cyclophosphamide.<sup>94</sup> The interaction between simulation and experiment is becoming more significant, and most structural analyses *via* NMR or scattering techniques rely on some form of computational refinement step, *e.g.*, simulated annealing.<sup>95</sup> In this section we focus on the interaction between simulations and experiments in the area of NMR and diffraction techniques, where simulations are an integral part of the analysis and interpretation of the data. However, it should be noted that the use of MD to interpret ambiguous experimental results at a molecular level is not restricted to these two techniques, for example, it is also useful in the analysis of circular dichroism (CD) data. In this instance, MD can be used to establish correlations between different protein conformations and their CD spectra.<sup>96</sup>

### 2.4.1 NMR and MD: Structure and Dynamics

The molecular resolution of NMR lends itself to an interplay with MD simulation. The ability of simulations to produce reliable NMR data is highlighted by the strong collaboration between experimentalists and

simulators. An example of such collaboration is the NMRlipids project, which provides a platform for discussions of experimental and simulation data and can be accessed through an online repository.<sup>97</sup>

The lipid bond order parameters<sup>98</sup> are key structural quantities, which can be computed from the positions of the atoms obtained in an MD trajectory. The bond order parameters provide information on the ordering of the lipid hydrocarbon tails, which show distinctively different values depending on the membrane state, *e.g.* gel or fluid. The C–H bond order parameter is defined as:

$$S_{CD} = \frac{1}{2} \langle (3 \cos^2 \theta - 1) \rangle, \quad (2.9)$$

where  $\theta$  is the instantaneous angle between the C–H bond and the bilayer normal. This definition is strictly related to its experimental measurement that can be obtained from the quadrupolar splitting,  $\Delta\nu_Q$ , in  $^2\text{H}$  NMR:

$|S_{CD}| = \frac{2}{3} \frac{\hbar}{e^2 q Q} \Delta\nu_Q$ , where  $\hbar$  is Planck's constant,  $e$  is the elementary charge,  $q$  is the electric field gradient and  $Q$  is the deuteron quadrupole moment. Alternatively, it can be accessed from dipolar splitting in  $^{13}\text{C}$  NMR.

In AA simulations, the order parameter can be calculated immediately from its definition (see eqn (2.9)). However, it can also be related to the C–C bond vectors<sup>99</sup> in UA simulations. Several tools and software packages are available for its determination, such as the VMD plugin MEMBPLUGIN,<sup>100</sup> or the open source Python library MDAnalysis, which enables the analysis of trajectories generated with several MD simulation software packages.<sup>101</sup> Agreement with experiment is in general good for most force fields, although several differences still persist in the description of the head group region due to difficulties in describing the water–lipid interface.

Regarding proteins, there are also a range of properties that provide a source of comparison between computer simulations and NMR experiments.  $J$ -couplings can be converted into dihedral angles using the Karplus relations<sup>102</sup> and compared with average backbone dihedral angles obtained in simulations to analyse secondary structure. In addition, the Nuclear Overhauser Effect (NOE) can probe the spatial relationship between hydrogens, giving insight on tertiary structure. Order parameters (indicated with  $S_2$ ) give dynamic information on protein flexibility based on the angular distributions of bond vector orientations. These properties are often used to parametrise<sup>103</sup> and validate<sup>104</sup> force fields.

The power of NMR comes from its ability to study structure and dynamics simultaneously. MD simulations are well suited to capturing the conformational flexibility of molecules and they can be used to generate an ensemble of structures which can then be compared with the experimental data. This idea has prompted the development of an integrated NMR-MD approach that has been realised in the past years. The state of the art of such approach is discussed in a recent review by Mocci and Laaksonen.<sup>105</sup>

Along these lines, Lindorff-Larsen *et al.*<sup>106</sup> have pioneered a new method for structure determination, called dynamic ensemble refinement. In this approach, the data from NMR experiments ( $S_2$  and NOE distances) are used as restraints to guide the MD simulations. Rather than applying the restraints to every configuration in the trajectory, the restraints are imposed on the ensemble as a whole. This idea is realised by modifying the restraints during the simulation to ensure that on average, the ensemble will reproduce the NMR data. The advantage of this simulation approach is that it can capture both structural and dynamical information simultaneously since the restraints are applied on the structure, *via* NOEs, and dynamics, *via*  $S_2$ , in the same simulation. This simulation method offers an efficient route to resolve protein structures. The ensemble approach has been used to visualise transient complexes,<sup>107</sup> study enzyme dynamics,<sup>108</sup> describe the structures of intrinsically disordered proteins,<sup>109,110</sup> probe correlated motions in proteins,<sup>111,112</sup> and explore the structures of membrane proteins.<sup>113</sup>

#### 2.4.2 Structure of Biomolecules and Diffraction: Solving the Phase Problem with MD

X-Ray diffraction was used to resolve the first protein structures<sup>114</sup> and is still the most commonly used tool for protein structure determination.<sup>115</sup> Lipid membranes are also commonly probed using scattering techniques.<sup>116</sup> However, in biologically relevant environments, membranes are in a fluid state and the lack of order makes structure determination *via* scattering difficult. In general, scattering techniques cannot provide an unambiguous result due to the so-called ‘phase problem’. Scattering experiments measure the intensity of scattered radiation,  $I(q)$ , which is related to the square modulus of the form factor  $F(q)$ , as a function of the wave vector  $q$ .  $F(q)$  is connected to the electron density profile,  $\rho_e(r)$ , in X-ray scattering (or neutron scattering length density,  $\rho_{NSLD}(r)$ , in neutron scattering) *via* the Fourier transform over the system volume,  $V = r^3$ :  $F(q) = \int_V \rho_e(r) e^{iq \cdot r} dV$ . However, obtaining the density profiles from an inverse Fourier transform of the form factor is not possible because experiments only yield the modulus of the form factor, so its amplitude is known, but not its phase. To overcome this problem, scattering data must be fit to a model structure, whose density function depends on fitting parameters which can be derived *via* comparison with experimental form factors.

Various simple models of lipid membrane structure have been proposed for this purpose.<sup>117–119</sup> MD can be of great help in this instance, by proposing structures that reduce the number of fitting parameters. An example of this idea is the H2 model, proposed by Klauda *et al.*<sup>120</sup> Alternatively, it is possible to directly compare form factors derived from simulations with the raw experimental data obtained in reciprocal space.<sup>121</sup> For instance, to calculate  $F(q)$  of membranes using MD, the

number density profile,  $N_\alpha(z)$ , perpendicular to the membrane has to be calculated.  $N_\alpha(z)$  can be easily obtained from a simulation by dividing the simulation box into a number of slices running parallel to the membrane surface (along  $z$ ) and counting the number of atoms of type  $\alpha$  in each slice. Then, to compare with X-ray data,  $N_\alpha(z)$  is multiplied by the number of electrons in each atom to obtain  $\rho_e(r)$  (or multiplied by the neutron scattering length to get  $\rho_{\text{NSLD}}(r)$  to compare with neutron data) and finally Fourier transformed to give  $F(q)$ . The process of generating X-ray and neutron data from simulations has been automated *via* the SIMtoEXP software package.<sup>122</sup> This software uses as input the number densities of each atom type,  $N_\alpha(z)$ . It then calculates  $\rho_e(r)$  and  $\rho_{\text{NSLD}}(r)$  and the form factors,  $F_e(q)$  and  $F_{\text{NSLD}}(q)$ . Raw experimental data can also be input alongside simulation data, or obtained from the growing repository of experimental data incorporated in the software.

## 2.5 Future Directions

Since the pioneering simulations of bovine pancreatic trypsin inhibitor in the late 1970s,<sup>93</sup> the size of biological structure that we are able to simulate has increased steadily. The ribosome,<sup>123</sup> with over 2 million atoms, offers a paradigmatic example of the system sizes that can be studied with state of the art supercomputers and using atomistic descriptions. However, the computer simulation of biomolecules is not only a problem of size, but also time. Important biological processes span a large range of timescales, from ns to ms, the latter being of particular significance in protein conformational changes. The need to reach both long simulation times and large system sizes has motivated the development of coarse-grained force fields (see Section 2.2). We discuss below examples of biomolecular simulations that are at the edge of what we can currently do using state of the art software and hardware, and use these examples as a springboard to delineate ideas for future directions.

The use of tera/petascale computers and highly parallelised molecular dynamics software (LAMMPS,<sup>124</sup> NAMD<sup>125</sup> or GROMACS<sup>126</sup>) are contributing to the rapid development of computational biology. An important application of MD methods is the investigation of proteins and the conformational changes they undergo. Being able to predict the three-dimensional structure of proteins still represents a big challenge in computational biology. This is connected to the huge difference in the number of configurations of the unfolded and native states, and the large entropy penalty in moving from the former to the latter state. Recent simulations have brought us a little bit closer to solving this problem. Taking advantage of hardware and software developments, which enabled simulations of up to the ms timescale, Shaw *et al.* performed atomistic molecular simulations of 12 fast-folding proteins. They were able to identify a single dominant route for the folding process and estimate the folding free energy barriers of  $\sim 5k_B T$ , which can be simulated using standard techniques. Higher energy barriers

require the use of advanced techniques, such as metadynamics or parallel tempering.<sup>127</sup>

Direct and accelerated dynamics simulations will play a role in the study of drug-protein interactions. In a recent simulation study this problem was investigated using an unbiased approach, where the drug (dasatinib) was able to freely sample the entire protein (Src kinase) surface, eventually finding its binding site.<sup>128</sup> Such simulations, which are very time consuming (typical simulations involved several tens of microseconds), are expected to play an important role in the development of computational approaches for drug screening.

Atomistic and coarse-grained computer simulations of biological membranes consisting of a single component are almost routine nowadays. Even for such simple biological structures, however, there are still challenges that need to be resolved. Part of this problem is connected to the differences in the results that can be obtained by using different structural probes, X-rays and neutrons,<sup>79</sup> *e.g.* in estimates of areas per lipid, which are commonly used to fit force fields. The computer simulation of very large systems is likely to expose more clearly the lack of thermodynamic consistency of some definitions of the area per lipid as well as the implications this has on the determination of mechanical properties of bilayers (see Section 2.3.4). One reason why simulations have focused on single component membranes, and binary and ternary mixtures to a lesser extent, is the large body of experimental studies on these types of model membranes. However, the plasma membrane (PM) is an asymmetric, multicomponent mixture containing dozens of components that confer the PM its unique properties as a barrier for the cell. Only very recently, a model of the PM has been considered using a coarse-grained force field and more than 50 components, including saturated and unsaturated lipids and cholesterol molecules, which included in total about 0.5 million particles.<sup>20</sup> These simulations provided a direct visualisation of flip-flop dynamics of cholesterol over 10's of microseconds, the asymmetric composition of the inner and outer leaflets of the bilayer, and the preference of cholesterol to accumulate in one of the leaflets. These simulations open a route to study the incorporation of proteins into the bilayer and to study their behaviour in the asymmetric environment of the PM.

The investigation of transmembrane proteins is of utmost importance to understand the function of biological membranes. Many biological functions and cell collective behaviour, *e.g.* adhesion, rely on the behaviour of proteins at membranes and how these proteins interact with different complexes.<sup>129</sup> Transmembrane proteins are also important therapeutic targets for many diseases. Recent coarse-grained simulation studies of integrins highlighted the importance of protein-lipid interactions and membrane restructuring, a result that has implications for signal transduction across the PM.<sup>130</sup>

Caffrey and Feigenson,<sup>131</sup> showed that the maximum activity of  $\text{Ca}^{2+}$ -ATPases, which are involved in muscular relaxation, is strongly



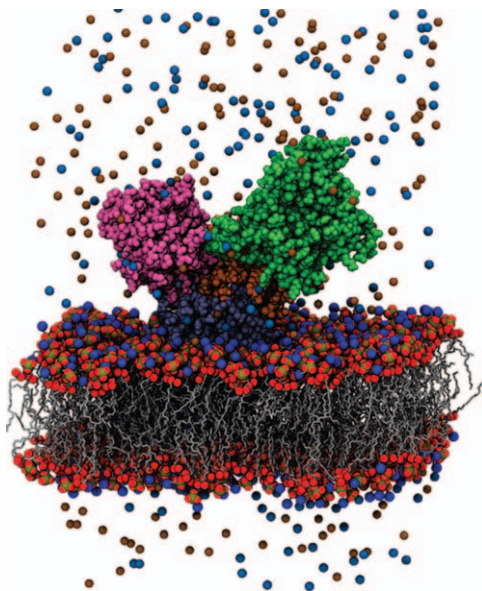


**Figure 2.5** Molecular dynamics simulation of a  $\text{Ca}^{2+}$ -ATPase molecular motor at a POPC bilayer and in a 0.1 M NaCl aqueous solution (blue and brown spheres). The water molecules have been removed for clarity. The colour regions in the protein indicate the actuator domain (pink), the nucleotide binding domain (green) and the phosphorylation domain (orange). Snapshot reproduced from ref. 133 with permission from the Royal Society of Chemistry.

dependent on the bilayer lipid composition. Atomistic computer simulations of these proteins (see Figure 2.5) in bilayer membranes<sup>132,133</sup> have shown that the protonation state of the  $\text{Ca}^{2+}$  binding sites is important in determining the stability of the ions in the protein channel. The simulations were also employed to quantify the deformation of the membrane around the protein and complementary studies using non-equilibrium thermodynamics theory have shed light on the thermodynamic activity of the molecular motor and its dependence on the membrane composition.<sup>134</sup> MD simulations can play an important role to advance the description of the molecular mechanisms leading to rotation and motion of ATPase molecular machines, and to understand the energy conversion from chemical energy to ion translocation or motion. Some efforts in this direction can be found in the literature.<sup>135</sup> However, being able to model the enzymatic pathway of molecular machines remains a hard challenge to computer simulations,<sup>136</sup> fundamentally due to the very long-time scales involved. Advancing in this question will, however, bring parallel advances in our ability to interpret experimental data, *e.g.*, in crystallographic studies.<sup>137</sup>

An important avenue in computer simulations is the investigation of lipid–nanomaterial interactions.<sup>138</sup> The interest in this problem is in part driven by the large number of particles that are released to the atmosphere



as a by-product of *e.g.* combustion processes, and the harmful effect these particles may have in the lungs. Coarse-grained models have been employed to study the interactions of carbon nanoparticles with DOPC bilayers<sup>139</sup> and DPPC lipid monolayers.<sup>140</sup> The DPPC simulations show that the lipids trap the nanoparticles, with the latter having a stronger affinity for the liquid expanded phase. Extension of these studies to atomistic models of lipid monolayers<sup>141</sup> would be highly desirable.

The ribosome is a very important pharmaceutical target,<sup>142</sup> as it performs the vital role of transcription of RNA into proteins. A solvated ribosome involves more than 2 million atoms making molecular dynamics simulations of this structure very challenging. However, simulations of this organelle have been successfully implemented,<sup>123,143</sup> providing important insight into translocation processes that are not accessible to experimental techniques. Future simulations in this area should provide important atomistic information to assist the development of drugs that specifically target bacteria.<sup>144</sup>

The computer simulation of the most primitive forms of life, viruses, has become a popular subject in recent years. In a pioneering study the complete satellite tobacco virus, consisting of 1 million atoms, was simulated using all atom molecular dynamics simulations.<sup>145</sup> More recent studies<sup>146</sup> have shown that the capsid of the virus is not stable, while the RNA core is stable. This work raises interesting questions about the virus assembly mechanism. Similarly, large scale atomistic molecular dynamics simulations have enabled the study of the small poliovirus capsid.<sup>147</sup> These simulations indicate that the capsid exchanges water molecules relatively quickly with an  $8 \text{ ns}^{-1}$  exchange rate, an effect that may contribute to the resistance of the capsid to high pressures.

Coarse-grained models have been developed to investigate the stability of virus capsids (STMV, SPMV and STNV) in the microsecond timescale,<sup>32,148</sup> as well as the dengue virion envelope<sup>149</sup> and the influenza A virion.<sup>22</sup> Coarse-grained models offer an appealing approach to investigate the assembly of these systems and to refine the experimental structures resolved using crystallography and NMR techniques.<sup>150</sup> Molecular dynamics methods using virus capsid models, both atomistic and coarse-grained, should provide a powerful tool to complement current approaches to develop antivirals and also to understand infection pathways.

## 2.6 Conclusion

Computer simulations of biological systems have undergone dramatic changes thanks to increasing computational power and the development of force fields that allow a significant increase of size and length scales, at the same time preserving the necessary level of accuracy. The introduction of advanced algorithms to sample rare events has also opened new avenues to investigate processes that are too slow to be observed using standard molecular dynamics techniques.<sup>144</sup> These advances have allowed us to simulate

large biological structures, like the ribosome or entire viruses, and bring us closer to the implementation of molecular dynamics simulations in therapeutic applications.

The fundamentals of molecular dynamics are well established and there are a wide range of force fields that provide results in qualitative and quantitative agreement with available experimental data. However, some fundamental questions in the simulation studies remain. One such question concerns how to best quantify the mechanical properties and areas of lipid bilayers. Recent work has exposed the lack of thermodynamic consistency of widely used definitions of the area per lipid, and the associated errors in the computation of mechanical properties. These errors increase with the size of the bilayer membrane. Hence, discrepancies between definitions may become more evident in the future as it becomes easier to simulate extremely large systems.

The next steps in the computational simulation of membranes will likely focus on the development of more realistic models, namely, models that incorporate asymmetry in composition, as well as truly multicomponent systems. These simulations are necessary to understand the behaviour of transmembrane proteins and how the local membrane composition influences the protein behaviour. The development of accurate force fields offers some hope that very large simulations targeting systems with dozens of components can be performed. However, a hard challenge in the simulation of biological systems is to reach longer timescales, *e.g.*, ms, which are needed to sample processes that involve large cooperative motions, or the assembly of very complex structures. It is not clear how the timescales can be expanded significantly using the coarse-grained route without losing relevant biological information. More work in this direction is highly desirable.

It is also expected that the future will bring a stronger integration between simulations and experiments, enabling these two approaches to be used in a more concerted way. Recent developments linking NMR and MD simulation are promising and highlight the advantages of an integrated approach. Simulations should also play an important role in structure prediction, hence complementing techniques, such as X-ray and neutron scattering. Indeed, MD simulations can help to bridge different experimental techniques, so that we can take full advantage of the power of NMR to interrogate three dimensional conformations and dynamics of molecules, and of scattering techniques to probe the structural features of complex biomolecule aggregates.

The simulation of molecular mechanisms associated with energy conversion in biomolecules and the conversion of this energy into motion, conformational changes or ion translocation, represents a major challenge to computer simulations. Such problems involve the full complexity of modelling, namely, simulating bond formation and breaking, evaluating energy transfer in complex geometries (proteins) and being able to explore conformational changes over timescales that are not within reach with our current hardware and software capabilities. The impressive advances in the

simulation of organelles is an indication that such problems may be solved in the long term, but probably they will involve new revolutions in both computing and MD algorithms.

## Acknowledgements

We acknowledge EPSRC for funding (EP/J003859/1, EP/J017566/1) and a PhD Studentship from Imperial College. We are also thankful for funding from EU 7th Framework Programme ITN-SNAL.

## References

1. L. Verlet, *Phys. Rev.*, 1967, **159**, 98.
2. M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids*, Clarendon Press; Oxford University Press, Oxford, 1989.
3. H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola and J. R. Haak, *J. Chem. Phys.*, 1984, **81**, 3684.
4. W. G. Hoover, *Phys. Rev. A*, 1985, **31**, 1695.
5. S. Nosé, *Mol. Phys.*, 1984, **52**, 255.
6. H. C. Andersen, *J. Chem. Phys.*, 1980, **72**, 2384.
7. G. Bussi, D. Donadio and M. Parrinello, *J. Chem. Phys.*, 2007, **126**, 014101.
8. M. Parrinello and A. Rahman, *Appl. Phys.*, 1981, **52**, 7182.
9. S. Nosé and M. L. Klein, *Mol. Phys.*, 1983, **50**, 1055.
10. B. Hess, H. Bekker, H. J. C. Berendsen and J. G. E. M. Fraaije, *J. Comput. Chem.*, 1997, **18**, 1463.
11. H. A. Lorentz, *Ann. Phys. (Berlin, Ger.)*, 1881, **248**, 127.
12. P. H. Hünenberger and J. A. McCammon, *J. Chem. Phys.*, 1999, **110**, 1856.
13. J. Domański, P. J. Stansfeld, M. S. P. Sansom and O. Beckstein, *J. Membrane Biol.*, 2010, **236**, 255.
14. G. Mallocci, A. V. Vargiu, G. Serra, A. Bosin, P. Ruggerone and M. Ceccarelli, *Molecules*, 2015, **20**, 13997.
15. D. Fritz, K. Koschke, V. A. Harmandaris, N. F. A. van der Vegt and K. Kremer, *Phys. Chem. Chem. Phys.*, 2011, **13**, 10412.
16. M. G. Saunders and G. A. Voth, *Annu. Rev. Biophys.*, 2013, **42**, 73.
17. W. G. Noid, J. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das and H. C. Andersen, *J. Chem. Phys.*, 2008, **128**, 244114.
18. S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman and A. H. de Vries, *J. Phys. Chem. B.*, 2007, **111**, 7812.
19. S. J. Marrink, H. J. Risselada and A. E. Mark, *Chem. Phys. Lipids*, 2005, **135**, 223.
20. H. I. Ingólfsson, M. N. Melo, F. J. van Eerden, C. Arnarez, C. A. Lopez, T. A. Wassenaar, X. Periole, A. H. de Vries, D. P. Tieleman and S. J. Marrink, *J. Am. Chem. Soc.*, 2014, **136**, 14554.
21. C. Arnarez, S. J. Marrink and X. Periole, *Chem. Sci.*, 2016, **7**, 4435.

22. T. Reddy, D. Shorthouse, D. L. Parton, E. Jefferys, P. W. Fowler, M. Chavent, M. Baaden and M. S. P. Sansom, *Structure*, 2015, **23**, 584.
23. M. Orsi and J. W. Essex, *PLoS One*, 2011, **6**, e28637.
24. I. R. Cooke, K. Kremer and M. Deserno, *Phys. Rev. E*, 2005, **72**, 011506.
25. H. Taketomi, Y. Ueda and N. Gō, *Int. J. Pept. Protein Res.*, 1975, **7**, 445.
26. J. Karanicolas and C. L. Brooks III, *Protein Sci.*, 2002, **11**, 2351.
27. J. W. Carter, C. M. Baker, R. B. Best and D. De Sancho, *J. Phys. Chem. B*, 2013, **117**, 13435.
28. S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A. E. Dawid and A. Kolinski, *Chem. Rev.*, 2016, **116**, 7898.
29. P. J. Stansfeld and M. S. P. Samson, *J. Chem. Theory Comput.*, 2011, **7**, 1157.
30. E. Lyman, F. M. Ytreberg and D. M. Zuckerman, *Phys. Rev. Lett.*, 2006, **96**, 028105.
31. M. Praprotnik, L. Delle Site and K. Kremer, *J. Chem. Phys.*, 2005, **123**, 224106.
32. G. Ayton, S. G. Bardenhagen, P. McMurtry, D. Sulsky and G. A. Voth, *J. Chem. Phys.*, 2001, **114**, 6913.
33. W. J. Mortier, S. K. Ghosh and S. Shankar, *J. Am. Chem. Soc.*, 1986, **108**, 4315.
34. S. Patel and C. L. Brooks III, *J. Comput. Chem.*, 2004, **25**, 1.
35. S. O. Yesylevskyy, L. V. Schäfel, D. Sengupta and S. J. Marrink, *PLoS Comput. Biol.*, 2010, **6**, e100081.
36. C. M. Baker, V. M. Anisimov and A. D. MacKerell, *J. Phys. Chem. B*, 2011, **115**, 580.
37. J. W. Ponder, C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, R. A. DiStasio, M. Head-Gordon, G. N. I. Clark, M. E. Johnson and T. Head-Gordon, *J. Phys. Chem. B*, 2010, **114**, 2549.
38. J. Chen, C. L. Brooks III and J. Khandogin, *Curr. Opin. Struct. Biol.*, 2008, **18**, 140.
39. R. M. Levy, L. Y. Zhang, E. Gallicchio and A. K. Felts, *J. Am. Chem. Soc.*, 2003, **125**, 9523.
40. M. Carballo-Pacheco, I. Vancea and B. Strodel, *J. Chem. Theory Comput.*, 2014, **10**, 3163.
41. A. de Ruiter, S. Boresch and C. Oostenbrink, *J. Comput. Chem.*, 2013, **34**, 1024.
42. M. Nilges, G. M. Clore and A. M. Gronenborn, *FEBS Lett.*, 1988, **239**, 129.
43. D. S. Goodsell and A. J. Olson, *Proteins*, 1990, **8**, 195.
44. J. de Joannis, P. S. Coppock, F. Yin, M. Mori, A. Zamorano and J. T. Kindt, *J. Am. Chem. Soc.*, 2011, **133**, 3625.
45. A. Patriksson and D. van der Spoel, *Phys. Chem. Chem. Phys.*, 2008, **10**, 2073.
46. N. V. Buchete and G. Hummer, *Phys. Rev. E*, 2008, **77**, 030902.
47. J. D. Chodera and F. Noé, *Curr. Opin. Struct. Biol.*, 2014, **25**, 135.

48. D. Shukla, Y. Meng, B. Roux and V. S. Pande, *Nat. Commun.*, 2014, **5**, 3397.
49. Folding@home, <http://folding.stanford.edu> (accessed August 2016).
50. I. C. Lin and M. E. Tuckerman, *J. Phys. Chem. B*, 2010, **114**, 15935.
51. G. M. Torrie and J. P. Valleau, *J. Comput. Phys.*, 1977, **23**, 187.
52. S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman and J. M. Rosenberg, *J. Comput. Chem.*, 1992, **13**, 1011.
53. A. Laio and M. Parrinello, *Proc. Natl. Acad. Sci.*, 2002, **99**, 12562.
54. C. Camilloni and M. Vendruscolo, *J. Am. Chem. Soc.*, 2014, **136**, 8982.
55. C. Jarzynski, *Phys. Rev. Lett.*, 1997, **78**, 2690.
56. G. Hummer and A. Szabo, *Proc. Natl. Acad. Sci.*, 2001, **98**, 3658.
57. Y. Xu, J. Shen, X. Luo, I. Silman, J. L. Sussman, K. Chen and H. Jiang, *J. Am. Chem. Soc.*, 2003, **125**, 11340.
58. J. A. Lundbæk, *J. Phys.: Condens. Matter*, 2006, **18**, S1305.
59. H. P. Duwe, J. Kaes and E. Sackman, *J. Phys. (France)*, 1990, **51**, 945.
60. L. Chernomordik, M. M. Kozlov and J. Zimmergerg, *J. Membr. Biol.*, 1995, **146**, 1.
61. W. Helfrich, *Z. Naturforsch.*, 1973, **28**, 693.
62. S. A. Safran, *Statistical Thermodynamics of Surfaces, Interfaces and Membranes*, Addison-Wesley, Reading, 1994.
63. A. E. Mark and S. J. Marrink, *J. Phys. Chem. B*, 2001, **105**, 6122.
64. V. A. Harmandaris and M. Deserno, *J. Chem. Phys.*, 2006, **125**, 204905.
65. C. Loison, M. Mareschal, K. Kremer and F. Schmidt, *J. Chem. Phys.*, 2003, **119**, 13138.
66. E. Lindahl and O. Edholm, *Biophys. J.*, 2000, **79**, 426.
67. F. P. Buff, R. A. Lovett and F. Stillinger Jr, *Phys. Rev. Lett.*, 1965, **15**, 621.
68. R. Evans, *Adv. Phys.*, 1979, **28**, 143.
69. P. Tarazona, E. Chacón and F. Bresme, *J. Chem. Phys.*, 2013, **139**, 094902.
70. R. Goetz, G. Gompper and R. Lipowsky, *Phys. Rev. Lett.*, 1999, **82**, 221.
71. J. Neder, B. West, P. Nielaba and F. Schmid, *J. Chem. Phys.*, 2010, **132**, 115101.
72. W. K. den Otter and W. J. Briels, *J. Chem. Phys.*, 2003, **118**, 4712.
73. O. Farago and P. Pincus, *J. Chem. Phys.*, 2004, **120**, 2934.
74. G. Brannigan and F. L. H. Brown, *Biophys. J.*, 2006, **90**, 1501.
75. S. J. Marrink, A. H. de Vries and A. E. Mark, *J. Phys. Chem. B*, 2004, **108**, 750.
76. Q. Waheed and O. Edholm, *Biophys. J.*, 2009, **97**, 2754.
77. W. K. den Otter and S. A. Shkulipa, *Biophys. J.*, 2007, **93**, 423.
78. M. C. Watson, E. G. Brandt, P. M. Welch and F. L. H. Brown, *Phys. Rev. Lett.*, 2012, **109**, 028102.
79. J. F. Nagle, *Faraday Discuss.*, 2013, **161**, 11.
80. L. Fernandez-Puente, I. Bivas, M. D. Mitov and P. Méléard, *Europhys. Lett.*, 1994, **28**, 181.
81. M. Hu, D. H. de Jong, S. J. Marrink and M. Deserno, *Faraday Discuss.*, 2013, **161**, 365.

82. D. Boal, *Mechanics of the Cell*, Cambridge University Press, Cambridge, 2002.
83. W. Rawicz, K. C. Olbrich, T. McIntosh, D. Needham and E. Evans, *Biophys. J.*, 2000, **79**, 328.
84. E. Chacón, P. Tarazona and F. Bresme, *J. Chem. Phys.*, 2015, **143**, 034706.
85. B. Jójárt and T. A. Martinek, *J. Comput. Chem.*, 2007, **28**, 2051.
86. L. Janosi and A. A. Gorfe, *J. Chem. Theory Comput.*, 2010, **6**, 3267.
87. D. Poger and A. E. Mark, *J. Chem. Theory Comput.*, 2010, **6**, 325.
88. J. H. Irving and J. G. Kirkwood, *J. Chem. Phys.*, 1950, **18**, 817.
89. W. Noll, *Indiana Univ. Math. J.*, 1955, **4**, 627.
90. R. B. Lehoucq and A. Von Lilienfeld-Toal, *J. Elasticity*, 2010, **100**, 5.
91. J. M. Vanegas, A. Torres-Sánchez and M. Arroyo, *J. Chem. Theory Comput.*, 2014, **10**, 691.
92. C. Das, M. G. Noro and P. D. Olmsted, *Biophys. J.*, 2009, **97**, 1941.
93. J. A. McCammon, B. R. Gelin and M. Karplus, *Nature*, 1977, **267**, 585.
94. D. Trzesniak and W. F. van Gunsteren, *Protein Sci.*, 2006, **15**, 2544.
95. B. Mao, R. Tejero, D. Baker and G. T. Montelione, *J. Am. Chem. Soc.*, 2014, **136**, 1893.
96. A. Glättli, X. Daura, D. Seebach and W. F. van Gunsteren, *J. Am. Chem. Soc.*, 2002, **124**, 12972.
97. The NMRlipids project, <http://nmrlipids.blogspot.co.uk> (accessed August 2016).
98. A. I. I. Tyler, J. A. Clarke, J. M. Seddon and R. V. Law, *Solid State NMR of Lipid Model Membranes*, Methods Molecular Biology series, Springer Science, 2015, vol. 1232, p. 227.
99. J. Douliez, A. Ferrarini and E. Dufourc, *J. Chem. Phys.*, 1998, **109**, 2513.
100. R. Guixà-González, I. Rodríguez-Espigares, J. M. Ramírez-Anguita, P. Carrió-Gaspar, H. Martínez-Seara, T. Giorgino and J. Selent, *Bioinformatics*, 2014, **30**, 1478.
101. N. Michaud-Agrawal, E. J. Denning, T. B. Woolf and O. Beckstein, *J. Comput. Chem.*, 2011, **32**, 2319.
102. M. Karplus, *J. Chem. Phys.*, 1959, **30**, 11.
103. X. Daura, K. Gademann, B. Jaun, D. Seebach, W. F. van Gunsteren and A. E. Mark, *Angew. Chem., Int. Ed.*, 1999, **38**, 236.
104. O. F. Lange, D. van der Spoel and B. L. de Groot, *Biophys. J.*, 2010, **99**, 647.
105. F. Mocci and A. Laaksonen, *Nucl. Magn. Reson.*, 2015, **44**, 592.
106. K. Lindorff-Larsen, R. B. Best, M. A. Depristo, C. M. Dobson and M. Vendruscolo, *Nature*, 2005, **433**, 128.
107. C. Tang, J. Iwahara and G. M. Clore, *Nature*, 2006, **444**, 383.
108. S. Esteban-Martin, R. B. Fenwick, J. Áden, B. Cossins, C. W. Bertoncini, V. Guallar, M. Wolf-Watz and X. Salvatella, *PLoS Comput. Biol.*, 2014, **10**, 1.
109. M. R. Jensen, M. Zweckstetter, J. R. Huang and M. Blackledge, *Chem. Rev.*, 2014, **114**, 6632.

110. J. Silvestre-Ryan, C. W. Bertoncini, R. B. Fenwick, S. Esteban-Martin and X. Salvatella, *Biophys. J.*, 2013, **104**, 1740.
111. R. B. Fenwick, S. Esteban-Martín and X. Salvatella, *Eur. Biophys. J.*, 2011, **40**, 1339.
112. R. B. Fenwick, L. Orellana, S. Esteban-Martín, M. Orozco and X. Salvatella, *Nat. Commun.*, 2014, **5**, 1.
113. M. Tang, A. E. Nesbitt, L. J. Sperling, D. A. Berthold, C. D. Schwieters, R. B. Gennis and C. M. Rienstra, *J. Mol. Biol.*, 2013, **425**, 1670.
114. J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff and D. C. Phillips, *Nature*, 1958, **181**, 662.
115. Protein Data Bank statistics, [http://www.rcsb.org/pdb/static.do?p=general\\_information/pdb\\_statistics/index.html](http://www.rcsb.org/pdb/static.do?p=general_information/pdb_statistics/index.html), (accessed August 2016).
116. A. I. I. Tyler, R. V. Law and J. M. Seddon, *X-ray Diffraction of Lipid Model Membranes*, Methods in Molecular Biology series, Springer Science, 2015, vol. 1232, p. 199.
117. M. H. F. Wilkins, A. E. Blaurock and D. M. Engelman, *Nat. New Biol.*, 1971, **230**, 72.
118. M. C. Wiener, R. M. Suter and J. F. Nagle, *Biophys. J.*, 1989, **55**, 315.
119. M. C. Wiener and S. H. White, *Biophys. J.*, 1991, **59**, 174.
120. J. B. Klauda, N. Kučerka, B. R. Brooks, R. W. Pastor and J. F. Nagle, *Biophys. J.*, 2006, **90**, 2796.
121. R. W. Benz, F. Castro-Román, D. J. Tobias and S. H. White, *Biophys. J.*, 2005, **88**, 805.
122. N. Kučerka, J. Katsaras and J. F. Nagle, *J. Membr. Biol.*, 2010, **235**, 43.
123. L. V. Bock, C. Blau, G. F. Schröde, I. I. Davydo, N. Fischer, H. Stark, M. V. Rodnina, A. C. Valiana and H. Grubmüller, *Nat. Struct. Mol. Biol.*, 2013, **20**, 1390.
124. S. Plimpton, *J. Comput. Phys.*, 1995, **117**, 1.
125. J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé and K. Schulten, *J. Comput. Chem.*, 2005, **26**, 1781.
126. S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess and E. Lindahl, *Bioinformatics*, 2013, **29**, 845.
127. G. Bussi, F. L. Gervasio, A. Laio and M. Parrinello, *J. Am. Chem. Soc.*, 2006, **128**, 13435.
128. Y. Shan, E. T. Kim, M. P. Eastwood, R. O. Dror, M. A. Seeliger and D. E. Shaw, *J. Am. Chem. Soc.*, 2011, **133**, 9181.
129. V. A. Jackson, S. Mehmood, M. Chavent, P. Roversi, M. Carrasquero, D. Del Toro, G. Seyit-Bremer, F. M. Ranaivoson, D. Comoletti, M. S. P. Sansom, C. V. Robinson, R. Klein and E. Seiradake, *Nat. Commun.*, 2016, **7**, 11184.
130. A. C. Kalli, T. Rog, I. Vattulainen, I. D. Campbell and M. S. P. Samson, *J. Membr. Biol.*, 2016, DOI: 10.1007/s00232-016-9908-z.
131. M. Caffrey and G. W. Feigenson, *Biochemistry*, 1981, **20**, 1949.



132. Y. Sugita, N. Miyashita, M. Ikeguchi, A. Kidera and C. Toyoshima, *J. Am. Chem. Soc.*, 2005, **127**, 6150.
133. A. Lervik, F. Bresme and S. Kjelstrup, *Phys. Chem. Chem. Phys.*, 2012, **14**, 3543.
134. A. Lervik, F. Bresme, S. Kjelstrup and J. M. Rubi, *Biophys. J.*, 2012, **103**, 1218.
135. Y. Ito and M. Ikeguchi, *Adv. Exp. Med. Biol.*, 2014, **805**, 411.
136. M. Karplus and J. Kuriyan, *Proc. Natl. Acad. Sci.*, 2005, **102**, 6679.
137. J. V. Møller, C. Olesen, A. L. Winther and P. Nissen, *Q. Rev. Biophys.*, 2010, **43**, 501.
138. A. E. Nel, L. Mädler, D. Velegol, T. Xia, E. M. V. Hoek, P. Somasundaran, F. Klaessig, V. Castranova and M. Thompson, *Nat. Mater.*, 2009, **8**, 543.
139. J. Wong-Ekkabut, S. Baoukina, W. Triampo, I. M. Tang, D. P. Tieleman and L. Monticelli, *Nat. Nanotechnol.*, 2008, **3**, 363.
140. N. Nisoh, M. Karttunen, L. Monticelli and J. Wong-ekkabut, *RSC Adv.*, 2015, **5**, 11676.
141. D. Mohammad-Aghaie, E. Mace, C. A. Sennoga, J. M. Seddon and F. Bresme, *J. Phys. Chem. B*, 2010, **114**, 1325.
142. J. Poehlsgaard and S. Douthwaite, *Nat. Rev. Microbiol.*, 2005, **3**, 870.
143. K. Y. Sanbonmatsu and C. S. Tung, *J. Phys.: Conf. Ser.*, 2006, **46**, 334.
144. J. R. Perilla, B. C. Goh, C. K. Cassidy, B. Liu, R. C. Bernardi, T. Rudack, H. Yu, Z. Wu and K. Schulten, *Curr. Opin. Struct. Biol.*, 2015, **31**, 64.
145. P. L. Freddolino, A. S. Arkhipov, S. B. Larson, A. McPherson and K. Schulten, *Structure*, 2006, **14**, 437.
146. J. R. Perilla, J. A. Hadden, B. C. Goh, C. G. Mayne and K. Schulten, *J. Phys. Chem. Lett.*, 2016, **7**, 1836.
147. Y. Andoh, N. Yoshii, A. Yamada, K. Fujimoto, H. Kojima, K. Mizutani, A. Nakagawa, A. Nomoto and S. Okazaki, *J. Chem. Phys.*, 2014, **141**, 165101.
148. A. Arkhipov, P. L. Fredolino and K. Schulten, *Structure*, 2006, **14**, 1767.
149. T. Reddy and M. S. P. Sansom, *Structure*, 2016, **24**, 375.
150. S. C. Harvey, A. S. Petrov, B. Devkota and M. B. Boz, *Methods Enzymol.*, 2011, **487**, 513.

## CHAPTER 3

# *Designing Chemical Tools with Computational Chemistry*

SILVIA RINALDI AND GIORGIO COLOMBO\*

Istituto di Chimica del Riconoscimento Molecolare, CNR, Via Mario Bianco 9, 20131 Milano, Italy

\*Email: g.colombo@icrm.cnr.it

## 3.1 Introduction

Proteins carry out their functions and participate in biochemical networks by switching among a limited number of structural sub-states, which favour the adaptation of the protein to different partners, environmental requirements and modulate reactivity in the case of enzymes.<sup>1–3</sup>

The dynamic interconversion among different sub-states allows biomolecules to fine-tune their functions in response to varying conditions. Changes in conformational and dynamic states are induced by several biochemical factors, including ligand-binding.<sup>1,4</sup> This realisation opens up for chemical biologists the perspective of designing and synthesising new molecular entities (NMEs).<sup>5,6</sup> These chemical tools reveal new biological functions or modulate specific cellular pathways by exploiting structural features or by tweaking the conformational dynamics of the proteins they target. In this framework, the new fundamental challenge for chemical discovery consists of identifying privileged structures capable of selectively interfering with key functional sub-states of proteins involved in signalling pathways, thus allowing to tune, and not only inhibit, entire signalling cascades that control cell life.<sup>7</sup> In chemical and molecular design terms, this represents a shift in paradigm that poses a number of new challenges and opportunities.

---

Chemical Biology No. 3

Computational Tools for Chemical Biology

Edited by Sonsoles Martín-Santamaría

© The Royal Society of Chemistry 2018

Published by the Royal Society of Chemistry, [www.rsc.org](http://www.rsc.org)

In this context, computational chemistry methods, ranging from classical structure-based techniques to the approaches integrating bioinformatics, from conformational analyses to recent theoretical progresses in the identification of functional sub-states through the analysis of protein internal dynamics and energetics, play a key role, especially when aptly merged with experimental drug design and validation methods.<sup>8</sup> Developing and validating new computational methods for chemical biology will allow both the chemistry and biology communities to enhance their understanding of complex biomolecular systems and increase the efficiency with which appropriate chemical modulators of signalling pathways can be identified. Importantly, this might have several consequences, leading to, for instance, the development of global pathway regulators with unique therapeutic potential, or to small molecules that modulate the functions of enzymes required for biotechnological applications of, *e.g.*, whole bacterial cells.

The validation of new molecular design concepts and experimental analysis of the resulting molecules with tests ranging from binding studies to the *in vivo* and *in vitro* characterisation of the modulator effects is required to make computational chemical biology an actual tool applicable in different laboratory contexts.<sup>9</sup>

From a general perspective, recent scientific advances stimulated by the genomic and post-genomic revolution have translated into a vast resource of information to analyse, that goes from three-dimensional structural repositories to databases of chemical and biological information. Among the most important milestones for chemical tool discovery and design were the unveiling of the genome, proteome and key functionally related super-families of targets for different organisms and pathogens. This wealth of data was expected to start a new era in molecular discovery, drug design and disease treatment. Chemistry was at the centre of the action, asked to mine the structural and functional information on proteins and translate it into novel effectors of their activity. Despite considerable advances in the know-how of molecular design, the heralded new era has had a slow start.

In the important case of disease treatment, for instance, it has clearly emerged that, with the exception of a few monogenic pathologies (*i.e.*, driven by one single factor), the problem is more complex than the simple identification of a single new target and the consequent design of a drug that may individually modulate the target activity. It has been proposed that the limitations of the 'one target, one drug' rule may explain the low numbers of newly discovered active drugs in recent years.<sup>10,11</sup> Better knowledge of fundamental genes and proteins currently suggests that signalling pathways in biological systems operate as modular networks,<sup>12</sup> in which a limited number of proteins act as *hubs* connecting different pathways. Different regulatory cascades can then be converged onto *hub proteins* whose inhibition/modulation can affect multiple signalling pathways commensurate with the administration of multiple drugs.

In the case of synthetic biology, the ability to regulate metabolic pathways by the controlled administration of small molecules that finely regulate the

activity and interactions of their targets may pave new ways for biotechnological applications, *e.g.*, in bacteria used for the production of intermediates of high added value products.

Exploiting protein 'connectivity maps', both at the level of structure and dynamics, can thus provide better tools and approaches for molecular discovery, improving both the low yield, elevated costs, and high risk of failures of traditional drug screening,<sup>13</sup> and generating new opportunities for industrial applications.

In parallel, a novel model of protein binding and complexation has emerged, which involves pre-existing equilibria among different conformational states and the role of dynamics in facilitating and triggering structural transitions.<sup>14</sup> In this view, perturbations induced by ligand-binding (as well as by covalent modification or binding to a partner-protein) lead to the activation of specific conformational states that meet functional requirements. Allosteric events are the prime mechanisms by which fine conformational regulation can be achieved: modifications at one site are propagated through the structure. The end result may be a shift in the structural population with the activation of dynamic states that encode specific functions. Understanding the chemical determinants of molecular recognition and their consequences on conformational dynamics is thus an essential requirement to integrate systems biology and the investigation of protein interaction networks into new concepts of chemical biology and molecular design.<sup>15</sup>

Besides the aforementioned advances, one important aspect to consider when thinking about the design and exploitation of chemical tools for modern chemical biology is the complexity of layers of structural and functional regulation that characterise cell life. Indeed, one should take into account that each cell possibly contains a mix of isoforms of a certain protein, whose properties and roles are differentiated by its interactor(s) (other proteins, membranes, substrates *etc.*), the conformation it populates, and the very cellular compartment where the protein is found. It must be noted that this complexity deeply influences the biological effect of any designed chemical tool as well as the cellular or *in vivo* phenotype that may be observed. Indeed, one may imagine that a small molecule targeting a subset of states of a certain protein will determine a phenotype that could potentially be distinct from the one obtainable by another small molecule targeting other subsets of states, or different isoforms.

The main chemical challenges generated by these concepts entail the development of novel methodologies for (1) the investigation and characterisation of the recognition properties of targets that may generate new pharmacological profile(s) and network regulation; (2) molecular discovery and design, based on the explicit characterisation of functional dynamics and its influence on possible interaction sites of the target protein; (3) the development of novel classes of compounds that are able to directly modulate the distribution of functional sub-states through rationally selected functional groups. Meeting these challenges will contribute directly to moving the role of computational chemistry in the investigation of complex

biomolecular systems to a new level. The overall perspective is the definition of a new set of rules for the rational discovery and development of new chemical tools able to modulate (and not only disrupt) cell functions.

In this general conceptual framework, the present chapter will focus on describing examples of the application structure- and computation-based methods to design small-molecule chemical tools that can be used to advance our understanding of the complexities of cell life, with perspective applications outside strictly defined chemical biology.

## 3.2 Structure Based Approaches for Chemical Biology

To start tackling the structural complexity underlying cell regulation, we will first concentrate on examples of selective targeting of different isoforms of the same protein.<sup>16–18</sup>

The most direct approach to develop small molecule chemical tools to help unveil the unknown or different roles of related protein isoforms starts from the analysis of their conformational differences. This may reveal pockets for ligand binding that change among targets or even change depending on the conformational states. By interacting with different structures of otherwise similar proteins, suitably designed ligands may select and consequently perturb distinct functionally oriented conformations. These events may have various consequences, at the single protein, protein–protein interaction, metabolic and phenotypic levels.

One paradigmatic realisation of these concepts is represented by the work of the Chiosis group in the development of paralog-specific inhibitors of the Hsp90 family of proteins.<sup>19</sup> These homodimeric molecular chaperones are highly conserved from bacteria to eukaryotes and oversee the correct folding and conformational maturation of several client proteins (mostly involved in signal transduction and regulatory cascades).<sup>20–23</sup> Different paralogs in the cell control protein folding in various compartments: in human cells, cytosolic Hsp90<sup>24</sup> exists in the alpha and beta isoforms, in the endoplasmatic reticulum (ER) Grp94<sup>25,26</sup> is the major representative of the family, while the mitochondrial form is called TRAP1 and its roles and interactions are currently a hot research subject.<sup>27</sup>

Hsp90 functions depend on ATP binding and hydrolysis, and impact cell proliferation, survival, and stress response.<sup>28</sup> This chaperone thus plays a pivotal role for controlling the intersections of multiple signalling networks associated with cell viability.<sup>29–31</sup> At the same time, its deranged activity has been associated with the development of several types of cancers. In fact, overexpression and up-regulation of Hsp90 activity is known to promote tumour cell survival and adaptation. Accordingly, a prominent strategy for the development of cancer chemotherapeutics is focused on the inhibition of Hsp90 in tumour cells. Indeed, clinical trials are currently being performed for several inhibitors targeting the N-terminal ATP-binding site.<sup>32</sup> The full exploitation of such drug-leads in the clinical setting is however hampered by dosage, tolerance and solubility problems.<sup>32–34</sup> Current

inhibitors cannot distinguish between the different Hsp90 isoforms, making their utility sometimes questionable. Indeed, it has been shown that some cancers express less TRAP1 than do their normal tissue counterparts, suggesting that TRAP1 function in mitochondria of normal and transformed cells is more complex. In some cases, TRAP1 has even been proposed to have a tumour suppressor role. It is clear from these simple examples that the availability of isoform-specific inhibitors (or modulators) would offer a chance to overcome the limitations and toxicities of currently used drugs, and would provide novel information on the mechanistic roles of the different chaperones at a very fine level of detail.<sup>35</sup>

The Chiosis group used a combined screening and structural modelling approach to identify ligands based on a common purinic scaffold that showed a 100-fold higher binding selectivity for the endoplasmic reticulum member Grp94 over the cytosolic Hsp90. Structural and computational analyses showed that such ligands bind to a novel pocket in Grp94 close to the ATP-binding site, which is not found in the cytosolic paralogs (Figure 3.1). Binding of the ligands turned out to freeze the local conformation of Grp94, a conformation that the authors define as inaccessible to the cytosolic Hsp90. Most interestingly, this pocket is also different from that of NECA and BnIm,<sup>16,36</sup> the previously available Grp94 inhibitors.

A fundamental aspect of this work is that the observed structure-based paralog-selectivity of the new molecules is maintained also at the cellular level. Through a series of carefully conducted chemical biology experiments, the authors discovered a Grp94-dependent mechanism for the regulation of the protein HER2, an important cancer promoter. Although HER2 was previously known to be a client of Hsp90, the fine molecular details of its dependence on Grp94 were not. The designed specific agents were also used to delineate the roles of Grp94 in multiple myeloma.<sup>37</sup> Similarly, Stamos and

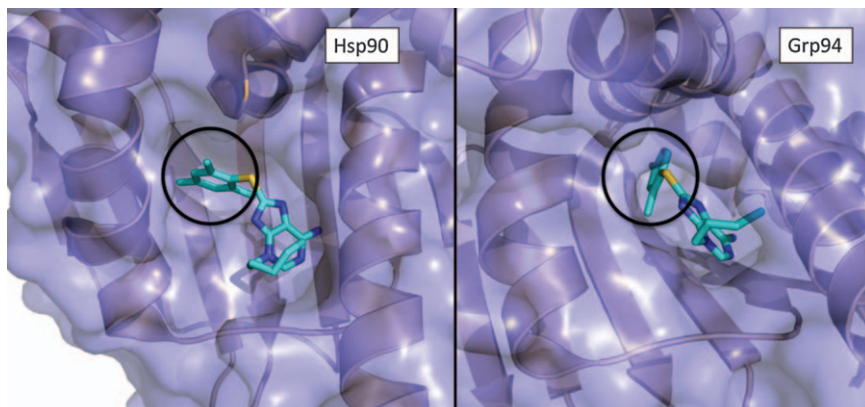


**Figure 3.1** Structural representations of the complex between Hsp90 and Grp94 with the small molecule inhibitor PU-H54 indicate different modes of binding and a different shape of the binding pockets.

co-workers exploited a structure-based drug design strategy to optimise a benzolactam series of Hsp90  $\alpha/\beta$  inhibitors to achieve >1000-fold selectivity versus GRP94 and TRAP1. These agents were used to demonstrate a role for cytosolic Hsp90 in promoting the clearance of mutant huntingtin protein and to propose cytosolic Hsp90 as a potential target in Huntington disease.<sup>38</sup>

These first examples, while centered on the specific case of molecular chaperones, are fully general in spirit and show the potentialities of modelling structural approaches to discover new small molecules that in turn spur novel ways to investigate important cellular pathways. Importantly, such new molecules allow to modify the target protein properties in its endogenous environment, without perturbing the proteome by protein deletion, over-expression or down regulation.

It is important to note at this point that the advancements of docking methods and the extension of the chemical space covered by compound libraries, combined with the availability of carefully curated open access resources make it possible to screen libraries of millions of molecules against several targets in a matter of days to weeks. While still limited by a number of factors (energy functions are strongly dependent on the parameterization strategies, they cannot calculate affinities accurately, especially for very similar compounds, and consequently they often fail at quantitatively rank-ordering molecules), modern docking approaches can separate ensembles of likely from ensembles of unlikely ligands, often with relevant hit rates above 10%.<sup>18</sup> This is now possible thanks to improvements in libraries, target quality, and methods. On the methodological side, important advances are represented by the improvements in the description of interaction terms with the addition of *e.g.*, polarisation, inclusion of water molecules, consideration of protonation states, treatment of metal centres, sampling approaches often derived from high level physics studies. A great push to the computational discovery of new small molecule chemical tools to probe the activities of new targets has come from the growth of publically available chemical libraries. The most prominent example is represented by the ZINC database, which contains over 100 million molecules, which have been stored and curated in three-dimensional, parameterised, dockable structures, often representing multiple accessible conformations. Finally, it should be mentioned that potential chemical space coverage may be further increased by the use of fragment based approaches, whereby small fragments representing the principal chemical functionalities, with typical molecular weights around 250, are used to probe binding sites.<sup>39</sup>

### 3.3 Structural Dynamics as a Source of Novel Chemical Tools

The examples of computational discovery of new chemical entities discussed so far are based on the use of single, static structures of the protein targets. However, proteins are not static entities. On the contrary, they are highly



dynamic molecules that can access a number of different (ensembles of) states available on their free energy landscape. Indeed, proteins carry out functions and participate in interaction networks by dynamically switching amongst diverse conformational sub-states. This allows biomolecules to tune their function for different conditions. In this context, perturbations induced by ligand-binding, covalent modifications, or protein-protein interactions can lead to the activation of specific conformational states. Careful analysis of dynamic properties may thus provide novel opportunities to design molecules that specifically target a certain dynamic sub-state or modulate the relative populations of different conformational ensembles with possible downstream effects at the biochemical, cellular and phenotypical levels.<sup>40,41</sup>

Allostery defines one of the most relevant and fine-tuned control mechanisms of cell processes: it represents regulation at a distance, whereby the affinity of a biomolecule towards a primary binding is modulated by binding an 'effector' at a different distant position, namely the allosteric site. This modulation may cause increase or decrease of protein activity and its downstream effects in the cell.<sup>7,42-44</sup>

The availability of a general framework to design molecules that meet the specific structural/dynamical constraints required to modulate a certain function is thus a clear necessity and opportunity towards innovative therapeutics and chemical tools. A full understanding of the roles of different sub-states of a molecular system will allow a more rational design of chemical biology experiments apt to modulate a specific interaction. This will ultimately result in a major control of the response obtained by any system in which the interaction is involved. In this context, one of the principal predicaments requires the identification and characterisation the protein sites that are not immediately evident from the analysis of the crystal structure of the protein that can be targeted to interfere with specific conformations: such a task can be pursued through an understanding of internal dynamics and allosteric coordination between different parts of the protein.<sup>45,46</sup> This aspect should then be complemented by the analysis of the receptor's conformational evolution and response to a given ligand. The interaction can in fact result in a mutual impact on structural properties whereby protein flexibility determines the ligand conformation while the ligand pose dictates the conformation of the binding site.<sup>47,48</sup>

As an example of the combination of the two aforementioned concepts, we will present work from our own experience. In an effort aimed at discovering new allosteric modulators of the Hsp90 molecular chaperone, and to include dynamic information in the discovery of new allosteric sites, we started by analysing the extent to which distinct subdomains of the *full-length* protein are mechanically coupled, and hence capable of propagating signals that, upon ligand-binding to a specific site, cause conformational responses in distal regions. MD investigations of the apo, ATP- and ADP-bound states of *full-length* Hsp90 based on the analysis of

correlated motions revealed that positive long-range cross-correlations could extend well beyond local elements and intra-domain regions in the ATP-state, indicating the presence of a diffuse interaction network. The functional consequence is the stabilisation of the closed ‘tense’ state by favouring the Hsp90 dimer complex to move as a coherent rigid body. In the ADP-state the positive correlation was lost and the two protomers moved in an anti-correlated way. This indicated collective motions towards opposite directions and a consequent opening of the clamp. In this context, a general theoretical model was developed to identify the ‘hotspots’ active in the allosteric communication between a certain binding site and distal regions of the protein implied in function. The cross-talk between the N- and C-domains was investigated by defining the communication propensity (CP) between any two residues as a function of fluctuation of their distance components.<sup>49–51</sup> CP is defined as:

$$CP_{ij} = \langle (d_{ij} - \langle d_{ij} \rangle)^2 \rangle \quad (3.1)$$

Where averages are taken along the MD trajectory.

CP describes a communication time; therefore, low CP values are related to efficiently communicating residues. Hotspots for allosteric transduction were identified by calculating for each residue the fraction of all other residues that have high communication efficiency with it at distances larger than increasing cutoffs.<sup>49</sup> This analysis illuminated different allosteric pathways that selectively depend upon the ligand identity: specific clusters of residues participate in the signal transduction from the N-terminal binding site (NTD) to the C-terminal binding site (CTD). In the ATP-bound, active state, long-range communication from the binding site was directed to residues at the CTD interface (Figure 3.2A). The latter region appears to be involved in client-protein binding, and has been shown to be the target of post-translational modifications. Evidence for efficient molecular communication between NTD residues and CTD regions in the activated state is of special importance given the growing interest in developing novel and specific inhibitors targeting allosteric Hsp90 regions. The hotspot C-terminal region was subjected to structural investigation and one pocket (Pocket A) was consistently detected in all representative MD conformations, located at the dimer interface and suitable to accommodate small compounds that interact directly with the hotspot allosteric residues. The information on signal transduction, on the conformational states spanned by hotspots in Pocket A, and the analysis of their chemical properties were used to develop pharmacophore models for virtual screening of drug databases. The pharmacophore recapitulated the complementary interactions necessary to guarantee productive binding with the putative allosteric site, and was used to screen the NCI repository. Out of 290 000 compounds, filtering with the pharmacophore returned 36 hits, corresponding to 0.01% of the database. These compounds showed affinity for the Hsp90 C-terminal domain, had effects on the viability of cancer cells and no effect on normal cells, induced degradation of specific Hsp90 client proteins, and could

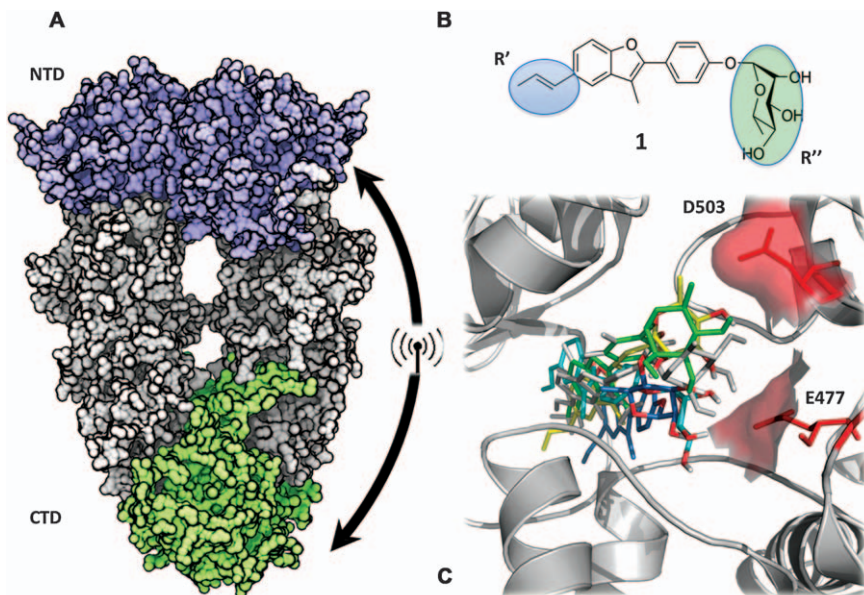


**Figure 3.2** General scheme of the use of MD simulations in the study of allostery and in allosteric drug discovery.

disrupt Hsp90 association with co-chaperones.<sup>45</sup> Experimental tests showed that the selected molecules could inhibit several important protein–protein interactions and demonstrated the ability to interrupt biological pathways fundamental to cancer cell proliferation. Since it is mechanically connected with the ATP-processing active site, the discovered C-terminal site may represent a point of allosteric regulation of Hsp90 and the designed small molecules may allosterically change the functional properties of the protein. Experiments, as well as new calculations, were then carried out to investigate this intriguing point. Defining the principles of ligand recognition and allostery is actually important for understanding the molecular links between small-molecule binding and observed effects. Moreover, the rational search for allosteric modulators of protein functions represents a relatively young field. Allosteric modulators may in fact even augment the enzymatic or binding properties of their targets generating new information on the role of the selection of a certain state in a signalling pathway. From the drug design point of view, allosteric activators can in principle modulate receptor function while still allowing the possibility of orthosteric agonist or antagonist binding. In this context, allosteric ligands can act as affinity modulators by changing the affinity of the orthosteric ligand for the receptor, as well as induce conformational changes that in turn modulate the efficacy of the orthosteric ligand in determining cellular responses. In this frame of thought, a library of small molecules was designed based on the *O*-aryl rhamnoside **1** (Figure 3.2B) to target the

allosteric site located 65 Å away from the active site, and the members of the library were shown to *activate* Hsp90 ATPase. The characterization of the stereochemical and dynamic properties of the allosteric site constituted the basis for the optimization of the allosteric ligands, generating compounds able to accelerate the chaperone enzymatic activity up to six times.<sup>52,53</sup>

The binding determinants of **1** in the allosteric site were first characterised through docking calculations combined to multiple structure analysis extracted from MD simulations. Indeed, considering the flexibility of the protein and of the allosteric pocket,<sup>54</sup> we selected an ensemble approach to characterise chaperone–ligand interactions. The minimum energy pose of the **1**/Hsp90 complex was used as input for long timescale MD simulations, including ATP at the NTD. The aim was to identify the hotspots of the allosteric site where key functional groups on the ligand best complement the receptor, taking the dynamic exchange between the binding partners explicitly into account. Structural cluster analysis of the resulting trajectory showed that the first ten clusters recapitulated 95% of the protein structural variability, with a relatively variable RMSD between visited pocket conformations, revealing the conformational diversity induced by **1**. Such differences, due to the cross-talk between the protein and the ligand, is not to be expected *a priori* and shows the importance of including dynamics in describing molecular recognition events. The structures obtained by cluster analysis were used as targets for re-docking **1**. The resulting structural ensembles were employed to generate a consensus model of Hsp90 residues and functional groups on **1** that define the most relevant stabilising contacts (Figure 3.2C). Two areas where functional group diversification on the lead could translate into a modulated response of the chaperone were identified: the carbohydrate moiety (R'', Figure 3.2B), which in the model points towards a negatively charged area (Figure 3.2C), and the propenyl-group (R', Figure 3.2B), which occupies a hydrophobic pocket. To take these properties into account, we designed glycomimics of the first ligand, as well as non-glyco derivatives. Interestingly, all designed compounds were shown to have an activating effect on Hsp90 enzymatic activity and conformational dynamics, proving that they were indeed genuine allosteric modulators of the chaperone.

Importantly, the stimulatory activities observed on the isolated protein were found to determine cytotoxic effects in cancer cell lines, which are known to depend on a well-balanced Hsp90 regulation. In particular, designed activators were proven to have an effect on the viability of cancer cells made resistant to clinically used inhibitors, indicating the possibility of anticancer interventions based on novel mechanisms of action.

The important message emerging from these combined endeavours is that perturbing a system not only by artificial inhibition but also by acceleration can interfere with protein–protein recognition processes, substrate processing and as consequence be potentially used to ask new questions on the roles of protein kinetic modulation in therapeutic interventions.

### 3.4 Combining Bioinformatics, Chemoinformatics and Structural Information to Explore Protein Functions

As stated in the previous paragraphs, one of the main goals for computational chemistry in the post-genomic era is to use small molecules to facilitate the characterisation of unknown biological functions. The analysis of the vast amount of available data using the tools of chemo- and bio-informatics provides new opportunities for the identification of protein functions. In this context, work by the Shoichet group has developed a new approach that quantitatively groups and relates proteins based on the chemical similarity of their ligands (the Similarity Ensemble Approach, SEA; <http://sea.bkslab.org>, see Figure 3.3).<sup>55</sup> Their work starts with 65 000 drug ligands that are grouped in different sets for hundreds of protein targets,

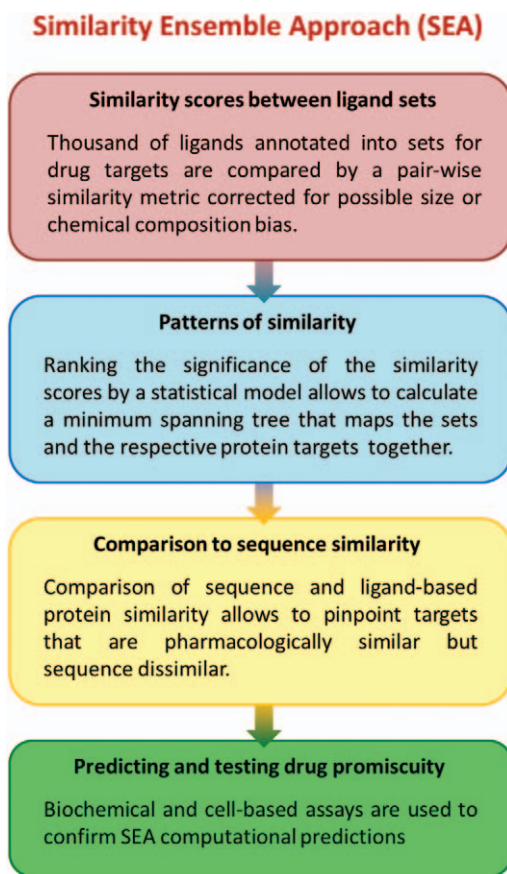


**Figure 3.3** Workflow showing how the statistical ensemble approach (SEA) and other related methods work.

and a similarity score between each set is calculated based on ligand topology. A statistical model is used to rank the significance of the resulting similarity scores. This method is similar in spirit to the BLAST E-value typically used to relate protein sequence similarity, whereby the sets are compared without size or chemical composition bias and allowing for a correction that accounts for the chemical similarity one might expect between ligand sets at random. Based on these results a minimum spanning tree maps the sets together.<sup>55</sup>

The most relevant (and fascinating) aspect of the results obtained is that, although the maps are connected only by chemical similarities, biologically meaningful clusters appeared. In the first application, interestingly, links among unexpected targets also emerged, among them that methadone, emetine and loperamide (Imodium) may antagonize muscarinic M3,  $\alpha 2$  adrenergic and neurokinin NK2 receptors, respectively. Very importantly, these predictions could subsequently be confirmed *via* biochemical and cell-based experiments. As the authors clearly point out, 'Relating receptors by ligand chemistry organizes biology to reveal unexpected relationships that may be assayed using the ligands themselves'.<sup>55</sup> This is indeed a very important point, as it shows that computational approaches can help researchers ask new questions, explore unknown relationships and discover novel uses for already known compounds. The latter counts particularly in the search for new applications of already known drugs, whose safety and toxicity profiles have already been characterised, which can aptly speed up the search for molecules that are immediately portable to clinical application. Another appealing aspect of SEA is that unexpected similarities between biological targets revealed through ligand similarities can provide useful information on the off-target activities of the ligands themselves. In this context, it is important to note that pharmacological clustering of protein targets is an emergent property of the approach: no biological information is in fact used to relate ligands, yet there are enough connections among the different ligand sets to link them, in some cases through sequential linkages, in coherent maps of pharmacologically interesting chemical space. This brings about valuable information on the clusters of targets that differs from that one might obtain from comparing protein structures or already known biological information. Importantly, the discovered connections among the ligand sets have implications for understanding the effects of bioactive molecules, and lead to testable hypotheses that are then verified experimentally. One important caveat is that such procedure would be prone to generate lots of artefacts if the control for random similarity among ligands is not introduced.

The observation that similar bioactive molecules are more likely to share similar targets is also at the basis of the method introduced by Michielin and Zoete. An interesting development of their approach, implemented in the publicly available server <http://www.swisstargetprediction.ch>, is the inclusion of the biophysical bases for molecular recognition as a parameter for similarity evaluation: indeed, ligand shape, conformations and electrostatic



potential define whether a molecule is fit to bind a certain site, besides chemical similarity based only on 2D molecular descriptors.<sup>56</sup> Therefore the authors combine both 2D and 3D similarity measures, the latter including 20 different conformations, to define a similarity score between ligands which is shown to significantly increase target prediction accuracy. Such scores are used to rank predicted targets. This method has been carefully validated on sets of known protein–ligand interactions, primarily retrieved from the ChEMBL database version 16 using stringent criteria on affinity and binding to remove ambiguous cases. An interesting feature of SwissTargetPrediction is the possibility to carry out predictions in five different organisms, as well as to map predictions based on protein homology, both within and between organisms.

### 3.5 Deep Networks and Big Data in the Discovery of New Drugs and Chemical Tools

The unprecedented availability of computer power and the continuous increase in the possibility to process enormous amounts of (heterogeneous) data is ushering the start of a new era, in which computer science approaches will play a critical part in the discovery of new molecules and the prioritisation of their applications as drugs, chemical tools, or biotech products. In this context, machine learning methods, and in particular deep learning,<sup>57</sup> have recently provided vivid examples of how successful informatics can be at the discovery and prioritisation of bioactive compounds. In general, such methods involve a three-step scheme whereby *first*, a set of molecular descriptors apt to capture the fundamental chemical and stereo-electronic properties related to the required function is selected; *second*, a suitable scoring function is used to compare/rank the molecules; *third*, a machine learning approach is applied to extract the relevant features able to proficiently correlate the presence of a certain set of molecular features to biological activity or inactivity.

Originated in the 1960s to make computers imitate the way the brain works and learns, machine-learning methods have been widely applied in software development and engineering problems. In the drug discovery realm, until recently, such methods have focused mainly on cheminformatics applications to unveil the relationships between the molecular descriptors used and the measured activity of the compounds to obtain qualitative classifiers or quantitative structure–activity relationship (QSAR) models. Feature extraction from the descriptor patterns is the decisive step in the model development process.<sup>58,59</sup> Common architectures involve principle component analysis, clustering, decision trees, multivariate linear regression, support vector machines (SVMs) and other related neural network approaches.<sup>60</sup> In these contexts, one single feature transformation step is used to project the feature into a suitable space for defining activity. While successful in many applications, this shallowness of the network (one single



layer of transformation lays between the molecular representation and the function descriptor) may represent a limitation when applied to more complex systems and data. A conceptual shift from focusing on the importance of selecting the descriptor to the engineering of the architecture of the machine-learning approach and of the training algorithms, has determined the breakthrough of ‘deep learning’ or deep neural network approaches.<sup>57</sup>

Deep networks feature multiple hidden layers, in contrast to the single one used to map chemical feature to function described above. Through this difference, they are capable of computing layers of adaptive non-linear features that capture increasingly complex data patterns with each additional layer. Such advances have been particularly impacting in the areas of pattern recognition and computer vision (see Figure 3.4).

The first important application of the technology in drug discovery dates to 2013, when a competition on the prediction of drug properties and activities organized by Merck was won by a deep learning application developed in academia, with an improvement of about 15% even over Merck’s proprietary systems.<sup>61</sup>

A collaboration between the Pande group at Stanford and Google research led to a paper showing that the exploitation of bigger sets of data covering

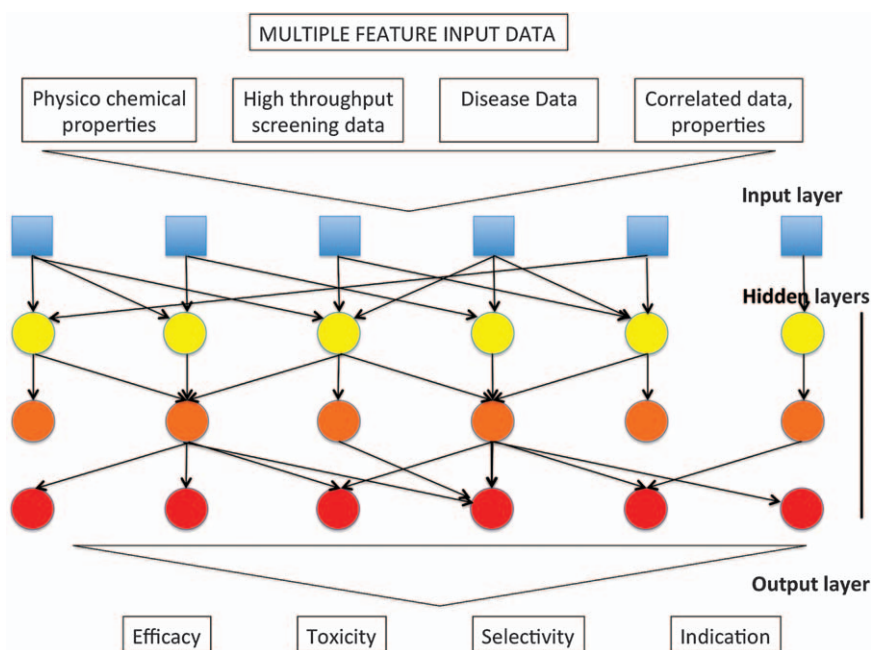


**Figure 3.4** Schematic of a deep learning neural network with one input layer  $x$ , several hidden layers and one output layer. In the hidden layers, different combinations of the output values of the preceding layers are calculated to produce the final property prediction.

more biological processes resulted in an increased number of bioactive drugs found in the screening step. The authors analysed a large amount of data for a number of different target processes, reaching a scale 18 times larger than previous applications. In particular, they used a total of 37.8 M data points across more than 200 distinct biological processes, showing the ability to carefully probe the sensitivity of these models to a variety of changes in model structure and input data. In the paper, the authors showed that the big data/deep learning approach works measurably better for discovering possible drugs than does analyzing smaller datasets and/or building models specifically targeting a single task.<sup>62</sup> The result is particularly encouraging, considering that given the costs of processes in discovery for drug companies (in the order of billions of dollars), any improvement may help saving substantial amounts of money.

Early in 2016, a publication from Insilico Medicine, Inc located at the Emerging Technology Centers at Johns Hopkins University, Datalytic Solutions and Mind Research Network presented a deep learning application to predict the therapeutic use of a large number of drugs using gene expression data obtained from high-throughput experiments on human cell lines.<sup>63,64</sup> Importantly, the authors introduced novel measures of differential signalling pathway activation score for a large number of processes to reduce the dimensionality of the data, while at the same time retaining biological relevance. These scoring parameters were then used to train the deep networks.

Specifically, researchers used the perturbation samples of 678 drugs across A549, MCF-7 and PC-3 cell lines from the Library of Integrated Network-Based Cellular Signatures (LINCS) project of the National Institutes of Health, NIH. They linked these data to 12 categories of therapeutic use from the MeSH (Medical Subject Headings) maintained by the National Library of Medicine of the NIH. To train the deep network, the researchers utilised gene level transcriptomic data and transcriptomic data processed using a pathway activation scoring algorithm, taking into account the effects of different concentrations of the drugs, administered for 6 and 24 h. The achieved accuracy in predicting the correct therapeutic application for each drug exceeded 54%. Most interestingly, some of the misclassified drugs were found to have more than a single application, illuminating the possibility for applying these methods in drug repurposing.<sup>64</sup>

One important aspect of this work is that, while the participating teams are part of private companies, they decided to release the data in the public domain describing a new methodology and the impact it can have in market oriented applications.

### 3.6 Conclusions and Perspectives

In this chapter, we have provided a general framework for the roles that computational methods can play in the discovery of small molecules that can eventually regulate biological pathways at different levels of

organisation, ranging from the structural to the phenotypical ones. The exploitation of novel methods for the physicochemical based revelation of cryptic or allosteric binding sites, for the definition of unknown protein pharmacologies, and for more efficient screening of small molecules driven by data on potential biological activities is providing unprecedented opportunities in the search of novel chemical tools for chemical biology and therapeutic applications.

With the continuous rise in computer power, algorithm design, and big data availability the future of computational chemistry for the discovery of new molecules is promising; the integration of different aspects, including characterisation of the functional dynamics of the systems involved, is most likely to play an increasingly important role in the development of novel pharmacological and biotechnological entities.

## References

1. R. G. Smock and L. M. Gierasch, *Science*, 2009, **324**, 198–203.
2. N. M. Goodey and S. J. Benkovic, *Nat. Chem. Biol.*, 2008, **4**, 474–482.
3. D. D. Boehr, R. Nussinov and P. E. Wright, *Nat. Chem. Biol.*, 2009, **5**, 789–796.
4. J. F. Swain and L. M. Gierasch, *Curr. Opin. Struct. Biol.*, 2006, **16**, 102–108.
5. J. A. Zorn and J. A. Wells, *Nat. Chem. Biol.*, 2010, **6**, 179–188.
6. M. E. Hill, D. J. MacPherson, P. Wu, O. Julien, J. A. Wells and J. A. Hardy, *ACS Chem. Biol.*, 2016, **11**, 1603–1612.
7. R. Nussinov and C.-J. Tsai, *Cell*, 2013, **153**, 293–305.
8. B. K. Shoichet, W. P. Walters, H. Jiang and J. Bajorath, *J. Med. Chem.*, 2016, **59**, 4033–4034.
9. L. Shrestha, H. J. Patel and G. Chiosis, *Cell Chem. Biol.*, 2016, **23**, 158–172.
10. A. L. Hopkins, *Nat. Biotechnol.*, 2007, **25**, 1110–1111.
11. A. L. Hopkins, *Nat. Chem. Biol.*, 2008, **4**, 682–690.
12. B. Vogelstein and K. W. Kinzler, *Nat. Med.*, 2004, **10**, 789–799.
13. J. van der Greef and R. N. McBurney, *Nat. Rev. Drug Discovery*, 2005, **4**, 961–967.
14. I. Bahar, C. Chennubhotla and D. Tobi, *Curr. Opin. Struct. Biol.*, 2007, **17**, 633–640.
15. P. Beltrao, C. Kiel and L. Serrano, *Curr. Opin. Struct. Biol.*, 2007, **17**, 378–384.
16. P. D. Patel, P. Yan, P. M. Seidler, H. J. Patel, W. Sun, C. Yang, N. S. Que, T. Taldone, P. Finotti, R. A. Stephani, D. T. Gewirth and G. Chiosis, *Nat. Chem. Biol.*, 2013, **9**, 677–684.
17. A. Panjkovich and X. Daura, *BMC Struct. Biol.*, 2010, **10**, 9.
18. J. J. Irwin and B. K. Shoichet, *J. Med. Chem.*, 2016, **59**, 4103–4120.
19. L. Neckers, M. Mollapour and S. Tsutsumi, *Trends Biochem. Sci.*, 2009, **34**, 223–226.

20. L. Whitesell, R. Bagatell and R. Falsey, *Curr. Cancer Drug Targets*, 2003, **5**.
21. M. Taipale, D. F. Jarosz and S. Lindquist, *Nat. Rev. Mol. Cell Biol.*, 2010, **11**, 515–528.
22. L. Whitesell and S. L. Lindquist, *Nat. Rev. Cancer*, 2005, **5**, 761–772.
23. J. S. Isaacs, W. P. Xu and L. Neckers, *Cancer Cell*, 2003, **3**, 213–217.
24. M. M. U. Ali, S. M. Roe, C. K. Vaughan, P. Meyer, B. Panaretou, P. W. Piper, C. Prodromou and L. H. Pearl, *Nature*, 2006, **440**, 1013–1017.
25. D. Eletto, D. Dersh and Y. Argon, *Semin. Cell Dev. Biol.*, 2010, **21**(5), 479–485.
26. D. E. Dollins, J. J. Warren, R. M. Immormino and D. T. Gewirth, *Mol. Cell*, 2007, **28**, 41–56.
27. A. Rasola, L. Neckers and D. Picard, *Trends Cell Biol.*, 2014, **24**, 455–463.
28. P. Workman, *Curr. Cancer Drug Targets*, 2003, **3**, 297–300.
29. J. Plescia, W. Salz, F. Xia, M. Pennati, N. Zaffaroni, M. G. Daidone, M. Meli, T. Dohi, P. Fortugno, Y. Nefedova, D. I. Gabrilovich, G. Colombo and D. C. Altieri, *Cancer Cell*, 2005, **7**, 457–467.
30. D. C. Altieri, L. R. Languino, J. B. Lian, J. L. Stein, I. Leav, A. J. van Wijnen, Z. Jiang and G. S. Stein, *J. Cell. Biochem.*, 2009, **107**, 845–852.
31. D. C. Altieri, *Mol. Cancer Ther.*, 2006, **5**, 478–482.
32. J. B. Trepel, M. Mollapour, G. Giaccone and L. Neckers, *Nat. Rev. Cancer*, 2010, **10**, 537–549.
33. J. Kaur and R. Ralhan, *Int. J. Cancer*, 2000, **85**, 1–5.
34. A. K. McCollum, C. J. TenEyck, B. Stensgard, B. W. Morlan, K. V. Ballman, R. B. Jenkins, D. O. Toft and C. Erlichman, *Cancer Res.*, 2008, **68**, 7419–7427.
35. S. Yoshida, *et al.*, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, E1604–E1612.
36. H. J. Patel, P. D. Patel, S. O. Ochiana, P. Yan, W. Sun, M. R. Patel, S. K. Shah, E. Tramentozzi, J. Brooks, A. Bolaender, L. Shrestha, R. Stephani, P. Finotti, C. Leifer, Z. Li, D. T. Gewirth, T. Taldone and G. Chiosis, *J. Med. Chem.*, **58**, 3922–3943.
37. Y. Hua, S. White-Gilbertson, J. Kellner, S. Rachidi, S. Z. Usmani, G. Chiosis, R. Depinho, Z. Li and B. Liu, *Clin. Cancer Res.*, 2013, **19**, 6242–6251.
38. J. T. Ernst, T. Neubert, M. Liu, S. Sperry, H. Zuccola, A. Turnbull, B. Fleck, W. Kargo, L. Woody, P. Chiang, D. Tran, W. Chen, P. Snyder, T. Alcacio, A. Nezami, J. Reynolds, K. Alvi, L. Goulet and D. Stamos, *J. Med. Chem.*, 2014, **57**, 3382–3400.
39. H. Chen, X. Zhou, A. Wang, Y. Zheng, Y. Gao and J. Zhou, *Drug Discovery Today*, 2015, **20**, 105–113.
40. M. De Vivo, M. Masetti, G. Bottegoni and A. Cavalli, *J. Med. Chem.*, 2016, **59**, 4035–4061.
41. G. Morra, A. Genoni, M. A. C. Neves, K. M. J. Merz and G. Colombo, *Curr. Med. Chem.*, 2010, **17**, 25–41.
42. J. A. Hardy and J. A. Wells, *Curr. Opin. Struct. Biol.*, 2004, **14**, 706–715.
43. R. Nussinov, C.-J. Tsai and J. Li, *J. Am. Chem. Soc.*, 2014, **136**, 17692–17701.

44. C.-J. Tsai and R. Nussinov, *PLoS Comput. Biol.*, 2014, **10**, e1003394.
45. G. Morra, M. A. C. Neves, C. J. Plescia, S. Tsutsumi, L. Neckers, G. Verkhivker, D. C. Altieri and G. Colombo, *J. Chem. Theory Comput.*, 2010, **6**, 2978–2989.
46. A. del Sol, C.-J. Tsai, B. Ma and R. Nussinov, *Structure*, 2009, **17**, 1042–1050.
47. M. G. Lerner, A. L. Bowman and H. A. Carlson, *J. Chem. Inf. Model.*, 2007, **47**, 2358–2365.
48. A. L. Bowman, M. G. Lerner and H. A. Carlson, *J. Am. Chem. Soc.*, 2007, **129**, 3634–3640.
49. G. Morra, G. M. Verkhivker and G. Colombo, *PLoS Comput. Biol.*, 2009, **5**, e1000323.
50. G. Morra, R. Potestio, C. Micheletti and G. Colombo, *PLoS Comput. Biol.*, 2012, **8**, e1002433.
51. G. Morra, A. Genoni and G. Colombo, *J. Chem. Theory Comput.*, 2014, **10**, 5677–5689.
52. S. Sattin, J. Tao, G. Vettoretti, E. Moroni, M. Pennati, A. Loperuolo, L. Morelli, A. Bugatti, A. Zuehlke, M. Moses, T. Prince, T. Kijima, K. Beebe, M. Rusnati, L. Neckers, N. Zaffaroni, D. Agard, A. Bernardi and G. Colombo, *Chemistry*, 2015, **21**, 13598–13608.
53. G. Vettoretti, E. Moroni, S. Sattin, J. Tao, D. Agard, A. Bernardi and G. Colombo, *Sci. Rep.*, 2016, **6**, 23830.
54. E. Moroni, H. Zhao, B. S. Blagg and G. Colombo, *J. Chem. Inf. Model.*, 2014, **54**, 195–208.
55. M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernberger, J. J. Irwin and B. K. Shoichet, *Nat. Biotechnol.*, 2007, **25**, 197–206.
56. D. Gfeller, O. Michielin and V. Zoete, *Bioinformatics*, 2013, **29**, 3073–3079.
57. Y. LeCun, Y. Bengio and G. Hinton, *Nature*, 2015, **521**, 436–444.
58. G. Schneider and S.-S. So, Landes Bioscience, Georgetown, 2001.
59. M. Reutlinger and G. Schneider, *J. Mol. Graphics Modell.*, 2012, **34**, 108–117.
60. J. E. Bajorath, Humana Press, Totowa, 2004.
61. J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl and V. Svetnik, *J. Chem. Inf. Model.*, 2015, **55**, 263–274.
62. S. Kearnes, K. McCloskey, M. Berndl, V. Pande and P. Riley, *J. Comput.-Aided Mol. Des.*, 2016, DOI: 10.1007/s10822-016-9938-8.
63. P. Mamoshina, A. Vieira, E. Putin and A. Zhavoronkov, *Mol. Pharm.*, 2016, **13**, 1445–1454.
64. A. Aliper, S. Plis, A. Artemov, A. Ulloa, P. Mamoshina and A. Zhavoronkov, *Mol. Pharm.*, 2016, **13**, 2524–2530.

## CHAPTER 4

# *Computational Design of Protein Function*

MARC GARCIA-BORRÀS,<sup>a</sup> KENDALL N. HOUK<sup>a</sup> AND  
GONZALO JIMÉNEZ-OSÉS<sup>\*b</sup>

<sup>a</sup> Department of Chemistry and Biochemistry, University of California, Los Angeles, California CA 90095-1569, USA; <sup>b</sup> Departamento de Química, Centro de Investigación en Síntesis Química, Universidad de La Rioja, 26006 Logroño, La Rioja, Spain

\*Email: gonzalo.jimenez@unirioja.es

## 4.1 Introduction

The chemical reactions that are central to metabolism are accelerated by protein catalysts, which ultimately makes life possible. Through billions of years of evolution, enzymes have achieved the capability to catalyse chemical transformations that are too slow to be measured under normal conditions. The most proficient enzymes can accelerate reactions with turnover rates as fast as the diffusion of reactants to the catalyst. Moreover, enzymes are highly selective in enhancing the rate of specific processes and reactants. These protein catalysts often depend on cofactors or coenzymes, and are frequently sensitive to the environment (temperature, pH, solvent, salinity, *etc.*), although are able to function over a large spectrum of conditions. Such features make them ideal constituents for complex and tightly regulated metabolic pathways. These highly efficient, selective, and biodegradable catalysts can be synthesised in large amounts by microorganisms to be used for industrial applications.

In 1946–1950 years after the term ‘enzyme’ was coined by physiologist Wilhelm Kühne, Linus Pauling hypothesised that enzymes are ‘closely complementary in structure to the activated complex for the reaction catalysed’.<sup>1</sup> Since the first high-resolution crystal structure of an enzyme was solved,<sup>2</sup> several factors have been proposed to explain the observed rate enhancements, ranging from non-covalent transition state (TS) stabilisation (restriction of motion, desolvation, electrostatic, *etc.*) to covalent bonding (hydrogen bonds, formation of transient intermediates, ionic interactions, *etc.*). The importance of pre-organisation in chemical catalysis is now well-recognised.<sup>3,4</sup> The most proficient<sup>5</sup> enzymes enter into the reaction by altering the structure of the TS and thus change the free-energy profile with respect to the uncatalysed reaction in solution.<sup>6</sup> This ‘covalent hypothesis’ explains why the vast majority of enzymes show TS binding constants—and subsequent accelerations—orders of magnitude higher than can be expected from non-covalent stabilisation.

Catalytic proficiency  $[K_{\text{tx}}^{-1} = (k_{\text{cat}}/K_{\text{M}})/k_{\text{uncat}}]$  is defined as the binding constant between the transition state and the enzyme. Of note,  $K_{\text{tx}}^{-1}$  spans 21 orders of magnitude (from  $10^8$  to  $10^{29} \text{ M}^{-1}$ )<sup>5,7–11</sup> with an average value of  $10^{16.0 \pm 4.0} \text{ M}^{-1}$ .<sup>12</sup> This value corresponds to an average  $\Delta G$  value for TS binding of  $22 \text{ kcal mol}^{-1}$ , but can range up to  $38 \text{ kcal mol}^{-1}$ ; these values are much higher than a noncovalent TS binding free energy of around  $15 \text{ kcal mol}^{-1}$ .

The importance of pre-organisation and electrostatics in enzyme catalysis has been demonstrated.<sup>4,13,14</sup> The enzyme active sites provide a pre-organised polar environment that stabilises the TS relative to the bound state more so than in the solution environment. It has been also shown that pre-organisation involves a reduction in the folding energy and results in an inversed activity/stability correlation.

The ability to design functional biocatalysts that can operate under varying conditions to perform chemical reactions not found in nature, and which can be used for the large-scale production of chemicals in a proficient, selective, ‘green’ and non-toxic manner, would revolutionise synthetic and materials chemistry.<sup>15</sup> Some desirable characteristics for industrial application of tailored enzymes would be: high catalytic turnover and substrate-, chemo-, regio- or stereo-selectivity, as well as a tolerance towards elevated temperatures, organic solvents, and chemical degradation.

However, the rational design of enzymes is a tremendous challenge for both chemistry and biochemistry. Most of the enormous complications associated to this problem is that we do not understand the rules of protein folding completely. Nevertheless, computational algorithms have become increasingly reliable for identifying amino acid sequences compatible with a target tertiary structure. A milestone in solving the inverse protein folding problem<sup>16–19</sup> was reached with the design and successful experimental validation of the structure of the 93-residue  $\alpha/\beta$  protein Top7.<sup>20</sup> This demonstrated that it is possible to use computational methods to predict sequences



that would produce an stable arbitrary fold. Since this possibility was demonstrated, protein structure prediction methods and rules<sup>21,22</sup> have become exceedingly successful in predicting very complex supramolecular architectures such as amyloids,<sup>23</sup> megadalton icosahedral complexes,<sup>24,25</sup> unnatural folds<sup>26,27</sup> and repeats,<sup>28,29</sup> *etc.* In light of this promising achievements, the next great challenge is to systematically create functional proteins that can promote non-natural chemical reactions.

## 4.2 The ‘Inside-out’ Design Protocol

### 4.2.1 Description of the Method

In parallel to the directed evolution approach,<sup>30–35</sup> the computer-assisted engineering of enzymes has been traditionally done using bioinformatics.<sup>36,37</sup> *De novo* computational design of functional proteins<sup>38</sup> bloomed when Baker and Houk developed the so-called ‘*inside-out*’ approach. The main idea behind this protocol is to utilise biochemical building blocks (amino acids, cofactors, *etc.*) to produce catalysts for non-biological processes that can be prepared by microbiological techniques.

The huge advancement in computer power has allowed the development and testing of improved structure prediction and conformational search algorithms. Quantum mechanical (QM) methods are nowadays capable of predicting the exact optimal arrangement of functional groups that maximise the binding and stabilisation of the transition states for any desired reaction. Catalytic conversion of non-natural chemicals into the adequate products would be possible, if a protein able to fold into this predicted necessary three-dimensional geometry can be designed. To override predicting the stability of new sequences from scratch, only well-characterised stable protein folds are used to incorporate the designed active. Also, as much of the native active site components as possible are preserved for the non-natural reaction or substrate.

In the first step of the ‘*inside-out*’ design protocol (Figure 4.1), QM calculations are performed to generate three-dimensional arrangements of functional groups found in protein sidechains that are optimal for stabilising the TS of the targeted reaction, typically through hydrogen bonds.<sup>39</sup> A *theozyme* (short for theoretical enzyme) is typically constructed from an array of amino acid side chains (commonly acting as general acids/bases) and backbone amides; incorporation of unnatural amino acids and cofactors allows expanding the chemical space. Various *theozyme* motifs differing on the nature of their functional groups are usually generated for a given reaction. The extent of catalysis is estimated based of the computed energy profile of each motif. These motifs are further diversified geometrically by producing an ensemble of conformations without disrupting the catalytic interactions.

The Rosetta<sup>40,41</sup> force field is highly optimised to predict amino acid sequences that will produce a thermodynamically stable fold for a given

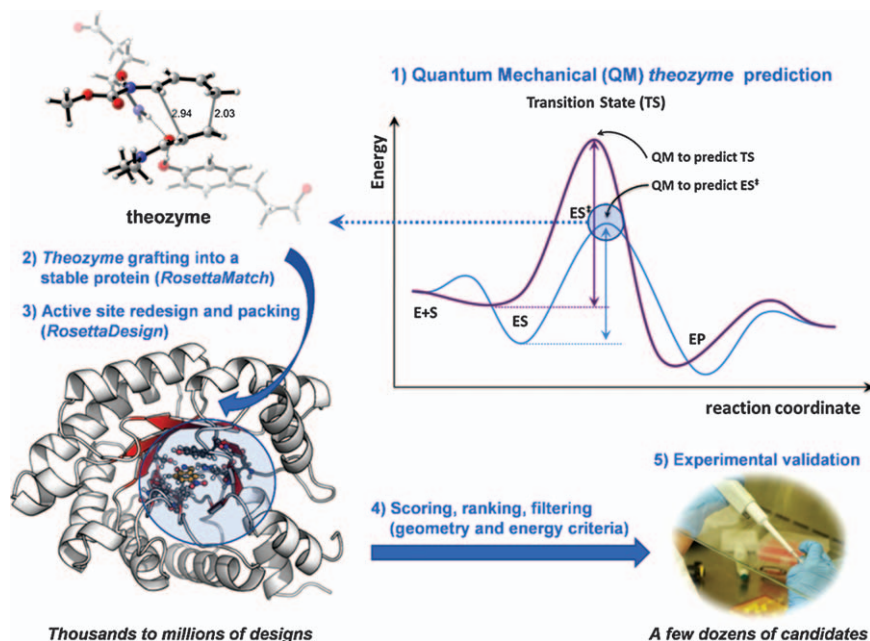


**Figure 4.1** Key steps in the computational 'inside-out' design protocol. A bimolecular Diels–Alder reaction is shown as an example.

(generally natural) fold.<sup>20,42</sup> Such predictions are generally less dependable when an unnatural *theozyme* geometry is grafted onto the protein scaffold. By definition, a *theozyme* is inherently a very unstable structure in which chemical bonds are partially created or cleaved. Catalytically relevant interactions between the substrate and the protein may work against protein stability. Thus, atomic properties such as bond lengths and atomic partial charges can be significantly different from that of the stable protein, and any interaction across *theozyme* and active site residues requires specialised parameters reflecting the nature of enzymatic transition state. Such parameters are under continuous development.<sup>41</sup>

Hence, the *RosettaMatch*<sup>43</sup> module searches the active sites of native proteins for backbone positions that can accommodate the three-dimensional arrangement in a *theozyme*, by sequentially attaching each side chain of the *theozyme* to the backbone of the protein scaffold. In an ideal match, the exact geometry of the *theozyme* can be realised. Deviations from the optimum geometry by just a few tenths of an angstrom and single-digit angles can lead to energetic penalties of up to 5 kcal mol<sup>-1</sup>, which translates to four orders of magnitude in terms of the reaction rate ( $k_{\text{cat}}$ ). Due to both the discrete nature of both the protein backbone and the primary matching algorithm, and the high computational cost associated with the mapping out of conformational space, matching is normally the bottleneck in the computational design protocol. This is particularly true when a *theozyme*

comprises three or more catalytic residues. The resulting matches are generally distorted from the *theozyme* geometry and necessitate some form of geometric filtering and ranking according to their *theozyme*-likeness.

As an alternative to *RosettaMatch*, *SABER* (Selection of Active/binding Sites for Enzyme Redesign)<sup>44</sup> searches the whole Protein Data Bank (PDB) for proteins with the appropriate catalytic functionality already in place. When a suitable active site is found, only those amino acid residues required to accommodate the new substrate in its transition state geometry are mutated.

After the *theozyme* has been attached to a scaffold protein, the *Rosetta-Design* module (now known as *Enzdes*)<sup>45</sup> is used to generate an optimal sequence/structure for the non-catalytic active site residues through rotamer sampling by Monte Carlo (MC) simulated annealing in multiple rounds, enforcing the *theozyme* geometry through distance and/or angle restraints. To ensure that the resultant protein sequence is intrinsically compatible with the *theozyme*, a last cycle of repacking and minimisation without restraints is commonly run.<sup>45</sup> Ideally, this process introduces amino acid residues that add interactions to stabilise the positions of the key catalytic residues, tune their  $pK_a$  values, and optimise transition state binding. In practice, each match that enters the active site design stage contains a *theozyme* that is already significantly distorted compared to the ideal QM TS geometry. Hence, even the highest ranked final designs normally differ quite considerably from the original *theozyme* geometry.

As a consequence, and prior to the experimental workup, the final set of designs are assessed towards their capability to stabilise the key catalytic residues, thus being ranked on the basis of knowledge-based criteria such as Rosetta energy, ligand-binding scores, hydrogen bonding, active site geometry, and packing scores.

Currently, it is standard that in the final stages of the enzyme design process, a group of experienced designers evaluate the quality or the feasibility of the designs based on their personal experiences as to what will create an active enzyme (the so-called 'endgame' stage). These decisions are highly arbitrary, are not encoded in the software, and include reversion of second-shell residues to the native amino acid, discarding mutations in specific positions of local domains that are thought to produce an unstable fold (Gly, Ser), poorly packable residues (Trp), *etc.*<sup>46</sup> An important improvement to the workflow in the future will be the minimisation of decisions dependent on the personal preferences and expertise of individual designers, in order to maximise the reliably and reproducibly of the designed enzymes.

Finally, MD simulations were found to be essential for assessing the structural integrity of a newly designed active site, frequently revealing design flaws that are intractable from static evaluations<sup>47,48</sup> and guiding additional/alternative mutations.<sup>49,50</sup> Analysis of the MD simulations enables the identification of structural weaknesses such as competing hydrogen-bonding contacts, water influx, unproductive or poor binding of the substrate and other sources for inactivity.<sup>51</sup>

### 4.2.2 Enzymes Designed Though the ‘Inside-out’ Approach: Kemp Eliminases

The groups of Houk, Baker, Mayo, Hilvert, DeGrado, and others have made notable advances in the highly multidisciplinary research area of biocatalysts engineering, and have demonstrated that computer-assisted design of enzymes for non-natural reactions is feasible.<sup>15</sup> Computer simulations based on quantum mechanics, Newtonian molecular mechanics and molecular dynamics principles are at the core of many enzyme design strategies, and have demonstrated superior performance for creating new types of catalysis with respect to random mutagenesis.<sup>52</sup> Using the formerly described ‘inside-out’ protocol, the Houk and Baker groups jointly designed enzymes for three chemically unrelated non-natural reactions: Kemp elimination,<sup>50,53</sup> retroaldol,<sup>54</sup> and Diels–Alder<sup>47</sup> reactions.

A central target in the enzyme design field has been the Kemp elimination, <sup>50,53,55–64</sup> which constitutes a well-studied model system involving a biochemically relevant and simple proton abstraction from carbon centres (Figure 4.2). The rate of the Kemp elimination strongly depends strongly on the medium (*i.e.* solvent) when a carboxylate functions as the general base;

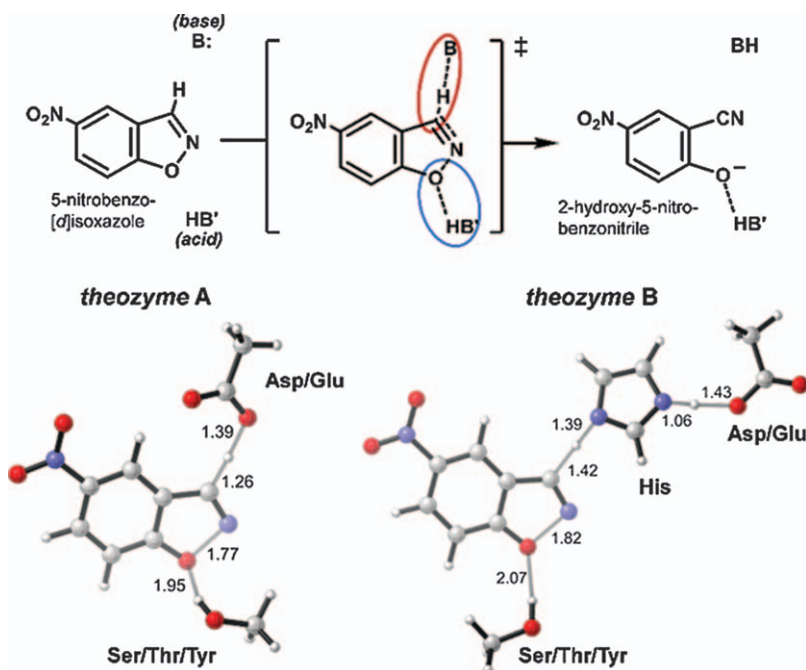


**Figure 4.2** The Kemp elimination reaction of 5-nitrobenzo[d]isoxazole, and the two theozymes that were employed.

Adapted from Houk *et al.*, *Angew. Chem., Int. Ed.*, 2013, 52, 5700–5725 with permission from John Wiley and Sons. Copyright © 2013 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.

rate accelerations of  $10^7$  can be achieved by using acetate as a base in a polar aprotic solvent such as acetonitrile compared to water.<sup>65</sup> An additional acceleration of  $10^6$  can be achieved through precise positioning of the donor and acceptor for this reaction,<sup>65,66</sup> yielding a theoretical limit for the rate enhancement of  $10^{13}$  for the Kemp elimination.

The Houk group has been actively involved in the design and perfection of Kemp eliminases by providing accurate computational structures and energies for the transition states comprising Asp/Glu (theozyme I) or the well-known catalytic dyad His-Asp/Glu (theozyme II) as general bases.

In the first Kemp eliminase design process, 8 of the 59 experimentally tested designs were active, although their measured activities were very modest.<sup>53,67</sup> Subsequent studies involving directed evolution greatly improved the catalytic efficiency of these initial computer-generated enzymes.<sup>50,62,63</sup> The directed evolution process, which utilised both global and local mutagenesis, generated a large increase in activity of up to two orders of magnitude.

Another Kemp eliminase, KE70, was optimised.<sup>50</sup> Rational mutations, predicted to be beneficial with calculations involving extensive backbone sampling were combined with random mutations. A greater than 400-fold improvement in the catalytic efficiency of the KE70 design was attained from nine rounds of mutation and selection. The improvement was due to higher  $k_{\text{cat}}$  values and lower  $K_{\text{M}}$  values, with the best variants having  $k_{\text{cat}}/K_{\text{M}}$  values greater than  $5 \times 10^4 \text{ s}^{-1} \text{ M}^{-1}$ .

In another design process, a single inactive design, HG1, was iteratively studied and redesigned using X-ray crystallography, MD simulations, and Rosetta Design.<sup>50</sup> The final Kemp eliminase from this study, HG3, was then subjected to directed evolution by the Hilvert and Mayo groups, which further increased the catalytic proficiency.<sup>63</sup> This evolved enzyme accelerates the reaction by  $6 \times 10^8$ , an improvement on HG3 by four orders of magnitude. The improved enzyme achieves a more precise realisation of the ideal geometry for catalysis (Figure 4.3). Hilvert and Mayo gave a detailed analysis of the evolution process, noting that the more proficient (evolved) enzyme has better shape-complementarity between the substrate and active site, eliminating unproductive binding, more efficient proton transfer, facilitated by the optimisation of the interactions between the catalytic base and the bound ligand, and better stabilisation of the transition state by the introduction of an oxyanion binder. This observation reinforces Pauling's paradigm—that is, the enzyme's ability to precisely complement the transition state of the reaction is critically important for its ability to accelerate the reaction. Along these lines, it should be emphasised that as revealed by calculations for certain transition structures, a deviation from the optimal geometry by just a tenth of an angstrom can lead to an energetic penalty of up to  $5 \text{ kcal mol}^{-1}$ , which translates to four orders of magnitude in  $k_{\text{cat}}$ .

Notably, the directed evolution process, which generated a large increase in activity, utilised both global and local mutagenesis. This contrasts with the protocol used in the first two design processes, where mutations were

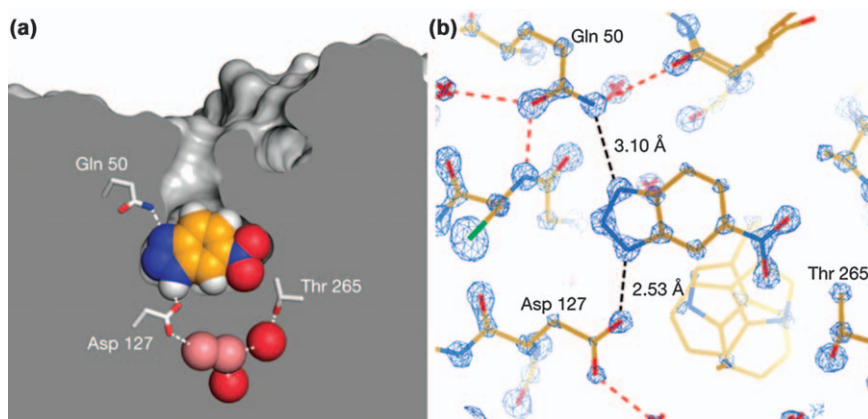


**Figure 4.3** Crystal structure of HG3.17 complexed with 6-nitrobenzotriazole. (a) Cut-away view of the active site showing the snug fit of the ligand within the binding pocket. The ligand is shown in space-filling representation (carbon, orange; nitrogen, blue; oxygen, red; hydrogen, white). Critical residues are shown as sticks (carbon, white). Ordered water molecules in the active site are shown as spheres (red, full occupancy; salmon, partial occupancy). (b) Electron density for the inhibitor interacting with Asp 127 and Gln 50. The  $2F_o - 2F_c$  map is contoured at  $5.5\sigma$ . Dashed lines indicate hydrogen bonds.

Reprinted by permission from Macmillan Publishers Ltd: *Nature* (ref. 63), copyright 2013.

only made near the active site. Related to this, the key role of remote mutations on the catalytic activity of engineered acyltransferases (LovD) used for the industrial production of simvastatin, has been demonstrated by our group through microsecond molecular dynamics simulations.<sup>68</sup> These simulations revealed that remote mutations located as far as 15–20 Å from the active site altered the conformational dynamics of the catalytic residues, eliminating the need for allosteric modulation by the acyl carrier domain of the LovD protein partner (LovF). MD simulations have been shown to be an important tool for discovering relevant conformational correlation between distal parts of proteins.<sup>69–71</sup>

## 4.3 QM/MM Approaches to Enzyme Design

### 4.3.1 Description of the Methods

First introduced in 1976 by Warshel and Levitt,<sup>72</sup> mixed quantum mechanical/molecular mechanical (QM/MM) methods allow treating of quantum relevant processes in complex classical environments. QM/MM techniques are extensively used strategies for studying complex biomolecular systems.<sup>73</sup> They allow the study of those chemical processes occurring in realistic biological systems such as proteins or protein–protein and protein–DNA



adducts, solvated in biologically representative environments (explicit solvent molecules and counter ions or cell membranes) that can comprise between 10 000–100 000 atoms, and that require the application of quantum mechanics to properly describe the processes that are occurring at an atomistic level (*e.g.* bond forming or bond cleaving processes).

The QM/MM approach is simple: it consists in the treatment of a small portion of the biochemical system using QM while the rest of the system is described using a less rigorous but computationally more efficient molecular mechanics (MM) methods. MM methods use classical mechanics to model atom–atom interactions, and the energy of these interactions is computed using simple potential energy functions called force fields (FF).<sup>74</sup> Typically, in biological systems QM/MM methods are applied to analyse and study chemical reactions in which bonds are formed an/or broken, explicitly considering the effects of the protein backbone and environment, the presence of explicit solvent molecules and their influence on the reaction profile.

Nowadays, one of the most successful and extended versions of QM/MM approaches are the molecular orbital self-consistent field (MO-SCF).<sup>75–77</sup> Such methods are based on solving the Schrödinger equation using a global Hamiltonian that is divided in three different components. The majority of the current QM/MM approaches are based on an additive scheme in which the system is described by a single hybrid Hamiltonian: (1) the QM part ( $H_{\text{QM}}$ ) to describe the small region of the system where the chemistry is taking place; (2) a classical treatment ( $H_{\text{MM}}$ ) for the rest of the system and solvent molecules; and (3) a coupling term ( $H_{\text{QM/MM}}$ ) to describe the interaction between the QM portion and the MM region, which includes the electrostatic perturbation that the QM wave function suffers due to the polarisation of the explicit solvent and surrounding environment. Using this representation, the total energy of the system is given by the direct addition of the eigenvalues of the partial Hamiltonians:

$$E = E_{\text{QM}} + E_{\text{MM}} + E_{\text{QM/MM}} \quad (4.1)$$

There exists another type of QM/MM approach based on a subtractive scheme, where the QM calculation is performed on an isolated system and the influence of the environment is estimated at the lower level (MM), considering the difference between the energies obtained from the calculation of the entire system and only the QM part at the MM level. Thus, within this approach the total energy of the system can be described by:

$$E = E_{\text{QM}}(\text{QM}) + E_{\text{MM}}(\text{QM} + \text{MM}) - E_{\text{MM}}(\text{QM}) \quad (4.2)$$

Current QM/MM applications on biomolecular systems combine the utilisation of the computationally efficient density functional theory (DFT) to describe the QM region and different FF to describe the MM part of the system. Typically, QM/MM methods have been extensively used to unravel and computationally analyse enzymatic reaction mechanisms. For example, QM/MM calculations have been widely applied to understand (metallo-)enzyme catalysed reaction mechanisms and the effects of



previously proposed mutations on the reaction paths.<sup>78–86</sup> Selected examples of such studies have been adequately covered thorough this book. However, in this section we will describe different examples where QM/MM approaches have been directly applied to propose new mutations in rational computational enzyme design.

### 4.3.2 Engineered Butyrylcholinesterase for Cocaine Detoxification

Recent applications of QM/MM strategies have showed the potential of QM/MM methods to provide useful insights to qualitatively characterise key residues in the catalytic steps of enzymatic reactions and quantify the effect of new mutations on the reaction rate ( $k_{\text{cat}}$ ).

For example, Zhan *et al.* proposed a systematic computational design protocol based on QM/MM scoring to catalyse a reaction of biomedical relevance such is the detoxification of cocaine by butyrylcholinesterase (BChE).<sup>87–91</sup> The cocaine metabolism pathway corresponds to its hydrolysis by BChE, although the reaction is not very efficient. Thus, BChE mutants with higher catalytic activity towards the (–)-cocaine active enantiomer can be designed to offer an alternative and promising protein–drug for the treatment of cocaine abuse and dependence. The cocaine hydrolysis reaction mechanism (Figure 4.4) involves first the nucleophilic attack of Ser198 on the ester carbon to form a tetrahedral intermediate (TI), then the deprotonation of His438 by the leaving alcohol group, followed by the nucleophilic attack of a water molecule assisted by proton transfer to His438 to form a second TI, and finally the breakdown of the TI with the carboxylic acid leaving and the enzyme recovering its resting state. Using subtractive QM/MM ONIOM calculations, Zhan and Gao<sup>92</sup> studied first these steps in order to characterise the rate-limiting step of the reaction. They include the substrate, the catalytic triad (S198, H438 and E325) and the oxyanion hole (G116, G117 and A199 backbones) in the QM region, and found similar energy barriers for the first (Ser198 nucleophilic attack) and third (water nucleophilic attack) steps. Then, in order to improve the hydrogen bond interactions between the oxyanion hole and the transition state for the first step (TS1), they proposed different mutations based on previous designs to improve reactivity (A328W/Y332A)<sup>93</sup> and they carried out MD simulations using the TS structure of the first step (TS1) to validate the performance of these mutations and to propose new ones (A199S/S287G/A328W/Y332G).<sup>94</sup> The simulations indicated that the new proposed A199S/S287G/A328W/Y332G BChE variant should have a significantly lower energy barrier for the step one of the process and offer higher catalytic efficiency because of the enhanced hydrogen bonding interactions between the enzyme active site and the TS1. Their predictions were confirmed by experimental kinetic measures, which indicated that the catalytic efficiency of the latter mutant showed a 456-fold improvement with respect the WT BChE against (–)-cocaine.

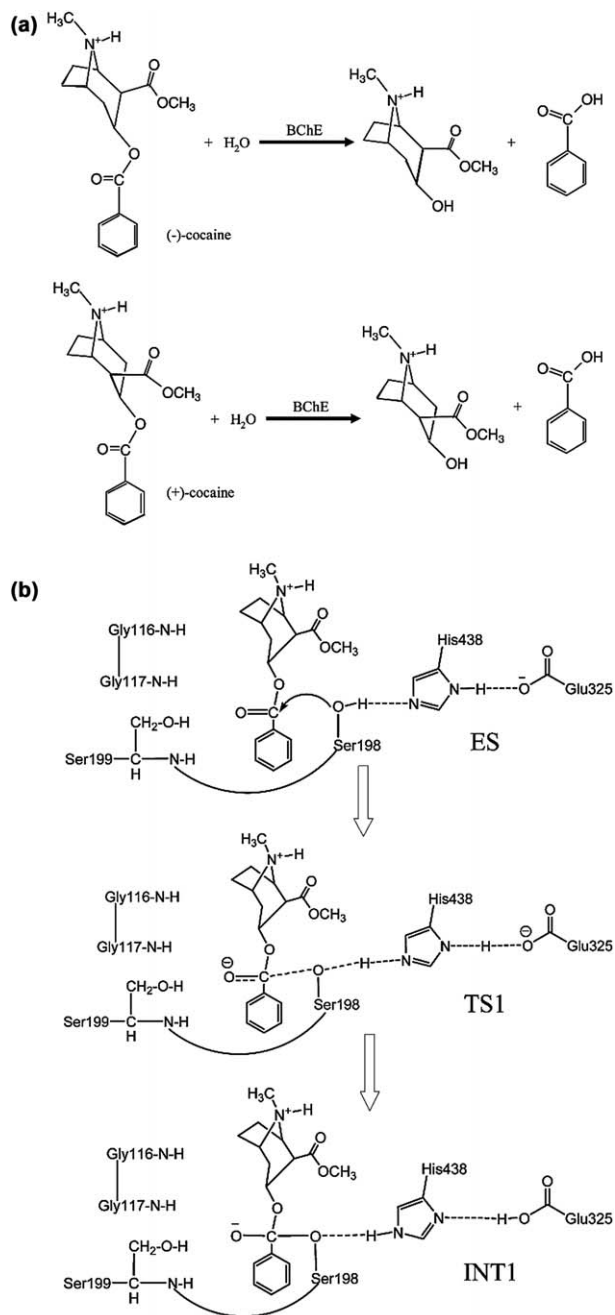
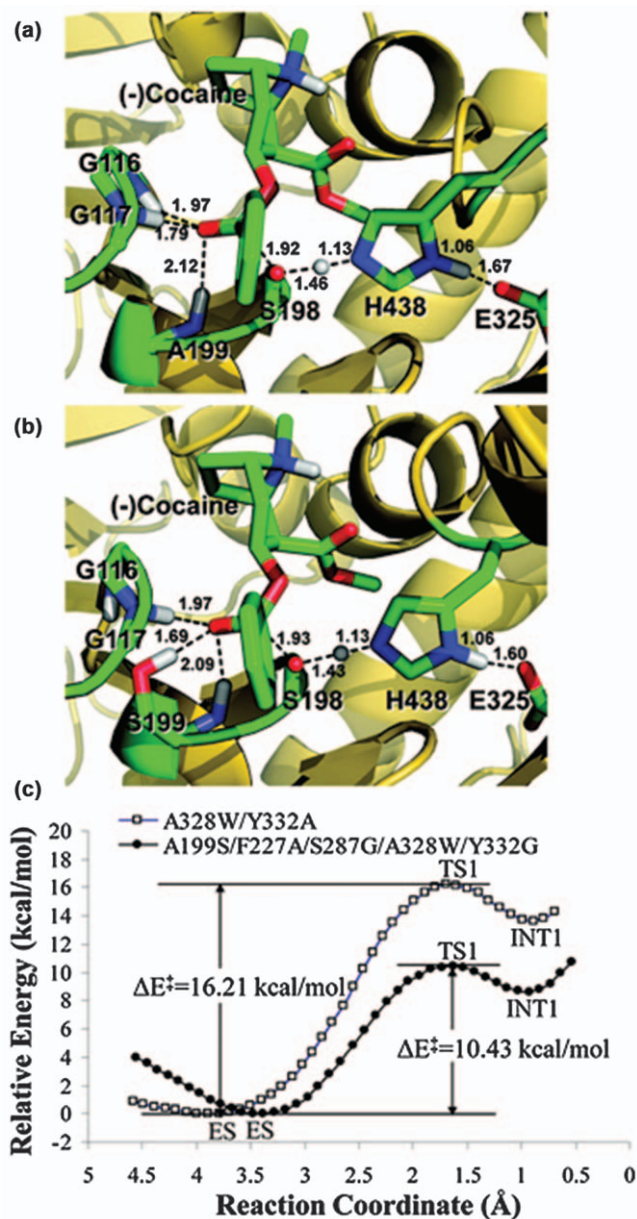


**Figure 4.4** Hydrolysis reactions of (–)-cocaine and (+)-cocaine: (a) reactants and products of the reactions; (b) schematic representation of the first reaction step for (–)-cocaine hydrolysis catalysed by a BChE mutant including A199S mutation.

Reprinted with permission from Zheng *et al.*, *J. Am. Chem. Soc.*, 2008, **130**, 12148–12155. Copyright (2008) American Chemical Society.

In order to improve their computational design protocol, the authors proposed a systematic methodology to account for the effects of the new introduced mutations on the energy barriers of BChE catalysed cocaine hydrolysis.<sup>95</sup> Their virtual screening starts by first defining a database of hypothetical mutants of BChE to be considered. Then, it follows by



estimating the total interaction energy (TIE) at MM level using MD simulations between the whole oxyanion hole of the enzyme and all (–)-cocaine atoms in the TS1 structure and each hypothetical mutant in its rate-determining transition state, TS1 in this case. If the estimated TIE, which includes the energetic contributions from both the hydrogen-bonding and non-bonding interactions at MM level, of a particular mutant improves the computed TIE for the reference enzyme (WT BChE, or previous mutant) then the practical effects of the new proposed mutations on the energy barrier for the rate-limiting step are evaluated by performing more accurate QM/MM calculations (combining DFT B3LYP functional and 6-31G(d) basis set to treat QM region with AMBER force field to describe the MM region) to finally select the best designs that will be tested *in vitro* and/or *in vivo*.

Using this strategy, the authors designed a new BChE mutant (A199S/F227A/S287G/A328W/Y332G) that had a QM/MM computed barrier for the first step of 4.4 kcal mol<sup>–1</sup> (compared to the 16.2 kcal mol<sup>–1</sup> barrier for the initial A328W/Y332A BChE mutant), and experimental site-directed mutagenesis and kinetic characterisation indicated a 2000-fold improved efficiency for (–)-cocaine hydrolysis (Figure 4.5). All together, these results show how QM/MM can provide detailed descriptions of key steps on enzymatic reaction mechanism and how it can be used to design efficient enzymes.

### 4.3.3 Electron Transfer Reactions Catalysed by Metalloproteins

QM/MM methods can be used to directly study and quantify the barrier height of an enzyme-catalysed reaction as described previously, but these methods can be also applied to analyse different quantum-derived properties of enzymes related to their catalytic activity.<sup>80,83,85</sup> Many (metallo)-enzyme catalysed reactions involve electron transfer (ET) processes, and QM/MM calculations have been recently applied to model and study electron

---

**Figure 4.5** Results from the QM/MM reaction coordinate calculations at the B3LYP/6-31G\*: Amber level. (a) Optimised geometry of the transition state (TS1) for the first reaction step of (–)-cocaine hydrolysis catalysed by A328W/Y332G BChE. (b) Optimised geometry of the transition state (TS1) for the first reaction step of (–)-cocaine hydrolysis catalysed by A199S/F227A/S287G/A328W/Y332G BChE. (c) Calculated potential energy surface along the reaction coordinate ( $R1 - R2 + R3$ ) of the enzymatic hydrolysis of (–)-cocaine. R1 is the length of the C···O transition bond between the hydroxyl oxygen of S198 and carbonyl carbon of (–)-cocaine benzoyl ester. R2 refers to the length of the O···H transition bond in the S198 hydroxyl group. R3 represents the length of the N···H transition bond between the hydroxyl hydrogen of S198 and the nitrogen of H438 side chain.

Reprinted with permission from F. Zheng *et al.*, *J. Am. Chem. Soc.*, 2008, **130**, 12148–12155. Copyright (2008) American Chemical Society.

transfer pathways in proteins and to propose new mutations to enhance the ET efficiency. For example, Guallar *et al.* proposed the e-pathway methodology, a QM/MM-based strategy to map and characterise the electron transfer pathways from donor to acceptor residues and/or cofactors in proteins.<sup>96–98</sup> In the e-pathway approach, the amino acids in the ET region between the donor and acceptor are described by QM while the rest of the system, including the donor and acceptor residues/cofactors, are treated at MM level. The method assumes that the unpaired electron, which is being transferred, has left the donor site but has not yet arrived at the acceptor site. Thus, both donor and acceptor sites are considered in their oxidised form in the MM region. The method strategy is based on iteratively modifying the QM region following the evolution of the spin density of the unpaired electron, moving residues from the QM region to the MM one in the successive iterative QM/MM calculations. The spin density analysis of each QM/MM calculation indicates which residue(s) contribute to the highest electron affinity molecular orbital. Then, in the following iteration, the identified residues are removed from the QM region and therefore the description of this orbital is no longer possible and a next ‘host’ for the electron can be characterised. The iterative process continues until the donor and acceptor can be connected through a direct pathway joining all the characterised molecular orbitals with short through-space jumps (about 4–5 Å). These orbitals identify those residues playing a major role in the ET process.

The e-pathway strategy has been used to study ET in (metallo-)enzymes, and for enzyme design. Vidal-Limón *et al.* used the QM/MM e-pathway approach to identify the residues involved in the suicide inactivation of the heme peroxidase P450-BM3.<sup>99</sup> The suicide inactivation is a common phenomenon observed in heme peroxidase, in which the heme is inactivated due to self-oxidation mediated by highly oxidizing enzyme intermediates formed during the catalytic cycle. The authors identified key oxidisable residues in the second coordination sphere of the heme cofactor, and based on their findings they proposed different mutations to include residues less oxidisable at these positions, which were then experimentally evaluated by site-directed mutagenesis. The residues W96 and F405 were predicted by the QM/MM e-Pathway to hold high spin density, and thus subsequent mutations were proposed based on these results (W96A, F405L, W96A/F405L). Site-directed mutagenesis confirmed that the computationally proposed variants are indeed up to 260 times more stable against H<sub>2</sub>O<sub>2</sub> inactivation than P450BM3 21B3, exhibiting similar reaction rates.

Guallar *et al.* engineered a Laccase oxidoreductase for the oxidation of arylamines.<sup>100</sup> Laccases are oxidases that are constituted by 4 copper ions, a trinuclear cluster that reduce molecular oxygen to water inside the protein matrix, and the remaining copper ion close to the surface where substrates are oxidised. The relatively low redox potential that these enzymes exhibit lowers the activity with respect to other oxidoreductases, and the authors proposed a computational strategy based on the combination of efficient conformational sampling of the substrate binding and simple QM/MM

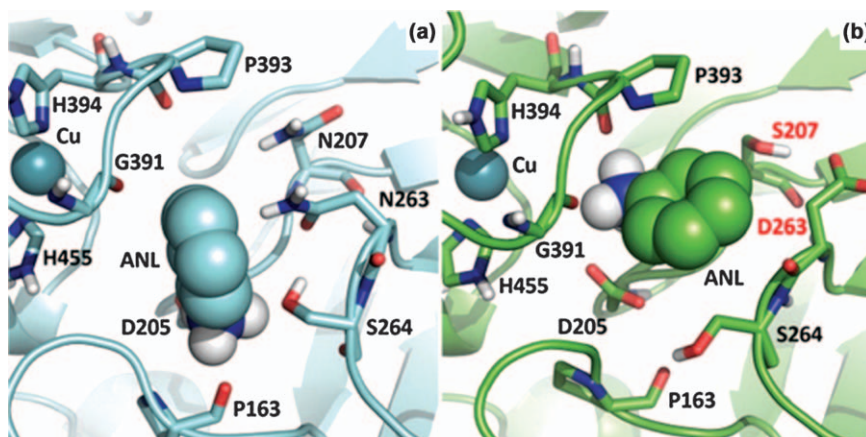


**Figure 4.6** Representative minima for aniline interaction with (a) parent laccase and (b) double mutant.

Reprinted with permission from G. Santiago *et al.*, *ACS Catal.*, 2016, 6, 5415–5423. Copyright (2016) American Chemical Society.

evaluations of different active site mutants to enhance the activity of Laccases towards the oxidation of different arylamines, such as aniline, or *p*-phenylenediamine. In particular, they used the protein landscape exploration (PELE)<sup>101,102</sup> MC-based algorithm to sample the protein–substrate interactions and most favourable substrate binding modes. Then, using QM/MM calculations they analysed the amount of spin density localised on the substrate. They used this quantity to score the reactivity of the different proposed variants, since it has been previously shown that there exist a direct relationship between the amount of spin density on the donor and its reduction/oxidation potential in ET proteins<sup>103</sup> and small aromatic compounds.<sup>104</sup> Using this computational approach, the authors proposed two-point mutations, N207 and N263, that are in direct contact with the substrate. In order to improve the electrostatic interactions of the active site with the substrate, they proposed N207S/N263D mutations which improved the amount of spin density localised on the different substrates considered (aniline, *N,N*-dimethyl-*p*-phenylenediamine and *p*-phenylenediamine) (Figure 4.6). The double-mutant computationally predicted by means of QM/MM computed spin densities was experimentally validated by site-directed mutagenesis, showing an enhancement of two-fold with respect to the WT laccase.<sup>100</sup>

## 4.4 Summary and Outlook

Computational methods have evolved to a point at which it is now possible to produce functional proteins able to catalyse chemical reactions with no natural precedence. So far, their rate accelerations are comparatively modest. However, they display features typically found only in nature's



enzymes—such as substrate specificity, stereoselectivity and shifted  $pK_a$  values. These features can be substantially improved, when computational design is combined with directed evolution.<sup>105–107</sup> Directed evolution has gradually become dominant on industrial processes,<sup>108</sup> as exemplified by the recent development of a transaminase to synthesise the diabetes drug sitagliptin.<sup>109</sup> However, laboratory evolution generally requires proteins with a basal level of the desired function. When this initial activity is not available in natural enzymes, computational enzyme design has proved useful to generate it. So far the ‘inside-out’ approach has produced catalysts for fairly simple uni- and bimolecular reactions, although the construction of more involved active sites for multistep transformations<sup>110</sup> and reengineered substrate specificities<sup>111</sup> are in progress. With the continued advance of computational methods and computer power, algorithms are being developed that will permit incorporation of more complex and robust active sites into proteins, as well as accurate and facile physics-based simulations, with which catalyst candidates can be tested and repaired prior to the experimental workup.

Enzymes provide finely tuned microenvironments for efficient and selective catalyses—sometimes of multiple substrates or of multiple steps, or requiring the presence of cofactors. Given the exponential development of hardware and algorithm capabilities, the future of computational enzyme design is very promising. The active development of quantum mechanical methods and classical force fields ensures that the performance and precision of physical calculations continue to improve. Ideally, all of the features common to effective enzymes need to be simulated rapidly and accurately to design enzymes routinely. At present, their simultaneous application has not been achieved, but the time when this can be accomplished is within sight.

## Acknowledgements

We are grateful to D. G. I. MINECO/FEDER (projects CTQ2015-70524-R and RYC-2013-14706 to G. J.-O.) and the National Science Foundation (CHE-1361104 to K. N. H.) for financial support. Computing resources were provided by CESGA, UR (Beronia cluster), UCLA Academic Technology Services/Institute for Digital Research and Education and IDRE and XSEDE (NSF OCI-1053575).

## References

1. L. Pauling, *Chem. Eng. News*, 1946, **24**, 1375–1377.
2. C. C. F. Blake, D. F. Koenig, G. A. Mair, A. C. T. North, D. C. Phillips and V. R. Sarma, *Nature*, 1965, **206**, 757–761.
3. A. Warshel, *J. Biol. Chem.*, 1998, **273**, 27035–27038.
4. M. Delgado, S. Görlich, J. E. Longbotham, N. S. Scrutton, S. Hay, V. Moliner and I. Tuñón, *ACS Catal.*, 2017, **7**, 3190–3198.
5. A. Radzicka and R. Wolfenden, *Science*, 1995, **267**, 90–93.



6. X. Y. Zhang and K. N. Houk, *Acc. Chem. Res.*, 2005, **38**, 379–385.
7. D. R. Edwards, D. C. Lohman and R. Wolfenden, *J. Am. Chem. Soc.*, 2012, **134**, 525–531.
8. A. Radzicka and R. Wolfenden, *J. Am. Chem. Soc.*, 1996, **118**, 6105–6109.
9. R. Wolfenden, X. Lu and G. Young, *J. Am. Chem. Soc.*, 1998, **120**, 6814–6815.
10. E. A. Taylor, D. R. J. Palmer and J. A. Gerlt, *J. Am. Chem. Soc.*, 2001, **123**, 5824–5825.
11. R. A. R. Bryant and D. E. Hansen, *J. Am. Chem. Soc.*, 1996, **118**, 5498–5499.
12. K. N. Houk, A. G. Leach, S. P. Kim and X. Zhang, *Angew. Chem., Int. Ed.*, 2003, **42**, 4872–4897.
13. D. J. Cram, *Angew. Chem. Int. Ed. Engl.*, 1986, **25**, 103–104.
14. A. J. T. Smith, R. Müller, M. D. Toscano, P. Kast, H. W. Hellinga, D. Hilvert and K. N. Houk, *J. Am. Chem. Soc.*, 2008, **130**, 15361–15373.
15. G. Kiss, N. Çelebi-Ölçüm, R. Moretti, D. Baker and K. N. Houk, *Angew. Chem., Int. Ed.*, 2013, **52**, 5700–5725.
16. C. Pabo, *Nature*, 1983, **301**, 200.
17. J. W. Ponder and F. M. Richards, *J. Mol. Biol.*, 1987, **193**, 775–791.
18. J. Bowie, R. Luthy and D. Eisenberg, *Science*, 1991, **253**, 164–170.
19. K. Yue and K. A. Dill, *Proc. Natl. Acad. Sci. U. S. A.*, 1992, **89**, 4163–4167.
20. B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard and D. Baker, *Science*, 2003, **302**, 1364–1368.
21. Y.-R. Lin, N. Koga, R. Tatsumi-Koga, G. Liu, A. F. Clouser, G. T. Montelione and D. Baker, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, E5478–E5485.
22. G. Bhardwaj, V. K. Mulligan, C. D. Bahl, J. M. Gilmore, P. J. Harvey, O. Cheneval, G. W. Buchko, S. V. S. R. K. Pulavarti, Q. Kaas, A. Eletsky, P.-S. Huang, W. A. Johnsen, P. Greisen, Jr., G. J. Rocklin, Y. Song, T. W. Linsky, A. Watkins, S. A. Rettie, X. Xu, L. P. Carter, R. Bonneau, J. M. Olson, E. Coutsiyas, C. E. Correnti, T. Szyperski, D. J. Craik and D. Baker, *Nature*, 2016, **538**, 329–335.
23. R. Gallardo, M. Ramakers, F. De Smet, F. Claes, L. Khodaparast, L. Khodaparast, J. R. Couceiro, T. Langenberg, M. Siemons, S. Nyström, L. J. Young, R. F. Laine, L. Young, E. Radaelli, I. Benilova, M. Kumar, A. Staes, M. Desager, M. Beerens, P. Vandervoort, A. Luttun, K. Gevaert, G. Bormans, M. Dewerchin, J. Van Eldere, P. Carmeliet, G. Vande Velde, C. Verfaillie, C. F. Kaminski, B. De Strooper, P. Hammarström, K. P. R. Nilsson, L. Serpell, J. Schymkowitz and F. Rousseau, *Science*, 2016, **354**, DOI: 10.1126/science.aah4949.
24. J. B. Bale, S. Gonen, Y. Liu, W. Sheffler, D. Ellis, C. Thomas, D. Cascio, T. O. Yeates, T. Gonen, N. P. King and D. Baker, *Science*, 2016, **353**, 389–394.
25. Y. Hsia, J. B. Bale, S. Gonen, D. Shi, W. Sheffler, K. K. Fong, U. Nattermann, C. Xu, P.-S. Huang, R. Ravichandran, S. Yi, T. N. Davis, T. Gonen, N. P. King and D. Baker, *Nature*, 2016, **535**, 136–139.

26. P.-S. Huang, K. Feldmeier, F. Parmeggiani, D. A. Fernandez Velasco, B. Hocker and D. Baker, *Nat. Chem. Biol.*, 2016, **12**, 29–34.
27. T. M. Jacobs, B. Williams, T. Williams, X. Xu, A. Eletsky, J. F. Federizon, T. Szyperski and B. Kuhlman, *Science*, 2016, **352**, 687–690.
28. T. J. Brunette, F. Parmeggiani, P.-S. Huang, G. Bhabha, D. C. Ekiert, S. E. Tsutakawa, G. L. Hura, J. A. Tainer and D. Baker, *Nature*, 2015, **528**, 580–584.
29. S. E. Boyken, Z. Chen, B. Groves, R. A. Langan, G. Oberdorfer, A. Ford, J. M. Gilmore, C. Xu, F. DiMaio, J. H. Pereira, B. Sankaran, G. Seelig, P. H. Zwart and D. Baker, *Science*, 2016, **352**, 680–687.
30. C. A. Tracewell and F. H. Arnold, *Curr. Opin. Chem. Biol.*, 2009, **13**, 3–9.
31. N. J. Turner, *Nat. Chem. Biol.*, 2009, **5**, 567–573.
32. F. H. Arnold and A. A. Volkov, *Curr. Opin. Chem. Biol.*, 1999, **3**, 54–59.
33. P. A. Romero, A. Krause and F. H. Arnold, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, E193–E201.
34. C. Jäckel, P. Kast and D. Hilvert, *Annu. Rev. Biophys.*, 2008, **37**, 153–173.
35. S. Bershtein and D. S. Tawfik, *Curr. Opin. Chem. Biol.*, 2008, **12**, 151–158.
36. J. D. Stewart, *Advances in Applied Microbiology*, Academic Press, 2006, vol. 59, pp. 31–52.
37. T. Furuya and K. Kino, *Appl. Microbiol. Biotechnol.*, 2010, **86**, 991–1002.
38. V. Nanda and R. L. Koder, *Nat. Chem.*, 2010, **2**, 15–24.
39. D. J. Tantillo, J. Chen and K. N. Houk, *Curr. Opin. Struct. Biol.*, 1998, **2**, 743–750.
40. F. DiMaio, A. Leaver-Fay, P. Bradley, D. Baker and I. Andre, *PLoS One*, 2011, **6**, e20450.
41. R. Das and D. Baker, *Annu. Rev. Biochem.*, 2008, **77**, 363–382.
42. A. Korkegian, M. E. Black, D. Baker and B. L. Stoddard, *Science*, 2005, **308**, 857–860.
43. A. Zanghellini, L. Jiang, A. M. Wollacott, G. Cheng, J. Meiler, E. A. Althoff, D. Rothlisberger and D. Baker, *Protein Sci.*, 2006, **15**, 2785–2794.
44. G. R. Nosrati and K. N. Houk, *Protein Sci.*, 2012, **21**, 697–706.
45. F. Richter, A. Leaver-Fay, S. D. Khare, S. Bjelic and D. Baker, *PLoS One*, 2011, **6**, e19230.
46. L. G. Nivón, S. Bjelic, C. King and D. Baker, *Proteins: Struct., Funct., Bioinf.*, 2014, **82**, 858–866.
47. J. B. Siegel, A. Zanghellini, H. M. Lovick, G. Kiss, A. R. Lambert, J. L. St. Clair, J. L. Gallaher, D. Hilvert, M. H. Gelb, B. L. Stoddard, K. N. Houk, F. E. Michael and D. Baker, *Science*, 2010, **329**, 313.
48. G. Kiss, D. Rothlisberger, D. Baker and K. N. Houk, *Protein Sci.*, 2010, **19**, 1760–1773.
49. S. Lindert, J. Meiler and J. A. McCammon, *J. Chem. Theory Comput.*, 2013, **9**, 3843–3847.
50. H. K. Privett, G. Kiss, T. M. Lee, R. Blomberg, R. A. Chica, L. M. Thomas, D. Hilvert, K. N. Houk and S. L. Mayo, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, 3790–3795.

51. A. R. H. Narayan, G. Jiménez-Osés, P. Liu, S. Negretti, W. Zhao, M. M. Gilbert, R. O. Ramabhadran, Y.-F. Yang, L. R. Furan, Z. Li, L. M. Podust, J. Montgomery, K. N. Houk and D. H. Sherman, *Nat. Chem.*, 2015, **7**, 653–660.
52. Y. Kipnis and D. Baker, *Protein Sci.*, 2012, **21**, 1388–1395.
53. D. Rothlisberger, O. Khersonsky, A. M. Wollacott, L. Jiang, J. DeChancie, J. Betker, J. L. Gallaher, E. A. Althoff, A. Zanghellini, O. Dym, S. Albeck, K. N. Houk, D. S. Tawfik and D. Baker, *Nature*, 2008, **453**, 190–195.
54. L. Jiang, E. A. Althoff, F. R. Clemente, L. Doyle, D. Röthlisberger, A. Zanghellini, J. L. Gallaher, J. L. Betker, F. Tanaka, C. F. Barbas, D. Hilvert, K. N. Houk, B. L. Stoddard and D. Baker, *Science*, 2008, **319**, 1387–1391.
55. O. Khersonsky, G. Kiss, D. Rothlisberger, O. Dym, S. Albeck, K. N. Houk, D. Baker and D. S. Tawfik, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, 10358–10363.
56. A. N. Alexandrova, D. Röthlisberger, D. Baker and W. L. Jorgensen, *J. Am. Chem. Soc.*, 2008, **130**, 15907–15915.
57. M. P. Frushicheva, J. Cao and A. Warshel, *Biochemistry*, 2011, **50**, 3849–3858.
58. I. V. Korendovych, D. W. Kulp, Y. Wu, H. Cheng, H. Roder and W. F. DeGrado, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**, 6823–6827.
59. G. Kiss, D. Röthlisberger, D. Baker and K. N. Houk, *Protein Sci.*, 2010, **19**, 1760–1773.
60. O. Khersonsky, D. Rothlisberger, O. Dym, S. Albeck, C. J. Jackson, D. Baker and D. S. Tawfik, *J. Mol. Biol.*, 2010, **396**, 1025–1042.
61. M. P. Frushicheva, J. Cao, Z. T. Chu and A. Warshel, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 16869–16874.
62. O. Khersonsky, D. Rothlisberger, A. M. Wollacott, P. Murphy, O. Dym, S. Albeck, G. Kiss, K. N. Houk, D. Baker and D. S. Tawfik, *J. Mol. Biol.*, 2011, **407**, 391–412.
63. R. Blomberg, H. Kries, D. M. Pinkas, P. R. E. Mittl, M. G. Grutter, H. K. Privett, S. L. Mayo and D. Hilvert, *Nature*, 2013, **503**, 418–421.
64. A. Li, B. Wang, A. Ilie, K. D. Dubey, G. Bange, I. V. Korendovych, S. Shaik and M. T. Reetz, *Nat. Commun.*, 2017, **8**, 14876.
65. A. J. Kirby, *Acc. Chem. Res.*, 1997, **30**, 290–296.
66. Y. Hu, K. N. Houk, K. Kikuchi, K. Hotta and D. Hilvert, *J. Am. Chem. Soc.*, 2004, **126**, 8197–8205.
67. D. S. Kemp, *Nature*, 1995, **373**, 196–197.
68. G. Jiménez-Osés, S. Osuna, X. Gao, M. R. Sawaya, L. Gilson, S. J. Collier, G. W. Huisman, T. O. Yeates, Y. Tang and K. N. Houk, *Nat. Chem. Biol.*, 2014, **10**, 431–436.
69. G. R. Bowman, X. Huang and V. S. Pande, *Methods*, 2009, **49**, 197–201.
70. G. R. Bowman, V. A. Voelz and V. S. Pande, *Curr. Opin. Struct. Biol.*, 2011, **21**, 4–11.

71. G. R. Bowman and P. L. Geissler, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, 11681–11686.
72. A. Warshel and M. Levitt, *J. Mol. Biol.*, 1976, **103**, 227–249.
73. H. Lin and D. Truhlar, *Theor. Chem. Acc.*, 2007, **117**, 185–199.
74. A. D. Mackerell, *J. Comput. Chem.*, 2004, **25**, 1584–1604.
75. A. Warshel and R. M. Weiss, *J. Am. Chem. Soc.*, 1980, **102**, 6218–6226.
76. H. M. Senn and W. Thiel, *Angew. Chem., Int. Ed.*, 2009, **48**, 1198–1229.
77. M. Orozco, *Chem. Soc. Rev.*, 2014, **43**, 5051–5066.
78. S. Martí, V. Moliner and I. Tuñón, *J. Chem. Theory Comput.*, 2005, **1**, 1008–1016.
79. C. R. W. Guimarães, M. Udier-Blagović and W. L. Jorgensen, *J. Am. Chem. Soc.*, 2005, **127**, 3577–3588.
80. A. Altun, S. Shaik and W. Thiel, *J. Comput. Chem.*, 2006, **27**, 1324–1337.
81. A. Altun, S. Shaik and W. Thiel, *J. Am. Chem. Soc.*, 2007, **129**, 8978–8987.
82. C. L. Stanton, I. F. W. Kuo, C. J. Mundy, T. Laino and K. N. Houk, *J. Phys. Chem. B*, 2007, **111**, 12573–12581.
83. S. Shaik, S. Cohen, Y. Wang, H. Chen, D. Kumar and W. Thiel, *Chem. Rev.*, 2009, **110**, 949–1017.
84. M. Altarsha, T. Benighaus, D. Kumar and W. Thiel, *J. Am. Chem. Soc.*, 2009, **131**, 4755–4763.
85. R.-Z. Liao and W. Thiel, *J. Chem. Theory Comput.*, 2012, **8**, 3793–3803.
86. G. Hou and Q. Cui, *J. Am. Chem. Soc.*, 2013, **135**, 10457–10469.
87. D. Q. Gao, H. Cho, W. C. Yang, Y. M. Pan, G. F. Yang, H. H. Tai and C. G. Zhan, *Angew. Chem., Int. Ed.*, 2006, **45**, 653–657.
88. F. Zheng, W. Yang, L. Xue, S. Hou, J. Liu and C.-G. Zhan, *Biochemistry*, 2010, **49**, 9113–9119.
89. L. Xue, M.-C. Ko, M. Tong, W. Yang, S. Hou, L. Fang, J. Liu, F. Zheng, J. H. Woods, H.-H. Tai and C.-G. Zhan, *Mol. Pharmacol.*, 2011, **79**, 290–297.
90. Y. Pan, D. Gao, W. Yang, H. Cho and C.-G. Zhan, *J. Am. Chem. Soc.*, 2007, **129**, 13537–13543.
91. J. Liu and C.-G. Zhan, *J. Chem. Theory Comput.*, 2012, **8**, 1426–1435.
92. C.-G. Zhan and D. Gao, *Biophys. J.*, 2005, **89**, 3863–3872.
93. H. Sun, Y.-P. Pang, O. Lockridge and S. Brimijoin, *Mol. Pharmacol.*, 2002, **62**, 220–224.
94. Y. Pan, D. Gao, W. Yang, H. Cho, G. Yang, H.-H. Tai and C.-G. Zhan, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 16656–16661.
95. F. Zheng, W. Yang, M.-C. Ko, J. Liu, H. Cho, D. Gao, M. Tong, H.-H. Tai, J. H. Woods and C.-G. Zhan, *J. Am. Chem. Soc.*, 2008, **130**, 12148–12155.
96. F. Wallrapp, D. Masone and V. Guallar, *J. Phys. Chem. A*, 2008, **112**, 12989–12994.
97. F. H. Wallrapp, A. A. Voityuk and V. Guallar, *PLoS Comput. Biol.*, 2013, **9**, e1002990.
98. S. Saen-Oon, M. F. Lucas and V. Guallar, *Phys. Chem. Chem. Phys.*, 2013, **15**, 15271–15285.
99. A. Vidal-Limón, S. Águila, M. Ayala, C. V. Batista and R. Vazquez-Duhalt, *J. Inorg. Biochem.*, 2013, **122**, 18–26.

100. G. Santiago, F. de Salas, M. F. Lucas, E. Monza, S. Acebes, Á. T. Martinez, S. Camarero and V. Guallar, *ACS Catal.*, 2016, **6**, 5415-5423.
101. K. W. Borrelli, A. Vitalis, R. Alcantara and V. Guallar, *J. Chem. Theory Comput.*, 2005, **1**, 1304-1311.
102. B. P. Cossins, A. Hosseini and V. Guallar, *J. Chem. Theory Comput.*, 2012, **8**, 959-965.
103. K. Artz, J. C. Williams, J. P. Allen, F. Lendzian, J. Rautter and W. Lubitz, *Proc. Natl. Acad. Sci. U. S. A.*, 1997, **94**, 13582-13587.
104. D. Meisel and P. Neta, *J. Am. Chem. Soc.*, 1975, **97**, 5198-5203.
105. A. Romero-Rivera, M. Garcia-Borras and S. Osuna, *Chem. Commun.*, 2017, **53**, 284-297.
106. S. Osuna, G. Jiménez-Osés, E. L. Noey and K. N. Houk, *Acc. Chem. Res.*, 2015, **48**, 1080-1089.
107. E. L. Noey, N. Tibrewal, G. Jiménez-Osés, S. Osuna, J. Park, C. Bond, D. Cascio, J. Liang, X. Zhang, G. Huisman, Y. Tang and K. N. Houk, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, E7065-E7072.
108. U. T. Bornscheuer, G. W. Huisman, R. J. Kazlauskas, S. Lutz, J. C. Moore and K. Robins, *Nature*, 2012, **485**, 185-194.
109. C. K. Savile, J. M. Janey, E. C. Mundorff, J. C. Moore, S. Tam, W. R. Jarvis, J. C. Colbeck, A. Krebber, F. J. Fleitz, J. Brands, P. N. Devine, G. W. Huisman and G. J. Hughes, *Science*, 2010, **329**, 305-309.
110. F. Richter, R. Blomberg, S. D. Khare, G. Kiss, A. P. Kuzin, A. J. T. Smith, J. Gallaher, Z. Pianowski, R. C. Helgeson, A. Grjasnow, R. Xiao, J. Seetharaman, M. Su, S. Vorobiev, S. Lew, F. Forouhar, G. J. Kornhaber, J. F. Hunt, G. T. Montelione, L. Tong, K. N. Houk, D. Hilvert and D. Baker, *J. Am. Chem. Soc.*, 2012, **134**, 16197-16206.
111. S. R. Gordon, E. J. Stanley, S. Wolf, A. Toland, S. J. Wu, D. Hadidi, J. H. Mills, D. Baker, I. S. Pultz and J. B. Siegel, *J. Am. Chem. Soc.*, 2012, **134**, 20513-20520.

## CHAPTER 5

# *Computational Enzymology: Modelling Biological Catalysts*

LAURA MASGRAU,<sup>\*a,b</sup> ÀNGELS GONZÁLEZ-LAFONT<sup>a,c</sup> AND  
JOSÉ M. LLUCH<sup>\*a,c</sup>

<sup>a</sup> Institut de Biotecnologia i de Biomedicina (IBB), Universitat Autònoma de Barcelona, 08193 Bellaterra (Cerdanyola del Vallès), Barcelona, Spain;

<sup>b</sup> Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, 08193 Bellaterra (Cerdanyola del Vallès), Barcelona, Spain;

<sup>c</sup> Departament de Química, Universitat Autònoma de Barcelona, 08193 Bellaterra (Cerdanyola del Vallès), Barcelona, Spain

\*Email: Laura.Masgrau@uab.cat; JoseMaria.Lluch@uab.cat

## 5.1 Introduction

Most biochemical reactions depend on enzyme catalysis and understanding how enzymes ‘work’ at the molecular level is a fundamental problem that can help with the design of new drugs (or more generally, inhibitors) and novel catalysts. Enzymes are proficient catalysts that achieve typical rate accelerations of  $10^8$ – $10^{12}$  (with some examples reaching nearly  $10^{20}$ -fold),<sup>1</sup> as compared to the uncatalysed (*e.g.* in solution, if available) reactions. This great rate enhancement is usually accompanied by great regio- and stereo-specificity and a high substrate (or reaction) specificity. Computational enzymology uses molecular modelling and computational chemistry methods to study at the molecular level enzyme activity. Elucidation of the reaction mechanism, structural characterisation of the transition state(s) and possible intermediates, calculation of energy barriers and reaction rates, identification of the factors (*e.g.* key enzyme–substrate interactions) that

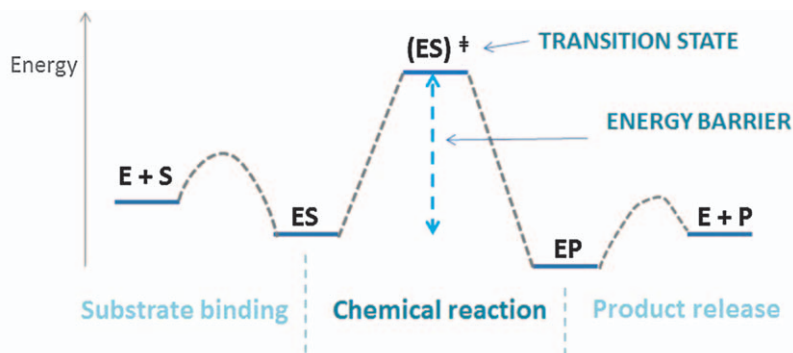


**Figure 5.1** General energetic scheme for the main processes involved in an enzymatic catalytic cycle.

contribute to the enzymatic reaction or analysing/predicting the effect of mutations, are some examples of aspects that can be addressed. Figure 5.1 shows a general scheme illustrating the main processes involved in an enzyme catalytic cycle.

In this chapter, we will focus on the main current methods used to study the chemical process of an enzymatic reaction; that is, the step (or steps) where chemical bonds are being broken and formed. Due to space limitations, the chapter focuses on ground-state reactivity and some theoretical approaches like the empirical valence bond (EVB) or the Car-Parrinello based methods are not commented on here but can be found elsewhere.<sup>2,3</sup> Of the covered contents, the general framework and basic methods will be briefly presented first, whereas the last part of the chapter is dedicated to more sophisticated approaches that can be of interest to more experienced readers.

## 5.2 General Framework

### 5.2.1 The Transition State and the Energy Barrier

Computational techniques are a powerful tool to study enzymatic reactions and a very good complement to experiment. Available experimental data generally consists on rate constants ( $k_{\text{cat}}$ ) and catalytic efficiencies ( $k_{\text{cat}}/K_M$ ) for wild type and even for some mutant enzymes, the latter helping in the identification of key residues for binding or reaction. These parameters are obtained by fitting the measured data assuming a pre-established mechanistic scheme (*e.g.* like the largely used Michaelis–Menten equation). Within the framework of transition state theory,<sup>4</sup> this experimental rate constant can be used to derive a phenomenological free energy barrier (or phenomenological free energy of activation,  $\Delta G_{\text{act,phen}}$ ) following eqn (5.1):

$$k_{\text{cat}}(T) = \frac{k_B T}{h} \exp\left(-\frac{\Delta G_{\text{act,phen}}(T)}{RT}\right) \quad (5.1)$$



where  $k_B$ ,  $h$  and  $R$  are the Boltzmann, Planck and ideal gas constants, respectively, and  $T$  is the temperature.  $\Delta G_{\text{act,phen}}$  corresponds to an apparent free energy difference between the transition state (also named activated complex, in which the bonds involved in the chemical transformation are partially broken or formed) and the reactants, and in enzyme-catalysed reactions typically falls between 10 and 25 kcal mol<sup>-1</sup>.

Accordingly, computational enzymology studies calculate an estimation of the activation free energy for the reaction of interest following a given reaction mechanism, and use the experimentally-derived  $\Delta G_{\text{act,phen}}$  as a reference to validate the model and/or the mechanism. If the agreement is satisfactory, further analysis can then be carried out (examples will be given along this chapter).

The reason for using the free energy barrier and not the rate constant directly as a reference is the exponential dependence that exists between the two, which converts a small error in the barrier ( $\pm 1$  kcal mol<sup>-1</sup>, which is within the accuracy of the quantum mechanics computational methods) in a large error ( $\sim$  one order of magnitude) in  $k_{\text{cat}}$ . However, and as will be explained along this chapter, the calculation of an activation free energy directly comparable to  $\Delta G_{\text{act,phen}}$  can be very time consuming and it is not always required. Thus, in some applications, the enzyme reaction is satisfactorily studied by calculating potential energy barriers, and even when the free energy ones are calculated, they usually do not account for all possible terms (like quantum vibrational energies, re-crossing or hydrogen tunnelling, which will be commented in Section 5.6 of this chapter).

The modelling of the enzyme-catalysed chemical reaction can be divided in three main steps: (1) Obtaining a valid enzyme-substrate (solvated) complex that represents the (ensemble of) Michaelis complex or reactants, and that will be the starting point to study the reactivity (Section 5.3). (2) Calculation of the energy barrier for the reaction, either by potential energy methods (Section 5.4) or by free energy ones (Section 5.5). (3) Analysis, in parallel with the two previous steps.

## 5.2.2 Quantum Mechanics Molecular Mechanics (QM/MM) Methods

### 5.2.2.1 General Aspects

An enzyme-catalysed reaction consists of the rearrangement of nuclei and electrons (breaking/making of chemical bonds, electron transfer *etc.*) that takes place in a complex environment (enzyme and solvent, at least) formed by thousands of atoms. As seen in other chapters, molecular mechanics (MM) force fields (FF) are very powerful to study big biomolecular systems. Roughly, within the FF each atom type is represented by a sphere, with an assigned point charge, linked to other atoms by harmonic (bonded) forces and that interacts with the rest of atoms through non-bonded forces (electrostatic and van der Waals interactions). As electrons are not accounted

explicitly, chemical reactivity (*e.g.* description of the TS) cannot be studied with traditional MM methods. However, quantum mechanical (QM) methods consider the distribution of electrons in a molecule explicitly, and are widely used to study the structure and energetics of chemical reactions. However, due to the computational cost, QM calculations are typically limited to few hundreds of atoms. In order to study biochemical reactions, a multi-scale approach that combines both methodologies have become the method of choice: that is, the QM/MM approach. In the QM/MM calculations the system is divided into two regions: the QM region (treated at a QM level of theory), which includes at least the atoms more closely involved in the chemical reaction; and the MM region (represented with a FF), which includes the rest of the system (see Figure 5.2(a)).

Some general aspects of QM/MM methods will be outlined here, but the reader can find more details in recent reviews by other authors.<sup>2,5-7</sup> QM/MM methods can be classified in two general approaches, the subtractive and the additive scheme. In the subtractive scheme the total energy is calculated as the sum of the MM energy for the entire system plus the QM energy of the QM region (model system on inner layer) and, to avoid ‘double-counting’, the MM energy of the QM region is then subtracted. The ONIOM methods implemented in Gaussian follow this scheme.<sup>6</sup> Originally, only *mechanical embedding* was used, that is, the interaction between the atoms in the inner and outer layers is calculated classically with the FF parameters (including point charges), which will need to be defined for the atoms in the QM region and will remain the same in going from reactants to products. Subsequent methods included different kinds of improvements, notably, the *electrostatic embedding* scheme (see below). Nowadays, most implemented QM/MM approaches use the additive scheme, in which the total energy ( $E^{\text{total}}$ ) is calculated as:

$$E^{\text{total}} = E_{\text{QM region}}^{\text{QM}} + E_{\text{MM region}}^{\text{MM}} + E^{\text{QM-MM}} \quad (5.2)$$

where  $E_{\text{QM region}}^{\text{QM}}$  is the QM energy of the QM atoms,  $E_{\text{MM region}}^{\text{MM}}$  is the MM energy of only the atoms in the MM region, and  $E^{\text{QM-MM}}$  is the coupling between the QM and MM subsystems.  $E^{\text{QM-MM}}$  formally contains three terms: (1) the QM-MM van der Waals interaction; (2) the QM-MM interaction corresponding to covalent bonds in the QM/MM boundary. Up to here, the QM-MM coupling is calculated with molecular mechanics potentials and parameters (the FF parameters involving a QM atom are usually taken from the equivalent MM atom, although it may be necessary to define new ones). And (3) the QM-MM electrostatic interaction, which is typically introduced by performing the QM calculation in the presence of the environmental MM point charges (*electrostatic embedding*), so that the electronic distribution of the QM region can be polarised by the environment. This makes that, in practice, the electrostatic coupling is usually contained in the first term of eqn (5.2) in what could be renamed as the  $E_{\text{QM region}}^{\text{QM(MM)}}$  term. The next improvement in the treatment of QM-MM electrostatic interactions would be the use of a polarisable FF able to adapt to changes in the

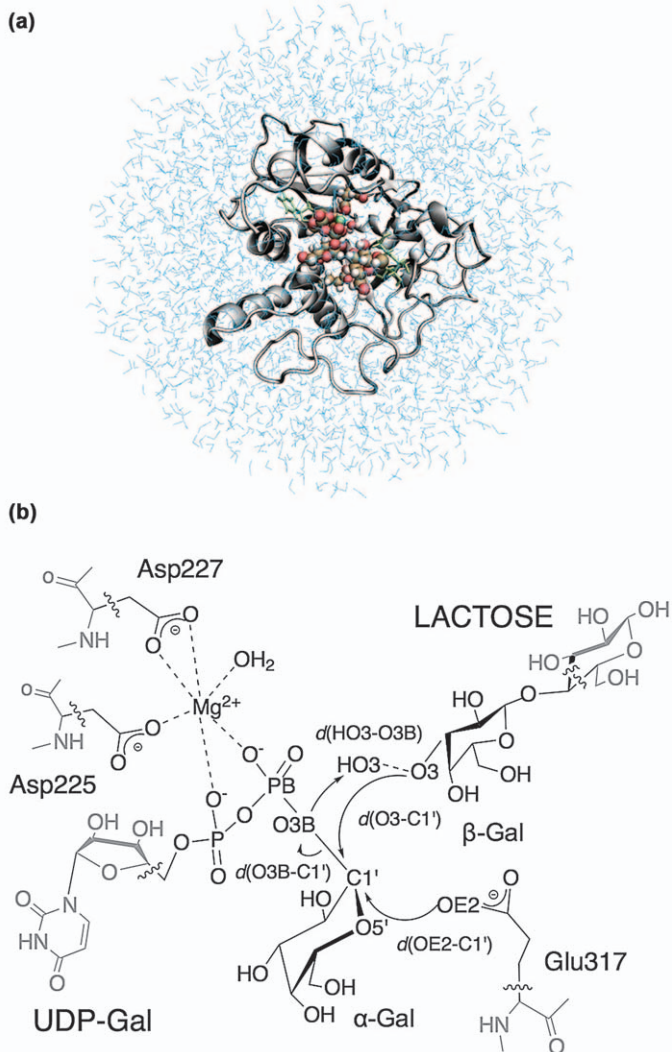


**Figure 5.2** Example of QM/MM partition of the enzyme–substrate system for  $\alpha 3GalT$ . (a) Molecular representation showing as spheres the atoms in the QM region. The rest of the system is treated at the MM level (remaining substrate atoms are shown as green tubes, water molecules as blue lines and the enzyme as grey cartoon). (b) Partial scheme of the active site, with QM atoms shown in black and MM ones in grey. The wavy lines indicate the covalent bonds at the boundary.

electronic distribution (*polarised embedding*), but its use in QM/MM calculations is not widespread yet.

In QM/MM calculations, the treatment of the boundary between the QM and MM regions requires special considerations in order to have the QM–MM interaction well described, especially when the separation cuts through

a covalent bond. The arising issues have been discussed extensively elsewhere,<sup>5</sup> and only some practical comments will be made in the following subsection.

### 5.2.2.2 Practical Issues

**5.2.2.2.1 The QM/MM Boundary and Partition.** In many QM/MM studies of biochemical reactions the frontier between the QM and MM regions needs to be placed in a covalent bond. This is for example the case for protein residues directly involved in catalysis (in this situation usually the full side-chain of the residue is included in the QM region), or for very large substrates or cofactors (see Figure 5.2(b) for an example). An adequate description of the van der Waals and bonding QM–MM interactions is necessary, but the most difficult term to treat is the electrostatic one, as the proximity of the MM point charges to the QM zone may introduce some artefacts like over polarisation of the wave function. Three main approaches have been developed to treat the boundary located at a covalent bond: the link atom, the boundary-atom and the frozen localised-orbital schemes. As general rules: the QM region should be as extended as possible; the boundary should be as far as possible from the zone where the changes in the electronic distribution occur (the reactive atoms); the bond that is cut should be apolar and a single bond (typically a C–C bond); the division should take into account the MM-groups definition to avoid unbalanced charges; and, still, testing different QM/MM partitions is advisable.

**5.2.2.2.2 The Choice of the QM Method.** The existing quantum mechanical electronic structure methods can be divided in three main types. (1) *Semi-empirical* methods (like MNDO, AM1, PM3 and related), which have the one- and two-electron integrals parametrised and thus are computationally not expensive, although their accuracy is limited. (2) *Ab initio* methods (like HF, MP2 or CCSD), with post Hartree Fock methods that introduce electron correlation that give good accuracies but are computationally very expensive. (3) *Density functional theory* (DFT) based methods (like B3LYP, PBE, M05 and many others), which can give accuracies comparable to the ones obtained with the *ab initio* ones but at less computational cost.

One way to take advantage of the range of methods available is to use a dual level approach. That is, one (low) level of theory is used to optimise the geometries and a higher level of theory, also with a larger basis set, is used in single-point energy calculations. Some of the accuracy limitations of semi-empirical methods can also be overcome by using specific reaction parameters (SRP)<sup>8</sup> that are fitted to higher level results of model systems for each particular reaction. Other approaches introduce a (dual-level) correction energy term as a function of the reaction coordinate, for which gradients can be calculated and thus applied in molecular dynamics

simulations;<sup>9</sup> or a different level of QM theory for the QM (gas-phase) energy and the QM-MM interactions.<sup>10</sup> The SCC-DFT semi-empirical method also performs much better (in general) than others.

## 5.3 Building the Model(s) of the Enzyme-Substrate Complex(es)

One of the characteristics of enzyme-substrate (E:S) systems is the high dimensionality so that the reactant complex of the reactivity studies does not correspond to a single structure but an ensemble of atomic configurations. Possible ways to take this into account in the calculations of the energy barrier will be discussed all along this chapter, starting from the preparation of the initial E:S structural model(s).

### 5.3.1 Starting Structure and System Setup

In order to model an enzyme-catalysed reaction, a reliable structure of the reactive E:S complex (or of a reaction intermediate or product) is required. Protein coordinates are usually obtained from high resolution (<2 Å) X-ray crystallographic structures, but the coordinates of the substrate are rarely found as it reacts too fast. Enzyme complexes with inhibitors that are substrate or transition state analogues are more often available from the Protein Data Bank ([www.rcsb.org](http://www.rcsb.org)). The substrate coordinates can then be built derived from the ligand's ones. Alternatively, protein-ligand docking techniques (introduced in other chapters of this book) can be used to predict the binding pose of the substrate inside the active site of the enzyme. It needs to be stressed here that conformational changes of the enzyme may occur during its catalytic cycle, especially upon ligand binding/unbinding. Thus, it is very important that our structural model corresponds to the reactive state of interest of the enzyme.

Structural data from the PDB needs to be checked for several things: missing coordinates (typically those of hydrogen atoms but maybe also some side-chains); missing loops (that can be modelled by homology modelling techniques); multiple side-chains conformations for one residue or non-desired molecule; protonation state of acidic residues (this can be done by visual inspection of possible interactions, or by using  $pK_a$  prediction software like the widely used PROPKA). Finally, once the E:S complex is built (which may also contain crystallographic water molecules), solvent water molecules are added to properly solvate the system (and counterions if full neutralisation is wanted).

### 5.3.2 Molecular Dynamics Simulations

Once the initial full E:S complex is built a geometry optimisation is carried out to relax possible clashes. This is usually done at the MM level and it is

followed by a molecular dynamics simulation, at least to equilibrate the solvent with the E:S complex. Force field parameters for the substrates (if they are not a biomolecule) can nowadays be obtained from generalised FF compatible with the main FFs (*e.g.* with the Antechamber or CGenFF codes). In principle, having an equilibrated system where the enzyme, solvent and substrate have been relaxed and allowed to adapt to each other during an MD simulation is highly desirable. Moreover, long-time MDs on a highly dimensional potential energy surface serve to explore the reactant valley corresponding to a substrate–enzyme complex, which as mentioned, involves an ensemble of nuclear configurations. The MDs are preferably run using periodic boundary conditions, although for subsequent QM/MM studies the system may be truncated to a reduced sphere, computationally less costly, with constrains/restrains in the outer boundary region. Some uncertainties accompany the simulations, like the simulation time that should be used or if the FFs are suitable enough to describe, with the accuracy required for the following QM/MM calculations, the non-bonding interactions between enzyme and substrate for long simulation times. Some studies of enzyme reactivity still skip this step, arguing that the X-ray crystal structure represents itself an average structure of the enzyme.

During a long MD simulation, structural fluctuations can lead to different E:S complexes presenting different catalytic characteristics. The structural differences can be of many sources (*e.g.* loop or side-chain conformation, binding mode or conformation of the substrate). If the interconversion among all the low-energy configurations of that extremely rugged, with a huge number of stationary points, E:S reactant valley is quite fast at physiological temperatures, the local equilibrium assumption can be adopted and a unique free energy profile can be calculated (Section 5.5). If not, further considerations will need to be done, which will be commented in Section 5.7.

## 5.4 Potential Energy Methods

If one snapshot (or several) from the molecular dynamics simulation of the solvated E:S complex is (are) selected to proceed with the reactivity studies on the high-dimensional potential energy surface of the catalytic system, one (or multiple) minimum-energy paths might be calculated. In this section, we describe some of the main methods to calculate the potential energy barrier (that is the reaction barrier without zero-point, thermic or entropic effects) of a catalytic reaction step associated with one reactant E:S configuration along with several analysis that can be carried out along the reaction path. How to introduce the effect of the different E:S configurations in the calculation of the energy barrier of the catalytic reaction will be commented in Sections 5.5 and 5.7.

### 5.4.1 Reaction Path Calculation

In order to calculate the potential energy barrier, the structure of the transition state corresponding to the reaction/mechanism of interest has to be



determined or, at least, approximated (see below). This is usually done by calculating a reaction path (potential energy profile) that connects reactants and products as a function of a reaction coordinate ( $\xi$ ) involving the bond distances that participate in the chemical transformation. As an example, from Figure 5.2(b),  $\xi = [d(\text{O3B}-\text{C1}') - d(\text{O3}-\text{C1}') - d(\text{HO3}-\text{O3B})]$  has been used to model a nucleophilic attack of O3 to C1', with O3B acting as the base catalyst; whereas the  $\xi = [d(\text{O3B}-\text{C1}') - d(\text{OE2}_{\text{E317}}-\text{C1}')]$  or  $\xi = [d(\text{O3B}-\text{C1}') - d(\text{OE2}_{\text{E317}}-\text{C1}') - d(\text{HO3}-\text{O3B})]$  would describe the nucleophilic attack of Glu317 on C1'.<sup>11</sup> If a scan calculation is done, the system is driven from reactants to products in constrained QM/MM optimisations along the defined reaction coordinate (typically in steps of 0.05–0.2 Å). The maximum(a) and local minima that may appear along the resulting potential energy profile are then taken as initial guesses for transition state(s) search or intermediate optimisation, respectively. Note that some QM/MM packages do not have implemented an algorithm for saddle point or transition state optimisation. In those cases, this initial guess is taken as an approximate model of the transition state.

One difficulty associated in calculating such reaction paths for large systems by scanning a reaction coordinate is the complexity to maintain the distant groups from the reactive zone in the same local minimum all along the reaction-path sequence. If special techniques are not used the optimisation protocol may lead to hysteresis effects. One way to overcome this problem is to run calculations forward and backward between each pair of stationary points in a reaction mechanism until the energy differences are constant. However, this process might be very time consuming depending on the QM level of theory used and the number of chemical-reaction steps of the catalysed process.

Moreover, in condensed phases in general and in enzymes in particular, it is not so obvious how to define the most adequate reaction coordinate. In fact, reaction-path techniques have been specially adapted to QM/MM optimisations to take special care regarding the selection of the relevant coordinates to be included in the path definition and the smoothness of the path. Chain-of-states methods, that include the string and the nudged-elastic-band (NEB) methods, have become very popular and efficient algorithms to directly determine reaction paths of enzymatic reactions in terms of minimum energy paths (MEPs) when little is known about the details of the most likely reaction path. In those methods, the whole path is represented by a sequence of structures that are optimised in a concerted fashion. In the NEB formulation, the reaction path is represented as a discretised series of nodes ('images') in configuration space between given initial (*e.g.* reactant) and final (*e.g.* product) states. Spring interactions are added between adjacent images to form a chain. The components of the spring potentials perpendicular to the band are removed to avoid corner-cutting effects, and the components of the molecular forces along the chain are projected out to avoid that the images slide towards the closest stable state. In the improved-tangent NEB with a climbing image formulation,<sup>12</sup> the



climbing image is free of spring forces but it is maximised in the direction of the reaction path and minimised in all other directions, and therefore converges to a transition state structure. For example, in the QM/MM study of the phosphoryl transfer reaction catalysed by *N*-acetyl-L-glutamate kinase (NAGK), 20 different reaction paths were optimised using the NEB algorithm implemented in the pDynamo package.<sup>13</sup> The initial structures were snapshots taken from MD trajectories initiated from four different crystal structures corresponding to reactants or products of the catalytic process. The NEB calculations were started off with a small number of intermediate structures but gradually increased until the energy profiles converged. The number of structures per path depended on the path length and ruggedness, and it ranged from 19 to 57 images, with an average of 33.

An alternative method that closely resembles the NEB algorithm is the string method<sup>14</sup> that defines the MEP as a continuous curve or string in configuration space. In practice, though, it is also necessary to discretise the string in a set of images which are optimised according to the force of the potential energy at that point, and then those optimised images are redistributed in one step along the string using interpolation techniques. The consecutive application of the NEB and string methods has also been used to locate the MEP of several enzymatic reactions.<sup>15</sup> It is noteworthy that neither of those chain-of-states methods requires the calculation of the Hessian matrix. Generally, those 'chain-of-states' reaction-path approaches are based on collective reaction coordinates or collective variables (CVs) that include the contribution of several different coordinates. The advantage of collective reaction coordinates is that, if they are properly selected, they naturally include environmental effects, thus assuring that the environment is relaxed along the reaction coordinate.

### 5.4.2 Transition State Localisation

Different algorithms for the location of stationary points and, in particular, for the location of transition state structures on QM/MM PESs for systems with thousands of atoms have been developed that take advantage of the QM/MM partition. One of the most efficient optimisation protocols is the so-called microiterative method. This methodology splits the system in two parts: a core zone, in which an accurate second-search is done (often with the P-RFO method<sup>16</sup>); and an environment that is kept minimised with a cheap first-order method (usually the L-BFGS algorithm<sup>17</sup>). The separation makes the sum of the expenses of the two procedures considerably lower than that of a single global search. In QM/MM techniques the iteration inside the core (usually the QM region) of any geometry optimisation is referred to as the 'macro-iteration' and the contributions of the surrounding MM environment are given the annotation of 'micro-iterations'. The two basic techniques for geometry optimisation within the microiterative approach are to either relax the macro and micro components at the same time, the so called adiabatic approach, or alternatively to follow a diabatic

methodology whereby the core is frozen as the environment is optimised and *vice versa*. The most suitable alternating frequency between the two processes (minimisation of the environment and the Newton step in the core) has been studied along with the analysis of how big the core must be to reach convergence with fewer iterations.<sup>18</sup> Those microiterative approaches may be used with Cartesian coordinates but, for instance, in the ChemShell interface<sup>19</sup> the hybrid delocalise coordinates (HDLCs) are used in combination with the microiterative optimisation algorithm. In this coordinate scheme, the system is divided into fragments or 'residues'. Internal coordinates are used to define the intramolecular degrees of freedom in each fragment whereas Cartesian coordinates are employed to define the relative position of the different residues. In this way, the computationally demanding internal-Cartesian coordinates transformation is only needed within those fragments. The dimer method is another algorithm specially designed to locate transition state structures on multidimensional QM/MM PESs.<sup>20</sup>

### 5.4.3 Analysis

The basic results obtained in modelling biochemical reactions are the structures and energies related to the reactive process: reaction mechanism (number of steps, rate-limiting one), characterisation of the transition state(s) and of possible intermediates, energy barrier and energy of reaction, and enzyme–substrate interactions that favour/unfavour the reaction. If alternative mechanisms have been proposed, the energy barrier is calculated for each possibility and compared, so that the preferred one (the one presenting the lowest energy barrier, and thus, the highest rate) is determined. The same kind of comparison can be made for different substrates or for mutated enzymes in order to rationalise differences in rate, mechanism, substrate specificity or regioselectivity, for example, and to propose further modifications to achieve a desired goal (*e.g.* alternative substrates, mutant enzymes).<sup>21</sup> It has to be noted that although the interpretation and analysis of the results is first mentioned in this section, these analysis are not restricted to the case where potential energy methods are used.

Multiple types of analysis can be carried out to gain further insight into the catalytic process. Typically, these involve calculation of magnitudes like atomic charges, spin densities (when applicable), electrostatic or van der Waals interactions. The purpose is identifying the key factors (*e.g.* the role of the protein residues, water molecules, metal ions, possible structural distortions of the substrate, structural fluctuations of the enzyme *etc.*) that contribute to the enzymatic reaction. Many examples can be found in the literature, but only a few cases can be commented on here and throughout the chapter.

#### 5.4.3.1 Analysis of Electrostatic Contributions

It is widely accepted that electrostatic effects have an important role in enzyme catalysis, mainly through a better electrostatic stabilisation of the transition state as compared to the reactant. Thus, the analysis of how

individual residues (or water molecules) interact with the substrate at different points of the reaction path is very useful to identify key actors. In QM/MM calculations different procedures have been described to analyse these electrostatic interactions. A simple analysis that allows to qualitatively estimate the electrostatic contributions consists of having the residue of interest in the MM region and carrying out two single-point energy calculations: one with the FF point charges and another one with the charges of this residue set to zero. The difference in the QM–MM electrostatic term between the two calculations provides an estimation of the electrostatic interaction of this residue with the QM zone (typically the substrates). If this is done at the transition state and at the reactant, the electrostatic contribution of this residue to the barrier height can be estimated.

This type of analysis was used for example to study the role of active site residues in the reaction mechanism of retaining glycosyltransferases (ret-GTs), which transfer a monosaccharide from an activated donor (*e.g.* UDP-Galactose) to an acceptor molecule with net retention of the configuration at the anomeric carbon (C1).<sup>22</sup> The presence of a nucleophilic residue on the  $\beta$ -face of the sugar in some ret-GTs had been taken as indicative of a double-displacement mechanism through formation of a glycosyl-enzyme covalent complex. In the absence of such residue, which seems to be the case for most ret-GTs, an alternative controversial mechanism had been proposed in which the acceptor molecule attacks by the same side where the UDP leaving group (front-side attack), with formation of an oxocarbenium intermediate. QM(DFT)/MM calculations were performed to compare both mechanisms in  $\alpha$ 3GalT, which has a Glu (Glu317) as putative nucleophile. The results suggested that both mechanisms may be possible (with differences in barrier heights that prevent concluding a favoured one). Most importantly, a key role in transition state stabilisation (an oxocarbenium with a positive charge increase at the anomeric centre of 0.32 a.u. as compared to the reactant) was revealed for Glu317 in reaction through the front-side mechanism (Figure 5.3), which could also explain the experimental mutagenesis results.<sup>11</sup> In the case of the EXTL2 glycosyltransferase, the structurally analogous Asp246 was also found to provide oxocarbenium stabilisation and to compensate for the presence of a positively charged Arg (that destabilises oxocarbenium formation) involved in important interactions with the acceptor. For EXTL2, the front-side mechanism was favoured (5.5 kcal mol<sup>-1</sup> lower energy barrier) over the nucleophilic attack by Asp246.<sup>23</sup> Thus, it seems that the actual general mechanism followed by this family of enzymes is a front-side nucleophilic attack. Electrostatic stabilisation of the oxocarbenium transition state or intermediate by residues on the  $\beta$ -face is used by some of these enzymes to facilitate catalysis; in limit cases of very strong stabilisation, the residue and substrate may collapse and form a covalent complex intermediate.

#### 5.4.3.2 Analysis of Steric Effects

In the study by Saura *et al.*<sup>24</sup> on the H<sub>13</sub>/H<sub>10</sub> abstraction reactions from arachidonic acid (AA) by the (Fe<sup>III</sup>–OH)<sup>-</sup> cofactor in 15-lipoxygenase, the

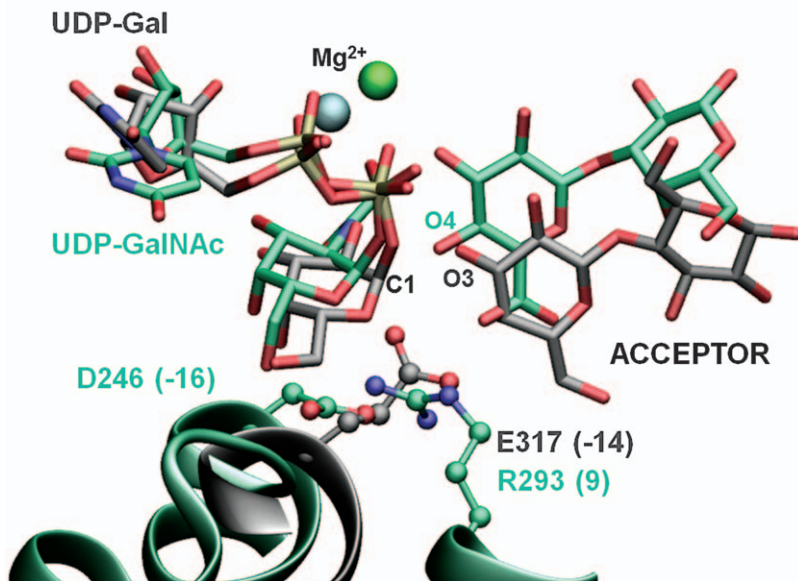


**Figure 5.3** Molecular representation of part of the  $\alpha$ 13GalT (grey) and  $\alpha$ 14-EXTL2 (green) active sites in their corresponding reactant complexes. The donor and acceptor substrates are shown in liquorice, the divalent cation as a sphere and residues on the  $\beta$ -face of the transferred sugar in ball-and-sticks. The estimated electrostatic contributions of these residues to the stabilisation of the substrates when the oxocarbenium species are formed (UDP-C1 distance of  $\sim 3$  and  $\sim 3.5$  Å for  $\alpha$ 13GalT and EXTL2, respectively) are given in parentheses (in kcal mol $^{-1}$  and with respect to reactants).

molecular origin of the enzyme regioselectivity was revealed by the analysis of both the active site geometrical evolution and the most relevant molecular interactions along the reaction pathways. It was shown that during either H-abstraction process AA progresses from a non-planar structure around C $_{13}$  or C $_{10}$  at the reactants to a planar AA-pentadienyl radical due to conjugation in a  $\pi$  system of five electrons delocalised over five carbon atoms (from C $_{11}$  to C $_{15}$  in the H $_{13}$ -abstraction and from C $_{8}$  to C $_{12}$  in the H $_{10}$ -abstraction). When the two H $_{13}$ - and H $_{10}$ -abstractions, initiated from the same initial MD structure and from similar initial  $d(\text{H}_{13}-(\text{OH}^- - \text{Fe}^{\text{III}}))$  and  $d(\text{H}_{10}-(\text{OH}^- - \text{Fe}^{\text{III}}))$  distances, were compared, it was concluded that the larger the geometrical change and, especially, the greater the steric hindrance that AA experiments from the non-planar structure at reactants to the planar AA-pentadienyl radical at products, the larger the potential energy barrier of the corresponding H-abstraction reaction. Therefore, it is the H $_{10}$ -abstraction the one that presents a higher potential energy barrier (30.9 kcal mol $^{-1}$ ) in front of the H $_{13}$ -abstraction (22.5 kcal mol $^{-1}$ ) because AA suffers more structural reorganisation along the H $_{10}$ -abstraction pathway, and those substrate motions are highly impeded by the steric hindrance of two conserved bulky residues, Leu597 and Ile663 (Figure 5.4). Consequently,

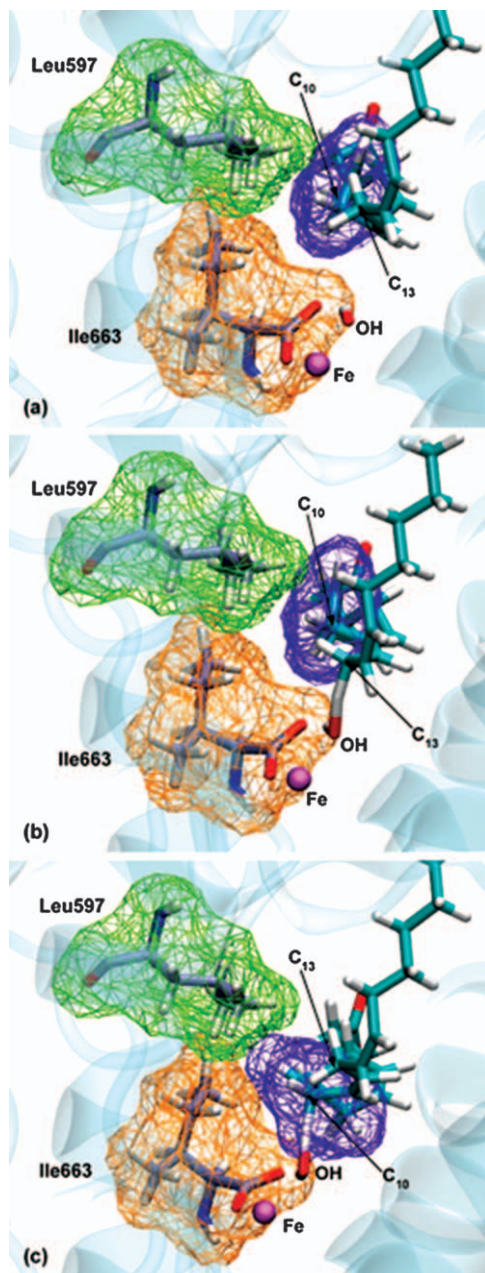


**Figure 5.4** Representation of the van der Waals surfaces of Ile663 (orange), Leu597 (green), and all atoms from C7 to C10 of AA (blue). Atoms of AA, OH, and Ile663 are highlighted by using the liquorice drawing method. These pictures correspond to the potential energy reaction path calculated from the same initial MD structure: (a) reactant; (b) TS of H<sub>13</sub> abstraction; (c) TS of H<sub>10</sub> abstraction. Reprinted with permission from P. Saura, R. Suardíaz, L. Masgrau, J. M. Lluch and À. González-Lafont, *ACS Catal.*, 2014, 4, 4351–4363. Copyright (2014) American Chemical Society.

those steric interactions were proposed to be the molecular origin of the exquisite regioselectivity of 15-lipoxygenase.

## 5.5 Free Energy Simulations

In order to obtain the activation free energy, an extensive sampling of the configurational space would need to be done. From the local equilibrium condition (see Section 5.3.2), an ensemble average over the paths leads to a single free energy barrier. Thus, when free energy methods are used to calculate the free energy barrier, molecular dynamics simulations are introduced along the reaction path. The activation free energy can be obtained from the calculation of the potential of mean force (PMF) along a reaction coordinate  $\xi(\vec{r}^{(N)})$ .<sup>25</sup> If  $P(z)$  is the classical mechanical probability density evaluated at  $\xi = z$  in a isothermal-isobaric ( $NPT$ ) ensemble, where  $N$  is the number of particles,  $P$  is the pressure and  $T$  is the temperature,

$$P(z) = \frac{1}{A} \frac{1}{h^{3N}} \int e^{-[H(\vec{r}^{(N)}, \vec{p}^{(N)}) + PV] / k_B T} \delta[\xi(\vec{r}^{(N)}) - z] d\vec{r}^{(N)} d\vec{p}^{(N)} \quad (5.3)$$

where  $A$  is a normalisation constant,  $h$  is Planck's constant,  $\vec{r}^{(N)}$  and  $\vec{p}^{(N)}$  denote the set of  $3N$  atomic Cartesian coordinates and conjugate momenta, respectively,  $H$  is the classical nuclear Hamiltonian (whose potential part is obtained by means of QM/MM calculations),  $V$  is the volume,  $k_B$  is Boltzmann constant, and  $\delta$  is a Dirac's delta function (the additive term  $PV$  in the exponential of the integral is absent in a canonic ( $NVT$ ) ensemble). Only nuclear configurations corresponding to  $\xi = z$  contribute to integral (eqn (5.3)). Then, the PMF at  $\xi = z$  is defined by

$$W(z) = -k_B T \ln P(z) \quad (5.4)$$

and the classical mechanical activation free energy profile,  $\Delta G_{\text{act}}(z)$ , is obtained by

$$\Delta G_{\text{act}}(z) = G_{\text{act}}(z) - G(z_R) = W(z) - W(z_R) - G_{R,T,\xi} + C(z) \quad (5.5)$$

where  $z_R$  is the reactant value of  $\xi$ ,  $G_{R,T,\xi}$  is the free energy corresponding to the reaction-coordinate motion at  $z_R$ , and  $C(z)$  is a correction term that arises from the Jacobian of the transformation from a locally rectilinear reaction coordinate (a linear combination of atomic Cartesian coordinates) to a generalised curvilinear reaction coordinate  $\xi$  (a non-linear combination).  $C(z)$  is identically zero for the rectilinear case and often small otherwise, and it is usually neglected. Note that in eqn (5.5)  $G_{R,T,\xi}$  appears because  $G(z_R)$ , but not  $G_{\text{act}}(z)$ , includes the contribution of the reaction-coordinate motion. However, eqn (5.3) involves integration over the momentum along the reaction coordinate. On the other hand,  $G_{R,T,\xi}$  is often not taken into account, in such a way that  $\Delta G_{\text{act}}(z)$  is obtained in practice just as a PMF difference. The free energy barrier is  $\Delta G_{\text{act}}(z^*)$ , where  $z^*$  turns out to be the



value of the reaction coordinate corresponding to the maximum of  $W(z)$  (the transition state).

The evolution of the system from the reactant valley to the product valley through the high-energy region (transition state) separating them is a rare event, which could be followed by a complete sampling of the configurational space by means of molecular dynamics simulations. This way, the PMF corresponding to the reaction could be determined using eqn (5.3) and eqn (5.4). In practice, the integral in eqn (5.3) is estimated as the fraction of structures with  $\xi = z$  that appears along the simulation. However, the exponential dependence of  $P(z)$  on  $W(z)$  shows that, when there is a free energy difference of a few  $k_B T$ ,  $P(z)$  for the higher energy regions is reduced considerably, only a few sample points would be obtained in those regions along the molecular dynamics simulation, and its calculation would involve very large statistical errors. The worst situation corresponds to the transition state, where the free energy is maximum. Unfeasible long molecular dynamics simulations would be required to fix this problem. Two methods are generally used to circumvent that trouble: Umbrella sampling (US) method and free energy perturbation (FEP) theory.

### 5.5.1 Umbrella Sampling Method

The umbrella sampling method<sup>26</sup> explores the configurational space using an artificial biasing potential energy function  $U'$ , which is added to the actual potential energy  $U$ . Then the configurations are selected employing a non-physical, non-Boltzmann distribution. So, the system is confined around some region of the configurational space and an enhanced sampling is achieved. After the simulation, the effect of the non-Boltzmann sampling has to be removed. The true probability density is obtained from the biased probability density  $P'(z)$  by

$$P(z) = \frac{P'(z)e^{\frac{U'(z)}{k_B T}}}{\left\langle e^{\frac{U'(z)}{k_B T}} \right\rangle_{H+U'}} = P'(z)e^{\frac{U'(z)-g}{k_B T}} \quad (5.6)$$

where  $\langle \rangle_{H+U'}$  indicates average according to the non-Boltzmann sampling. The biasing potential has to be chosen in such a way that this average converges easily.  $g$  is the free energy difference coming from the adding of the biasing potential.

When calculating free energy differences between two states, a good overlap between the distributions of configurational points corresponding to both states is needed. Then, a complete calculation requires a number of separate simulations, each confining the sampling around a different region of the configuration space (e.g. one point of the reaction path). In practice, this is like calculating a reaction path scan (Section 5.4.1) but running a



QM/MM MD simulation at each step of the reaction coordinate. To this aim, the reaction coordinate  $\xi$  is divided into intervals (windows), with a specific value  $z_i$  and a biasing potential  $U'_i$  assigned to each window  $i$ , trying that the sampling within each window is as uniform as possible. The biasing potential often takes the harmonic form

$$U'_i(z) = \frac{1}{2}K(z - z_i)^2 \quad (5.7)$$

where  $K$  is a suitable force constant.

Combining eqn (5.4) and eqn (5.6) the PMF for each window  $i$  can be obtained through

$$W(z) = -k_B T \ln P'_i(z) - U'_i(z) + g_i \quad (5.8)$$

$P'_i(z)_i$  is calculated as normalised histogram of the values of  $z$  occurring during the simulation within window  $i$ . In principle, the free energy constants  $g_i$  are unknown. Ultimately, the whole PMF can be obtained joining the overlapping unbiased probability densities of the set of windows along the reaction coordinate.

### 5.5.1.1 Weighted Histogram Analysis Method

The weighted histogram analysis method (WHAM)<sup>27-29</sup> has been the most used method to combine the contributions of the different windows. WHAM calculates the values  $g_i$  iteratively and provides the total  $P(z)$  along the reaction coordinate. The free energies are obtained self-consistently solving

$$e^{-\frac{g_i}{k_B T}} = \frac{\sum_{m=1}^{N_w} \sum_{k=1}^{n_m} \frac{e^{-\frac{U'_i(\xi(\vec{r}_{k,m}^{(N)}))}{k_B T}}}{\sum_{j=1}^{N_w} n_j e^{\frac{g_j}{k_B T}} e^{-\frac{U'_j(\xi(\vec{r}_{k,m}^{(N)}))}{k_B T}}}}{\sum_{j=1}^{N_w} n_j e^{\frac{g_j}{k_B T}} e^{-\frac{U'_j(\xi(\vec{r}_{k,m}^{(N)}))}{k_B T}}} \quad (5.9)$$

where  $\vec{r}_{k,m}^{(N)}$  stands for the atomic Cartesian coordinates of configuration  $k$  in the simulation in the window  $m$  ( $1 \leq m \leq N_w$ ),  $N_w$  is the number of windows,  $n_m$  is the number of configurations generated in window  $m$  and  $g_1 = 0$  by definition.

After convergence of the set of  $g_i$ ,

$$P(z) = \frac{\sum_{m=1}^{N_w} \sum_{k=1}^{n_m} \frac{\delta(\xi(\vec{r}_{k,m}^{(N)}) - z)}{\sum_{j=1}^{N_w} n_j e^{\frac{g_j}{k_B T}} e^{-\frac{U'_j(\xi(\vec{r}_{k,m}^{(N)}))}{k_B T}}}}{\sum_{j=1}^{N_w} n_j e^{\frac{g_j}{k_B T}} e^{-\frac{U'_j(\xi(\vec{r}_{k,m}^{(N)}))}{k_B T}}} \quad (5.10)$$

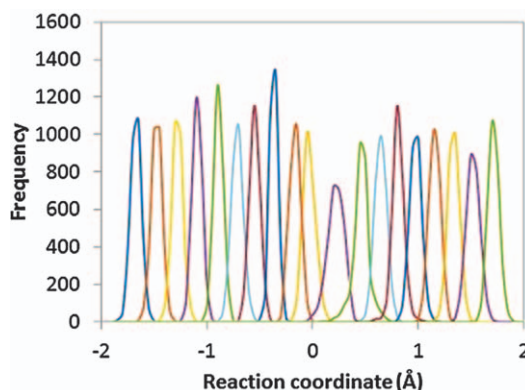


**Figure 5.5** Overlapping biased configurational distributions corresponding to the hydrogen abstraction reaction from arachidonic acid catalysed by the enzyme 8R-Lipoxygenase.

In practice, the Dirac's delta function in eqn (5.10) is calculated defining  $\delta(\xi(\vec{r}^N) - z) = 1$  if  $\xi(\vec{r}^N) \in \text{bin centered at } z$ , and zero otherwise. From here, eqn (5.4) provides the complete PMF along the reaction coordinate. It has to be noted that without a good overlap between the configurational distributions in adjacent windows the iterative cycle in WHAM fails to converge. An example of a set of sufficiently overlapping biased configurational distributions is shown in Figure 5.5, which corresponds to the hydrogen abstraction reaction from arachidonic acid catalysed by the enzyme 8R-Lipoxygenase.

The key assumption behind the WHAM analysis is that each simulation samples fully the corresponding window, in such a way that the corresponding equilibrium probability densities are reached. However, deviations from the equilibrium assumption can happen if the runs in each window are too short or the windows are too wide, so containing regions separated by barriers.

### 5.5.1.2 Dynamic Histogram Analysis Method

Recently, a new dynamic histogram analysis method (DHAM),<sup>30</sup> a generalisation of WHAM, was developed. DHAM reduces the errors coming from the WHAM analysis in the case of poorly converged simulations. Both WHAM and DHAM unbiased multiple biased simulations to get the underlying PMF. However, DHAM takes into account the dynamics and time sequence, and does not require that equilibrium is reached in each window. It can be also shown that WHAM is a special limiting case of DHAM.

DHAM builds the transition-count matrices  $T^{(m)}$  for each umbrella sampling simulation in the window  $m$ . The trajectories of the different biasing windows are projected into bins along the discretised reaction coordinate, so obtaining discrete time series of the bin indices. An element

of the matrix,  $T_{ji}^{(m)}$ , stands for the integer number of the observed transitions from bin  $i$  to bin  $j$  in simulation  $m$ . An unbiased Markov matrix of transition probabilities between the bins is determined from those matrix elements using a maximum-likelihood approach. Then, the unbiased equilibrium  $P(z)$  (*i.e.* the unbiased probability density corresponding to each bin) is calculated as the eigenvector of the Markov matrix corresponding to eigenvalue 1. One advantage of DHAM is that the Markov matrix also contains kinetic information and then can be used to estimate rate constants directly from umbrella sampling simulations.

The hydrogen abstraction from arachidonic acid catalysed by the human enzyme 15-lipoxygenase-2 has been the first enzyme reaction whose kinetics was determined directly from umbrella sampling simulations using the DHAM method.<sup>31</sup> The WHAM and DHAM free energy barriers turned out to be very similar. The rate constant with DHAM was obtained as the inverse of the average of the relaxation times at different lag times. The relaxation times were calculated as  $-\ln\lambda_2$ , where  $\lambda_2$  is the second largest eigenvalue of the corresponding Markov matrix. Since using the DHAM method the barrier-crossing times can be estimated, at least at an approximate level, directly from the global analysis of local umbrella sampling trajectories, the effect of the dynamical re-crossing (see Section 5.6) is incorporated. It was shown that the dynamical re-crossing at the canonical variational transition state was negligible for that reaction.

### 5.5.1.3 Two-dimensional Free Energy Surfaces

In some cases, several bonds are broken and formed simultaneously and then a two-dimensional free energy surface (2D FES) can be useful to analyse the reaction mechanism. The  $\beta$ -oxidation of butyryl-coenzyme A by short-chain acyl-CoA dehydrogenase is an example.<sup>32</sup> This oxidation occurs formally by abstraction of the  $\alpha$ -proton and  $\beta$ -hydride of the substrate. The 2D FES was built as a function of both  $z_1$ , the reaction coordinate of the proton transfer, and  $z_2$ , the reaction coordinate of the hydride transfer (both of them taken as the antisymmetric combination of the distances corresponding to the bond being broken and the bond being formed). The umbrella sampling method was used performing a series of molecular dynamics simulations including a biasing harmonic potential centered at each point (window) of the 2D grid ( $z_1, z_2$ ). The 2D FES was built by using WHAM and this free energy surface revealed that the mechanism is stepwise, with the proton transfer (free energy barrier height of 18 kcal mol<sup>-1</sup>) preceding the hydride transfer (which imposes a total free energy barrier height of 22 kcal mol<sup>-1</sup> for the overall reaction).

It is worth noting that transition state theory is not directly related to a 2D FES, but to 1D PMFs. So, although rate constants cannot be calculated from a 2D FES, these free energy surfaces can shed light on the reaction mechanism. After that, the suitable 1D PMFs for each of the reaction steps (proton transfer and hydride transfer in this case, as a function of the reaction

coordinates  $z_1$  and  $z_2$ , respectively) can be built. The 1D PMF energy barriers not necessarily match the ones extracted from the 2D FES, because an 1D PMF is not just a cut through the 2D FES but rather implies averaging over one more degree of freedom. The size of the discrepancy between them will depend on each particular system. In this present case, there was reasonable agreement.

Another interesting example is the methyl transfer reaction from *S*-adenosylmethionine to the nucleophilic oxygen atom of catecholate catalysed by catechol *O*-methyltransferase (COMT).<sup>33</sup> This reaction proceeds from a charged reactant to a neutral product. Then a large electrostatic coupling with the environment takes place. However, the different behaviour of the environment in the catalysed and non-catalysed reactions can be clearly seen comparing the corresponding 2D PESs (see Figure 5.6) built as a function of a solute ( $q$ ) and a solvent or environment ( $s$ ) degrees of freedom.

From reactant to product,  $q$  (an antisymmetric combination of the breaking and forming bond distances) changes from negative to positive values, whereas  $s$  (a measure of the electrostatic potential created by the environment) evolves from positive to smaller values. The position of the saddle point is very similar in both free energy surfaces ( $q^{\text{TS}} = 0.13$  and  $0.08 \text{ \AA}$  and  $s^{\text{TS}} = 67$  and  $78 \text{ kcal mol}^{-1} |e^{-1}|$ , in the enzyme and in aqueous solution, respectively). However, a minimum appears at the reactant region for the enzyme (at around  $s^{\text{Reactant}} = 85 \text{ kcal mol}^{-1} |e^{-1}|$ , close to the value corresponding to the saddle point), but not in aqueous solution (where  $s^{\text{Reactant}}$  must be higher than  $140 \text{ kcal mol}^{-1} |e^{-1}|$ ). These values show that the solvent environment must suffer an important reorganisation around the two reacting fragments in going from reactant to product in aqueous solution, but conversely, the enzyme is pre-organised, in such a way that the electrostatic properties of the active site at the Michaelis complex of COMT are already close to those corresponding to the saddle point. The transition state ensembles are also depicted in Figure 5.6.

### 5.5.2 Free Energy Perturbation Theory

Free energy perturbation (FEP) is a very popular method<sup>26</sup> for calculating the free energy difference between a reference or unperturbed system with a Hamiltonian  $H_0$  and a target or perturbed system characterised by a Hamiltonian  $H_1$ . As the sake of example, we will present here the theory corresponding to a canonical ensemble (NVT). So, the free energy difference between the perturbed 1 and the reference 0 systems can be classically written as

$$\Delta A_{0 \rightarrow 1} = A_1 - A_0 = -k_B T \ln \frac{\int e^{-\frac{H_1(\vec{r}^N, \vec{p}^N)}{k_B T}} d\vec{r}^N d\vec{p}^N}{\int e^{-\frac{H_0(\vec{r}^N, \vec{p}^N)}{k_B T}} d\vec{r}^N d\vec{p}^N} = -k_B T \ln \left\langle e^{-\frac{(H_1 - H_0)}{k_B T}} \right\rangle_0 \quad (5.11)$$

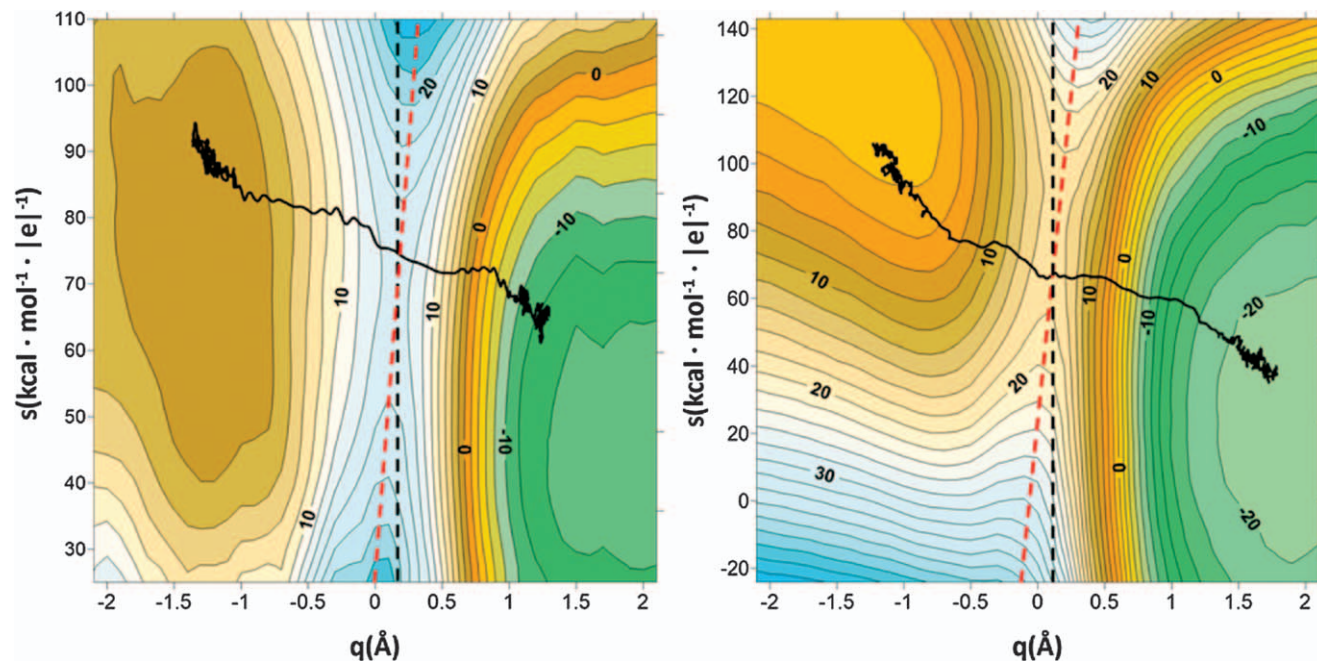


**Figure 5.6** 2D FESs (isocontour lines are drawn each  $2 \text{ kcal mol}^{-1}$ ) corresponding to the enzymatic reaction (left) and in solution (right). Averaged reactive trajectories are shown as a continuous line. TS surfaces obtained using the solute (1D) or the solute and the solvent coordinates (2D) are indicated by dashed lines in black and red, respectively. Reprinted with permission from R. García-Meseguer, K. Zinovjev, M. Roca, J. J. Ruiz-Pernía and I. Tuñón, *J. Phys. Chem. B*, 2015, 119, 873–882. Copyright (2014) American Chemical Society.

where  $\langle \rangle_0$  denotes an ensemble average over the canonical ensemble sampled from the equilibrated reference state. So, at every different configuration of the equilibrated unperturbed system the corresponding Hamiltonian is switched to the one corresponding to the perturbed system, and the differences (perturbation energies)  $\Delta H = H_1 - H_0$  are collected to calculate the ensemble average. Using Cartesian coordinates, the kinetic term is separable from the potential term and cancels out when the difference is made, in such a way that  $\Delta H = \Delta U$ .

The FEP can be computed in two opposite directions, forward ( $0 \rightarrow 1$ ) and backward ( $1 \rightarrow 0$ ). In principle,  $\Delta A_{0 \rightarrow 1}$  should match  $-\Delta A_{1 \rightarrow 0}$ , but usually hysteresis effects appear due to finite sampling errors. In general, the smaller the difference between the results along both directions, the better the convergence of the calculations. In order to minimise the error, the free energy is usually taken as the arithmetic mean of  $\Delta A_{0 \rightarrow 1}$  and  $|\Delta A_{1 \rightarrow 0}|$ .

When the perturbation is large (like in going from reactant to product states), a multi-window calculation is carried out. So, several intermediate systems between the reference (*e.g.* reactants) and the target (*e.g.* products) systems are constructed, each one with the Hamiltonian  $H_i$ , and each  $\Delta A_{i \rightarrow i+1}$  is computed by sampling at equilibrated system  $i$ . If there are  $N - 2$  intermediate systems, the total change of free energy is obtained by

$$\Delta A = \sum_{i=1}^{N-1} \Delta A_{i \rightarrow i+1} = -k_B T \sum_{i=1}^{N-1} \ln \left\langle e^{-\frac{(H_{i+1} - H_i)}{k_B T}} \right\rangle_i \quad (5.12)$$

The intermediate Hamiltonians are sometimes considered as a function of a parameter  $\lambda$ , such that  $\lambda = 0$  and  $\lambda = 1$  for the unperturbed and the target systems, respectively. Often the dependence is given by a linear function

$$H(\lambda_i) = \lambda_i H_1 + (1 - \lambda_i) H_0 \quad (5.13)$$

To improve the efficiency of the calculations, the forward and the backward FEPs are calculated at once. So, the sampling at the equilibrated system  $i$  is used to compute both  $\Delta A_{i \rightarrow i+1}$  and  $\Delta A_{i \rightarrow i-1}$ . This is the double-wide sampling.

### 5.5.2.1 Application to QM/MM Simulations (QM/MM FEP Method)

A variant of FEP can be used to restrict the sampling to the computationally cheaper force field part.<sup>34,35</sup> In the case of QM/MM simulations, the MM region can be sampled while the QM region is kept fixed. The perturbation consists of a geometry step of the QM atoms (for instance, along a suitable reaction coordinate).

For a reaction process, a reaction coordinate  $\xi$  depending only on the QM positions has to be defined. A set of discrete windows (intermediate systems) along  $\xi$  is also defined, each window being characterised by a specific value  $z_i$  of  $\xi$ . For each window  $i$ ,  $\xi$  is constrained to the corresponding value  $z_i$ ,

while all other QM and MM degrees of freedom are optimised. So, a minimum-energy path is built from reactants (reference system) to products (target system).

The free energy difference between two consecutive windows  $i+1$  and  $i$  (that corresponds to the difference between  $W(z_{i+1})$  and  $W(z_i)$ ) can be partitioned into a QM part and an environment part, which includes the contribution for the MM region and the QM-MM interactions:

$$\Delta A^{i \rightarrow i+1} = \Delta A_{\text{QM}}^{i \rightarrow i+1} + \Delta A_{\text{env}}^{i \rightarrow i+1} \quad (5.14)$$

The QM part accounts for four contributions:

$$\Delta A_{\text{QM}}^{i \rightarrow i+1} = \Delta U_{\text{QM}}^{i \rightarrow i+1} + \Delta E_{\text{QM,ZPE}}^{i \rightarrow i+1} + \Delta E_{\text{QM,th}}^{i \rightarrow i+1} - T \Delta S_{\text{QM}}^{i \rightarrow i+1} \quad (5.15)$$

where the four terms are the differences in potential energy, the zero-point energy correction, the thermal contribution to the internal energy, and the entropic contribution, respectively, between the QM regions of the two consecutive windows.

$U_{\text{QM}}$  is calculated from the QM wave function obtained by electrostatic embedding (that is, the MM point charges polarise the QM region) as the expectation value of the pure electronic QM Hamiltonian without the MM point charges. In other words,  $U_{\text{QM}}$  is computed as the QM electronic energy in the electrostatic embedding scheme minus the electrostatic interaction ( $E_Q$ ) between the QM and MM regions in order to avoid double counting (see below) of  $E_Q$ . The other three terms in eqn (5.15) require harmonic QM/MM frequency calculations for the QM region (including all MM atoms linked to QM atoms that are frozen during the molecular dynamics sampling). Note that the harmonic frequencies required at non-stationary points should correspond to vibrational modes orthogonal to the reaction coordinate.

The environmental part can be obtained as follows. In the window  $i$  a molecular dynamics sampling is carried out keeping frozen the optimised QM coordinates of this window. So, the MM atoms move under the forces created by the QM region. For each configuration along the molecular dynamics simulation, the perturbation energy (using Cartesian coordinates) is computed as

$$\Delta U_{\text{pert}}^{i \rightarrow i+1} = U_{\text{QM/MM}}(\vec{R}_{\text{QM}}^{i+1}, \vec{R}_{\text{MM}}^i) - U_{\text{QM/MM}}(\vec{R}_{\text{QM}}^i, \vec{R}_{\text{MM}}^i) \quad (5.16)$$

where  $\vec{R}$  stands for the nuclei coordinates, the subscripts QM and MM correspond to the QM and MM regions, respectively, and the superscripts  $i$  or  $i+1$  indicate the windows  $i$  or  $i+1$ , respectively.  $U_{\text{QM/MM}}$  includes all energy contributions coupling the QM and the MM regions, that is, the van der Waals interactions, the force field terms of the junctions, and the electrostatic interactions ( $E_Q$ ). Since  $E_Q$  is included here, it has to be excluded from the term  $U_{\text{QM}}$  (see above). That is, for each generated configuration, the perturbation is achieved changing the QM coordinates from positions of



window  $i$  to positions of window  $i + 1$ , but maintaining frozen the MM positions of that configuration.

The perturbation energies are introduced in the exponents of eqn (5.12) (as mentioned,  $\Delta H = \Delta U$ ) to calculate the corresponding exponential averages and thus  $\Delta A_{\text{env}}^{i \rightarrow i+1}$ . Using eqn (5.14),  $\Delta A^{i \rightarrow i+1}$  can then be obtained by adding  $\Delta A_{\text{QM}}^{i \rightarrow i+1}$ , and the summation of all those values provides  $\Delta A$  along the calculated path. Forward ( $i \rightarrow i + 1$ ) and backward ( $i + 1 \rightarrow i$ ) FEPs are usually calculated, and the convergence is checked.

$E_Q$  is in principle computed from a QM electron density obtained from full self-consistent field iterations in each molecular dynamic step along the simulations. However, we can introduce two approximations:

- (1) The QM electron density is frozen in each window. That is, the changes in the polarisation of the QM electron density due to the varying MM coordinates along the molecular dynamics sampling for a given window are neglected.
- (2) The QM density is represented by electrostatic-potential derived point charges (ESP charges).

This way, the QM calculations are avoided in the sampling runs. Then, when highly computationally demanding *ab initio* or DFT QM methods are employed, most of the computational time is spent on generating the minimum-energy path.

As the sake of example, in Figure 5.7 the forward and backward free energy profiles calculated along the pathway corresponding to the hydrogen

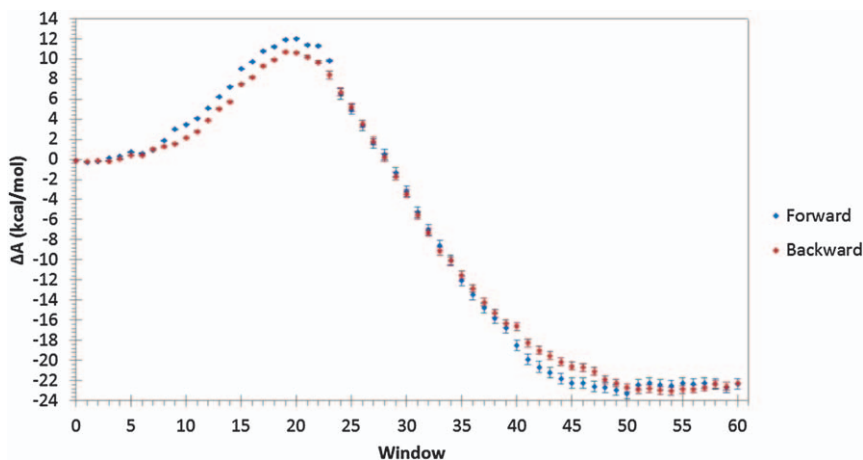


**Figure 5.7** Forward and backward free energy profiles vs. the reaction window along the  $H_{11}$ -abstraction from linoleic acid catalysed by the enzyme 15-lipoxygenase-1.

Reprinted with permission from J. Soler, P. Saura, D. García-López, L. Masgrau, J. M. Lluch and À. González-Lafont, *J. Phys. Chem. B*, 2016, **120**, 1950–1960. Copyright (2015) American Chemical Society.

(H<sub>11</sub>)—abstraction from linoleic acid catalysed by the enzyme 15-lipoxygenase-1 have been plotted as a function of the reaction window.<sup>36</sup> It has to be mentioned that the three last terms of eqn (5.15) were not included in those calculations.

As can be seen, the forward and backward free energy profiles agree quite well, so indicating a good convergence of the calculations. The maxima of both profiles permit to get the corresponding free energy barriers: 12.1 kcal mol<sup>-1</sup> (forward), 10.8 kcal mol<sup>-1</sup> (backward) and 11.4 kcal mol<sup>-1</sup> (average).

### 5.5.3 String Method: Minimum Free Energy Paths

In this section, another approach to introduce the free energy contributions in the calculation of energy barriers for enzymatic reactions is presented. This methodology is based on using the concept of CVs combined with building up a string reaction pathway to directly locate the minimum free energy path (MFEP) on the free energy surface without having to fully determine that surface.<sup>37</sup> The MFEP can be assumed to be lying at the center of a ‘reaction tube’ where most of the reactive trajectories go through, so the rest of the free energy surface is not explicitly needed to explain the enzyme reactivity. The NEB and string methods have been generalised to determine free energies. Working on the free energy space is easier because the free energy is less rugged than the underlying PES. As indicated above, the string method uses several replicas of the system that correspond to different, equidistant images between reactants and products basins in a given reaction coordinate space. When these replicas are allowed to relax, maintaining the equidistant parametrisation, this ‘string’ eventually converges to the MFEP with only the potential of mean force needed as an input. One of the greatest advantages of the string method and related techniques is that the computational cost is practically independent of the number of dimensions used, making affordable the study of very complicated processes. Assuming that the set of selected collective coordinates is representative of the reaction mechanism, it can be shown that the MFEP is the most likely path of transition between two free energy local minima. Then the isocommittor hypersurfaces can be calculated from the MFEP and the transition region can be identified *a priori*, defined as the region in the committor 1/2 surface with significant probability (that is, the region where the probability for going to reactants and to products is the same).<sup>38</sup> In the zero-temperature string method version a multidimensional harmonic constraint is used during the molecular dynamics in the CV space, so that each replica never goes far from some point on the string. The update of the path is done by following the approximated gradient of the free energy at these points. Thus, the MFEP is still a local approach not covering the free energy landscape in the orthogonal directions to itself. Once the string is converged the free energy profile along a given path CV can be calculated using whatever technique: umbrella sampling, metadynamics, *etc.* Note that the free energy along the MFEP is not a PMF because the degrees of freedom

orthogonal to the path are not sampled in the zero-temperature string method.

A combination of an *on the fly* version of the string method<sup>39</sup> with a set of CVs was used to locate the MFEPs of the reaction catalysed by guanidinoacetate methyltransferase.<sup>40</sup> This enzyme catalyses the methylation of guanidinoacetate by *S*-adenosyl-*L*-methionine, a complex reaction that involves a methyl transfer and a proton transfer. Using the four interatomic distances involved in the forming/breaking of chemical bonds as the set of CVs, the methodology converges to two different MFEPs when the underlying PES is described at the AM1/MM level. One MFEP corresponds to a reaction path in which the methyl is transferred first and the proton transfer takes place later, whereas along the second MFEP the methyl and proton transfers are reversed, both mechanisms corresponding to asynchronous stepwise processes. When those two converged strings were projected on a two-dimensional FES, it was verified that they really were MFEPs (see Figure 5.8). The umbrella sampling protocol was then used to obtain the corresponding PMFs (and the activation free energies) along the MFEPs as a function of a CV that measures the advance of the reaction process. The free energy profiles along the MFEPs are, in this case, a very good approximation to the PMFs because the curvature orthogonal to the path remains essentially constant. According to the corrected PMFs at the M06-2X level, the first mechanism, in which methyl transfer precedes proton transfer, is clearly

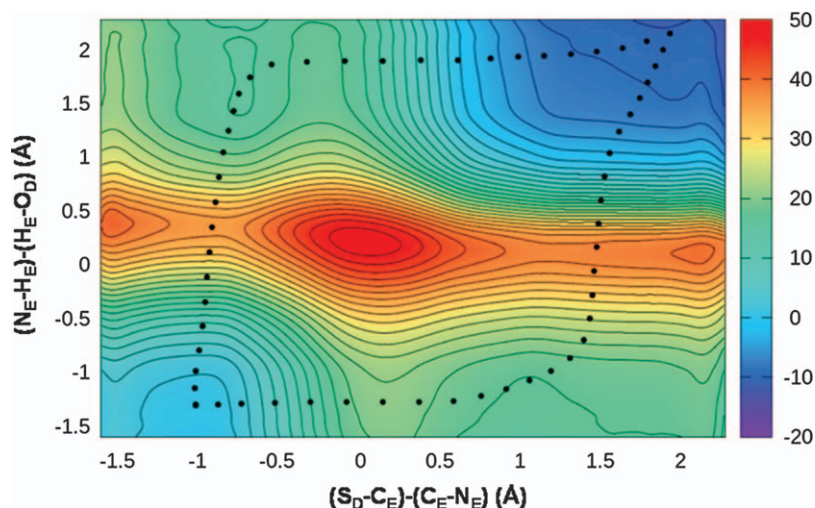


**Figure 5.8** Free energy surface for the guanidinoacetate methyltransferase catalysed reaction obtained as a function of the two antisymmetric combinations of distances defining the proton and the methyl transfer. Isocontour lines are drawn in  $\text{kcal mol}^{-1}$ . Projected on the surface (black dots) are the two converged MFEPs obtained with the string method.

Reprinted with permission from K. Zinovjev, J. J. Ruiz-Pernía and I. Tuñón, *J. Chem. Theory Comput.*, 2013, 9, 3740–3749. Copyright (2013) American Chemical Society.

favoured presenting a PMF barrier of about  $21.9 \text{ kcal mol}^{-1}$ , while in the alternative mechanism, where the proton transfer precedes methyl transfer, the PMF barrier is about  $33.2 \text{ kcal mol}^{-1}$ , and both mechanisms are now described as asynchronous but concerted. In a recent study, the same *on the fly* string method to obtain the MFEP on a multidimensional FES was used to unravel the still-controversial features of the reaction mechanism for cytosine C<sub>5</sub>-methylation in M. HhaI.<sup>41</sup>

In order to build up the reactions paths from global features of the free energy landscape a new approach was developed: the finite-temperature string method.<sup>42,43</sup> This method locates what it is called the transition principal curve as the reaction path for the catalytic mechanism. A principal curve is such that its intersection with each of the hyperplanes perpendicular to itself in some appropriate metric coincides with the averaged position of the system in these planes. In case of the finite-temperature version, the system is maintained in one of the Voronoi cells, defined by a series of points along the path. In this way, the system is free to move in directions orthogonal to the path and the principal curve can be recovered. The main advantage of the principal curve is that it is smooth even if the underlying free energy surface is rugged. In this method, the sampling within the cells depends on a delicate balance between their size and the free energy slope. However, in the case of enzymatic reactions important differences between the principal curve and the MFEP are not expected, especially when adequate collective variables are employed.<sup>40</sup>

Hummer and co-workers combined an adapted finite-temperature string method with US simulations to analyse the mechanism of several catalytic reactions.<sup>44</sup> The strategy in this case consisted of updating the string by sampling the most relevant phase-space regions using the US technique at points along the string in CVs to generate ensembles of configurations along it. Following this algorithm, the string also converges to the MFEP. The PMF can be obtained afterwards as a function of a given reaction coordinate without further simulations. Another alternative strategy is due to Roux and co-workers who used a protocol to update the string *on the fly* by means of swarms of trajectories (SoT) that refine the transition path also in the multidimensional space of a given set of CVs.<sup>45,46</sup>

The umbrella-sampling-like implementation of the adapted finite-temperature string method was employed to carry out a theoretical study of the nicotinamide adenine dinucleotide phosphate (NADP<sup>+</sup>) reduction to produce NADPH catalysed by flavoenzyme ferredoxin-NADP<sup>+</sup> reductase (FNR) in photosynthesis.<sup>47</sup> That reduction of NADP<sup>+</sup> to NADPH is proposed to take place by hydride transfer from the N<sub>5</sub> position of the flavin ring of FNR to the C<sub>4</sub> position of the nicotinamide ring (see Figure 5.9(a)). The aim of that paper was to understand the mechanistic role of Tyr303 because its substitution by a Ser practically deactivated the capacity of the enzyme to reduce NADP<sup>+</sup> as indicated by experimental measurements. The proposal was that Tyr303 might not be directly involved in the hydride transfer process but as it is located between the flavin ring of FNR and the nicotinamide

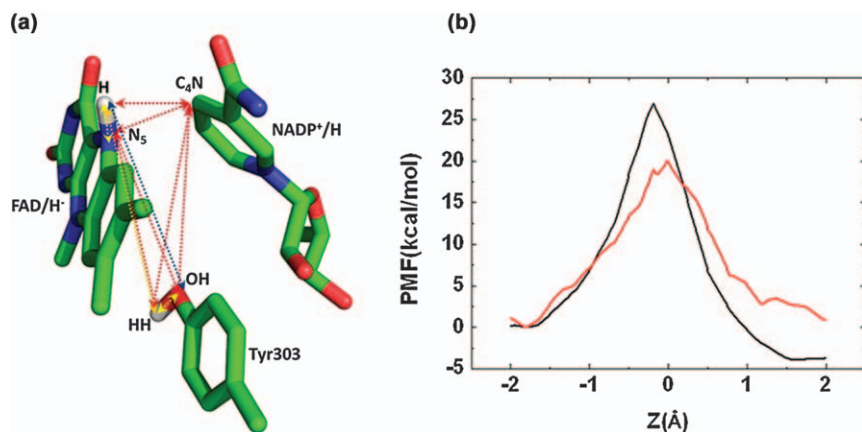


**Figure 5.9** (a) Interatomic distances included as collective coordinates in the definition of the string; (b) AM1/MM classical mechanical PMF (red line) with the amide group of the nicotinamide ring in the *cis* conformation, obtained from string simulations including the dihedral angle of that amide group as one of the collective coordinates. Dual-level QM/MM PMF (black line) based on the B3LYP-D energy correction. Reprinted with permission from I. Lans, M. Medina, E. Rosta, G. Hummer, M. Garcia-Viloca, J. M. Lluch and À. González-Lafont, *J. Am. Chem. Soc.*, 2012, **134**, 20544–20553. Copyright (2012) American Chemical Society.

ring of  $\text{NADP}^+$  could prevent the interaction between both rings. For this reason, during hydride transfer Tyr303 side-chain must be displaced from its position participating in this way in the nuclear motions that are significant in going from reactants to products. The initial string connecting the reactant and product structures was defined as the curve in the ten-dimensional space of CVs that connected the 45 images taken from a previously determined 1D AM1/CHARMM22 potential energy profile. In that set of ten collective coordinates several relevant interatomic distances and a dihedral angle related to the Tyr303 residue were incorporated. An AM1/CHARMM22 MD simulation was ran for each image using the US method with harmonic restraining potentials centered at each image and acting on the collective coordinates. Fitting of high-order polynomial functions over the obtained average collective coordinates along the string was employed to construct an updated string that was again divided in 45 segments. This procedure was repeated until the change in the restrained position of the collective coordinates fell below a given threshold. From the sampled restrained dynamics data, a dual-level PMF based on B3LYP-D energy corrections along the reaction coordinate  $\xi$  (defined as the difference between the  $\text{N}_5\text{--H}$  distance and the  $\text{H--C}_4\text{N}$  distance) and using a histogram-free implementation of the WHAM equations was obtained. The results show that the configurations generated by the sampling along the string (combined with umbrella sampling) have a greater statistical weight. When those

configurations are binned as a function of the  $z$  values of the reaction coordinate the correct PMF (with lower barriers) is obtained (see Figure 5.9(b)). Those calculated similar free energy barriers for both the forward and reverse reactions are in good agreement with the experimental hydride transfer rate constants reported for this system. In contrast, if  $z$  is used to directly guide the sampling low probability configurational regions are visited that cause hysteresis effects giving too high free energy barriers. Moreover, the Tyr303 displacement is smooth along the string free energy pathway whereas it is abruptly forced when the string is not used.

## 5.6 Calculation of the Reaction Rate Constant

As a matter of fact, the actual physical property that can be experimentally measured in a direct way for an enzymatic reaction is not the free energy barrier, but the reaction rate constant. The phenomenological free energy of activation is a logarithmic transformation of the experimental rate constant by means of eqn (5.1). However, the right dependence of the rate constant on the free energy barrier is more complicated than what is seen in eqn (5.1). Therefore,  $\Delta G_{\text{act,phen}}$  includes in general additional contributions and its direct comparison with the free energy barrier can be misleading when other factors, like re-crossing and tunnelling (and zero-point energy effects, depending on the approach used to calculate the free-energy barrier), are significant. So, the best kinetic quantitative comparison, if needed, between experiment and theory should be done by means of the reaction rate constant.

### 5.6.1 Ensemble-averaged Variational Transition State Theory with Multi-dimensional Tunnelling (EA-VTST/MT)

The more complete theoretical scheme to calculate reaction rate constants is likely ensemble-averaged variational transition state theory with multi-dimensional tunnelling<sup>48,49</sup> or methods based on it. EA-VTST/MT is based on canonical variational transition state theory, and includes non-classical nuclear quantum effects, namely, zero-point energy, tunnelling and dynamical re-crossing corrections. When hydrogen motion is involved, nuclear quantum effects become important: a large H/D kinetic isotope effect (KIE), indicative of quantum nuclear tunnelling, has been reported for several enzyme-catalysed  $\text{H}^\bullet/\text{H}^+/\text{H}^-$ -transfers, with soybean lipoxygenase-1 (KIE  $\sim 80$ ) showing the largest value; the inclusion of quantum vibrational energies (zero-point energies) typically reduces the barrier height by  $\sim 3 \text{ kcal mol}^{-1}$  for  $\text{H}^\bullet/\text{H}^+/\text{H}^-$ -transfer. However, transition state theory assumes that once trajectories originating on the reactant side of the transition state dividing surface cross it toward the product side, they do not return to the reactant side *via* re-crossing through that dividing surface. In



addition, all such reactants to products crossing trajectories must have been initiated at the reactant side. When this assumption does not hold, a re-crossing transmission coefficient may be used to account for its breakdown (calculated re-crossing transmission coefficients are usually between 0.6 and 1.0).

The complete dynamics approach requires three stages, but just the two first stages are usually carried out. The methodology has been recently reviewed and only the general approach will be explained here. In step 1 of stage 1 a classical mechanical (CM) potential of mean force (PMF),  $W^{\text{CM}}$ , is determined, using umbrella sampling molecular dynamics (or any other statistical mechanics method), as a function of a suitable distinguished reaction coordinate  $\xi$ . All atoms of the entire  $N$ -atom system are treated on an equal footing during the dynamical simulation. The classical mechanical activation free energy profile,  $\Delta G_{\text{act}}^{\text{CM}}(z)$ , is obtained from the CM PMF by

$$\Delta G_{\text{act}}^{\text{CM}}(z) = W^{\text{CM}}(z) - W^{\text{CM}}(z_R) - G_{R,T,\xi}^{\text{CM}} \quad (5.17)$$

Eqn (5.17) is actually eqn (5.5) with  $C(z)=0$ . The superscript has been added to underline that this is a nuclear classical mechanical calculation so far. A quantal vibrational correction,  $\Delta W_{\text{vib}}(z)$ , to the computed PMF is obtained in step 2 of stage 1 from instantaneous normal-mode analysis of an  $N_1$ -atom primary zone (for simplicity,  $N_1$  usually involves from 20 to 100 atoms). Then, the quasi-classical PMF,  $W^{\text{QC}}(z)$ , which contains quantised vibrations for the primary zone atoms except for the degree of freedom corresponding to  $z$ , is obtained by

$$W^{\text{QC}}(z) = W^{\text{CM}}(z) + \Delta W_{\text{vib}}(z) \quad (5.18)$$

The quasi-classical activation free energy profile,  $\Delta G_{\text{act}}^{\text{QC}}(z)$ , is obtained by

$$\Delta G_{\text{act}}^{\text{QC}}(z) = W^{\text{QC}}(z) - [W^{\text{CM}}(z_R) + G_{R,T,\xi}^{\text{CM}} + \Delta W_{\text{vib},R}] \quad (5.19)$$

where  $\Delta W_{\text{vib},R}$  is the value of the quantal vibrational correction at the reactant, including both the reaction coordinate and the vibrational modes orthogonal to it. The quassiclassical free energy barrier is  $\Delta G_{\text{act}}^{\text{QC}}(z_*^{\text{QC}})$ , where  $z_*^{\text{QC}}$  is the value of the reaction coordinate corresponding to the maximum of  $W^{\text{QC}}(z)$ , which doesn't necessarily coincide with the value of  $z$  at the  $W^{\text{CM}}$  maximum.

In stage 2, the calculation of the transmission coefficient in the static secondary zone (SSZ) approximation is carried out. This secondary zone includes  $N_2 = N - N_1$  atoms. For each member of the quasi-classical transition state ensemble, the secondary-zone atoms are frozen, the saddle point structures of the primary-zone atoms are located by treating them as a cluster embedded in the potential field of SSZ, and an isoinertial minimum-energy path through the  $3N_1$ -dimensional primary zone is determined. For each individual reaction path  $i$  the net transmission coefficient  $\gamma_i$  is obtained



as the product of two factors: the quassiclassical transmission coefficient  $\Gamma_i$ , which corrects the rate constant for dynamical recrossing, and the semi-classical transmission coefficient  $\kappa_i$ , which includes both tunneling and nonclassical reflection. The average net transmission coefficient is

$$\gamma = \langle \Gamma_i \kappa_i \rangle \quad (5.20)$$

and the value of the rate constant determined with the EA-VTST/MT method turns out to be

$$k^{\text{EA-VTST/MT}}(T) = \gamma(T) \frac{k_B T}{h} \exp \left( - \frac{\Delta G_{\text{act}}^{\text{QC}}(z_{\star}^{\text{QC}}, T)}{RT} \right) \quad (5.21)$$

By combining eqn (5.1) and (5.22) we have

$$\Delta G_{\text{act,phen}}(T) = -RT \ln(\gamma(T)) + \Delta G_{\text{act}}^{\text{QC}}(z_{\star}^{\text{QC}}, T) \quad (5.22)$$

where  $\Delta G_{\text{act}}^{\text{QC}}(z_{\star}^{\text{QC}}, T)$  includes quantised vibrational modes and a possible shift in the transition state location as compared to the classical one.

EA-VTST/MT has been applied to accurately calculate primary and secondary KIEs of enzyme-catalysed  $\text{H}^\bullet/\text{H}^+/\text{H}^-$  transfers (in wild-type and mutant enzymes) and to rationalise their values or unusual temperature dependence observed experimentally.<sup>50,51</sup> For example, in enzymes like soybean lipoxygenase-1,<sup>52</sup> methylamine and aromatic amine dehydrogenases<sup>53</sup> or methylmalonyl-CoA mutase,<sup>54</sup> the inclusion of nuclear tunnelling by multidimensional approaches in the calculation of the rate constants, was necessary to explain the large measured primary H/D KIEs ( $\sim 81$ , 18, 55 and 50, respectively); tunnelling enhanced the light isotope rates by up to two orders of magnitude. In the case of the hydride transfer catalysed by *E. coli* dehydrofolate reductase (DHFR), the contribution from the quantised bound vibrations (a reduction of the activation free energy by 2.5 and 1.9 kcal mol<sup>-1</sup> for H and D, respectively) was the main responsible for the observed primary KIE of  $\sim 3$ .<sup>55</sup> However, for the predicted (and afterwards confirmed by experiment) secondary KIE, tunnelling was the main cause of the deviation from the quasi-classical value. The weak temperature dependence of the KIE measured for this hydride transfer in the range 278–318 K was attributed to two factors: a shift in the location of the quasi-classical transition state with  $T$  (that is, the transition state corresponds to another ensemble of structures), with the associated changes in the quantised vibrational energies of the isotopes; and an unusual  $T$  dependence of the tunnelling transmission coefficient,  $\kappa$ , as a consequence of a more symmetric and thinner barrier at higher  $T$ .<sup>56</sup> And as last example, we mention a study aiming at analyzing the role of protein dynamics in DHFR catalysis. It was concluded that using a <sup>15</sup>N/<sup>13</sup>C/<sup>2</sup>H-substituted enzyme neither the quasi-classical free energy barrier nor tunnelling was altered, whereas the re-crossing was.<sup>57</sup> This result was interpreted as a lack of

significant participation of promoting motions in driving tunnelling (a matter strongly debated in the field).

## 5.7 Further Considerations about the Relationship between the Activation Free Energy and the Extension of the Sampling of the Configurational Space

As introduced in Section 5.3.2, the conformational landscape of an E:S system is very diverse and different E:S configurations may present quite different catalytic characteristics.<sup>7,24,58–60</sup> Depending on the time-scale of those fluctuations or conformational changes, two extreme scenarios can be envisaged: if they are fast (local equilibrium assumption), they may be properly sampled and introduced in the activation free energy by (some of) the free energy methods described in Section 5.5 (notice that not all of them perform structural sampling of the QM atoms). As these methods use QM/MM MD (starting from a single E:S structure coming from the (MM) MD), conformational sampling of the E:S complex and along the path is intrinsically introduced in the free energy barrier calculation. But, how much sampling is considered to be enough when the E:S complex flexibility and its effect on enzymatic catalysis are factors to be included? Due to the computational cost, the QM/MM simulation times are by now relatively short (*e.g.* 100 ps or 10 ps per window in an umbrella sampling simulation using semi-empirical or DFT QM methods, respectively). Therefore, the sampling of the configurational space, although introduced, may still be limited. In that sense, enhanced sampling techniques (*e.g.* replica exchange MD) have been developed trying to increase the short accessible time window that limits the configurational sampling within the unique selected reactant valley, and that also might limit the statistical accuracy of the calculated thermodynamic/kinetic properties. The answer to the above question is that the required amount of conformational sampling is highly dependent on the system, the problem of interest and the methodology used.

At the other end, there are the slow fluctuations or conformational changes, which will not be taken into account by standard applications of the free-energy methods. In those cases, potential energy methods might be used as an alternative to obtain the energy barriers. As indicated above, several E:S structures generated by the equilibrium MDs can be selected and multiple minimum-energy paths can be built up. Even corresponding to the same reaction step, those reaction paths can present rather different potential energy barriers. Then, a mean potential energy barrier is calculated to compare with the experimental phenomenological activation free energy barrier. Different possibilities exist for averaging barriers and the accuracy of the different formulations has been analysed.<sup>61</sup> Assuming that each snapshot reacts with a local rate constant that fulfils the conventional transition state theory, and that the free energy and the potential energy barriers have

similar values, the average potential energy barrier that would correspond to the overall rate constant could be obtained as follows:

$$\Delta E_{AV}^{\ddagger} = -RT \ln \left\{ \frac{1}{n} \sum_{i=1}^n \exp \left( -\frac{\Delta E_i^{\ddagger}}{RT} \right) \right\} \quad (5.23)$$

where  $R$  is the gas constant,  $T$  is the temperature,  $\Delta E_i^{\ddagger}$  is the potential energy barrier for each snapshot, and  $n$  is the number of snapshots generated. This exponential average expression has been shown to give a reasonable estimate of the barrier, as long as the different snapshots are well chosen. Another issue is how to determine the minimum number of enzymatic conformations required for the convergence of the average energy barrier.

There are different options though for selecting those E:S pre-reactive conformations. Obviously, the best choice would be to randomly select a number of points in time along an extensive unrestrained MD trajectory of the E:S complex. However, this is computationally very costly and another kind of selection must be done: (1) Those structures can be selected by carrying out a geometrical clustering of the E:S conformations. In this case, two main problems must be taken into account: (a) the probability of appearance of those clusters of pre-reactive geometries is not known; (b) the energetics of the reaction process might be biased by the filter imposed in the selection of the E:S structures; (2) Another option could consist of manually picking ‘transition-state-like’ geometries from the MD trajectory. In this case, the same two disadvantages of the previous protocol remain but, in addition, the ‘transition-state-like’ definition is ambiguous.

For example, in the study mentioned above, Saura *et al.*<sup>24</sup> used, amongst others, a distance criteria of  $d(\text{H}_{13}-(\text{OH}^- - \text{Fe}^{\text{III}})) \leq 3.0 \text{ \AA}$  to select structures for the study of  $\text{H}_{13}$  abstraction from arachidonic acid (AA) by the  $(\text{Fe}^{\text{III}} - \text{OH}^-)$  cofactor in 15-lipoxygenase. Snapshots were selected from three replica trajectories of 8 ns each having average  $d(\text{H}_{13}-(\text{OH}^- - \text{Fe}^{\text{III}}))$  of  $5.96 \pm 1.33$ ,  $6.49 \pm 1.13$  and  $5.52 \pm 0.90 \text{ \AA}$ . The distance criteria are such that the complexes selected are suitable, in principle, for catalysis. After QM/MM minimisation, the  $d(\text{H}_{13}-(\text{OH}^- - \text{Fe}^{\text{III}}))$  varies from 2.47 to 4.02  $\text{\AA}$ , which may be considered not a big difference in the context of MD applications, but is likely to be significant for catalysis. The corresponding calculated potential energy barriers span from 18.6 to 30.4  $\text{kcal mol}^{-1}$ . Assuming an equal probability of the starting states, the exponential averaged value of 19.6  $\text{kcal mol}^{-1}$  was obtained for the  $\text{H}_{13}$ -abstraction energy barrier in 15-LOX. An equivalent approach was applied to study  $\text{H}_{10}$ -abstraction. The exponential averaged energy barrier of 23.6  $\text{kcal mol}^{-1}$  for  $\text{H}_{10}$ -abstraction is in agreement with the experimental regioselectivity, as 15-LOX is a lipoxygenase that preferentially abstracts  $\text{H}_{13}$  and oxygenates on  $\text{C}_{15}$ . In this case, the analysis showed that the great dispersion obtained in the calculated potential energy barriers is due to the flexibility of enzyme residues but especially to the great number of positions the flexible substrate can adopt

inside such large active site. The different E:S binding-modes for the 15-LOX-AA system (and the associated reaction paths) correspond to different configurational regions that could not be properly sampled running a QM/MM MD simulation within a window of a free energy profile associated to a unique reaction path. The calculation of several reaction paths (and the use of eqn (5.23) or analogous) may be then nowadays the best practical way to take into account the contribution of that ensemble of reactive structures distributed in so different regions of the configurational space. Even better, although it would be computationally very expensive, the free energy barrier (e.g. PMF-based methods) should be calculated for each reaction path.

Ribeiro *et al.*<sup>59</sup> used a similar analysis to determine the range of barrier heights obtained for catalysis by HIV-1 protease. They selected 40 conformations of the E:S complex from replica MDs and obtained activation free energies (zero-point energies and entropic corrections were calculated for the QM region based on normal mode analysis) spanned from 14.5 to 51.3 kcal mol<sup>-1</sup>. The authors identified three main causes for that: an alternative mechanism for two of the snapshots, a different conformation of the active centre (particularly, the orientation of one proton and two key distances), and variations in the electrostatic field created by the enzyme due to its structural fluctuations. The exponential averaged value for the activation free energy was 16.5 kcal mol<sup>-1</sup>, and the study concluded that the thermal fluctuations they observe are fundamental for the enzymatic rate constant, as reaction will take place from transient enzyme conformations (that can be very few) associated.

Finally, it is interesting to comment that the existence of multiple reactive E:S conformations (with very slow interconversion rates) presenting different reactivity, has been related to the different rates that are found experimentally in single-molecule experiments, that is, when the activity of an individual enzyme molecule is measured. In a recent computational study on L-lactate dehydrogenase, 2D PMFs were calculated for two different E:S states differing in the conformation of a protein loop. The study concluded that the two states were reactive and that the different associated rates would be in agreement with the broad range of rate constants measured in single-molecule experiments.

More systematic studies are necessary to better assess which of these fluctuations or conformational changes are actually being introduced by free energy simulation methods (especially those that do not fall within the two extreme scenarios described), how much sampling is needed, and how multiple reactant valleys have to be taken into account.

## References

1. B. G. Miller and R. Wolfenden, *Annu. Rev. Biochem.*, 2002, **71**, 847.
2. S. C. L. Kamerlin and A. Warshel, *Faraday Discuss.*, 2010, **145**, 71.
3. E. Brunk and U. Rothlisberger, *Chem. Rev.*, 2015, **115**, 6217.
4. H. Eyring and A. E. Stearn, *Chem. Rev.*, 1939, **24**, 253.

5. H. M. Senn and W. Thiel, *Angew. Chem., Int. Ed.*, 2009, **48**, 1198.
6. L. W. Chung, W. M. C. Sameera, R. Ramozzi, A. J. Page, M. Hatanaka, G. P. Petrova, T. V. Harris, X. Li, Z. F. Ke, F. Y. Liu, H. B. Li, L. N. Ding and K. Morokuma, *Chem. Rev.*, 2015, **115**, 5678.
7. R. Lonsdale, J. N. Harvey and A. J. Mulholland, *Chem. Soc. Rev.*, 2012, **41**, 3025.
8. À. González-Lafont, T. N. Truong and D. G. Truhlar, *J. Phys. Chem.*, 1991, **95**, 4618.
9. J. J. Ruiz-Pernía, E. Silla, I. Tuñón, S. Marti and V. Moliner, *J. Phys. Chem. B*, 2004, **108**, 8427.
10. S. Marti, V. Moliner and I. Tuñón, *J. Chem. Theory Comput.*, 2005, **1**, 1008.
11. H. Gomez, F. Mendoza, J. M. Lluch and L. Masgrau, *Adv. Protein Chem. Struct. Biol.*, 2015, **100**, 225.
12. G. Henkelman, B. P. Uberuaga and H. Jonsson, *J. Chem. Phys.*, 2000, **113**, 9901.
13. M. Sánchez-Martínez, E. Marcos, R. Tauler, M. Field and R. Crehuet, *J. Phys. Chem. B*, 2013, **117**, 14261.
14. E. Weinan, W. Q. Ren and E. Vanden-Eijnden, *Phys. Rev. B*, 2002, **66**, 52301.
15. A. Aleksandrov and M. Field, *Phys. Chem. Chem. Phys.*, 2012, **14**, 12544.
16. J. Baker, *J. Comput. Chem.*, 1986, **7**, 385.
17. D. C. Liu and J. Nocedal, *Math. Program.*, 1989, **45**, 503.
18. X. Prat-Resina, J. M. Bofill, À. González-Lafont and J. M. Lluch, *Int. J. Quantum Chem.*, 2004, **98**, 367.
19. 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, *J. Mol. Struct.: THEOCHEM*, 2003, **632**, 1.
20. A. Heyden, A. T. Bell and F. J. Keil, *J. Chem. Phys.*, 2005, **123**, 224101.
21. S. Adel, F. Karst, A. González-Lafont, M. Pekarova, P. Saura, L. Masgrau, J. M. Lluch, S. Stehling, T. Horn, H. Kühn and D. Heydeck, *Proc. Natl. Acad. Sci. U. S. A.*, 2016, **113**, E4266.
22. H. Gómez, J. M. Lluch and L. Masgrau, *J. Am. Chem. Soc.*, 2013, **135**, 7053.
23. F. Mendoza, H. Gómez, J. M. Lluch and L. Masgrau, *ACS Catal.*, 2016, **6**, 2577.
24. P. Saura, R. Suardíaz, L. Masgrau, J. M. Lluch and À. González-Lafont, *ACS Catal.*, 2014, **4**, 4351.
25. G. K. Schenter, B. C. Garrett and D. G. Truhlar, *J. Chem. Phys.*, 2003, **119**, 5828.
26. *Series in Chemical Physics: Free Energy Calculations. Theory and Applications in Chemistry and Biology*, ed. Ch. Chipot and A. Pohorille, Springer, New York, 2007, vol. 86, pp. 1–517.
27. S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman and J. M. Rosenberg, *J. Comput. Chem.*, 1992, **13**, 1011.

28. M. Souaille and B. Roux, *Comput. Phys. Commun.*, 2001, **135**, 40.
29. F. Q. Zhu and G. Hummer, *J. Comput. Chem.*, 2012, **33**, 453.
30. E. Rosta and G. Hummer, *J. Chem. Theory Comput.*, 2015, **11**, 276.
31. R. Suardiaz, P. G. Jambrina, L. Masgrau, À. González-Lafont, E. Rosta and J. M. Lluch, *J. Chem. Theory Comput.*, 2016, **12**, 2079.
32. T. D. Poulsen, M. Garcia-Viloca, J. L. Gao and D. G. Truhlar, *J. Phys. Chem. B*, 2003, **107**, 9567.
33. R. García-Meseguer, K. Zinovjev, M. Roca, J. J. Ruiz-Pernía and I. Tuñón, *J. Phys. Chem. B*, 2015, **119**, 873.
34. J. Kästner, H. M. Senn, S. Thiel, N. Otte and W. Thiel, *J. Chem. Theory Comput.*, 2006, **2**, 452.
35. J. M. Dieterich, H. J. Werner, R. A. Mata, S. Metz and W. Thiel, *J. Chem. Phys.*, 2010, **132**, 35101.
36. J. Soler, P. Saura, D. Garcia-Lopez, L. Masgrau, J. M. Lluch and À. González-Lafont, *J. Phys. Chem. B*, 2016, **120**, 1950.
37. L. Maragliano, A. Fischer, E. Vanden-Eijnden and G. Ciccotti, *J. Chem. Phys.*, 2006, **125**, 24106.
38. K. Zinovjev and I. Tuñón, *J. Comput. Chem.*, 2014, **35**, 1672.
39. L. Maragliano and E. Vanden-Eijnden, *Chem. Phys. Lett.*, 2007, **446**, 182.
40. K. Zinovjev, J. J. Ruiz-Pernía and I. Tuñón, *J. Chem. Theory Comput.*, 2013, **9**, 3740.
41. J. Aranda, K. Zinovjev, K. Swiderek, M. Roca and I. Tuñón, *ACS Catal.*, 2016, **6**, 3262.
42. E. Weinan, W. Q. Ren and E. Vanden-Eijnden, *J. Phys. Chem. B*, 2005, **109**, 6688.
43. E. Vanden-Eijnden and M. Venturoli, *J. Chem. Phys.*, 2009, **130**, 194103.
44. E. Rosta, M. Nowotny, W. Yang and G. Hummer, *J. Am. Chem. Soc.*, 2011, **133**, 8934.
45. A. C. Pan, D. Sezer and B. Roux, *J. Phys. Chem. B*, 2008, **112**, 3432.
46. M. Sánchez-Martínez, M. Field and R. Crehuet, *J. Phys. Chem. B*, 2015, **119**, 1103.
47. I. Lans, M. Medina, E. Rosta, G. Hummer, M. Garcia-Viloca, J. M. Lluch and À. González-Lafont, *J. Am. Chem. Soc.*, 2012, **134**, 20544.
48. C. Alhambra, J. Corchado, M. L. Sanchez, M. Garcia-Viloca, J. Gao and D. G. Truhlar, *J. Phys. Chem. B*, 2001, **105**, 11326.
49. J. Z. Pu, J. L. Gao and D. G. Truhlar, *Chem. Rev.*, 2006, **106**, 3140.
50. M. Garcia-Viloca, J. Gao, M. Karplus and D. G. Truhlar, *Science*, 2004, **303**, 186.
51. L. Masgrau and D. G. Truhlar, *Acc. Chem Res*, 2015, **48**, 431.
52. I. Tejero, M. Garcia-Viloca, À. González-Lafont, J. M. Lluch and D. M. York, *J. Phys. Chem. B*, 2006, **110**, 24708.
53. K. E. Ranaghan, L. Masgrau, N. S. Scrutton, M. J. Sutcliffe and A. J. Mulholland, *ChemPhysChem*, 2007, **8**, 1816.
54. A. Dybala-Defratyka, P. Paneth, R. Banerjee and D. G. Truhlar, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 10774.

55. M. Garcia-Viloca, D. G. Truhlar and J. L. Gao, *Biochemistry*, 2003, **42**, 13558.
56. J. Z. Pu, S. H. Ma, J. L. Gao and D. G. Truhlar, *J. Phys. Chem. B*, 2005, **109**, 8551.
57. L. Y. P. Luk, J. J. Ruiz-Pernía, W. M. Dawson, M. Roca, E. J. Loveridge, D. R. Glowacki, J. N. Harvey, A. J. Mulholland, I. Tuñón, V. Moliner and R. K. Allemann, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, 16344.
58. A. Lodola, J. Sirirak, N. Fey, S. Rivara, M. Mor and A. J. Mulholland, *J. Chem. Theory Comput.*, 2010, **6**, 2948.
59. A. J. M. Ribeiro, D. Santos-Martins, N. Russo, M. J. Rarnos and P. A. Fernandes, *ACS Catal.*, 2015, **5**, 5617.
60. K. Swiderek, I. Tuñón, S. Marti and V. Moliner, *ACS Catal.*, 2015, **5**, 1172.
61. A. M. Cooper and J. Kästner, *ChemPhysChem*, 2014, **15**, 3264.



## CHAPTER 6

# *Computational Chemistry Tools in Glycobiology: Modelling of Carbohydrate– Protein Interactions*

ALESSANDRA LACETERA,<sup>a</sup> M. ÁLVARO BERBÍS,<sup>a</sup>  
ALESSANDRA NURISSO,<sup>b</sup> JESÚS JIMÉNEZ-BARBERO<sup>c</sup> AND  
SONSOLES MARTÍN-SANTAMARÍA<sup>\*a</sup>

<sup>a</sup> Center for Biological Research, CIB-CSIC. Ramiro de Maeztu, 9, 28040-Madrid, Spain; <sup>b</sup> School of Pharmaceutical Sciences, University of Geneva, University of Lausanne, Rue Michel Servet 1, CH-1211 Geneva 4, Switzerland; <sup>c</sup> CIC-bioGUNE, Bizkaia Science and Technology Park, 48160-Derio, Spain

\*Email: smsantamaria@cib.csic.es

## 6.1 What are the Carbohydrates?

Carbohydrates (glycans, sugars, saccharides) are familiar compounds that we use, for example, in our daily diet.<sup>1</sup> However, when viewed at the atomic level, they are complex molecules whose chemical diversity is immense.<sup>2</sup> They exist in nature either as single entities or as part of glycoconjugates, essentially glycolipids and glycoproteins. But even within these classes, they show a vast chemical variability. The possibilities of branching at different hydroxyl groups, as well as the possible presence of pendant substituent groups, either neutral or charged, provide them with the possibility of

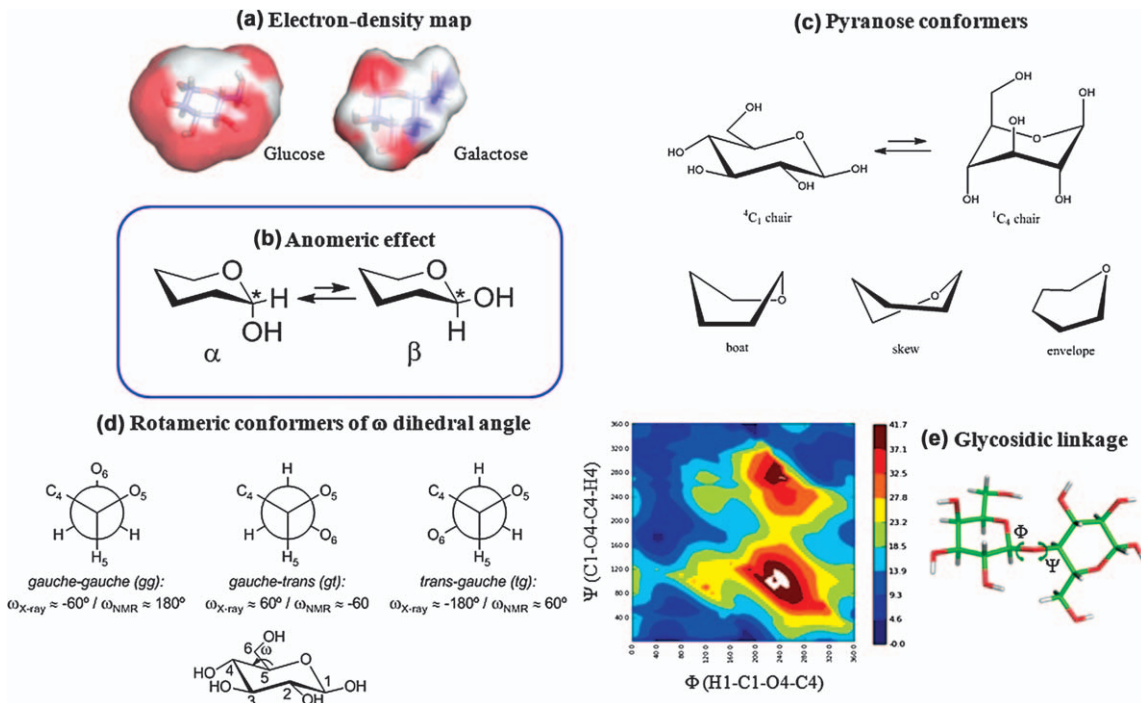


**Figure 6.1** (a) Electron density map of glucose and galactose. Polar patches are coloured in red and blue (high and low electron-density respectively); apolar patches are coloured in white. In contrast to glucose, alpha face (shown) of galactose is hydrophobic. (b) Preferential axial orientation ( $\alpha$ ) of electronegative substituents at the anomeric position ( $C^*$ ) determined by the anomeric effect. (c) Conformations of the monosaccharide  $\beta$ -D-glucopyranose: (up) regular chair conformations; (down) other possible conformers (boat, skew and envelope). (d) Newman projections (top) of the *gauche-gauche* (*gg*), *gauche-trans* (*gt*) and *trans-gauche* (*tg*) rotameric conformers of the  $\omega$  dihedral angle (defined as O5-C5-C6-O6 torsion). (e) Key torsion angles ( $\Phi$  and  $\Psi$ ) that define the conformation around the glycosidic linkages, and the energy diagram (energy in kcal mol $^{-1}$ ) calculated with the MM3 force-field, defining the energetically preferred conformations of a disaccharide (lactose).

encoding valuable information.<sup>2–4</sup> In turn, this variability generates a huge possibility of constitutional isomers and a plethora of conformations, encoding for diverse biological information through interactions with their biological partners. Moreover, their chemical properties, in terms of the presence of polar and non-polar patches (Figure 6.1), makes them exhibit peculiar stability and when acting as ligands for receptors, to be fairly adaptable to a variety of environments. They may adopt different shapes and display rather distinct conformational and dynamic properties. Glycan diversity depends not only on the monosaccharides, which they are composed of, but also on the glycosidic linkage and the anomeric effects (Figure 6.1), the orientation of all the torsional angles,<sup>5,6</sup> and the intramolecular hydrogen bonding between adjacent OH groups.<sup>7</sup> Therefore, it is of paramount importance to predict and characterise their conformational and structural properties in a systematic manner. Computational techniques have proven to be essential in this aspect.<sup>8</sup>

## 6.2 From Mono to Polysaccharides: An Overview of the Increasing Complexity

### 6.2.1 Monosaccharides

Carbohydrate molecules are built from small constituents, which are monosaccharides. Many different monosaccharides exist in nature. They may have a different number of carbon atoms—usually between five and nine—although other possibilities also exist. Due to the presence of different substituents for almost every carbon atom, these become stereogenic centers. Therefore, different sugars may exist (D and L), with identical chemical composition, but with distinct molecular recognition features when interacting with receptors. Even simple monosaccharides can adopt rather distinct three-dimensional shapes. They are usually presented as six-membered pyranose rings, although five-, seven-, and even eight-membered rings can be found. For the common pyranose rings, the regular geometry is one of the possible chair conformations, which minimises steric congestion and accounts for the anomeric effect. However, depending on the orientation, nature and number of substituents, other shapes may also form, from twisted boat geometries to alternative chairs or other possible conformations (*e.g.* skew or envelope, Figure 6.1). It has been postulated, however, that a rapid interconversion between the major chair conformation and some other minor geometries may take place.<sup>9</sup> One paradigmatic example is that of the sulfated iduronic acid (IdoA) ring that exists as an ensemble of different conformations in dynamic equilibrium, which is altered by substitution features<sup>10,11</sup> and interaction with biological partners.<sup>12,13</sup> Additionally, in solution, an important role is played by the water molecules that establish carbohydrate–water interactions, competing with intramolecular hydrogen bonds, and thus modulating the conformational behaviour and dynamics of carbohydrates.<sup>14</sup>

Regarding the conformation of the dihedral angle ( $\omega$ ) around C5-C6 in a pyranose ring ( $\omega$  formed by O5-C5-C6-O6 according to the X-ray definition or by H5-C5-C6-O6 according to the NMR definition), a combination of solvation, steric, and stereoelectronic (*gauche* effect) effects drives the observed geometries. It has been observed that for Glc-type configurations (equatorial OH-4), a combination of *gg* and *gt* rotamers coexist, while for Gal-type moieties (axial OH-4), this bond adopts the *gt* and *tg* geometries (Figure 6.1). This experimental behaviour has also been validated using computational protocols.<sup>15</sup>

### 6.2.2 Disaccharides: The Glycosidic Linkage and the *Exo-anomeric* Effect

A disaccharide is not the mere junction of two monosaccharides where the anomeric oxygen of one particular monosaccharide is attached to one secondary (or primary) hydroxyl group of the second sugar moiety, forming what is named the “glycosidic linkage” (Figure 6.1). The glycosidic linkage is an acetal linkage with particular chemical properties (sensitive to strong acid and basic conditions) and conformational features, which will be finally responsible of the binding (and biological) properties of the glycan.<sup>16</sup>

If the substituent at the anomeric center shows lone pair electrons, as an alkoxy group, retrodonation can proceed from the exocyclic oxygen to the sigma anti-bonding orbital of the intracyclic bond C-O. This is known as the *exo-anomeric* effect. In this case, the lone-pair orbital has to be anti-periplanar to the antibonding orbital of C-O, and consequently the alkyl substituent at the glycosidic oxygen adopts a *syn*-type orientation. The molecular orbital theory is also able to predict the modification of the bond length. In fact, for equatorially substituted pyranosides, the intracyclic C-O bond is longer than the exocyclic C-O bond.<sup>17</sup> Actually, these preferences lead to a *syn*-type conformation, although it has been observed that a minor proportion of *anti* conformation may take place in  $\beta$ -glycosides (*e.g.* in 1 $\rightarrow$ 2, 1 $\rightarrow$ 3 and 1 $\rightarrow$ 4 linkages) provided that the contiguous OH-2 displays an equatorial orientation (*i.e.* glucose or galactose). The *anti*-type geometry is also favoured by the *exo-anomeric* effect.<sup>18</sup>

The anomeric effect was first described in carbohydrates in the 1950s,<sup>19-21</sup> and describes the tendency of the electronegative substituents at the anomeric carbon (C1) of a pyranose ring to adopt the axial rather than the equatorial orientation, in contrast to what would be expected based solely on terms of steric factors. Therefore, both C1 isomers can co-exist but one of them is usually preferred. Normally for D-sugars, the  $\alpha$  isomer is more stable than expected. The origin of the anomeric effect has been a matter of historical controversy. Over the last decades, modelling approaches including *ab initio* calculations and molecular mechanics (MM) studies have allowed us to gain insights into all factors that come into play to produce this fascinating phenomenon.<sup>22,23</sup> On the basis of these studies, the origin of the anomeric effects are now understood as the result of multiple steric, electrostatic and stereoelectronic interactions,

although it is still the subject of virulent scientific debates.<sup>24</sup> In fact, in a recent example, a combination of quantum mechanics (QM) calculations with natural bond order analysis, chemical synthesis and NMR methods, has shown that it is possible to restore the anomeric effect for an acetal when replacing one of the oxygen atoms by a CF<sub>2</sub> group, providing key findings in chemical sciences.<sup>25</sup> On the one hand, this work strongly suggested the key relevance of the stereo-electronic component of the anomeric effect. On the other hand, the CF<sub>2</sub> analogue adopted the natural glycoside conformation, which might provide new avenues for the design of glycomimetics with sugar-based drug design.<sup>26</sup> Finally, it is important to mention that, although the anomeric effect is a well-known phenomenon for carbohydrate chemists it is not exclusive to the field of sugars, also appearing in many other types of molecules.<sup>25</sup>

Generally speaking, combined modelling/NMR approaches are mandatory for characterising the glycosidic linkage of a particular disaccharide or oligosaccharide (either  $\alpha/\beta$ , either 1  $\rightarrow$  2, 1  $\rightarrow$  3, 1  $\rightarrow$  4, or 1  $\rightarrow$  6).<sup>27,28</sup> All the particular features mentioned above for saccharides, besides their intrinsic dynamic properties, require the combination of rigorous experimental and theoretical protocols.<sup>29</sup>

### 6.2.3 Studying the Conformations Around the Glycosidic Linkage

Different modelling protocols can be employed to describe the conformational behaviour around a particular glycosidic linkage. To describe the behaviour of these torsional angles, the idea of the Ramachandran plot in peptides was extended to di- and even to oligo- and polysaccharides.<sup>30,31</sup> In analogy with the peptides, the relative orientation between two consecutive sugar residues can be defined by the  $\varphi$  and  $\psi$  torsional angles. In these maps, the energy values for all possible  $\varphi$  and  $\psi$  torsion values are systematically calculated and then plotted together, generating a simply projection of the potential energy surface, indicating in the minima the energetically allowed conformations of the disaccharide (Figure 6.1). For oligosaccharides, a map for every glycosidic linkage present in the molecule is necessary to get a suitable evaluation of the overall shape. Despite the simplification in considering each disaccharide moiety independent from the influence from the other saccharides, especially in the case of branched oligosaccharides, these maps still represent a good tool to help in the general understanding of the behaviour of these molecules. Maps of the most common disaccharidic species are available on-line at the Glyco3D website.<sup>32</sup> Obviously for long and heavily branched oligosaccharides, the systematic approach may lead to failures and it is advisable to resort to molecular dynamics (MD) simulations.

### 6.2.4 Oligosaccharides

Different oligosaccharides may adopt a variety of shapes and display rather distinct dynamic features. In fact, they may resemble chain-like structures,

as the LacNAc (Gal $\beta$ 1  $\rightarrow$  4GlcNAc) family, from LacNAc to diLacNAc (Gal $\beta$ 1  $\rightarrow$  4GlcNAc $\beta$ 1  $\rightarrow$  3Gal $\beta$ 1  $\rightarrow$  4GlcNAc) and triLacNAc (Gal $\beta$ 1  $\rightarrow$  4GlcNAc $\beta$ 1  $\rightarrow$  3Gal $\beta$ 1  $\rightarrow$  4GlcNAc $\beta$ 1  $\rightarrow$  3Gal $\beta$ 1  $\rightarrow$  4GlcNAc) or branched structures, with special mention to the blood group determinant oligosaccharides (the ABH-system), the core 2/4 *O*-glycans, and the GM1 oligosaccharide, the sugar part of the corresponding ganglioside. Their conformational properties, adaptability and recognition features have been and still are deeply analysed.<sup>33</sup>

Chemical methods have been employed to provide stable chemical analogues of oligosaccharides, with increased resistance to glycosyl hydrolases. Typically, either the glycosidic oxygen (inter-unit) or the endocyclic oxygen within the ring is substituted by other chemical element. Therefore, C-,<sup>34</sup> S-,<sup>35</sup> Se-,<sup>36</sup> and N<sup>37</sup>-glycosyl analogues have been synthesised to provide chemical probes for interaction studies or to prepare enzyme inhibitors. Given the different chemical properties of these analogues, their conformational properties have been carefully analysed and compared to those of their parent *O*-glycosides.<sup>38</sup> Computational studies have been directed towards the analysis of the flexibility of these new linked glycosides, and the preferences of the binding proteins to select the appropriate “binding pose” from the conformational assemble, not always being the most stable conformer in terms of potential energy.<sup>39–42</sup>

### 6.2.5 *N*-glycans

In glycoproteins, oligosaccharides are present as *N*-glycans. The different substitution patterns make them to be named as displaying antennas. In fact, depending of the degree of branched substituents, they are bi-antennary, tri-antennary and even tetra-antennary molecules. The different substitutions provide these molecules with different conformational and dynamic features,<sup>43</sup> which can be exploited for recognition. Additional core substitutions at certain positions might modify the dynamics of the different arms in a rather dramatic way. These aspects have been addressed in computational studies of *N*-glycans.<sup>43,44</sup> Novel NMR methods have also been applied to deduce their conformational behaviour.<sup>45–48</sup>

### 6.2.6 Polysaccharides

One ubiquitous class of oligo- and polysaccharides in nature are the glycosaminoglycans (GAGs), a term that encompasses a variety of sulfated saccharides. Key compounds are heparin and heparan sulfate, as well as hyaluronic acid.<sup>49</sup> The relevancy of GAGs is owed to their recognition by different receptors, triggering a plethora of biological responses, including inflammation, cell adhesion, and regulation of cell growth and proliferation.<sup>4,50</sup> The structure of these molecules has been extensively studied using a variety of experimental (preeminently in the shape of NMR restraints) and modelling methods.<sup>51,52</sup> Special mention should be given to ‘heparin pentasaccharide’ one of the carbohydrate-based drugs in clinic,

used as an anti-thrombotic agent.<sup>53,54</sup> The study of the conformational properties of this molecule is one paradigmatic example of the employment of computational methods and NMR, both in its free and bound state to a variety of receptors. As a token, the application of both protein- and ligand-detected NMR methods, assisted by computational protocols have allowed for the assessment of the binding mode of the heparin pentasaccharide and related GAGs to fibroblast growth factors,<sup>55</sup> as well as the conformation of the GAGs in the bound state.<sup>13</sup> These studies allowed identification of the bioactive structure of the complex, and revealing of a receptor-driven conformational selection of particular IdoA ring puckers.

Plant polysaccharides have many possible geometries, shapes and therefore, physical and chemical properties. Many of them have been extensively studied, such as pectins, which contain L-rhamnose and arabinose and have complex branched structures,<sup>56,57</sup> and others, such as chitin (GlcNAc $\beta$ 1  $\rightarrow$  4GlcNAc)<sub>n</sub>, which displays a linear shape.<sup>58</sup>

Glycolipids are amphiphilic molecules composed by a polar sugar part attached to a lipophilic lipid chain. They display rather distinct structural and conformational properties. In this regard, the conformational properties of different lipooligosaccharides (LOS) and lipopolysaccharides (LPS) from different sources can be accessed by modelling protocols, in combination with NMR methods.<sup>59,60</sup>

Nodulation (Nod) factors are lipochitooligosaccharides incorporating N-acetylglucosamines that mediate the recognition between *Rhizobium* bacteria and their legume symbionts. Further modifications, such as methylations, sulfations, and fatty acids, account for the recognition between the specific partners. The structure of Nod factors and their recognition properties by receptors have also been exhaustively modelled *in silico*.<sup>35,61,62</sup>

## 6.3 Computational Methodologies for the Study of Carbohydrates

Given the difficulty of obtaining proper crystals of carbohydrates beyond a particular (relatively small) size, molecular modelling has proven to be very useful, either alone or in combination with experimental techniques, usually solution-state NMR spectroscopy methods. In fact, the combination of modelling/NMR protocols has been extremely useful to deduce the conformational and dynamic properties of free and bound carbohydrate molecules.<sup>34,59,63</sup>

Molecular modelling of carbohydrates can be performed at different levels of complexity. Specific structural properties of carbohydrates can be extracted from QM methods, especially those related to electronic features, to study chemical reactions (for instance, by using hybrid QM/MM approaches), and to calculate force constants and atom charges to be used as force field parameters.<sup>64–66</sup> These methods can nowadays be employed even to deal with molecular recognition details. Indeed, *ab initio* calculations provide a very good precision on one hand, although on the other hand at a



high computational cost. A reasonable approach can consist of preliminary calculations by means of semi-empirical methods followed by more precise studies at *ab initio* level. Early studies used semi-empirical methods as AM1, PM3, and MNDO.<sup>67</sup> More recently, also PM5 and PM6 have been used for carbohydrates.<sup>68</sup> DFT has also been employed for conformational studies of carbohydrates using different basis sets,<sup>69–72</sup> and also solvation models.<sup>73</sup>

The most employed protocols for the study of carbohydrates are based on force fields, which are used for performing molecular mechanics (MM) and molecular dynamics (MD) simulations (Figure 6.2). Research in force field development has been very active for years, yielding several high quality force fields able to reproduce the dynamic and electronic properties of carbohydrates. They differ in how specific aspects are considered during the calculations.<sup>8,33</sup> In particular, MD simulations are one of the most useful tools to have insights into the conformational dynamics of the system, combined, if available, with NMR experimental information. Comparison of these force fields is commented in below.

The interaction of carbohydrates with their receptors can also be performed at different levels of complexity, also using force-field based methods (MM and MD simulations) or quantum mechanics methods (Figure 6.2).<sup>74,75</sup> Choosing the appropriate computational method is a key decision in a given study, and it will depend on the properties under investigation, and also on the starting experimental information we have in our hands. Usually, the availability of the 3D structure of the receptor (by means of X-ray crystallography, NMR spectroscopy or homology modelling) is mandatory. In some cases, the 3D structure of the carbohydrate/receptor complex is also

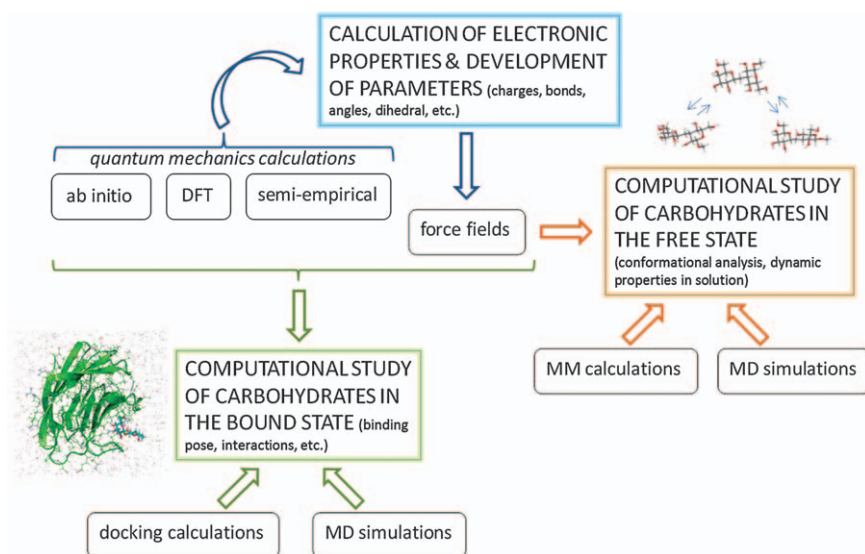


**Figure 6.2** Computational methodologies employed for the study of carbohydrates.

available,<sup>76</sup> although this is not frequently the case. Moreover, there are serious concerns with the available X-ray structures of sugar complexes regarding the shapes of the bound carbohydrates, which have been frequently described with incorrect geometries.<sup>77</sup>

Computational docking protocols are usually adopted to explore the possible interactions (and putative binding poses) of a given carbohydrate ligand to one particular receptor and to estimate the binding affinity of the complex. They normally consist in a conformational sampling of the ligand into the binding site of the receptor, usually rigid or partially flexible,<sup>78</sup> followed by the evaluation of the interaction energy of the predicted binding modes, by applying scoring functions. AutoDock is one of the most used molecular docking software<sup>79</sup> for the study of carbohydrate–protein interactions. Indeed, other programs have also been employed to deal with interactions involving carbohydrates, such as Glide, Vina, and Dock.<sup>49,80</sup> Since docking programs are not specifically developed to treat with carbohydrates, further refinement of the proposed docked poses may need to be accomplished by means of MD simulations in water. If available, experimental information (*i.e.* NMR/X-ray data, binding/affinity data, *etc.*) will also be taken into account in the calculations.

Despite the recent efforts in developing docking programs specific for predicting carbohydrate–protein interactions,<sup>78,81,82</sup> further refinement of the proposed docked poses may need to be accomplished by means of MD simulations. MD simulations are used to simulate how biological systems such as carbohydrates or carbohydrate–protein complexes evolve as a function of time under physiological conditions, *i.e.* by ensuring the full flexibility of the systems embedded in a solvated environment. From those calculations, structural, dynamic, and thermodynamic information can be estimated *in silico*.<sup>83</sup> The freely available molecular visualisation programs such as VMD<sup>84</sup> and UnityMol/SweetUnityMol<sup>85</sup> have been developed for allowing the representation of modelling results concerning carbohydrates, either alone or in their interactions with other biomolecules.

## 6.4 Force Fields for Carbohydrates

In order to study carbohydrate structures and the dynamic and energetic behaviour of carbohydrates and protein–carbohydrate complexes using molecular modelling techniques, appropriate force fields have been developed. Based on empirical parameters, they allow the reproduction *in silico* of the particular dynamic and stereo-electronic properties characterising carbohydrate systems, in both unbound and bound state.<sup>8,33</sup> In this scenario, it is very important the choice of a proper force field, able to reproduce the behaviour of carbohydrates in solution. Comparison of the performance of different force field for analysing carbohydrate molecules has been described.<sup>8,23,63</sup> Those force fields have been regularly used, in combination with structural experimental information, to gain insights into the dynamic

and energetic behaviour of carbohydrates and protein–carbohydrate complexes. Indeed, the user should be aware about the compatibility among force fields when modelling protein–carbohydrate complexes.

GLYCAM<sup>64</sup> is the most known force field, designed to add carbohydrate functionality to AMBER force field. QM data were employed to compute all torsion terms, derived in a hierarchical manner using fitting valence parameters. This is the only force field which identifies with the same atom type (Cg) the anomeric carbon (C1) in both anomers  $\alpha$  and  $\beta$ . The anomeric carbon shows a different partial charge, calculated with ensemble-averaged partial charge sets, in  $\alpha$  and  $\beta$  anomer. The feature to have just one C atom type facilitates the simulation of ring-flipping, having equilibrium between conformers with axial and equatorial substituents at the anomeric center. GLYCAM is suitable for mono- and oligosaccharides, for D and L enantiomers and for all glycosidic linkage possibilities. It is also the only force field to include parameters for N-glycosidic linkages. Great caution should be taken before mixing the GLYCAM parameter for carbohydrates with non-AMBER parameter sets. GLYCAM and CHARMM<sup>65,86</sup> turn the 1–4 non-bonded interactions off to correctly reproduce the rotation of the  $\omega$ -angle. This choice aims to simulate the correct population present for each rotamer of this angle, without being influenced by the possibility of O6 to interact with either O4 or O5.

In CHARMM, all the terms, including the glycosidic torsional terms were obtained with QM techniques, based on a hierarchical approach. In this force field, different atom types were introduced for  $\alpha$  and  $\beta$  anomers and rotational preferences of the exocyclic group were analysed for  $\alpha$ - and  $\beta$ -gluco and galactopyranose. All glycosidic linkages have been scanned separately. The final parameter set was validated *via* MD simulations of disaccharides in aqueous and crystal phases and compared to X-ray geometries, experimental solution and NMR *J*-coupling data. Originally, this force field was designed for proteins and then parameterised just for carbohydrates containing D-hexopyranose monosaccharides.

GLYCAM and CHARMM are both compatible with the classic TIP3P water model,<sup>15</sup> while GROMOS<sup>87</sup> is the only force field compatible with the SPC water model.<sup>88</sup> GROMOS has been optimised to enhance the stability of the <sup>4</sup>C<sub>1</sub> chair conformation and it is suitable only for unbranched hexopyranoses. No distinction is made between  $\alpha$  and  $\beta$  monomers in terms of charges and type of anomeric atom, and electrostatic–van der Waals 1–4 scaling factors are not introduced, so that the *gauche* effects on  $\omega$  angles are correctly reproduced.

Other two force fields very common for simulating carbohydrates are OPLS-AA-SEI (Scaling of Electrostatic Interactions)<sup>89</sup> and MM3. OPLS-AA-SEI was originally used for biomolecular force field and later extended for carbohydrates. This force field includes 1–4 and also for 1–5 and 1–6 non-bonded interactions. It is parameterised to reproduce *ab initio* calculation of energies of <sup>4</sup>C<sub>1</sub> pyranoses with explicit water. As the other force fields, except GLYCAM, only one carbon atom type is set for the anomeric carbon, with

unique charge sets. Non-bonded parameters are imported from OPLSAA force-field and charges are derived from acetals and standard alcohols. MM3<sup>90</sup> is widely used for minimisation and optimisation of oligosaccharide structures, as well as for conformational search. This force field utilises bond dipole moments rather than partial atomic charges, and reproduces adequately the fine details of the molecular structure, low-energy QM conformers and potential energy curves in the gas phase. However, it is not suitable for solution phase dynamics because it was not developed to run efficiently in explicit solvent.

## 6.5 Modelling Carbohydrate–Protein Interactions

Carbohydrates interact with different proteins (*i.e.* carbohydrate-binding proteins or lectins), playing roles of paramount importance in biological processes, such as cellular recognition, attachment and adhesion. From a global perspective, the knowledge of the interaction features through which a given saccharide is recognised by different lectins (*i.e.* galectins *vs.* C-type lectins *vs.* sigLeCs) is crucial from a biomedical perspective. Thus, the use of modelling protocols (conformational search, molecular docking, MD simulations) in this field is essential to dissect their molecular recognition properties.

Given the extensive presence of hydroxyl groups in glycan moieties, obviously their involvement in *hydrogen bond interactions* is recurrent in most of the sugar/lectin complexes. One paradigmatic example is that of the recognition of lactose derivatives by the members of the galectin family. In this case, an intricate network of intermolecular hydrogen bonds holds together the Gal and Glc moieties of the disaccharide, as shown in Table 6.1 (PDB-ID 1GZW) for the *hgal-1*/lactose complex. Fittingly, a key hydrogen bond between the properly positioned His44 residue and the axial OH4 of the Gal moiety provides the exquisite Gal-specificity of galectins. Interestingly, the stabilising hydrogen bonds are also complemented by CH– $\pi$  interactions with the non-polar face of the Gal moiety and the Trp68, which provide further strength and Gal-specificity to the recognition event. The X-ray structure of this type of complexes has served as framework to perform different computational approaches to understand the recognition features of glycans and glycomimetics by members of this family.<sup>78</sup> Key examples include the recognition of C-glycosyl derivatives<sup>34</sup> and thio and dithio glycosides.<sup>35</sup>

In order to have a complete vision of the ligand–protein interaction it is also important to take into account the role of water molecules. This is of special relevance, given the presence of a number of polar groups which could putatively participate in water-bridged hydrogen bonds with the protein. In particular, MD simulations have shown that in some of these glycomimetics, the absence of a given hydroxyl group at the sugar-like platform may be compensated by the presence of individual water molecules. This

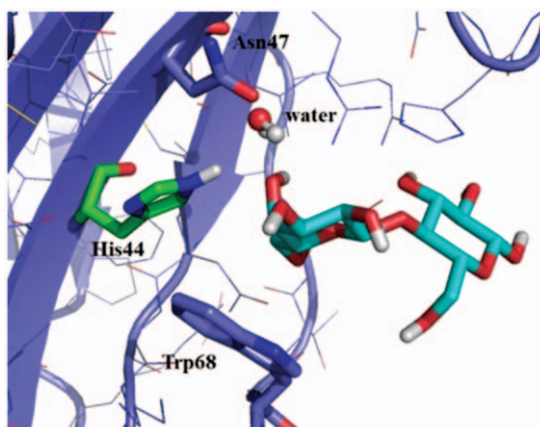
**Table 6.1** Illustrative examples of the main carbohydrate–protein interactions.**Interaction: H-bond**

*Macromolecule:*  
Human Galectin-1

*Ligand:* Lactose

*PDB-ID:* 1GZW

*Residues involved:*  
His44, Asn47, Trp68.  
One additional water molecule.

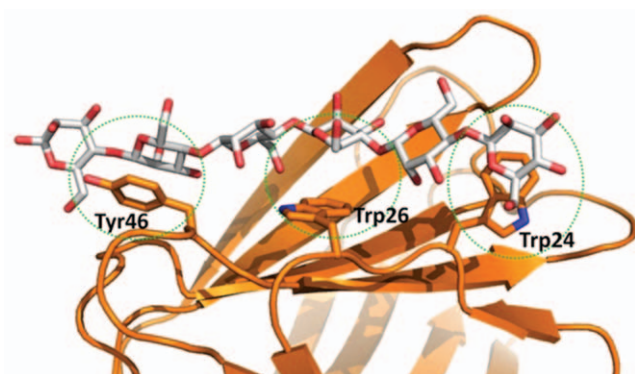
**Interaction: CH– $\pi$** 

*Macromolecule:*  
*Piromyces*  
carbohydrate-  
binding module

*Ligand:* Mannoheptaose

*PDB-ID:* 1GWL

*Residues involved:*  
Trp24, Trp26, Tyr46.

**Interaction: van der Waals**

*Macromolecule:*  
Hevein (truncated  
hevein of 32  
aminoacids)

*Ligand:* Chitobiose

*PDB-ID:* 1T0W

*Residues involved:*  
Tyr30.

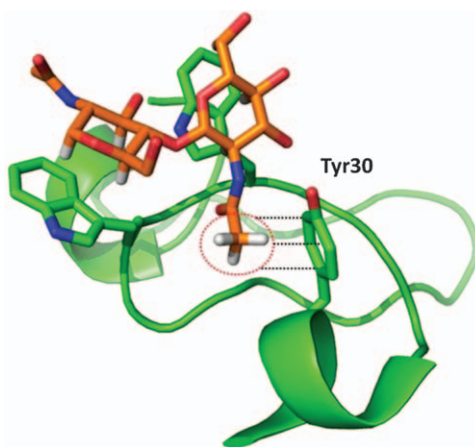


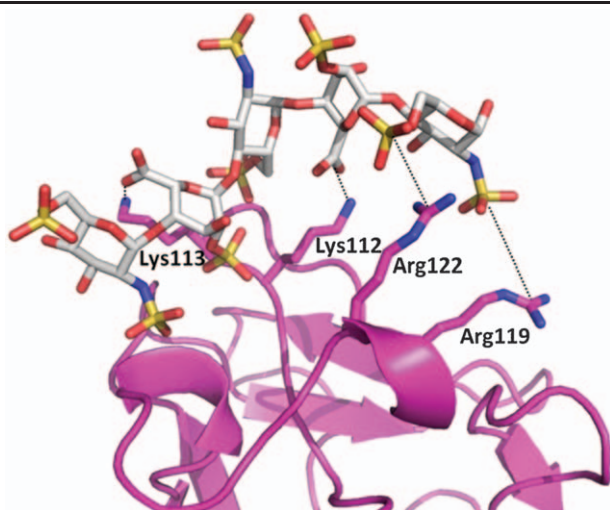
**Table 6.1** (Continued)**Interaction: ionic interaction**

*Macromolecule:*  
Fibroblast  
growth factor

*Ligand:* Heparin  
oligosaccharide

*PDB-ID:* 1AXM

*Residues involved:*  
Lys112, Lys113,  
Arg119, Arg122.

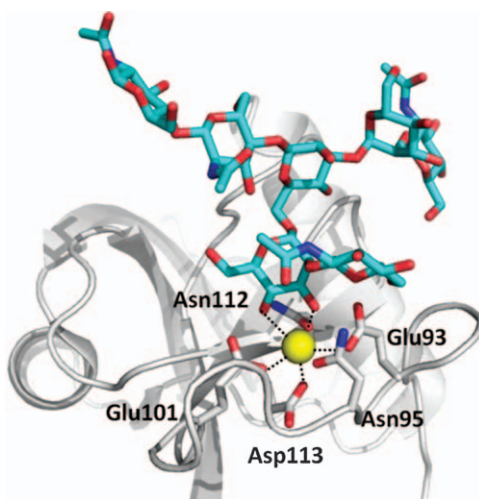
**Interaction:  $\text{Ca}^{2+}$  mediated**

*Macromolecule:*  
Codakine (C-type  
lectin)

*Ligand:* Biantennary  
nonasaccharide

*PDB-ID:* 2VUZ

*Residues involved:*  
Glu93, Asn95, Glu101,  
Asn112, Asp113.



evidence manifests that exposed sugar binding sites are usually well hydrated and the hydroxyl groups occupy lectin locations that were covered by water molecules in the free state, as evidenced by MD simulations.<sup>91</sup>

Important carbohydrate–protein interactions are represented by *CH– $\pi$  interactions*, which stabilise the binding (Table 6.1).<sup>92,93</sup> When binding lectins, they are very frequent, together with stacking interactions, typically involving a tryptophane residue, although tyrosine moieties are also fairly frequent. Exhaustive computational approaches to account for this interaction have been described for simple ligand–receptor models,<sup>94,95</sup> as well as for saccharide–lectin complexes.<sup>34,96</sup> Another example of CH– $\pi$  interactions



is illustrated in Table 6.1 (PDB-ID 1GWL), showing mannohexaose bound to one type of carbohydrate-binding module, a non-catalytic domain frequently contained in enzymes that act on polysaccharides, guiding the enzyme to the appropriate substrate.

Also GlcNAc residues are characterised by the clustering C–H groups below the plane of the ring, which enables the establishment of CH– $\pi$  interactions with aromatic residues of protein receptors, such as hevein domains.<sup>97–99</sup> The contribution of the CH– $\pi$  component to stabilise the binding has been dissected in both experimental<sup>38,100</sup> and computational approaches.<sup>101,102</sup> On the basis of these studies, this interaction is considered a major driving force for the binding of several types of carbohydrates to their receptors.<sup>34,92,103,104</sup> Nevertheless, the recognition involves the synergy involvement of different forces and factors, including solvent and entropy aspects.<sup>92</sup> Interestingly, the binding of GlcNAc oligomers to hevein is defined by a subtle balance between enthalpic and entropic effects, also including *van der Waals interactions* between non-polar substituents of the sugars (acetamide methyl groups) and aromatic amino acids. One example of this interaction is the hevein/chitobiose complex (PDB-ID 1T0W) shown in Table 6.1. The MD approach permitted to dissect the key interactions and, together with statistical analysis methods, the principal overall motions of these sugar–lectin complexes.<sup>97</sup>

The recognition of highly charged oligosaccharides, such as those from the heparin or heparan sulfate family, displays different features. In this case, where the sugar moieties are heavily decorated by anionic sulfated and carboxylate groups, usually no CH– $\pi$  interactions are observed. In contrast, structural experimental and MD simulations have shown that there are clusters of sulfate groups stemming out of the sugar platform pointing towards well positioned lysine chains at the protein to establish *ionic interactions*. Thus, important stabilising interactions take place that provide a high stabilisation to the complex. A well-known example is that of the recognition of heparin fragments by fibroblast growth factor (FGF) and their receptors (FGFR), which have been extensively studied by X-ray, NMR, and MD simulations. In this context, a recent work has provided a combined NMR/MD perspective of the dimerisation process that takes place upon the establishment of the symmetric ternary complex formed by heparin oligosaccharides with both FGF and FGFR.<sup>55</sup> Table 6.1 shows another example of ionic interactions, where a biologically active dimer of FGF is bound to a heparin decasaccharide (PDB-ID 1AXM).

In some cases, the participation of a  $\text{Ca}^{2+}$  cation is required to activate the protein–carbohydrate binding. The *calcium-mediated binding* of the typical trimannoside of mammalian glycoproteins bound to the mannose-binding protein (a C-type lectin) was analysed in a seminal work on the synergy of experimental and theoretical protocols.<sup>105</sup> The authors described the key sugar–cation–lectin interactions by employing a combined protocol based on NMR and AMBER–GLYCAM molecular dynamics simulations. In this way, the key geometry and energy features of the recognition process were



elucidated. One example of calcium-mediated binding is represented in Table 6.1 (PDB-ID 2VUZ).

Finally, here is a challenging example of the application of molecular modelling to the study of carbohydrate binding. The binding of the Man-9 N-glycan oligosaccharides, present on hemagglutinin (HA), to surface protein D (SP-D) has been explored using large scale MD simulations.<sup>106</sup> Since this molecular recognition process is related to viral infection, the MD simulation has provided a first atomic-level molecular view of an innate host defence mechanism mediated by a lectin through the inhibition of its viral glycoprotein target. This representative example can illustrate the usefulness of molecular modelling in elucidating molecular recognition mechanisms at the atomic level of the many biological roles that carbohydrates play.

## 6.6 Conclusions

In this chapter, the key structural features of carbohydrates and the stereo-electronic effects regulating their physiological behaviour have been reviewed and summarised. The most common strategies adopted for reproducing those features *in silico* have been also described and compared, highlighting advantages and limitations. Finally, interactions characterising protein-carbohydrate complexes have been summarised, reporting successful examples of recognition mechanism studies based on the combination of computational and experimental strategies. We suggest a regular visit to the Glycopedia<sup>107</sup> website for continuous updates in this field.

## Acknowledgements

This work was funded by grants from the Spanish Ministry for Economy and Competitiveness (MINECO CTQ2014-57141-R and CTQ2015-64597-C2-1-P), the Excellence fellowship program of the University of Geneva, the Swiss National Science Foundation (grant P300P3-158507), and the European Commission (TOLLerant H2020-MSCA-ETN-642157).

## References

1. A.-C. Eliasson, *Carbohydrates in Food*, CRC Press Taylor & Francis Group, 2016.
2. H. J. Gabius, H. C. Siebert, S. André, J. Jiménez-Barbero and H. Rüdiger, *ChemBioChem*, 2004, **5**, 740–764.
3. D. Solis, N. V. Bovin, A. P. Davis, J. Jiménez-Barbero, A. Romero, R. Roy, K. Smetana Jr. and H. J. Gabius, *Biochim. Biophys. Acta*, 2015, **1850**, 186–235.
4. A. Varki, R. D. Cummings, J. D. Esko, H. H. Freeze, P. Stanley, C. R. Bertozzi, G. W. Hart and M. E. Etzler, *Essential of Glycobiology*, Cold Spring Harbor Laboratory Press, 2009.
5. R. D. Cummings, *Mol. Biosyst.*, 2009, **5**, 1087–1104.

6. H. J. Gabius, S. André, J. Jiménez-Barbero, A. Romero and D. Solis, *Trends Biochem. Sci.*, 2011, **36**, 298–313.
7. S. S. Mallajosyula and A. D. MacKerell Jr., *J. Phys. Chem. B*, 2011, **115**, 11215–11229.
8. E. Fadda and R. J. Woods, *Drug Discovery Today*, 2010, **15**, 596–609.
9. B. M. Sattelle and A. Almond, *Glycobiology*, 2011, **21**, 1651–1662.
10. B. Mulloy, M. J. Forster, C. Jones, A. F. Drake, E. A. Johnson and D. B. Davies, *Carbohydr. Res.*, 1994, **255**, 1–26.
11. J. C. Munoz-Garcia, J. Lopez-Prados, J. Angulo, I. Diaz-Contreras, N. Reichardt, J. L. de Paz, M. Martin-Lomas and P. M. Nieto, *Chemistry*, 2012, **18**, 16319–16331.
12. M. Guerrini, M. Hricovini and G. Torri, *Curr. Pharm. Des.*, 2007, **13**, 2045–2056.
13. L. Nieto, A. Canales, G. Gimenez-Gallego, P. M. Nieto and J. Jiménez-Barbero, *Chemistry*, 2011, **17**, 11204–11209.
14. S. B. Engelsens, C. Monteiro, C. Herve, de Penhoat and S. Perez, *Biophys. Chem.*, 2001, **93**, 103–127.
15. K. N. Kirschner and R. J. Woods, *Proc. Natl. Acad. Sci. U. S. A.*, 2001, **98**, 10541–10545.
16. Y. C. Lee and R. T. Lee, *Acc. Chem. Res.*, 1995, **28**, 321–327.
17. D. Fries, S. Rao and M. Sundaralingam, *Acta Crystallogr., Sect. B: Struct. Sci.*, 1971, **27**, 994–1005.
18. P. Vidal, B. Vauzeilles, Y. Blériot, M. Sollogoub, P. Sinaÿ, J. Jiménez-Barbero and J. F. Espinosa, *Carbohydr. Res.*, 2007, **342**, 1910–1917.
19. J. T. Edward, *Chem. Ind.*, 1955, 1102–1104.
20. R. Lemieux and P. Chu, *J. Am. Chem. Soc.*, 1958, **133**, 31N.
21. D. R. Ferro, A. Provasoli, M. Ragazzi, B. Casu, G. Torri, V. Bossennec, B. Perly, P. Sinay, M. Petitou and J. Choay, *Carbohydr. Res.*, 1990, **195**, 157–167.
22. S. A. Samsonov and M. T. Pisabarro, *Carbohydr. Res.*, 2013, **381**, 133–137.
23. F. V. Toukach and V. P. Ananikov, *Chem. Soc. Rev.*, 2013, **42**, 8376–8415.
24. E. J. Cocinero, P. Carcabal, T. D. Vaden, J. P. Simons and B. G. Davis, *Nature*, 2011, **469**, 76–79.
25. B. Xu, L. Unione, J. Sardinha, S. Wu, M. Etheve-Quelquejeu, A. Pilar Rauter, Y. Blériot, Y. Zhang, S. Martin-Santamaria, D. Diaz, J. Jiménez-Barbero and M. Sollogoub, *Angew. Chem., Int. Ed.*, 2014, **53**, 9597–9602.
26. A. Lacetera, S. Galante, J. Jiménez Barbero and S. Martín Santamaría, *Glycans in medicinal chemistry. Reference Module in Chemistry, Molecular Sciences and Chemical Engineering*, 2016.
27. J. Dabrowski, T. Kozar, H. Grosskurth and N. E. Nifant'ev, *J. Am. Chem. Soc.*, 1995, **117**, 5534–5539.
28. J. L. Asensio and J. Jiménez-Barbero, *Biopolymers*, 1995, **35**, 55–73.
29. A. Imberty and S. Perez, *Chem. Rev.*, 2000, **100**, 4567–4588.

30. J. Jiménez-Barbero, O. Noble, C. Pfeffer and S. Perez, *New J. Chem.*, 1988, **12**, 941–946.
31. J. Jiménez-Barbero, C. Bouffar-Roupe, C. Rochas and S. Perez, *Int. J. Biol. Macromol.*, 1989, **11**, 265–272.
32. Glyco3D, <http://glyco3d.cermav.cnrs.fr>, (accessed March, 2017).
33. G. Widmalm, *Carbohydr. Res.*, 2013, **378**, 123–132.
34. P. Vidal, V. Roldos, M. C. Fernandez-Alonso, B. Vauzeilles, Y. Blériot, F. J. Cañada, S. André, H. J. Gabius, J. Jiménez-Barbero, J. F. Espinosa and S. Martin-Santamaria, *Chemistry*, 2013, **19**, 14581–14590.
35. S. Martin-Santamaria, S. André, E. Buzamet, R. Caraballo, G. Fernandez-Cureses, M. Morando, J. P. Ribeiro, K. Ramirez-Gualito, B. de Pascual-Teresa, F. J. Cañada, M. Menendez, O. Ramstrom, J. Jiménez-Barbero, D. Solis and H. J. Gabius, *Org. Biomol. Chem.*, 2011, **9**, 5445–5455.
36. T. Suzuki, H. Makyio, H. Ando, N. Komura, M. Menjo, Y. Yamada, A. Imamura, H. Ishida, S. Wakatsuki, R. Kato and M. Kiso, *Biorg. Med. Chem.*, 2014, **22**, 2090–2101.
37. E. M. Sanchez-Fernandez, R. Risquez-Cuadro, M. Aguilar-Moncayo, M. I. Garcia-Moreno, C. Ortiz Mellet and J. M. Garcia Fernandez, *Org. Lett.*, 2009, **11**, 3306–3309.
38. J. L. Asensio, H. C. Siebert, C. W. von Der Lieth, J. Laynez, M. Bruix, U. M. Soedjanaamadja, J. J. Beintema, F. J. Cañada, H. J. Gabius and J. Jiménez-Barbero, *Proteins*, 2000, **40**, 218–236.
39. J. F. Espinosa, F. J. Cañada, J. L. Asensio, M. Martín-Pastor, H. Dietrich, M. Martín-Lomas, R. R. Schmidt and J. Jiménez-Barbero, *J. Am. Chem. Soc.*, 1996, **118**, 10862–10871.
40. A. Garcia-Herrero, E. Montero, J. L. Munoz, J. F. Espinosa, A. Vian, J. L. Garcia, J. L. Asensio, F. J. Cañada and J. Jiménez-Barbero, *J. Am. Chem. Soc.*, 2002, **124**, 4804–4810.
41. L. Gonzalez, J. L. Asensio, A. Ariosa-Alvarez, V. Verez-Bencomo and J. Jiménez-Barbero, *Carbohydr. Res.*, 1999, **321**, 88–95.
42. E. Montero, M. Vallmitjana, J. A. Perez-Pons, E. Querol, J. Jiménez-Barbero and F. J. Cañada, *FEBS Lett.*, 1998, **421**, 243–248.
43. B. M. Sattelle and A. Almond, *Carbohydr. Res.*, 2014, **383**, 34–42.
44. A. Almond and J. O. Duus, *J. Biomol. NMR*, 2001, **20**, 351–363.
45. M. Rinnbauer, B. Ernst, B. Wagner, J. Magnani, A. J. Benie and T. Peters, *Glycobiology*, 2003, **13**, 435–443.
46. E. W. Sayers and J. H. Prestegard, *Biophys. J.*, 2000, **79**, 3313–3329.
47. A. Canales, A. Mallagaray, J. Perez-Castells, I. Boos, C. Unverzagt, S. André, H. J. Gabius, F. J. Cañada and J. Jiménez-Barbero, *Angew. Chem., Int. Ed.*, 2013, **52**, 13789–13793.
48. A. Canales, A. Mallagaray, M. A. Berbis, A. Navarro-Vazquez, G. Dominguez, F. J. Cañada, S. André, H. J. Gabius, J. Perez-Castells and J. Jiménez-Barbero, *J. Am. Chem. Soc.*, 2014, **136**, 8011–8017.
49. V. Gargiulo, M. A. Morando, A. Silipo, A. Nurisso, S. Perez, A. Imberty, F. J. Cañada, M. Parrilli, J. Jiménez-Barbero and C. De Castro, *Glycobiology*, 2010, **20**, 1208–1216.

50. R. Sasisekharan, R. Raman and V. Prabhakar, *Annu. Rev. Biomed. Eng.*, 2006, **8**, 181–231.
51. V. H. Pomin, *Prog. Biophys. Mol. Biol.*, 2014, **114**, 61–68.
52. T. R. Rudd, M. A. Skidmore, M. Guerrini, M. Hricovini, A. K. Powell, G. Siligardi and E. A. Yates, *Curr. Opin. Struct. Biol.*, 2010, **20**, 567–574.
53. A. S. Gallus and D. W. Coghlan, *Curr. Opin. Hematol.*, 2002, **9**, 422–429.
54. J. C. Reverter, *Drugs Today*, 2002, **38**, 185–194.
55. L. Nieto, A. Canales, I. S. Fernandez, E. Santillana, R. Gonzalez-Corrochano, M. Redondo-Horcajo, F. J. Cañada, P. Nieto, M. Martin-Lomas, G. Gimenez-Gallego and J. Jiménez-Barbero, *ChemBioChem*, 2013, **14**, 1732–1744.
56. J. K. Choi, B. H. Lee, C. H. Chae and W. Shin, *Proteins*, 2004, **55**, 22–33.
57. Y. Wu, L. Ai, J. Wu and S. W. Cui, *Int. J. Biol. Macromol.*, 2013, **56**, 76–82.
58. E. F. Franca, R. D. Lins, L. C. G. Freitas and T. P. Straatsma, *J. Chem. Theory Comput.*, 2008, **4**, 2141–2149.
59. R. Cighetti, C. Ciaramelli, S. E. Sestito, I. Zanon, L. Kubik, A. Arda-Freire, V. Calabrese, F. Granucci, R. Jerala, S. Martin-Santamaria, J. Jiménez-Barbero and F. Peri, *ChemBioChem*, 2014, **15**, 250–258.
60. R. Marchetti, A. Canales, R. Lanzetta, I. Nilsson, C. Vogel, D. E. Reed, D. P. Aucoin, J. Jiménez-Barbero, A. Molinaro and A. Silipo, *ChemBioChem*, 2013, **14**, 1485–1493.
61. L. Mulder, B. Lefebvre, J. Cullimore and A. Imberty, *Glycobiology*, 2006, **16**, 801–809.
62. P. Rouge, W. Nerinckx, C. Gough, J. J. Bono and A. Barre, *Adv. Exp. Med. Biol.*, 2011, **705**, 511–521.
63. N. Sapay, A. Nurisso and A. Imberty, *Methods Mol. Biol.*, 2013, **924**, 469–483.
64. K. N. Kirschner, A. B. Yongye, S. M. Tschampel, J. Gonzalez-Outeirino, C. R. Daniels, B. L. Foley and R. J. Woods, *J. Comput. Chem.*, 2008, **29**, 622–655.
65. O. Guvench, E. R. Hatcher, R. M. Venable, R. W. Pastor and A. D. Mackerell, *J. Chem. Theory Comput.*, 2009, **5**, 2353–2370.
66. R. D. Lins and P. H. Hunenberger, *J. Comput. Chem.*, 2005, **26**, 1400–1412.
67. M. Rahal-Sekkal, N. Sekkal, D. C. Kleb and P. Bleckmann, *J. Comput. Chem.*, 2003, **24**, 806–818.
68. J. Fanfrlik, A. K. Bronowska, J. Rezac, O. Prenosil, J. Konvalinka and P. Hobza, *J. Phys. Chem. B*, 2010, **114**, 12666–12678.
69. M. Appell, G. Strati, J. L. Willett and F. A. Momany, *Carbohydr. Res.*, 2004, **339**, 537–551.
70. A. Nowacki and B. Liberek, *J. Phys. Chem. A*, 2008, **112**, 7072–7079.
71. A. Nowacki, D. Walczak and B. Liberek, *Carbohydr. Res.*, 2012, **352**, 177–185.
72. I. Tvaroska, F. R. Taravel, J. P. Utille and J. P. Carver, *Carbohydr. Res.*, 2002, **337**, 353–367.
73. F. A. Momany and U. Schnupf, *Carbohydr. Res.*, 2011, **346**, 619–630.
74. T. Ishida, *J. Phys. Chem. B*, 2010, **114**, 3950–3964.

75. M. C. Miller, J. P. Ribeiro, V. Roldos, S. Martin-Santamaria, F. J. Cañada, I. A. Nesmelova, S. André, M. Pang, A. A. Klyosov, L. G. Baum, J. Jiménez-Barbero, H. J. Gabius and K. H. Mayo, *Glycobiology*, 2011, **21**, 1627–1641.
76. J. Agirre, G. J. Davies, K. S. Wilson and K. D. Cowtan, *Curr. Opin. Struct. Biol.*, 2016, **44**, 39–47.
77. J. Agirre, G. Davies, K. Wilson and K. Cowtan, *Nat. Chem. Biol.*, 2015, **11**, 303.
78. O. C. Grant and R. J. Woods, *Curr. Opin. Struct. Biol.*, 2014, **28c**, 47–55.
79. G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell and A. J. Olson, *J. Comput. Chem.*, 2009, **2009**, 2785–2791.
80. M. Agostino, C. Jene, T. Boyle, P. A. Ramsland and E. Yuriev, *J. Chem. Inf. Model.*, 2009, **49**, 2749–2760.
81. A. Kerzmann, J. Fuhrmann, O. Kohlbacher and D. Neumann, *J. Chem. Inf. Model.*, 2008, **48**, 1616–1625.
82. A. K. Nivedha, D. F. Thieker, S. Makeneni, H. Hu and R. J. Woods, *J. Chem. Theory Comput.*, 2016, **12**, 892–901.
83. S. A. Adcock and J. A. McCammon, *Chem. Rev.*, 2006, **106**, 1589–1615.
84. W. Humphrey, A. Dalke and K. Schulten, *J. Mol. Graphics Modell.*, 1996, **14**, 33–38.
85. Z. Lv, A. Tek, F. Da Silva, C. Empereur-Mot, M. Chavent and M. Baaden, *PLoS One*, 2013, **8**, e57990.
86. O. Guvench, S. N. Greene, G. Kamath, J. W. Brady, R. M. Venable, R. W. Pastor and A. D. Mackerell Jr., *J. Comput. Chem.*, 2008, **29**, 2543–2564.
87. C. Oostenbrink, A. Villa, A. E. Mark and W. F. Van Gunsteren, *J. Comput. Chem.*, 2004, **25**, 1656–1676.
88. H. Berendsen, J. Postma, W. Van Gunsteren and J. Hermans, in *Interaction Models for Water in Relation to Protein Hydration*, ed. B. Pullman and D. Reidel, Publishing Company, 1981, pp. 331–342.
89. D. Kony, W. Damm, S. Stoll and W. F. Van Gunsteren, *J. Comput. Chem.*, 2002, **23**, 1416–1429.
90. N. L. Allinger, Y. H. Yuh and J. H. Lii, *J. Am. Chem. Soc.*, 1989, **111**, 8551–8566.
91. S. Di Lella, V. Sundblad, J. P. Cerliani, C. M. Guardia, D. A. Estrin, G. R. Vasta and G. A. Rabinovich, *Biochemistry*, 2011, **50**, 7842–7857.
92. J. L. Asensio, A. Arda-Freire, F. J. Cañada and J. Jiménez-Barbero, *Acc. Chem. Res.*, 2013, **46**, 946–954.
93. M. F. Lopez-Lucendo, D. Solis, S. André, J. Hirabayashi, K. Kasai, H. Kaltner, H. J. Gabius and A. Romero, *J. Mol. Biol.*, 2004, **343**, 957–970.
94. G. Terraneo, D. Potenza, A. Canales, J. Jiménez-Barbero, K. K. Baldridge and A. Bernardi, *J. Am. Chem. Soc.*, 2007, **129**, 2890–2900.
95. S. Vandenbussche, D. Diaz, M. C. Fernandez-Alonso, W. Pan, S. P. Vincent, G. Cuevas, F. J. Cañada, J. Jiménez-Barbero and K. Bartik, *Chemistry*, 2008, **14**, 7570–7578.

96. G. Colombo, M. Meli, F. J. Cañada, J. L. Asensio and J. Jiménez-Barbero, *Carbohydr. Res.*, 2004, **339**, 985–994.
97. G. Colombo, M. Meli, F. J. Cañada, J. L. Asensio and J. Jiménez-Barbero, *Carbohydr. Res.*, 2005, **340**, 1039–1049.
98. M. C. Fernandez-Alonso, D. Diaz, M. A. Berbis, F. Marcelo, F. J. Cañada and J. Jiménez-Barbero, *Curr. Protein Pept. Sci.*, 2012, **13**, 816–830.
99. J. Jiménez-Barbero, F. Javier Canada, J. L. Asensio, N. Aboitiz, P. Vidal, A. Canales, P. Groves, H. J. Gabius and H. C. Siebert, *Adv. Carbohydr. Chem. Biochem.*, 2006, **60**, 303–354.
100. M. I. Chavez, C. Andreu, P. Vidal, N. Aboitiz, F. Freire, P. Groves, J. L. Asensio, G. Asensio, M. Muraki, F. J. Cañada and J. Jiménez-Barbero, *Chemistry*, 2005, **11**, 7060–7074.
101. M. C. Fernandez-Alonso, F. J. Cañada, J. Jiménez-Barbero and G. Cuevas, *J. Am. Chem. Soc.*, 2005, **127**, 7379–7386.
102. K. Ramirez-Gualito, R. Alonso-Rios, B. Quiroz-Garcia, A. Rojas-Aguilar, D. Diaz, J. Jiménez-Barbero and G. Cuevas, *J. Am. Chem. Soc.*, 2009, **131**, 18129–18138.
103. J. L. Asensio, F. J. Cañada, M. Bruix, C. Gonzalez, N. Khier, A. Rodriguez-Romero and J. Jiménez-Barbero, *Glycobiology*, 1998, **8**, 569–577.
104. M. Muraki, *Protein Pept. Lett.*, 2002, **9**, 195–209.
105. E. W. Sayers and J. H. Prestegard, *Biophys. J.*, 2002, **82**, 2683–2699.
106. B. C. Goh, M. J. Rynkiewicz, T. R. Cafarella, M. R. White, K. L. Hartshorn, K. Allen, E. C. Crouch, O. Calin, P. H. Seeberger, K. Schulten and B. A. Seaton, *Biochemistry*, 2013, **52**, 8527–8538.
107. Glycopedia, [www.glycopedia.eu](http://www.glycopedia.eu), (accessed March, 2017).

## CHAPTER 7

# *Molecular Modelling of Nucleic Acids*

HANSEL GÓMEZ,<sup>a,b</sup> JÜRGEN WALTHER,<sup>a,b</sup>  
LEONARDO DARRÉ,<sup>a,b</sup> IVAN IVANI,<sup>a,b</sup> PABLO D. DANS<sup>a,b</sup> AND  
MODESTO OROZCO<sup>\*a,b,c</sup>

<sup>a</sup> Institute for Research in Biomedicine (IRB Barcelona) – The  
Barcelona Institute of Science and Technology, 08028 Barcelona, Spain;

<sup>b</sup> Joint BSC-CRG-IRB Program in Computational Biology, Institute for  
Research in Biomedicine, Barcelona, Spain; <sup>c</sup> Department of Biochemistry  
and Molecular Biomedicine, Faculty of Biology, University of Barcelona,  
08028 Barcelona, Spain

\*Email: [modesto.orozco@irbbarcelona.org](mailto:modesto.orozco@irbbarcelona.org)

## 7.1 Introduction

Nucleic acids are flexible and highly charged polymers whose structure is defined by the balance between opposite forces: strong phosphate-phosphate electrostatic repulsions, attractive nucleobase-nucleobase interactions (stacking and hydrogen bonds), and solvent effects, which on one hand disfavour nucleobase interactions, but on the other, turn down electrostatic repulsion. Careful calibration of all these interactions is a great challenge for theoretical approaches, which have to face an additional challenge: the extreme size of these molecules. As an example, an average human chromosome (if extended) measures around 5 cm, but is determined by interactions between the building blocks at the angstrom scale. In other



words, the study of nucleic acids implies a multiscale navigation: from  $10^{-10}$  to  $10^{-2}$  m, *i.e.* an extreme challenge for any theoretical framework.<sup>1,2</sup>

We review here the current *state of the art* nucleic acid simulation methods, including a few recent relevant examples of the type of information that these methods can provide. The presentation of the methods will follow a sequential order depending on the level of resolution used: (i) quantum mechanical (QM) approaches, aiming to represent with electronic detail systems with up to a few nucleotides, (ii) hybrid QM/MM methods mixing QM description in a small part of the system and classical molecular mechanics (MM) description in the rest, which are intended to explore electronic details of a small portion of a nucleic acid in its molecular context, (iii) methods based on classical atomistic force fields coupled to molecular dynamics (MD) algorithms, which can represent reasonably large nucleic acids (up to a few hundreds of nucleobases) in its solvent environment, (iv) coarse-grain (CG) models, where atomistic detail is lost, allowing then the exploration of larger systems, and (v) mesoscopic models, which at the expense of an even larger simplification on the nucleic acid representation, aim to capture large nucleic acid polymers. We will discuss the main directions of advancement of the field, and the expectations for the future originated from the development of more powerful hardware, the improvement in simulation methods, the development of new multiresolution approaches and the implementation of experimental data as inputs to guide simulations.

## 7.2 QM Methods

QM is the fundamental branch of physics that better explains the nature and behaviour of matter on the atomic level, thus providing the best quantitative information on molecules and their interactions. The consideration of QM methods is strictly necessary to study processes depending on the electronic structure, including catalytic, photophysical or spectroscopic properties that cannot be described with classical MM.

### 7.2.1 Basic Methodological Description

Conventional QM approaches using wavefunction-based methods describe the motion of individual electrons and their application implies the solution of complicated mathematical expressions. These are typically known as *ab initio* (from first principles of quantum mechanics) methods and include a big number of approaches from simplest Hartree–Fock (HF) calculations, Moller–Plesset (MP), configuration interaction (CI), coupled cluster (CC) to multi-reference wavefunction methods like multi-configurational self-consistent field (MCSCF), also called post-HF methods. All the latest improve the HF description by considering electron-correlation effects, thus providing the most reliable calculations. However, the added accuracy implies a significant computational cost that makes impossible any application on big systems and accordingly, these high-precision methods can only be used in studies with reduced models of nucleic acids.

An alternative to these computationally demanding methods arrived with density functional theory (DFT), which is based on the assumption that any property of the system can be derived from the spatial distribution of electrons, *i.e.* the electron density. This led to simpler approaches that provide results of a similar quality to post-HF methods at a fraction of the computational cost. Taking all this together, DFT has emerged as a powerful tool that is widely used, alone or in combination with post-HF methods to provide high precision results.

### 7.2.2 Examples of Use

Some groups have systematically applied high-level QM calculations to study basic interactions on models of nucleic acids, as those by Hobza and Šponer. Their work has provided benchmark datasets that are widely used in force-field validations. Recent examples of works published by these groups include the studies of complicated nucleic acid motifs as the Sarcin–Ricin internal loop,<sup>3</sup> detailed analysis of backbone rotamers in DNA<sup>4</sup> and RNA<sup>5</sup> and the impact of ion polarisation on the stabilisation of certain quadruplexes.<sup>6,7</sup> Using similar approaches Phan's group characterised the guanine base stacking in G-quadruplexes<sup>8</sup> and Parker and co-workers estimate the energetic components of base stacking and their relation with the helical parameters<sup>9</sup> confirming that London's dispersion forces are the most important attractive components in base stacking, and that the backbone plays a non-negligible role in determining good stacking geometries, especially for RNA. Nawort and co-workers analysed the impact of the presence of 2'-thiouridine and its degradation products in tRNA on the fidelity of the translation process,<sup>10</sup> and Brovarets and Hovorun described the double proton transfer of the G–C Watson–Crick base pair, finding that it is a too rare event to have an important role in spontaneous DNA mutation.<sup>11</sup>

Even though these high-level calculations have demonstrated the power of QM to provide detailed information on nucleotide conformation and on specific interactions between nucleotides, they also highlight the limitations associated to the use of static reduced models of nucleic acids, the only manageable today by means of high-level QM approaches (including DFT). Some authors have used simpler, alternative QM approximations to consider entire (small) nucleic acid molecules from a dynamic point of view. A recent example is the Car–Parrinello molecular dynamics (CPMD) work by Arcella *et al.* where they studied the dynamic changes in covalent topology in a small DNA hairpin when subjected to electrospray ionisation mass spectrometry (MS-ESI).<sup>12</sup>

## 7.3 Hybrid QM/MM

### 7.3.1 Basic Methodological Description

The size and conformational complexity of systems as nucleic acids demands for methods capable of treating, in a dynamical way, up to several

thousand atoms and allowing for simulations over times scales that cannot be achieved if just QM methods are considered.<sup>1</sup> In that sense, hybrid QM/MM schemes have become very popular to study chemical processes depending on the electronic structure (*i.e.* where a QM description is needed) in large systems untreatable by pure QM methods. The basic idea of this methodology as conceived by Warshel and Levitt is to use a QM method for the chemically active region and a MM treatment for the surroundings.<sup>13</sup> In this way, the QM/MM approach enables the use of QM methods in biomolecules as solvated nucleic acids at a reasonable computational cost and the necessary accuracy.

When more extensive conformational sampling is required, as for computing free energy barriers, efficient QM methods are required. The latest explains the popularity of semi-empirical Hamiltonians that are often recalibrated to study reactions involving nucleic acids. In that sense, semi-empirical methods share the disadvantages of force fields since they perform best for the kind of systems that were considered in the parameterisation process, failing in other cases. Moreover, semi-empirical methods can reproduce kinetic properties in catalytic studies, but the obtained results should be taken with caution as calculated free energy surfaces can be distorted.<sup>14</sup>

### 7.3.2 Examples of Use

Among recent examples of semi-empirical QM/MM calculations applied to nucleic acids we should cite the study of the catalytic mechanism of the Human Flap Endonuclease (hFEN1),<sup>15</sup> or the work by Tuñón's group on the mechanism of action of N6-adenosine methyltransferase.<sup>16</sup> Other levels of theory have been considered using a QM/MM scheme in catalytic studies. An example is the study of intra-strand oxidative crosslink lesions in DNA where the CPMD method was applied.<sup>17,18</sup> On the other hand, Zhang and co-workers considered state-of-the-art QM(DFT)/MM Born–Oppenheimer MD calculations to study the catalytic mechanism of the nucleotidyl transfer reaction in human DNA polymerase kappa. The authors were able to explain the bypass of major benzo(a)pyrene-derived dG lesion by the enzyme after describing the activation of the 3'-OH primer terminus and the following associative nucleotidyl transfer reaction.<sup>19</sup> More recently, Rosta and co-workers used similar methodologies to study the different reactivities of  $\text{Mg}^{2+}$  and  $\text{Ca}^{2+}$  in the active site of Ribonuclease H1, and were able to explain the calcium inhibition mechanism of the phosphodiester cleavage reaction.<sup>20</sup> According to their conclusions,  $\text{Ca}^{2+}$  is less efficient in the activation of the nucleophilic water molecule, resulting in a dramatic loss of the catalytic activity in Ribonuclease H1. Also in a recent paper Molina *et al.* combined QM and QM/MM calculations with X-ray time-course data to describe the catalytic mechanism of a restriction enzyme that hydrolyses the double-stranded DNA.<sup>21</sup> The authors described a two-steps reaction mechanism, where the attacking water molecule is activated by the targeted phosphate group to later protonate the O3' atom and break the

phosphodiester bond. Palermo and co-workers have recently reviewed other examples of similar studies that focused on enzymatic processing of nucleic acids.<sup>22</sup>

There are fields others than reactivity where a QM level of description is required, typically in combination to MM, either using a QM/MM scheme, as explained before, or QM calculations on MM-derived ensembles. One of these fields is the study of charge transfer (CT) in nucleic acids, a process which might be relevant in DNA mutagenesis and repair, and that is of paramount importance in the field of nucleic acid nanotechnology (consult ref. 23 for a recent review). For example, Lech and co-workers studied the effective electronic coupling in different G-motives using the semi-empirical method INDO/S.<sup>24</sup> In this study, the analysis of 1000 G-tetrad models built from a MM MD-derived ensemble suggested that the G-tetrad orientation plays a key role in the hole transport within the  $\pi$ -stacks. Bacolla *et al.* also considered a MM-derived ensemble to perform multiple ionisation potential (IPs) calculations at the QM(DFT)/MM level and studied the sequence context-dependent mutagenesis at mononucleotide repeats (A-tracts and G-tracts), finding evidences on the coupling of CT and mutagenicity.<sup>25</sup>

Some recent advances have been made in the methodologies to better describe CT in biomolecules. It is worth mentioning the new multi-scale method proposed by Kubař and Elstner, which combines a non-adiabatic propagation scheme and a linear scaling QM approach in a QM/MM coupling framework.<sup>26</sup> This approach, which implies an unbiased description of the dynamics of electron transfer processes involving hundreds of quantum atoms, was successfully applied to describe the hole transfer in a solvated DNA oligonucleotide, opening new venues for the accurate representation of CT in nucleic acids.

Another area of nucleic acid research where a quantum description is strictly required is the prediction of spectroscopic properties. Some relevant works on this area can be found in the scientific literature, where a combination of QM (DFT and time-dependent DFT (TD-DFT)) and MM techniques is used to recover the UV absorption and electronic circular dichroism (ECD) spectra of different nucleic acid sequences, thus connecting the ECD signals to specific structural patterns. As an example, Spata and Matsika computed the UV absorption and ECD spectra of an adenine-based oligonucleotide, finding that mixing CT and excited states studies is essential to explain photophysics in DNA.<sup>27</sup> In another interesting study, Gatusso and co-workers combined classical MD simulations with QM/MM calculations to obtain the properties of excited states of single chromophores, then being able to model the ECD spectra of different double helix B-DNA sequences.<sup>28</sup>

Finally, the photophysical properties of nucleotides have also been under investigation mostly due to its impact in defining the UV-damage of DNA. As example, Zelený and co-workers performed QM(DFT)/MM surface hopping dynamic calculations to study the photophysical properties of cytosine and guanine.<sup>29</sup> According to their results, there are significant differences in the decay rates of the photo excited states corresponding to G and C.

Interestingly, while the photo-deactivation of cytosine is not influenced by the environment, major deactivation happens for G, which is explained by the dramatic reduction of the out-of-plane motions of the  $\text{NH}_2$  group when inserted in duplex DNA.<sup>29</sup> In another recent work, D'Abramo *et al.* studied the hyperchromic effect in  $(\text{dA})_n \cdot (\text{dT})_n$  at the QM(CASPT2/CASSCF)/MM level of theory.<sup>30</sup> They reported that denatured DNA exhibits an increase of the absorbance with respect to double-stranded DNA since in the latest context there is a lower delocalisation of the excitonic states. For further details on this interesting topic we recommend a very recent and extensive review by Improta and co-workers.<sup>31</sup>

Overall, pure QM and QM/MM methods are providing invaluable information on nucleic acids properties, impossible to obtain from other theoretical or experimental approach. However, these approaches suffer the limitations of the extreme cost of the QM calculations, limiting the size of the system to be studied and forcing the use of classical approaches (see below).

## 7.4 Atomistic Force-field Simulations

### 7.4.1 Basic Methodological Description

Force fields are a set of classical expressions that allow to obtain a simple energy functional relating the given configuration of the system with its energy. Within the force-field approximation, the electronic degrees of freedom are not implicitly considered, but introduced parametrically in a simplified way that requires previous knowledge of the studied system. For example, in a QM calculation of a water molecule, the input file contains just information on the nature of the atoms, and at most, a first guess of the structure of the molecule. The calculation takes care of finding the optimal electronic and nucleic spatial distributions, resulting in a nuclear geometry where O and H atoms are at around 0.96 Å, and the H–O–H angle is around 104.5 degrees. The final electron distribution shows a significant density of electrons located along the two O–H vectors, defining spontaneously the two O–H chemical bonds. Furthermore, the optimised electron density shows spontaneously the migration of electrons from the hydrogen to the oxygens, resulting in a polar molecule. As a consequence, if the QM simulation is repeated, but introducing not one, but two molecules of water, a strong electrostatic interaction between oxygens and hydrogens would be spontaneously found, leading to the formation of the  $\text{HOH} \cdots \text{OH}_2$  hydrogen bond, where the two oxygens would remain at around 2.8 Å. This balances the electrostatic attraction between the negatively charged oxygen and the positively charged hydrogen, as well as the short-range repulsion between the inner electronic shells of the interacting atoms. In summary, a QM calculation provides all the information of a system by just looking at its chemical composition. Force field based classical calculations require much more information. In the water example, to reproduce its geometry we need

to define the optimal O–H distance, and H–O–H angle, and some functional determining the dependence of the energy with perturbations on distance (stretching) and angle (bending). For a molecule with rotational degrees of freedom (for example methanol) we would need to add additional parameterised terms accounting for the torsional energy profile. Finally, to reproduce the calculation of the water dimer (which was trivial from QM calculations), we would need to add, at least, a Coulombic term to account for the polar water–water interactions, and a dispersion–repulsion functional (usually a van der Waals term) to represent inner shell and nuclei–nuclei interactions. This means that in addition to bond, angle and torsion parameters we would have to assign point charges to represent the charge distribution of the molecule, some measure of the inner shell volume (for example the van der Waals radii), and a measure of the deformability of such inner shell (the atom hardness). All in all, within the classical force field approximation the energy of a system for a given nucleic configuration ( $R$ ) is expressed as:

$$E(R) = E_{\text{stretching}} + E_{\text{bending}} + E_{\text{torsion}} + E_{\text{Coulomb}} + E_{\text{van de Waals}} \quad (7.1)$$

where each of these terms is computed by means of simple equations containing a myriad of parameters to represent, as well as possible, the properties of the molecules of interest. Determining simple, but accurate functional forms and parameters for the terms in eqn (7.1) constitute the greatest challenge in the area.

An intrinsic assumption of classical atomistic models is the ‘transferability’ of the force field so that the parameters defining a molecule (or a molecular subsystem) can be used to represent it in different environments. This implies, for example, that the parameters of water optimised to represent pure water solutions, are assumed to be valid to represent also water molecules in complex mixtures, or even the waters solvating macromolecules. Furthermore, when dealing with large molecules, such as proteins or nucleic acids, the ‘transferability’ assumption is extended to assume that the macromolecular force field can be defined by the addition of the individual force fields of building blocks (amino acids or nucleotides), which are refined for the isolated building blocks and are used unaltered for any kind of protein or nucleic acid.

To our knowledge, the first MD simulations of nucleic acids should be credited to Michael Levitt, who collected the first picoseconds of atomistic trajectories.<sup>32–34</sup> In such initial studies the problems to keep structures stable were evident. These problems highlight the challenges of the simulation of nucleic acids which are originated in the large magnitude of the nucleotide–nucleotide interactions, especially of the phosphate–phosphate repulsion, which generates strong forces producing instabilities in the structure. The introduction of explicit solvent during the seventies and eighties, as well as more accurate force fields and simulation protocols was not sufficient to keep nucleic acid stable in the sub-nanosecond time scale. Only when long-range electrostatic contribution was accurately introduced



and coupled to a new generation of force fields, calculations remained stable in the nanosecond time scale (for a historic view on the evolution of MD simulations of nucleic acids we address the reader to previous reviews).<sup>1,2,35–39</sup>

### 7.4.2 Force-field Refinements

The force fields developed in the late 90s (for example ref. 40) represented a breakthrough in the field of MD simulations of nucleic acids, as they allowed for the first time to collect equilibrated trajectories that sampled conformations of DNA and RNA not far from the experimental ones. However, around 10 years ago, the increase in computer power uncovers the existence of some force field errors which produced corrupted structures in the  $\sim 10$  ns time scale.<sup>41,42</sup> Re-parameterisation efforts yielded to force fields such as *parmbsc0*,<sup>42,43</sup> which have been extensively used for ten years to simulate a variety of nucleic acids in the multi-nanosecond and sub-microsecond time scale.<sup>35</sup> Unfortunately, as simulation approached to the microsecond regime, some errors in third generation force fields become evident: problems to reproduce some RNA folds, excessive terminal base-pair fraying, under twisting of B-DNA, or dramatic problems to reproduce several loop structures, among others.<sup>35,43–46</sup> The correction of these problems implied the specialisation of force fields, which are now RNA- or DNA-specific, and the re-parameterisation of several torsional terms (e.g. ref. 46).

Decades of developments and applications have selected the CHARMM and the AMBER families of force fields as the most prevalent ones for nucleic acid simulations. The latest version of CHARMM (*CHARMM36*) corrected several inaccuracies of previous versions, providing a reasonable representation of DNA duplexes, and canonical RNAs.<sup>47</sup> However, in our hands, when used for microsecond trajectories of B-DNA duplexes, the excessive end base-pair fraying leads to the corruption of a significant number of base pairs (see Figure 7.1). *CHARMM36*, as many other force fields, also has very serious problems to describe some non-helical structures of DNA and RNA (see Figure 7.2). Recently, MacKerell and co-workers developed a version of



**Figure 7.1** Average structures from the last 50 ns of 1  $\mu$ s simulations of Drew-Dickerson dodecamer with different force fields, compared to a high-resolution NMR structure (PDB: 1NAJ). P. D. Dans, I. Ivani, How accurate are accurate force-fields for B-DNS? *Nucleic Acids Research*, 2017, by permission of Oxford University Press.



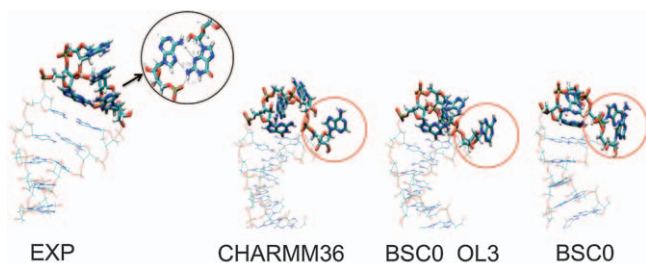


**Figure 7.2** Average structures from last 50 ns of 1  $\mu$ s simulations of an RNA hairpin (GAAA tetraloop) using different force fields. We show the disruption in the loop region with residue A9 opening, compared with an X-Ray structure (PDB: 1JJ2). Unpublished results.

the force field that incorporates a Drude's oscillator term accounting for polarisation effects (the energy contribution arising from the mutual relaxation of electron distribution of interacting particles).<sup>48</sup> The resulting force field (*CHARMM36pol*) has been implemented in NAMD, and in our hands implies an  $8\times$  increase in the computational cost of the simulation respect to the non-polarised version (e.g. for a DNA dodecamer as that shown in Figure 7.1 we obtained on average  $8\text{ ns day}^{-1}$  on 128 CPUs using the suggested implementation of Drude model in NAMD code).<sup>48,49</sup> In our opinion, *CHARMM36pol* is a milestone in the development of nucleic acids force fields, as for the first time polarisable force fields are able to reproduce some nucleic acids structures (like the B-DNA duplex) in the  $\sim 100\text{ ns}$  regime. However, more work is still needed to re-calibrate all the terms to the incorporation of polarisation, as *CHARMM36pol* leads to severe corruption of even canonical DNAs when simulations approach the microsecond regime (see Figure 7.1).

*Parmbsc0*,<sup>42,43</sup> a local refinement of parm99 was the dominant force field for a decade, and served as starting point for several improvements. Jurečka and co-workers refined in a sequential way different torsions such as the  $\chi$ ,  $\epsilon/\zeta$  and very recently  $\beta$ .<sup>44,45,50,51</sup> Some of these corrections are specific for DNA or RNA, while others may apply for both families of nucleic acid.<sup>44,45,51</sup> As expected, the resulting force fields lead to results of quality similar or superior to those of *parmbosc0* (see Figures 7.1 and 7.2), especially in the case of loops and some other unusual DNA structures. The RNA-specific versions available still have very serious problems to simulate some RNA motives, leading in some cases to severe corruptions of the structure (see Figure 7.2). In our hands, the latest refined version of Jurečka and co-workers' force field (OL15), which incorporates all the previous DNA refinements, provides high-quality results for B-DNA (see Figure 7.1).<sup>51</sup> Also in the DNA-realm the Barcelona developers of *parmbosc0* presented recently the *parmbosc1* force field,<sup>46</sup> which was tested in more than one hundred structures of DNA, providing results of extreme accuracy (in some cases of quality equal or superior to models derived from experimental restraints for a large variety of DNA motifs).<sup>46,52,53</sup>

Even close to convergence, we believe that classical non-polarisable DNA force fields still might require some refinement. For example, it has been demonstrated that the last-generation force fields are able to fold some basic DNA motives and distinguish between stable and unstable conformations of DNA,<sup>46</sup> but it is not clear if they will be able to reproduce in a systematic way observable properties like melting temperature. Furthermore, it is not clear whether or not currently used non-bonded terms, which have been unmodified for more than a decade, and that use very simple combination rules can finely reproduce DNA-protein interactions. In fact, some authors have pointed out intrinsic shortcomings in these terms that might need specific correction. For example, Chen and Garcia<sup>54</sup> and later Elcock's group<sup>55</sup> have suggested that nucleobase-nucleobase stacking is overestimated in water by around 1 kcal mol<sup>-1</sup>, proposing a linear scaling of the van der Waals term. Unfortunately, the resulting force field is not well balanced and produces suboptimal results,<sup>46,52,53</sup> but the issue of the potential inaccuracy of the stacking and its impact on the structure of DNA is under debate, as existing experimental evidence is quite contradictory.<sup>1,56,57</sup> Similar criticisms have been raised by Case and co-workers,<sup>58</sup> and by Cheatham's group<sup>59</sup> on the accuracy of the phosphate non-bonded term, which might affect the quality of DNA-protein interactions. It is likely that, as the range of applicability of MD simulations extends, and simulations reach multi-microsecond scales in a systematic way, more errors are going to emerge.

Force fields for RNA are still far in accuracy respect those for DNA, and intense effort is focused in this point. For example, Turner's group used NMR data on small systems to tune  $\chi$  torsion.<sup>60</sup> Bussi's group followed similar strategies to correct some of the current AMBER torsional terms.<sup>61</sup> Zgarbová and co-workers used high level QM to refine  $\chi$  torsion,<sup>62</sup> while Chen and García, additionally to  $\chi$  refinement, scaled base stacking to successfully fold RNA hairpins.<sup>54</sup> Cheatham's group discussed inaccuracies of current force fields based on exhaustive sampling of small loop structures and provided valuable data for re-parameterisation efforts.<sup>63</sup> Moreover, Shaw's group has embarked on an ambitious project to implement a variety of corrections into the RNA force field.<sup>64</sup> Considering the similarity between DNA and RNA, the poorer performance of RNA force fields might surprise a non-expert, but there is little doubt that RNA conformational space is much more complex and difficult to represent than that of DNA. Two reasons exist: (i) DNA structural universe is dominated by the double helix, while RNA samples much more complex conformations, increasing the complexity of the parameterisation process, (ii) the presence of the C2'OH group increases dramatically the interrelationships between the different degrees of freedom, adding more problems to the parameterisation.<sup>65</sup>

In any case, irrespective of the nucleic acids considered, it is very important to do not overestimate our expectations on classical 2-body force fields, as there are many intrinsic problems in the basic formalism that will always hamper their accuracy. For example, when nucleobases are stacked, they move from a polar to an apolar environment. This implies a

rearrangement of electron distribution that is completely ignored by two-body additive force fields. The use of atom-centered point charges is a cheap and very powerful strategy, but comparison of classical and QM potentials demonstrated their shortcomings in reproducing accurate electrostatic interactions.<sup>66</sup> The same criticism can be made on the use of rotationally-averaged charges, which neglect intramolecular polarisation effects coupled with structural movements.<sup>67</sup> Even if it is not discussed much, the ‘van der Waals’ term includes a mixture of poorly defined interaction terms, some of them anisotropic, *i.e.* impossible to reproduce by means of spherical potentials. Finally, ions are a major source of trouble in nucleic acids simulations, as they are difficult to parameterise (experimental data on ion solutions are typically quite inaccurate) and have an important impact in nucleic acid properties.<sup>68–72</sup> Refined two-body descriptions might be accurate for  $\text{Na}^+$  and  $\text{K}^+$ ,<sup>53</sup> which are the most prevalent cations in physiological conditions, but other ions can be much more difficult to represent. For example, MacKerell and co-workers have suggested that  $\text{Li}^+$  requires the use of polarisable force fields,<sup>69–71</sup> and it is clear that bivalent ions are a serious unsolved problem. For example,  $\text{Mg}^{2+}$  is crucial to stabilise some RNA folds, but its representation in classical simulations is very complex due to its complex coordination pattern and its tendency to transfer charge to DNA.

### 7.4.3 Recent Examples of Force-field Studies of Nucleic Acids

The number of atomistic simulations of nucleic acids has been continuously increasing since the late nineties. As an example, *parmbsc0*, one of the many force fields available to simulate nucleic acids was cited in around 200 scientific papers during 2015. It is impossible to summarise here all the studies where atomistic force fields have been used to explore different aspects of nucleic acids structure, properties and interactions. We will cite here only a few examples of the type of studies that have been published in the field. (See ref. 1, 35–39 for a more detailed discussion.)

The representation of physical properties of duplex DNA has been a recurrent topic of research in the field, not only for their intrinsic interest for the understanding DNA function, but because this information is crucial in the parameterisation of coarse-grain models (see below). We will cite here for its impact on the scientific community the works of the Ascona B-DNA consortium (ABC) that raised severe criticisms to the prevalent near-neighbour harmonic models of DNA deformation.<sup>41,73,74</sup> Their work showed the existence of bimodal distributions of some helical degrees of freedom, as well as the importance of, at least, tetramer effect.<sup>73,74</sup> Other authors have explored similar concepts finding hidden correlations between DNA polymorphism and ion exchange, which can explain induce fit mechanisms in DNA recognition.<sup>68</sup> Base flipping, a special type of anharmonicity, has been extensively studied by different groups,<sup>53,59,75–77</sup> due to its essential role in many biological activities of DNA, for example, in replication, transcription and repair. The response of DNA to mechanical perturbations has been

widely studied by different authors.<sup>1,2,35,36</sup> Examples are the study of how mechanical distortion induced by the presence of mismatches<sup>78</sup> or the binding of proteins<sup>79,80</sup> are transmitted along the DNA, the study of curvature,<sup>81,82</sup> the coupling of elemental degrees of freedom as a response to mechanical perturbation,<sup>68,83</sup> or the intrinsic limits of the elastic response to DNA deformation.<sup>81–83</sup> Quite surprisingly, despite some old studies<sup>84–87</sup> little has been done to systematically study the physical properties of RNA duplexes and DNA–RNA hybrids. However, it is clear that RNA duplexes are more rigid than the equivalent DNA duplexes and that DNA–RNA hybrids show an intermediate behaviour with a clear DNA–RNA asymmetry.<sup>86,87</sup> Moreover, it is well characterised that single stranded regions of RNA can act as hinge points for functionally-relevant movements in RNA.<sup>88</sup> However, there is a surprising lack of systematic studies on RNA physical properties, probably due to the still sub-optimal quality of RNA force fields.

Classical MD simulations have been largely used to study large conformational transitions and this field is expected to increase as new computers and codes approach the simulation regime to the millisecond scale. Examples of nucleic acids folding and unfolding have been reported, showing excellent results in those cases where folding landscape is simple and folding happens in the microsecond time scale.<sup>51,89–92</sup> The transitions between alternative states of duplex DNA have been a topic of interest for decades (for reviews on this topic see ref. 2, 35, 36), and as the biasing strategies are improved, simulations extended, and force fields refined, more accurate representations of different transitions such as: the Watson–Crick  $\leftrightarrow$  Hoogsteen,<sup>93,94</sup> the B  $\leftrightarrow$  A<sup>95</sup> and the B  $\leftrightarrow$  Z<sup>96</sup> transitions have been described. The impact of environment on nucleic acid properties has been also analysed in some details by means of atomistic MD simulations. For example, different studies have been focused on the impact of full dehydration on DNA structure,<sup>97–100</sup> the effect of amine cations on duplex structure and stability,<sup>101–103</sup> the pH-dependent role of chemical denaturants,<sup>89,104</sup> the structure of nucleic acids in apolar environments,<sup>105</sup> or in media of biotechnological interest, such as silica or grapheme.<sup>106,107</sup> All these studies illustrate the power of atomistic MD simulations, but caution is needed as force fields were calibrated to reproduce aqueous environments and their transferability to other solvents has not been always demonstrated.

MD simulations are deeply used to study, a variety of unusual nucleic acids, including different helical conformations of both DNA and RNA<sup>45,108–114</sup> and a myriad of non-helical RNA motifs such as loops, hairpins, ribozymes, riboswitches and transfer RNAs.<sup>54,63,115–122</sup> For its biological relevance, small RNA hairpins have been deeply studied as model elements of more complex structures.<sup>116–118</sup> Nevertheless, these systems stay challenging for most of force fields (see Figure 7.2). We believe that the community needs to create an accepted ontology of unusual nucleic acids structures, and use it to explore systematically the reliability of different force fields on unusual structures. *Big data* initiatives such as BigNASim<sup>123</sup>

constitute one step forward in this direction as they favour data sharing providing unbiased expert information on the expected quality of a given simulation. Protocols for nucleic acid parameterisation are quite well defined, which has encouraged researchers to perform simulations with nucleic acids containing non-coding nucleotides, for example those very prevalent in transfer RNAs,<sup>121</sup> those originated from UV damage, oxidative or deamination stress,<sup>124–128</sup> or simply those that can have biotechnological impact as they confer stability to certain structures or modify their physicochemical properties.<sup>129–131</sup> A specially exciting use of MD in this area is in the study of epigenetic variants of DNA, as these studies might clarify why apparently innocuous changes (for example cytosine methylation) can produce such dramatic alterations in DNA functionality.<sup>132–139</sup> Even with its obvious interest, this type of study needs to be taken with caution, as in some cases parameterisation can be more complex than anticipated.

A growing interest exists in studying protein–nucleic acid complexes, as proteins mediate many of the nucleic acid functionalities. There are a countless number of studies of protein–nucleic acids complexes. Some of these studies are centered in determining the basic thermodynamic and kinetic mechanisms of nucleic acid–protein recognition,<sup>140–146</sup> while others are centered in the description of especially relevant complexes, among them those involving DNA packing, repairing and editing of nucleic acids.<sup>147–153</sup> Although RNA force fields are robust enough to reliably represent some RNA–protein complexes,<sup>154</sup> for most complexes this remains a challenging task.<sup>155</sup> We can expect an even larger explosion of this type of projects in the future, but again caution is required, since the excellent ability of the last generation of atomistic force fields to describe naked nucleic acids might not be extrapolated to the representation of complexes, where small unbalances in the interaction energies between protein and nucleic acids can lead to significant errors. Community efforts to evaluate the level of accuracy of current force fields would be crucial in the advance of the field, as the works of the Ascona B-DNA consortium (see above) were instrumental to derive the last generation of DNA force fields.

## 7.5 The Coarse-grain Approach

When the time- or length-scales accessible to atomistic MD simulations fall short for the biological/technological questions that need to be addressed, the immediate next level of theory that can be used to reduce the computational cost is based on the simplification of the energy functional and on the reduction of degrees of freedom by grouping atoms in ‘beads’. The resulting models are called ‘coarse-grained’ (CG), an ambiguous term that refers to rather different modelling approaches. In the present section, the focus is placed on the family of *particle-based CG* (pbCG) models, which are typically used in combination with standard structural sampling methods.

### 7.5.1 Basic Methodological Description

The driving idea behind any CG approach is to simplify the system representation by neglecting dispensable information to answer a given question. For example, the specific interaction of a phosphate oxygen with a counter ion might be important in some types of studies, but in many others, it is not so relevant as long as the average distribution function of the ion around the phosphate group is properly reproduced, *i.e.* a correct representation of the effective electrostatic potential of the phosphate group is obtained. Thus, we might substitute specific atoms (phosphate atoms, counterions and water in this case) by effective beads whose interaction potentials are parameterised to reproduce the desired property. Such parameterisation can be done by choosing either a physically-based, statistically-based or mixed-based energy function that dictates the ensemble properties of the system. This short-cut favours computational efficiency in three possible ways: (i) it reduces the number of pair-wise interactions to compute, (ii) it allows for larger time steps for the integration of the equations of motion, and (iii) it facilitates diffusion in the energy landscape by reducing friction. In the nucleic acids pbCG modelling area several different simplification approaches have been proposed ranging from models using one to up to eight effective beads to represent each nucleotide. Table 7.1 summarises the most recent and/or used pbCG models available for DNA and RNA. For DNA, although highly simplified models have been proposed (five beads to represent four nucleotides in a base-pair step<sup>156</sup> or one bead per nucleotide<sup>157</sup>), 6–8 beads per nucleotide are most frequently used.<sup>158–163</sup> Thus, allowing more flexibility in the backbone (1–3 beads) and sugar (1–2 beads), and more interaction sites in the base (3–6 beads). In addition, the positioning of the beads in these more detailed models tends to be chosen so that they capture the atomistic positioning and connectivity. (Figure 7.3). In the case of RNA, the formalisms used are similar, but the motivation for pbCG development has been somewhat different. While in the case of DNA the main interest has been the study of the dynamical properties of the fibre, the main interest in RNA has been to predict folding (see next section for further details), which implies different parameterization processes used in pbCG models of DNA compared to RNA.

In terms of the description of the functional energy of the system (*i.e.* the parameterisation), two main approaches have proven successful: the top-down<sup>158,160–163,165–173</sup> and the bottom-up.<sup>156,157,159,174</sup> In the top-down approach a set of thermodynamic and/or structural measurements are defined as target metrics for fitting the parameters of a pre-defined set of effective beads. Alternatively, the bottom-up approach is formally based on the idea that the many-body potential of mean force (PMF) of a selected CG representation for a given system is completely determined by its atomistic representation and the coarse-graining map. As the entanglement of interactions in a many-body system is impossible to solve, in practice pair-wise PMF calculations are used in an iterative manner, by means of iterative inverse Boltzmann (*e.g.* Vercauteren's model<sup>157</sup>) or inverse Monte Carlo

**Table 7.1** Features of the most widely used or recently proposed pbCG models for nucleic acids (bb: ribose-phosphate backbone, R: purine and Y: pyrimidine).

| Name/dev. group                                                | Number of beads per nucleotide        | Max. num. nt                  | Type of potential                        | Environment                       | Compatible molecules                                   | Meant for:                                            |
|----------------------------------------------------------------|---------------------------------------|-------------------------------|------------------------------------------|-----------------------------------|--------------------------------------------------------|-------------------------------------------------------|
| <b>DNA</b>                                                     |                                       |                               |                                          |                                   |                                                        |                                                       |
| OxDNA <sup>169–172</sup> /<br>Ouldridge                        | Rigid body (with 5 interaction sites) | 12 400 (6200 bp)              | Top-down                                 | Implicit (+500 nM added salt)     | No                                                     | Dynamical properties (ssDNA; ddDNA).                  |
| SIRAH <sup>162,163</sup> /<br>Pantano                          | 6 beads (bb.: 3; base: 3)             | 208 (104 bp)                  | Top-down                                 | Explicit or implicit/Debye–Hückel | Water, ions and proteins                               | Dynamical properties (ssDNA; dsDNA).                  |
| 3SPN.0/1/2/<br>2C <sup>167,168,173,211,212</sup> /<br>de Pablo | 3 beads (bb.: 2; base: 1)             | 144 (ssDNA)<br>2980 (1490 bp) | Top-down                                 | Implicit/Debye–Hückel             | No                                                     | Dynamical properties (ssDNA; dsDNA; confined DNA).    |
| BioModi <sup>183</sup> /<br>Nguyen                             | 3 beads (bb.: 2; base: 1)             | 350 (ssDNA)<br>64 (32 bp)     | Top-down                                 | Implicit/Debye–Hückel             | Proteins and polymers                                  | Dynamical properties (ssDNA; dsDNA; crowded systems). |
| MARTINI <sup>158</sup> /<br>Marrink                            | 6/7 beads (bb.: 3; base: 3Y/4R)       | 40 (ssDNA)<br>200 (100 bp)    | Top-down/<br>bottom-up + elastic network | Explicit                          | Water, ions, lipids, carbohydrates, polymers, proteins | Dynamical properties (ssDNA; dsDNA).                  |
| UNRES <sup>213</sup> /<br>Scheraga                             | 6–8 beads (bb.: 2; base: 4C/5G, T/6A) | 60 (ssDNA)<br>120 (60 bp)     | Bottom-up/<br>top-down                   | Implicit/Debye–Hückel             | Proteins, DNA (coarser: NARES-2P)                      | Dynamical properties (ssDNA; dsDNA).                  |
| HiRe-DNA <sup>160</sup> /<br>Derreumaux                        | 6/7 beads (bb.: 5; base: 1Y/2R)       | 32 (16 bp)                    | Top-down                                 | Implicit                          | RNA                                                    | Folding.                                              |



**Table 7.1** (Continued)

| Name/dev. group                       | Number of beads per nucleotide                                        | Max. num. nt  | Type of potential      | Environment              | Compatible molecules | Meant for:                                        |
|---------------------------------------|-----------------------------------------------------------------------|---------------|------------------------|--------------------------|----------------------|---------------------------------------------------|
| Nordenskiöld <sup>156</sup>           | 5 beads (base pair step: phosphates: 4 beads; sugars + bases: 1 bead) | 400 (200 bp)  | Bottom-up              | Implicit + explicit ions | Ions                 | Dynamical properties (dsDNA).                     |
| Aksimentiev <sup>174</sup>            | 2 beads (bb.: 1; base: 1)                                             | 200 (ssDNA)   | Bottom-up/<br>Top-down | Implicit                 | No                   | Dynamical properties (dsDNA).                     |
| Vercauteren <sup>157</sup>            | 1 bead                                                                | 1000 (500 bp) | Bottom-up              | Implicit + explicit ions | Ions                 | Dynamical properties (linear and circular dsDNA). |
| Stachiewicz and Molski <sup>161</sup> | SIRAH scheme                                                          | —             | Top-down               | Explicit (MARTINI)       | Water and ions       | Dynamical properties (dsDNA).                     |
| <b>RNA</b><br>Ren2010 <sup>193</sup>  | 5 beads (bb: 2; base: 1)                                              | 27            | Top-down               | Implicit                 | No                   | Structure prediction. Dynamical properties.       |
| Ren2013 <sup>196</sup>                | 5 beads (bb: 2; base: 1)                                              | 122           | Top-down               | Implicit/ Debye–Hückel   | Mg <sup>2+</sup>     | Structure prediction. Dynamical properties.       |
| YUP <sup>188</sup> /<br>Harvey        | 1 bead (bb: 1)                                                        | 76            | Top-down               | Implicit                 | No                   | Structure prediction.                             |

|                                            |                                                                        |                                                                                   |                                          |                           |                                                        |                                                                     |
|--------------------------------------------|------------------------------------------------------------------------|-----------------------------------------------------------------------------------|------------------------------------------|---------------------------|--------------------------------------------------------|---------------------------------------------------------------------|
| NAST <sup>189</sup> /<br>Jonikas           | 1 bead (bb: 1)                                                         | 160                                                                               | Top-down                                 | Implicit                  | No                                                     | Structure prediction.                                               |
| OxRNA <sup>214</sup> /<br>Ouldridge        | Rigid body (with 5 interaction sites)                                  | $10^3$ (structure stability) $10^2$ (folding)                                     | Top-down                                 | Implicit                  | No                                                     | Folding. Dynamical properties.                                      |
| TOPRNA <sup>215</sup> /<br>Brooks          | 3 beads (bb: 2; base: 1)                                               | 12                                                                                | Top-down                                 | Implicit                  | No                                                     | Dynamical properties (topological constraints of secondary motifs). |
| SimRNA <sup>191</sup> /<br>Brujnicki       | 5 beads (bb: 2; base: 3) + a virtual point on the base + 3D cubic grid | <50                                                                               | Top-down                                 | Implicit                  | No                                                     | Folding dynamics and structure prediction.                          |
| RNAkb <sup>192</sup> /<br>Levitt           | 5 beads (bb: 2; base: 3)                                               | 76                                                                                | Top-down                                 | Implicit                  | No                                                     | Structure prediction.                                               |
| HiRe-RNA <sup>164</sup> /<br>Derreumaux    | 6/7 beads (bb:4; base: 1/2)                                            | 36                                                                                | Top-down                                 | Implicit                  | No                                                     | Folding and structure prediction.                                   |
| IfoldRNA <sup>202,203</sup> /<br>Dokholyan | 3 beads (bb: 2; base: 1)                                               | < 50 ( <i>ab initio</i> ) 161 using base-pairing and HRP experimental restraints. | Top-down                                 | Implicit/<br>Debye–Hückel | No                                                     | Folding dynamics and structure prediction.                          |
| FARNA <sup>190</sup> /<br>Das and Baker    | 1 bead (base: 1)                                                       | < 50 (unless including secondary structure information)                           | Top-down                                 | Implicit                  | No                                                     | Structural prediction.                                              |
| MARTINI <sup>178</sup> /<br>Marrink        | 6/7 beads (bb: 3; base: 3Y/4R)                                         | 4793 ( <i>Thermus thermophilus</i> 70S ribosome)                                  | Top-down/<br>bottom-up + elastic network | Explicit                  | Water, ions, lipids, carbohydrates, polymers, proteins | Dynamical properties (ssRNA; dsRNA).                                |

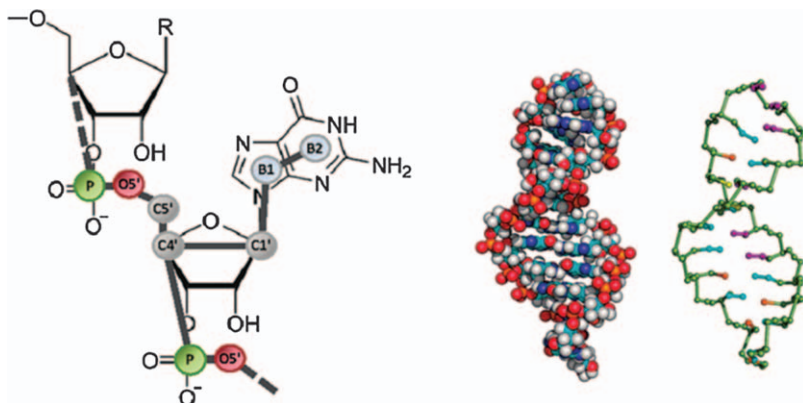


**Figure 7.3** Example of pbCG model (HiRe-RNA) showing the positioning and connectivity of the model.

Reprinted with permission from S. Pasquali and P. Derreumaux, *J. Phys. Chem. B*, 2010, **114**, 11957. Copyright 2010 American Chemical Society.

(e.g. Nordenskiöld<sup>156</sup>) procedures, to obtain the effective beads parameters. Often this parameterisation approach is supplemented with experimental data for further refinement.<sup>174</sup> For both approximations, the functional form of the energy is very simple: harmonic potentials to represent bonded interactions and Lennard-Jones and Coulomb-like potentials to represent non-bonded interactions.<sup>156,158–160,162,163,166</sup> Solvent screening is typically accounted for by continuum electrostatic models such as the Debye–Hückel approach, although some authors represent solvent by defining some kind of explicit water/ion beads.<sup>158,163</sup>

In some cases, the pbCG force field for nucleic acids has been developed in the context of more comprehensive, biologically (or technologically) oriented force fields, which means that compatibility with models for other bio-macromolecules is an additional requisite. A few examples are the widespread used MARTINI force field,<sup>175</sup> which in addition to proteins,<sup>176,177</sup> nucleic acids<sup>158,178</sup> and also carbohydrates,<sup>179</sup> has one of the most diverse set of models for lipids.<sup>180,181</sup> The UNRES model developed in Scheraga's group,<sup>182</sup> the unified BioModi CG force field,<sup>183</sup> and the SIRAH pbCG force field<sup>162,163,184</sup> are other examples. It is also worth mentioning some efforts for mixing atomistic and coarse grain models to study nucleic acids dynamics as, for example, the AMBER/SIRAH DNA representation proposed by Machado and co-workers<sup>185</sup> or the adaptive resolution approach (AdResS<sup>186</sup>) from Praprotnik's group, which combines the AMBER force field to represent the DNA with a mixture of TIP3P and their own CG model for water that interchange solvent resolution when close to the DNA.<sup>187</sup>

### 7.5.2 Coarse-grained Methods for Predicting RNA Structures

Despite the rather small chemical differences between DNA and RNA, the conformational diversity they present is notable. While DNA is found mainly as double stranded structures able to form extremely large assemblies, RNA

shows a great diversity of motifs which combine single stranded and double stranded segments glued by long range contacts that collapse the linear sequence into complex three-dimensional structures. This difference explains the separated routes taken for DNA or RNA CG development, as the final objectives of the models are different *i.e.* predicting duplex properties in DNA while predicting the structure in RNA. The development of RNA structure prediction methods has been fuelled by the increasing gap between the amount of RNA sequence information and the number of RNA structures solved by NMR, X-ray crystallography or cryo-electron microscopy. This has encouraged the development of RNA CG models ranging from highly simplified ones using only one bead per nucleotide like YUP,<sup>188</sup> NAST<sup>189</sup> or FARNA,<sup>190</sup> to more detailed models using up to six or seven beads per nucleotide like SimRNA,<sup>191</sup> RNAkb,<sup>192</sup> HiRe-RNA<sup>164</sup> or Ren *et al.* model<sup>193</sup> (see Table 7.1 for details). Irrespective of the resolution level, accurate results are difficult to obtain, with most predictions showing large RMSD values from the experimental structure,<sup>194</sup> which illustrates the complexity of the problem. Fortunately, when these approaches are supplemented with knowledge about the 2D and 3D contacts, the performance improves substantially. One example is the one bead per nucleotide NAST model developed by Jonikas and co-workers,<sup>189</sup> which is based on molecular dynamics simulations, and uses a knowledge-based statistical potential function supplemented with secondary structure information. The biggest strength of NAST is that it allows for the modelling of rather large, *e.g.* 160 nt, RNA molecules (as a reference consider that the average size of a tRNA molecule is  $\sim 75$  nt). Recently, NAST has been combined with tertiary contact data based on co-evolution.<sup>195</sup> This approach allows for the prediction of all-atom structures of representatives of five RNA families (70–120 nt with at least one long-range contact) with RMSD values of 7–10 Å, which is far better than the average RMSD obtained from pure *in silico* methods ( $\sim 20$  Å) for similarly long sequences.<sup>194</sup> In the case of short sequences ( $< 50$  nt), good predictions with RMSD values in the range of 2–4 Å are achieved by the most detailed models like HiRe-RNA,<sup>164</sup> the model proposed by Ren's group<sup>193</sup> or the very recent SimRNA.<sup>191</sup> The latter has also stood out for its ability to accurately predict both secondary structure and global conformation of pseudoknots, and, as a potential method to investigate intermediate states in the folding dynamics.<sup>191</sup> Also worth mentioning is a recent model developed in Ren's group where the energy functional was improved by including Debye–Hückel electrostatics and hydrogen-bond terms.<sup>196</sup> This new model not only has the ability to fold short sequences ( $< 30$  nt; RMSD  $\sim 4$  Å), but can also capture the structure of long sequence RNAs (*e.g.* the 5S ribosome) using NMR and SAXS restraints. In addition, it allows for the explicit treatment of ions, which was proven to aid in the folding of pseudoknots.<sup>196</sup> This is an example of a modelling trend that has received much attention in the last years, in which inclusion of experimental restraints on pbCG-based folding simulations is used to improve the conformational sampling, which is most important for long RNA predictions.<sup>197–199</sup> Besides NMR or SAXS restraints, other examples of guiding experimental inputs include

sequence covariance analysis,<sup>200</sup> and foot-printing approaches like selective 2'-hydroxyl acylation analysed by primer extension (SHAPE) for secondary structure prediction,<sup>201</sup> or hydroxyl radical probing method (HRP) that accounts for solvent accessibility.<sup>197</sup> From these experiments base-pairing and/or solvent accessibility information can be obtained and integrated into pbCG simulation protocols. That is the case of the iFoldRNAv2 engine<sup>202</sup> that combines a three-bead model representing the phosphate, sugar and nucleobase in RNA (see Table 7.1) with discrete molecular dynamics (dMD) simulations.<sup>203</sup> A striking improvement in the accuracy of this method is achieved when the pbCG model is supplemented with experimental base-pairing and HRP restraints, allowing the authors to go from a poor structural prediction (RMSD  $\sim 32$  Å without restraints) to a remarkable RMSD value of  $\sim 8$  Å for a complex system like the M-box riboswitch (161 nt).<sup>202</sup>

Following an alternative motivation for RNA pbCG development, that is, not aimed at structural prediction but to provide RNA models that preserve stable conformations in molecular simulations, an extension of the MARTINI forcefield to RNA has recently been published.<sup>178</sup> This model is analogous to the MARTINI DNA representation (6–7 beads per nucleotide), with slight bead type and position modifications to account for the presence of the ribose 2'-OH group and the loss of the methyl group in uracil as compared to thymine. Similar to its DNA counterpart, this model requires the usage of elastic networks to maintain RNA stable conformations, thus cannot be used to study processes involving extensive conformational changes. However, being developed in the framework of the comprehensive MARTINI forcefield, it allows for the simulations of intermolecular interactions between RNA and proteins, lipids or other stable nucleic acids motifs. This becomes even more appealing considering the explicit treatment of the solvent and ions inherent to the MARTINI forcefield. Although beyond the pbCG framework, and thus not described in any detail here, other powerful theoretical approaches for RNA structure prediction can be found in the literature. Some examples are the graph formalism by Schlick's group,<sup>204–207</sup> the probabilistic model (BARNACLE) devised by Hamelryck and co-workers,<sup>208</sup> the atomistic approaches followed by the Wales' group in the exploration of the free energy landscape of hairpins<sup>209</sup> or the PDB based 'stop-motion modelling' folding method from the group of Bussi.<sup>210</sup> All of these, and also the previously described methods, still face many challenges, for example, those derived from the interaction of RNA with ions, peptides or proteins, or the presence, in some cases, of a variety of non-coding nucleobases.

## 7.6 Mesoscopic Models

At the nuclear organisation level, eukaryotic DNA is present in the form of chromatin, where 147 base pairs of DNA are wrapped  $\sim 1.7$  times around a histone protein octamer forming the nucleosome, which connected by linker DNA define the chromatin fibre. The string of nucleosomes can reach different types and levels of compaction not just related to its intrinsic

organisation, but also due to constraints of the cellular environment. Compaction of chromatin is crucial in normal cellular functioning, it being one of the main mechanisms allowing the cell to react in front of external stimuli. This explains the magnitude of the effort, both experimental and theoretical focused in the understanding of chromatin structure and dynamics. Current knowledge assumes that structuration of chromatin is hierarchical: the nucleosome fibre is partially or totally compacted into a polymer of 30 nm width (~kb range), but other levels are dictated by cellular activity. It is clear that the overall level of compaction of chromatin is highly regulated and plastic, as less compact regions concentrate highly expressed genes.

Representation of the global chromatin structure is complex, and computational methods are becoming invaluable to bridge noisy and fuzzy low-resolution experiments with structural models that help us to visualise the internal structure of eukaryotic nuclei. In this chapter, we first want to introduce a basic methodological description of computational chromatin models. Then we look more into detail on representative models of the two resolution levels, nucleosome fibre models and chromosome models.

### 7.6.1 Basic Methodological Description

Because of the highly complex environment in the cell nucleus, and the large size of the DNA there, computational models cannot capture all factors involved in chromatin condensation and cannot achieve anything close to atomistic resolution. Also, currently a single computational model cannot cover yet the whole range of resolution levels of chromatin compaction. Two types of models exist covering chromatin properties at different length scales: (i) nucleosome fibre simulations in the kb range and (ii) chromosome simulations in the Mb–Gb range. No matter of the resolution scale of both model types known experimental properties of chromatin fibres are used to derive, parametrise and adapt the models. The shorter the simulated chromatin chain is the higher can be the level of detail applied to the model, and the smaller the need of experimental information. However, even for the simulation of a few dozens of nucleosomes experimental input is still required to obtain meaningful results.

In the nucleosome fibre models, known physical properties of the nucleosome core particle and linker DNA (derived from experimental data or from atomistic simulations in reduced systems) are used to derive DNA deformation parameters as well as coarse-grained electrostatic and steric potentials. It is common for example in nucleosome fibre models to base the coarse-grained representation of the nucleosome on its experimentally determined X-ray structure,<sup>216</sup> and the properties of the simulated nucleosome fibre can be cross-validated with other experiments.<sup>217</sup> In recent nucleosome fibre models parameters such as the ionic environment, the DNA linker length, the presence and absence of linker histones, or the existence of post-translational histone modifications can be included to resemble the *in vivo* situation as close as possible.<sup>149,218</sup>

In the case of chromosome models the parametrisation follows a different approach. While the basis of nucleosome fibre modelling is taking into account atomistic properties of its constituents the chromosome models are usually based on polymer physics. A monomer unit of a chromosome model can comprise from less than 1 kb up to several kb. Intra- and inter-chain interactions can be adjusted according to experiments mapping small-scale chromatin properties or the organisation of the whole genome such as Hi-C,<sup>219</sup> cryo-EM,<sup>220</sup> FISH<sup>221</sup> or STORM.<sup>222</sup>

## 7.6.2 Nucleosome Fibre Simulations

The nucleosome fibre models take the accurate atomistic description of the constituents of chromatin and transfers it into a coarse-grained model. This allows to probe the structure of the chromatin fibre at the kb level, while having a method that maintains solid roots in the rigorous formalism of physical atomistic models. Among the different approaches used we comment here the six-angle model, where the nucleosome is represented as a rigid object with three rotational degrees of freedom. Three angles describe the linker DNA path between two consecutive nucleosomes, the remaining three angles determine the nucleosome orientation relative to the incoming and outgoing linker DNA<sup>223</sup> (see Figure 7.4). Inter-nucleosome interactions

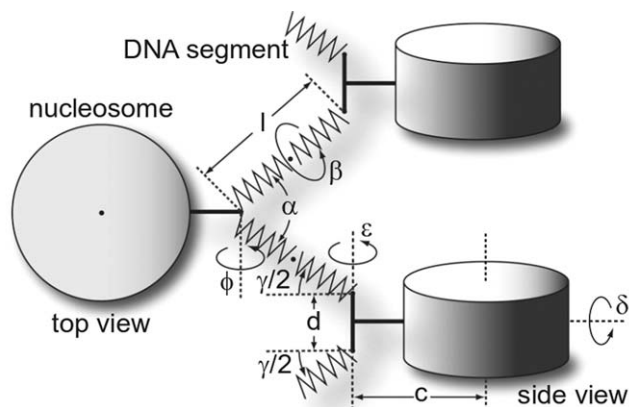


**Figure 7.4** Schematic representation of the six-angle model. The nucleosomes are represented as cylinders connected by linker DNA which is composed of elastic pieces (depicted as springs) of length  $l$ . The angles  $\delta$ ,  $\epsilon$  and  $\phi$  describe the orientation of the nucleosome with respect to the incoming and outgoing linker DNA while the angles  $\alpha$ ,  $\beta$  and  $\gamma$  describe the DNA path between two consecutive nucleosomes.

Reprinted from *Biophysical Journal*, Vol. 98, R. Stehr, R. Schöpflin, R. Ettig, N. Kepper, K. Rippe and G. Wedemann, 2010, 98, Exploring the Conformational Space of Chromatin Fibres and Their Stability by Numerical Dynamic Phase Diagrams, 1028, Copyright 2010, with permissions from Elsevier.



are represented by a spherocylindrically shaped potential. This model is among others used to relate the geometric phase space to the energy associated with a given chromatin conformation by numerical phase diagrams.<sup>224</sup> Another model represents the nucleosome core as an irregular surface with a set of Debye–Hückel charges along with flexible histone tails and linker histones.<sup>225</sup> The linker DNA is described as the histone tails and linker histones by a worm-like chain. With this model, for example, the impact of nucleosome repeat lengths, linker histones, or core histone modifications on the chromatin compaction has been determined.<sup>149,218</sup> Recent developments move towards a more sequence-dependent description of linker flexibility, which can be deduced from the use of elastic deformation parameters derived from atomistic simulations of all possible unique tetramers.

### 7.6.3 Chromosome Simulations

Chromosome simulations are usually done by means of physical polymer fibre models, or in some cases by even simpler models which assume an infinitely deformable continuum string, whose conformation is modified to fulfil known experimental restraints. There are several possibilities to introduce chromatin interactions known to exist by experimental measures in the calculation. One can, for example, assign specific attractive inter-chain interactions between monomers in certain regions with the same epigenetic mark,<sup>226</sup> capturing then the nature of the epigenomic domains characterised experimentally. Another possibility is to add diffusive binders which can interact (with the same or different affinity) with a fraction of the monomers in the polymer chain (see Figure 7.5).<sup>227</sup> This type of simulation can incorporate chromatin folding data obtained by chromosome conformational capture experiments<sup>219</sup> or FISH.<sup>221</sup> Similarly, other polymer models use spatial elastic constraints to promote random looping along the chromatin fibre to create a linearly compressed array of consecutive chromatin, which has been useful to reproduce 5C and Hi-C data on the organisation of the human mitotic chromosome.<sup>228</sup>

Chromatin simulation is the ultimate challenge for DNA simulation groups, as when dealing with structures that can reach centimetre scale, atomistic or even normal coarse-grain models cannot be used, and simpler multi-resolution methods need to be applied. Interesting results are being obtained by combining two models: (i) nucleosome fibre models that can be easily connected with pure physical models with roots in atomistic simulations and (ii) chromosome models that can be validated by recent experimental data on global eukaryotic nuclear structure. The field is still in its first stages of development, but we can envision a situation where the user can zoom into the genome starting at a coarse view onto the whole nucleus, then reaching single chromosome resolution at the Mb level up to nucleosome level, or why not even up to atomistic resolution.

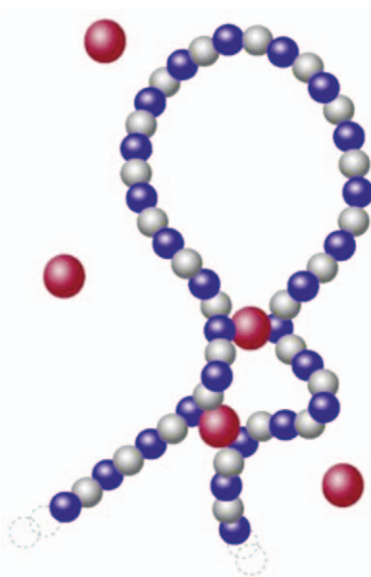


**Figure 7.5** The ‘strings and binders switch model’ uses a self-avoiding worm-like chain where a fraction of the monomers (blue) can act as a binding site for diffusive binders (red).

Reprinted from M. Barbieri, M. Chotalia, J. Fraser, L.-M. Lavitas, J. Dostie, A. Pombo, *et al.*, *Proc. Natl. Acad. Sci. U.S.A.*, 2012, **109**, 16173. With permission from PNAS.

## 7.7 Conclusions

We have reviewed some of the latest developments and applications of theoretical methods to study nucleic acids (NAs), from the detailed electronic structure of the basic structural units in RNA and DNA to the chromosome dynamics with mesoscopic approaches. The notorious improvement of both computation power and algorithms has allowed the addressing of increasingly complex studies regarding NA structure, dynamics, and function, in close integration with experimental data. Since the theoretical methods and computers are in constant evolution we can envision a near future where the multiscale nature of NAs would be efficiently tackled. Moreover, all these theoretical approaches will be harmonically integrated and applied to gain a holistic view of these complex biomolecules.

## Acknowledgements

This work has been supported by the Spanish Ministry of Science (BFU2014-61670-EXP and BIO2015-64802-R), the Catalan SGR, the Instituto Nacional de Bioinformatica, and the European Research Council (ERC SimDNA), the European Union’s Horizon 2020 research and innovation program under grant agreement no. 676556 (MuG). M.O. is an ICREA academia researcher.

L.D. is a SNI (Sistema Nacional de Investigadores; ANII, Uruguay) researcher. P.D.D. is a SNI and PEDECIBA (Programa de Desarrollo de las Ciencias Básicas) researcher. H.G. is a Juan de la Cierva postdoc and J.W. is a La Caixa Foundation Ph.D. fellow.

## References

1. P. D. Dans, J. Walther, H. Gómez and M. Orozco, *Curr. Opin. Struct. Biol.*, 2016, **37**, 29.
2. M. Orozco, A. Pérez, A. Noy and F. J. Luque, *Chem. Soc. Rev.*, 2003, **32**, 350.
3. H. Kruse, M. Havrila and J. Šponer, *J. Chem. Theory Comput.*, 2014, **10**, 2615.
4. A. Mládek, M. Krepl, D. Svozil, P. Čech, M. Otyepka and P. Banáš, *et al.*, *Phys. Chem. Chem. Phys.*, 2013, **15**, 7295.
5. A. Mládek, P. Banáš, P. Jurečka, M. Otyepka, M. Zgarbová and J. Šponer, *J. Chem. Theory Comput.*, 2014, **10**, 463.
6. K. Gkionis, H. Kruse, J. A. Platts, A. Mládek, J. Koča and J. Šponer, *J. Chem. Theory Comput.*, 2014, **10**, 1326.
7. J. Šponer, A. Mládek, N. Špačková, X. Cang, T. E. Cheatham 3rd and S. Grimme, *J. Am. Chem. Soc.*, 2013, **135**, 9785.
8. C. Lech, B. Heddi and A. Phan, *Nucleic Acids Res.*, 2013, **41**, 2034.
9. T. M. Parker, E. G. Hohenstein, R. M. Parrish, N. V. Hud and D. C. Sherrill, *J. Am. Chem. Soc.*, 2013, **135**, 1306.
10. E. Sochacka, R. H. Szczepanowski, M. Cypriak, M. Sobczak, M. Janicka and K. Kraszewska, *et al.*, *Nucleic Acids Res.*, 2015, **43**, 2499.
11. O. O. Brovarets and D. M. Hovorun, *J. Biomol. Struct. Dyn.*, 2014, **32**, 1474.
12. A. Arcella, J. Dreyer, E. Ippoliti, I. Ivani, G. Portella and V. Gabelica, *et al.*, *Angew. Chem., Int. Ed.*, 2015, **54**, 467.
13. A. Warshel and M. Levitt, *J. Mol. Biol.*, 1976, **103**, 227.
14. V. Mlýnský, P. Banáš, J. Šponer, M. W. Van Der Kamp, A. J. Mulholland and M. Otyepka, *J. Chem. Theory Comput.*, 2014, **10**, 1608.
15. J. Sgrignani and A. Magistrato, *ACS Catal.*, 2015, **5**, 3864.
16. J. Aranda, K. Zinovjev, M. Roca and I. Tuñón, *J. Am. Chem. Soc.*, 2014, **136**, 16227.
17. J. Garrec, C. Patel, U. Rothlisberger and E. Dumont, *J. Am. Chem. Soc.*, 2012, **134**, 2111.
18. P. Chandan, J. Garrec, C. Dupont and E. Dumont, *Biochemistry*, 2013, **52**, 425.
19. L. Lior-Hoffmann, L. Wang, S. Wang, N. E. Geacintov, S. Broyde and Y. Zhang, *Nucleic Acids Res.*, 2012, **40**, 9193.
20. E. Rosta, W. Yang and G. Hummer, *J. Am. Chem. Soc.*, 2014, **136**, 3137.
21. R. Molina, S. Stella, P. Redondo, H. Gómez, M. J. Marcaida and M. Orozco, *et al.*, *Nat. Struct. Mol. Biol.*, 2015, **22**, 65.

22. G. Palermo, A. Cavalli, M. L. Klein, M. Alfonso-Prieto, M. Dal Peraro and M. De Vivo, *Acc. Chem. Res.*, 2015, **48**, 220.
23. R. Venkatramani, S. Keinan, A. Balaeff and D. N. Beratan, *Coord. Chem. Rev.*, 2011, **255**, 635.
24. C. J. Lech, A. T. Phan, M. E. Michel-Beyerle and A. A. Voityuk, *J. Phys. Chem. B*, 2013, **117**, 9851.
25. A. Bacolla, X. Zhu, H. Chen, K. Howells, D. N. Cooper and K. M. Vasquez, *Nucleic Acids Res.*, 2015, **43**, 5065.
26. T. Kubař and M. Elstner, *J. R. Soc., Interface*, 2013, **10**, 20130415.
27. V. A. Spata and S. Matsika, *J. Phys. Chem. A*, 2014, **118**, 12021.
28. H. Gattuso, X. Assfeld and A. Monari, *Theor. Chem. Acc.*, 2015, **134**, 36.
29. T. Zelený, M. Ruckebauer, A. J. A. Aquino, T. Müller, F. Lankaš and T. Dršata, *et al.*, *J. Am. Chem. Soc.*, 2012, **134**, 13662.
30. M. D'Abramo, C. L. Castellazzi, M. Orozco and A. Amadei, *J. Phys. Chem. B*, 2013, **117**, 8697.
31. R. Improta, F. Santoro and L. Blancafort, *Chem. Rev.*, 2016, **116**, 3540.
32. M. Levitt, *Nature*, 1969, **224**, 759.
33. M. Levitt, *Cold Spring Harbor Symp. Quant. Biol.*, 1983, **47**, 251.
34. M. Hirshberg and M. Levitt in *Dynamics and the Problem of Recognition in Biological Macromolecules*, ed. O. Jardetzky and J.-F. Lefèvre, Springer, US, New York, 1996, ch. 13, p. 173.
35. A. Pérez, F. J. Luque and M. Orozco, *Acc. Chem. Res.*, 2012, **45**, 196.
36. M. Orozco, A. Noy and A. Pérez, *Curr. Opin. Struct. Biol.*, 2008, **18**, 185.
37. T. E. Cheatham 3rd, *Curr. Opin. Struct. Biol.*, 2004, **14**, 360.
38. A. D. MacKerell and L. Nilsson, *Curr. Opin. Struct. Biol.*, 2008, **18**, 194.
39. C. A. Laughton and S. A. Harris, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2011, **1**, 590.
40. T. E. Cheatham 3rd, P. Cieplak and P. A. Kollman, *J. Biomol. Struct. Dyn.*, 2012, **16**, 845.
41. D. L. Beveridge, G. Barreiro, K. S. Byun, D. A. Case, T. E. Cheatham 3rd and S. B. Dixit, *et al.*, *Biophys. J.*, 2004, **87**, 3799.
42. A. Pérez, I. Marchán, D. Svozil, J. Šponer, T. E. Cheatham 3rd, C. A. Laughton and M. Orozco, *Biophys. J.*, 2007, **92**, 3817.
43. A. Pérez, F. J. Luque and M. Orozco, *J. Am. Chem. Soc.*, 2007, **129**, 14739.
44. M. Zgarbová, F. J. Luque, J. Šponer, T. E. Cheatham 3rd, M. Otyepka and P. Jurečka, *J. Chem. Theory Comput.*, 2013, **9**, 2339.
45. M. Krepl, M. Zgarbová, P. Stadlbauer, M. Otyepka, P. Banáš and J. Koča, *et al.*, *J. Chem. Theory Comput.*, 2012, **8**, 2506.
46. I. Ivani, P. D. Dans, A. Noy, A. Pérez, I. Faustino and A. Hospital, *et al.*, *Nat. Methods*, 2016, **13**, 55.
47. K. Hart, N. Foloppe, C. M. Baker, E. J. Denning, L. Nilsson and A. D. MacKerell, *J. Chem. Theory Comput.*, 2012, **8**, 348.
48. A. Savelyev and A. D. MacKerell, *J. Comput. Chem.*, 2014, **35**, 1219.
49. W. Jiang, D. J. Hardy, J. C. Phillips, A. D. Mackerell, K. Schulten and B. Roux, *J. Phys. Chem. Lett.*, 2011, **2**, 87.

50. E. Fadrná, N. Špačková, J. Sarzyńska, J. Koča, M. Orozco and T. E. Cheatham 3rd, *et al.*, *J. Chem. Theory Comput.*, 2009, **5**, 2514.
51. M. Zgarbová, J. Šponer, M. Otyepka, T. E. Cheatham 3rd, R. Galindo-Murillo and P. Jurečka, *J. Chem. Theory Comput.*, 2015, **11**, 5723.
52. P. D. Dans, I. Ivani, A. Hospital, G. Portella, C. González and M. Orozco, *Nucleic Acids Res.*, 2017, **45**, 4217.
53. P. D. Dans, L. Danilāne, I. Ivani, T. Dršata, F. Lankaš and A. Hospital, *et al.*, *Nucleic Acids Res.*, 2016, **44**, 4052.
54. A. A. Chen and A. E. García, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, 16820.
55. R. F. Brown, C. T. Andrews and A. H. Elcock, *J. Chem. Theory Comput.*, 2015, **11**, 2315.
56. S. Bommarito, N. Peyret and J. SantaLucia, *Nucleic Acids Res.*, 2000, **28**, 1929.
57. K. M. Guckian, B. A. Schweitzer, R. X.-F. Ren, C. J. Sheils, D. C. Tahmassebi and E. T. Kool, *J. Am. Chem. Soc.*, 2000, **122**, 2213.
58. T. Steinbrecher, J. Latzer and D. A. Case, *J. Chem. Theory Comput.*, 2012, **8**, 4405.
59. R. Galindo-Murillo, D. R. Roe and T. E. Cheatham 3rd, *Nat. Commun.*, 2014, **5**, 5152.
60. I. Yildirim, H. A. Stern, J. D. Tubbs, S. D. Kennedy and D. H. Turner, *J. Phys. Chem. B*, 2011, **115**, 9261.
61. A. Gil-Ley, S. Bottaro and G. Bussi, *J. Chem. Theory Comput.*, 2016, **12**, 2790.
62. M. Zgarbová, M. Otyepka, J. Šponer, A. Mládek, P. Banáš and T. E. Cheatham 3rd, *et al.*, *J. Chem. Theory Comput.*, 2011, **7**, 2886.
63. C. Bergonzo, N. M. Henriksen, D. R. Roe and T. E. Cheatham 3rd, *RNA*, 2015, **21**, 1578.
64. S. Piana-Agostinetti, Oral Presentation at the Meeting: *RNA: Structure, Dynamics and Function*. 2016, May, Trieste, Italy.
65. L. Darré, I. Ivani, P. D. Dans, H. Gomez, A. Hospital and M. Orozco, *J. Am. Chem. Soc.*, 2016, **138**, 16355.
66. C. Alemán, M. Orozco and F. J. Luque, *Chem. Phys.*, 1994, **189**, 573.
67. M. Basma, S. Sundara, D. Calgan, T. Vernali and R. J. Woods, *J. Comput. Chem.*, 2001, **22**, 1125.
68. P. D. Dans, I. Faustino, F. Battistini, K. Zakrzewska, R. Lavery and M. Orozco, *Nucleic Acids Res.*, 2014, **42**, 11304.
69. A. Savelyev and A. D. MacKerell, *J. Phys. Chem. B*, 2015, **119**, 4428.
70. A. Savelyev and A. D. MacKerell, *J. Phys. Chem. Lett.*, 2015, **6**, 212.
71. A. Savelyev and A. D. MacKerell, *J. Chem. Theory Comput.*, 2015, **11**, 4473.
72. F. Pan, C. Roland and C. Sagui, *Nucleic Acids Res.*, 2014, **42**, 13981.
73. M. Pasi, J. H. Maddocks, D. Beveridge, T. C. Bishop, D. A. Case and T. E. Cheatham 3rd, *et al.*, *Nucleic Acids Res.*, 2014, **42**, 12272.
74. P. D. Dans, A. Pérez, I. Faustino, R. Lavery and M. Orozco, *Nucleic Acids Res.*, 2012, **40**, 10668.

75. M. Zgarbová, M. Otyepka, J. Šponer, F. Lankaš and P. Jurečka, *J. Chem. Theory Comput.*, 2014, **10**, 3177.
76. N. K. Banavali, *J. Am. Chem. Soc.*, 2013, **135**, 8274.
77. N. K. Banavali, *J. Phys. Chem. B*, 2013, **117**, 14320.
78. G. Rossetti, P. D. Dans, I. Gomez-Pinto, I. Ivani, C. Gonzalez and M. Orozco, *Nucleic Acids Res.*, 2015, **43**, 4309.
79. S. Kim, E. Broströmer, D. Xing, J. Jin, S. Chong and H. Ge, *et al.*, *Science*, 2013, **339**, 816.
80. T. Dršata, M. Zgarbová, N. Špačková, P. Jurečka, J. Šponer and F. Lankaš, *J. Phys. Chem. Lett.*, 2014, **5**, 3831.
81. M. Taranova, A. D. Hirsh, N. C. Perkins and I. Andricioaei, *J. Phys. Chem. B*, 2014, **118**, 11028.
82. J. Curuksu, M. Zacharias, R. Lavery and K. Zakrzewska, *Nucleic Acids Res.*, 2009, **37**, 3766.
83. T. Lionnet and F. Lankaš, *Biophys. J.*, 2007, **92**, L30.
84. A. Noy, A. Pérez, F. Lankaš, F. J. Luque and M. Orozco, *J. Mol. Biol.*, 2004, **343**, 627.
85. A. Pérez, A. Noy, F. Lankaš, F. J. Luque and M. Orozco, *Nucleic Acids Res.*, 2004, **32**, 6144.
86. A. Noy, A. Pérez, M. Márquez, F. J. Luque and M. Orozco, *J. Am. Chem. Soc.*, 2005, **127**, 4910.
87. A. Noy, F. J. Luque and M. Orozco, *J. Am. Chem. Soc.*, 2008, **130**, 3486.
88. K. Pachulska-Wieczorek, K. J. Purzycka and R. W. Adamiak, *Nucleic Acids Res.*, 2006, **34**, 2984.
89. A. Pérez and M. Orozco, *Angew. Chem., Int. Ed.*, 2010, **49**, 4805.
90. G. Portella and M. Orozco, *Angew. Chem., Int. Ed.*, 2010, **49**, 7673.
91. S. Kannan and M. Zacharias, *Biophys. J.*, 2007, **93**, 3218.
92. A. E. Garcia and D. Paschek, *J. Am. Chem. Soc.*, 2008, **130**, 815.
93. C. Yang, E. Kim and Y. Pak, *Nucleic Acids Res.*, 2015, **43**, 7769.
94. E. N. Nikolova, G. B. Goh, C. L. Brooks 3rd and H. M. Al-Hashimi, *J. Am. Chem. Soc.*, 2013, **135**, 6766.
95. R. Soliva, F. J. Luque, C. Alhambra and M. Orozco, *J. Biomol. Struct. Dyn.*, 1999, **17**, 89.
96. M. Moradi, V. Babin, C. Roland and C. Sagui, *Nucleic Acids Res.*, 2013, **41**, 33.
97. A. Arcella, G. Portella, M. L. Ruiz, R. Eritja, M. Vilaseca and V. Gabelica, *et al.*, *J. Am. Chem. Soc.*, 2012, **134**, 6596.
98. V. Gabelica in *Nucleic Acids in the Gas Phase*, ed. V. Gabelica, Springer Berlin Heidelberg, Heidelberg, 2014, ch. 1, p. 3.
99. M. Rueda, S. G. Kalko, F. J. Luque and M. Orozco, *J. Am. Chem. Soc.*, 2003, **125**, 8007.
100. M. Rueda, F. J. Luque and M. Orozco, *J. Am. Chem. Soc.*, 2006, **128**, 3608.
101. G. Portella, M. W. Germann, N. V. Hud and M. Orozco, *J. Am. Chem. Soc.*, 2014, **136**, 3075.
102. A. Sen, D. Sahu and B. Ganguly, *J. Phys. Chem. B*, 2013, **117**, 9840.

103. M. Nakano, H. Tateishi-Karimata, S. Tanaka and N. Sugimoto, *J. Phys. Chem. B*, 2014, **118**, 379.
104. G. Portella, M. Terrazas, N. Villegas, C. González and M. Orozco, *Angew. Chem., Int. Ed. Engl.*, 2015, **54**, 10488.
105. A. Arcella, G. Portella, R. Collepardo-Guevara, D. Chakraborty, D. J. Wales and M. Orozco, *J. Phys. Chem. B*, 2014, **118**, 8540.
106. B. Shi, Y. K. Shin, A. A. Hassanali and S. J. Singer, *J. Phys. Chem. B*, 2015, **119**, 11030.
107. A. D. Bobadilla and J. M. Seminario, *J. Phys. Chem. C*, 2013, **117**, 26441.
108. R. P. Singh, R. Blossey and F. Cleri, *Biophys. J.*, 2013, **105**, 2820.
109. M. Rebič, F. Mocci, A. Laaksonen and J. Uličný, *J. Phys. Chem. B*, 2015, **119**, 105.
110. B. Islam, M. Sgobba, C. A. Laughton, M. Orozco, J. Šponer and S. Neidle, *et al.*, *Nucleic Acids Res.*, 2013, **41**, 2723.
111. G. C. Shields, C. A. Laughton and M. Orozco, *J. Am. Chem. Soc.*, 1997, **119**, 7463.
112. E. Cubero, A. Aviñó, B. G. de la Torre, M. Frieden, R. Eritja and F. J. Luque, *et al.*, *J. Am. Chem. Soc.*, 2002, **124**, 3133.
113. E. Cubero, N. G. A. Abrescia, J. A. Subirana, F. J. Luque and M. Orozco, *J. Am. Chem. Soc.*, 2003, **125**, 14603.
114. I. Bešševová, P. Banáš, P. Kührová, P. Košinová, M. Otyepka and J. Šponer, *J. Phys. Chem. B*, 2012, **116**, 9899.
115. D. E. Condon, S. D. Kennedy, B. C. Mort, R. Kierzek, I. Yildirim and D. H. Turner, *J. Chem. Theory Comput.*, 2015, **11**, 2729.
116. P. Kührová, P. Banáš, R. B. Best, J. Šponer and M. Otyepka, *J. Chem. Theory Comput.*, 2013, **9**, 2115.
117. G. B. Goh, J. L. Knight and C. L. Brooks 3rd, *J. Phys. Chem. Lett.*, 2013, **4**, 760.
118. S. Halder, P. Kührová, P. Banáš, V. Spiwok, J. Šponer and P. Hobza, *et al.*, *J. Chem. Theory Comput.*, 2015, **11**, 3866.
119. M. Dubecký, N. G. Walter, J. Šponer, M. Otyepka and P. Banáš, *Biopolymers*, 2015, **103**, 550.
120. Y. Xu, A. D. MacKerell and L. Nilsson, *Bioorg. Med. Chem.*, 2016, **24**, 4826.
121. A. Saint-Leger, C. Bello, P. D. Dans, A. G. Torres, E. M. Novoa and N. Camacho, *et al.*, *Sci. Adv.*, 2016, **2**, e1501860.
122. N. Špačková and J. Šponer, *Nucleic Acids Res.*, 2006, **34**, 697.
123. A. Hospital, P. Andrio, C. Cugnasco, L. Codo, Y. Becerra and P. D. Dans, *et al.*, *Nucleic Acids Res.*, 2015, **44**, D272.
124. M. Kara and M. Zacharias, *Biophys. J.*, 2013, **104**, 1089.
125. M. Krepl, M. Otyepka, P. Banáš and J. Šponer, *J. Phys. Chem. B*, 2013, **117**, 1872.
126. N. A. Kuznetsov, C. Bergonzo, A. J. Campbell, H. Li, G. V. Mechetin and C. de los Santos, *et al.*, *Nucleic Acids Res.*, 2015, **43**, 272.
127. X. Guo, Z. Liu, Q. Song, L. Wang and D. Zhong, *J. Phys. Chem. B*, 2015, **119**, 3446.



128. E. Dumont, T. Dršata, C. F. Guerra and F. Lankaš, *Biochemistry*, 2015, **54**, 1259.
129. I. Faustino, C. Curutchet, F. J. Luque and M. Orozco, *Phys. Chem. Chem. Phys.*, 2014, **16**, 1101.
130. R. Soliva, R. Güimil García, J. R. Blas, R. Eritja, J. L. Asensio and C. González, *et al.*, *Nucleic Acids Res.*, 2000, **28**, 4531.
131. J. R. Blas, F. J. Luque and M. Orozco, *J. Am. Chem. Soc.*, 2004, **126**, 154.
132. N. Špačková, E. Cubero, J. Šponer and M. Orozco, *J. Am. Chem. Soc.*, 2004, **126**, 14642.
133. C. Bianchi and R. Zangi, *J. Phys. Chem. B*, 2013, **117**, 2348.
134. J. Yang, L. Lior-Hoffmann, S. Wang, Y. Zhang and S. Broyde, *Biochemistry*, 2013, **52**, 2828.
135. A. T. Carvalho, L. Gouveia, C. R. Kanna, S. K. Wärmländer, J. A. Platts and S. C. L. Kamerlin, *Epigenetics*, 2014, **9**, 1604.
136. P. M. D. Severin, X. Zou, H. E. Gaub and K. Schulten, *Nucleic Acids Res.*, 2011, **39**, 8740.
137. A. Pérez, C. L. Castellazzi, F. Battistini, K. Collinet, O. Flores and O. Deniz, *et al.*, *Biophys. J.*, 2012, **102**, 2140.
138. G. Portella, F. Battistini and M. Orozco, *PLoS Comput. Biol.*, 2013, **9**, e1003354.
139. I. Jimenez-Useche and C. Yuan, *Biophys. J.*, 2012, **103**, 2502.
140. G. Paillard and R. Lavery, *Structure*, 2004, **12**, 113.
141. Y. Yonetani and H. Kono, *J. Phys. Chem. B*, 2013, **117**, 7535.
142. S. Furini, P. Barbini and C. Domene, *Nucleic Acids Res.*, 2013, **41**, 3963.
143. B. Bouvier, K. Zakrzewska and R. Lavery, *Angew. Chem., Int. Ed.*, 2011, **50**, 6516.
144. B. Bouvier and R. Lavery, *J. Am. Chem. Soc.*, 2009, **131**, 9864.
145. H. Wang and C. A. Laughton, *Methods*, 2007, **42**, 196.
146. A. de Ruiter and B. Zagrovic, *Nucleic Acids Res.*, 2015, **43**, 708.
147. J. Erler, R. Zhang, L. Petridis, X. Cheng, J. C. Smith and J. Langowski, *Biophys. J.*, 2014, **107**, 2911.
148. D. Winogradoff, I. Echeverria, D. A. Potoyan and G. A. Papoian, *J. Am. Chem. Soc.*, 2015, **137**, 6245.
149. R. Collepardo-Guevara, G. Portella, M. Vendruscolo, D. Frenkel, T. Schlick and M. Orozco, *J. Am. Chem. Soc.*, 2015, **137**, 10205.
150. Y. Cai, K. Kropachev, M. A. Terzidis, A. Masi, C. Chatgililoglu and V. Shafirovich, *et al.*, *Biochemistry*, 2015, **54**, 4181.
151. T. C. Bishop, *Biophys. J.*, 2008, **95**, 1007.
152. C. Mura and J. A. McCammon, *Nucleic Acids Res.*, 2008, **36**, 4941.
153. A. Banerjee, W. Yang, M. Karplus and G. L. Verdine, *Nature*, 2005, **434**, 612.
154. M. Krepl, A. Cléry, M. Blatter, F. H. T. Allain and J. Šponer, *Nucleic Acids Res.*, 2016, **44**, 6452.
155. M. Krepl, M. Havrila, P. Stadlbauer, P. Banáš, M. Otyepka and J. Pasulka, *et al.*, *J. Chem. Theory Comput.*, 2015, **11**, 1220.

156. N. Korolev, D. Luo, A. P. Lyubartsev and L. Nordenskiöld, *Polymers*, 2014, **6**, 1655.
157. A. Naômé, A. Laaksonen and D. P. Vercauteren, *J. Chem. Theory Comput.*, 2015, **11**, 2813.
158. J. J. Uusitalo, H. I. Ingólfsson, P. Akhshi, D. P. Tieleman and S. J. Marrink, *J. Chem. Theory Comput.*, 2015, **11**, 3932.
159. M. Maciejczyk, A. Spasic, A. Liwo and H. A. Scheraga, *J. Chem. Theory Comput.*, 2014, **10**, 5020.
160. T. Cragolini, P. Derreumaux and S. Pasquali, *J. Phys. Chem. B*, 2013, **117**, 8047.
161. A. Stachiewicz and A. Molski, *J. Comput. Chem.*, 2015, **36**, 947.
162. P. D. Dans, A. Zeida, M. R. Machado and S. Pantano, *J. Chem. Theory Comput.*, 2010, **6**, 1711.
163. L. Darré, M. R. Machado, P. D. Dans, F. E. Herrera and S. Pantano, *J. Chem. Theory Comput.*, 2010, **6**, 3793.
164. S. Pasquali and P. Derreumaux, *J. Phys. Chem. B*, 2010, **114**, 11957.
165. M. R. Machado and S. Pantano, *J. Chem. Theory Comput.*, 2015, **11**, 5012.
166. D. M. Hinckley, G. S. Freeman, J. K. Whitmer and J. J. de Pablo, *J. Chem. Phys.*, 2013, **139**, 144903.
167. G. S. Freeman, D. M. Hinckley, J. P. Lequieu, J. K. Whitmer and J. J. de Pablo, *J. Chem. Phys.*, 2014, **141**, 165103.
168. R. Frederickx, T. Veld and E. Carlon, *Phys. Rev. Lett.*, 2014, **112**, 198102.
169. T. E. Ouldridge, *Mol. Phys.*, 2014, **113**, 1.
170. J. P. K. Doye, T. E. Ouldridge, A. A. Louis, F. Romano, P. Šulc and C. Matek, *et al.*, *Phys. Chem. Chem. Phys.*, 2013, **15**, 20395.
171. P. Šulc, T. E. Ouldridge, F. Romano, J. P. K. Doye and A. A. Louis, *Nat. Comput.*, 2014, **13**, 535.
172. T. E. Ouldridge, P. Šulc, F. Romano, J. P. K. Doye and A. A. Louis, *Nucleic Acids Res.*, 2013, **41**, 8886.
173. H. Li, Z. Wang, N. Li, X. He and H. Liang, *J. Chem. Phys.*, 2014, **141**, 044911.
174. C. Maffeo, T. T. M. Ngo, T. Ha and A. Aksimentiev, *J. Chem. Theory Comput.*, 2014, **10**, 2891.
175. S. J. Marrink and D. P. Tieleman, *Chem. Soc. Rev.*, 2013, **42**, 6801.
176. L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman and S.-J. Marrink, *J. Chem. Theory Comput.*, 2008, **4**, 819.
177. D. H. de Jong, G. Singh, W. F. D. Bennett, C. Arnarez, T. A. Wassenaar and L. V. Schäfer, *et al.*, *J. Chem. Theory Comput.*, 2013, **9**, 687.
178. J. J. Uusitalo, H. I. Ingólfsson, S. J. Marrink and I. Faustino, *Biophys. J.*, 2017, **113**, 246.
179. C. A. López, A. J. Rzepiela, A. H. de Vries, L. Dijkhuizen, P. H. Hünenberger and S. J. Marrink, *J. Chem. Theory Comput.*, 2009, **5**, 3195.

180. S. J. Marrink, A. H. de Vries and A. E. Mark, *J. Phys. Chem. B*, 2004, **108**, 750.
181. S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman and A. H. de Vries, *J. Phys. Chem. B*, 2007, **111**, 7812.
182. A. Liwo, M. Khalili and H. A. Scheraga, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 2362.
183. C. B. Markegard, I. W. Fu, K. A. Reddy and H. D. Nguyen, *J. Phys. Chem. B*, 2015, **119**, 1823.
184. L. Darré, M. R. Machado, A. F. Brandner, H. C. González, S. Ferreira and S. Pantano, *J. Chem. Theory Comput.*, 2015, **11**, 723.
185. M. R. Machado, P. D. Dans and S. Pantano, *Phys. Chem. Chem. Phys.*, 2011, **13**, 18134.
186. M. Praprotnik, S. Matysiak, L. D. Site, K. Kremer and C. Clementi, *J. Phys. Condens. Matter*, 2007, **19**, 292201.
187. J. Zavadlav, R. Podgornik and M. Praprotnik, *J. Chem. Theory Comput.*, 2015, **11**, 5035.
188. R. K. Z. Tan, A. S. Petrov and S. C. Harvey, *J. Chem. Theory Comput.*, 2006, **2**, 529.
189. M. A. Jonikas, R. J. Radmer, A. Laederach, R. Das, S. Pearlman and D. Herschlag, *et al.*, *RNA*, 2009, **15**, 189.
190. R. Das and D. Baker, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 14664.
191. M. J. Boniecki, G. Lach, W. K. Dawson, K. Tomala, P. Lukasz and T. Soltysinski, *et al.*, *Nucleic Acids Res.*, 2016, **44**, e63.
192. J. Bernauer, X. Huang, A. Y. L. Sim and M. Levitt, *RNA*, 2011, **17**, 1066.
193. Z. Xia, D. P. Gardner, R. R. Gutell and P. Ren, *J. Phys. Chem. B*, 2010, **114**, 13497.
194. C. Laing and T. Schlick, *J. Phys. Condens. Matter*, 2010, **22**, 283101.
195. C. Weinreb, A. J. Riesselman, J. B. Ingraham, T. Gross, C. Sander and D. S. Marks, *Cell*, 2016, **165**, 963.
196. Z. Xia, D. R. Bell, Y. Shi and P. Ren, *J. Phys. Chem. B*, 2013, **117**, 3135.
197. F. Ding, C. A. Lavender, K. M. Weeks and N. V. Dokholyan, *Nat. Methods*, 2012, **9**, 603.
198. C. M. Gherghe, C. W. Leonard, F. Ding, N. V. Dokholyan and K. M. Weeks, *J. Am. Chem. Soc.*, 2009, **131**, 2541.
199. C. A. Lavender, F. Ding, N. V. Dokholyan and K. M. Weeks, *Biochemistry*, 2010, **49**, 4931.
200. R. R. Gutell, A. Power, G. Z. Hertz, E. J. Putz and G. D. Stormo, *Nucleic Acids Res.*, 1992, **20**, 5785.
201. J. T. Low and K. M. Weeks, *Methods*, 2010, **52**, 150.
202. A. Krokhotin, K. Houlihan and N. V. Dokholyan, *Bioinformatics*, 2015, **31**, 2891.
203. F. Ding, S. Sharma, P. Chalasani, V. V. Demidov, N. E. Broude and N. V. Dokholyan, *RNA*, 2008, **14**, 1164.
204. N. Kim, N. Shiffeldrim, H. H. Gan and T. Schlick, *J. Mol. Biol.*, 2004, **341**, 1129.

205. N. Kim, K. N. Fuhr and T. Schlick in *Biophysics of RNA Folding*, ed. R. Russell, Springer-Verlag, New York, 2013, ch. 3, p. 23.
206. N. Kim, M. Zahran and T. Schlick, *Methods Enzymol.*, 2015, **553**, 115.
207. N. Baba, S. Elmetwaly, N. Kim and T. Schlick, *J. Mol. Biol.*, 2016, **428**, 811.
208. J. Frellsen, I. Moltke, M. Thiim, K. V. Mardia, J. Ferkinghoff-Borg and T. Hamelryck, *PLoS Comput. Biol.*, 2009, **5**, e1000406.
209. D. Chakraborty, R. Collepardo-Guevara and D. J. Wales, *J. Am. Chem. Soc.*, 2014, **136**, 18052.
210. S. Bottaro, A. Gil-Ley and G. Bussi, *Nucleic Acids Res.*, 2016, **44**, 5883.
211. D. M. Hinckley, J. P. Lequieu and J. J. de Pablo, *J. Chem. Phys.*, 2014, **141**, 035102.
212. J. J. de Pablo, *Annu. Rev. Phys. Chem.*, 2011, **62**, 555.
213. Y. He, M. Maciejczyk, S. Ołdziej, H. A. Scheraga and A. Liwo, *Phys. Rev. Lett.*, 2013, **110**, 098101.
214. P. Šulc, F. Romano, T. E. Ouldridge, J. P. K. Doye and A. A. Louis, *J. Chem. Phys.*, 2014, **140**, 235102.
215. A. M. Mustoe, H. M. Al-Hashimi and C. L. Brooks 3rd, *J. Phys. Chem. B*, 2014, **118**, 2615.
216. C. A. Davey, D. F. Sargent, K. Luger, A. W. Maeder and T. J. Richmond, *J. Mol. Biol.*, 2002, **319**, 1097.
217. J. C. Hansen, J. Ausio, V. H. Stanik and K. E. Van Holde, *Biochemistry*, 1989, **28**, 9129.
218. A. Luque, G. Ozer and T. Schlick, *Biophys. J.*, 2016, **110**, 2309.
219. E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragozcy and A. Telling, *et al.*, *Science*, 2009, **326**, 289.
220. F. Song, P. Chen, D. Sun, M. Wang, L. Dong and D. Liang, *et al.*, *Science*, 2014, **344**, 376.
221. J. Rouquette, C. Cremer, T. Cremer and S. Fakan, *Int. Rev. Cell Mol. Biol.*, 2010, **282**, 1.
222. M. A. Ricci, C. Manzo, M. F. García-Parajo, M. Lakadamyali and M. P. Cosma, *Cell*, 2015, **160**, 1145.
223. R. Stehr, N. Kepper, K. Rippe and G. Wedemann, *Biophys. J.*, 2008, **95**, 3677.
224. R. Stehr, R. Schöpflin, R. Ettig, N. Kepper, K. Rippe and G. Wedemann, *Biophys. J.*, 2010, **98**, 1028.
225. G. Arya, Q. Zhang and T. Schlick, *Biophys. J.*, 2006, **91**, 133.
226. D. Jost, P. Carrivain, G. Cavalli and C. Vaillant, *Nucleic Acids Res.*, 2014, **42**, 9553.
227. M. Barbieri, M. Chotalia, J. Fraser, L.-M. Lavitas, J. Dostie and A. Pombo, *et al.*, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, 16173.
228. N. Naumova, M. Imakaev, G. Fudenberg, Y. Zhan, B. R. Lajoie and L. A. Mirny, *et al.*, *Science*, 2013, **342**, 948.

# *Uncovering GPCR and G Protein Function by Protein Structure Network Analysis*

FRANCESCA FANELLI<sup>\*a,b</sup> AND ANGELO FELLINE<sup>a</sup>

<sup>a</sup> Department of Life Sciences, University of Modena and Reggio Emilia, Italy; <sup>b</sup> Center for Neuroscience and Neurotechnology, University of Modena and Reggio Emilia, Italy

\*Email: fanelli@unimo.it

## **8.1 Introduction**

Graph theory is a branch of discrete mathematics aimed at studying graphs, mathematical structures used to model pairwise relationships between objects. A complex system can, indeed, be described as a network (graph) made of a set of interacting elements (*e.g.* nodes), which are bound together by links (*e.g.* contacts).<sup>1</sup> Hyper-linked nodes, named hubs, are considered fundamental in maintaining the network. Recently, concepts and methods borrowed from graph theory have been used in structural biology. Indeed, the representation of biomolecular structures as networks of interacting amino acids/nucleotides is ever-increasingly employed to investigate and elucidate complex phenomena such as protein folding and unfolding, protein stability, the role of structurally and functionally important residues, protein–protein and protein–DNA interactions as well as intra-protein and inter-protein communication and allostery.<sup>1–38</sup> These studies rely on methods that differ in the set of graph construction rules. The graph-based

approach proposed by Vishveshwara *et al.*<sup>19</sup> and defined as protein structure network (PSN) has been recently implemented in the Wordom software.<sup>39</sup> It basically computes network features (*e.g.* nodes, hubs (*i.e.* hyper-connected nodes), links *etc.*) and shortest communication pathways from molecular dynamics (MD) trajectories (herein indicated as PSN-MD). The employment of MD trajectories instead of a single structure serves to provide a dynamic description of the network as links break and form with atomic fluctuations. A strategy has recently been developed to infer a dynamic structure network even when dealing with a single structure rather than a trajectory. In this case, a system's dynamics is inferred from the coarse grained elastic network model paired with normal mode analysis (ENM-NMA).<sup>40</sup> The approach is herein defined as PSN-ENM.

G protein-coupled receptors' (GPCRs) function rests on intramolecular and intermolecular structural communication.<sup>41</sup> Indeed, GPCRs regulate most aspects of cell activity by transmitting extracellular signals inside the cell.<sup>42,43</sup> GPCRs share an up-down bundle of seven transmembrane helices (H1–H7) connected by three intracellular (I1, I2, and I3) and three extracellular loops (E1, E2, and E3), an extracellular N-term and an intracellular C-term (Figure 8.1). Upon activation by extracellular signals, the receptors activate the  $\alpha$ -subunit in heterotrimeric guanine nucleotide binding proteins (G proteins) by catalysing the exchange of bound GDP for GTP, *i.e.* they act as guanine nucleotide exchange factors (GEFs). Thus, GPCRs are allosteric proteins that transform extracellular signals into promotion of nucleotide exchange in intracellular G proteins. Regulated protein–protein interactions are key features of many aspects of GPCR function and there is increasing evidence that these receptors act as part of multi-component units comprising a variety of signalling and scaffolding molecules.<sup>43,44</sup>

Heterotrimeric  $G\alpha$  proteins activated by GPCRs belong to the superfamily of Ras GTPases, molecular switches that control a variety of biological processes by cycling between OFF and ON states depending on the bound nucleotide.<sup>45–47</sup> The  $G\alpha$  proteins bear an  $\alpha$ -helical domain (H) in addition to the Ras-like domain shared by all members of the Ras superfamily. The G proteins deprived of the  $\alpha$ -helical domain are therefore named as small G proteins. The Ras-like or GTPase (G) domain represents the basic functional unit of G proteins being deputed to guanosine di-phosphate (GDP)/guanosine tri-phosphate (GTP) binding and the universal switch mechanism.<sup>47</sup> The switch-on process requires the release of the bound GDP and the subsequent binding of the GTP, an intrinsically slow process catalysed by Guanine nucleotide Exchange Factors (GEFs).<sup>45</sup> In the inactive GDP-bound state, the  $G\alpha$  proteins form membrane-associated  $\alpha\beta\gamma$  heterotrimers, forming one of nature's most important miniature (nano-) machines.<sup>48–50</sup>

The Ras-like domain holds a Rossmann fold, characterised by a three-layer ( $\alpha\beta\alpha$ ) sandwich architecture (Figure 8.2). The nucleotide docks into a binding site contributed by the  $\beta 1/\alpha 1$ ,  $\alpha 1/\beta 2$  ( $\alpha F/\beta 2$  in the  $G\alpha$  proteins),  $\beta 3/\alpha 2$ ,  $\beta 5/\alpha 4$ , and  $\beta 6/\alpha 5$  loops (*i.e.* G boxes 1–5 (G1–G5), Figure 8.2). Sequence

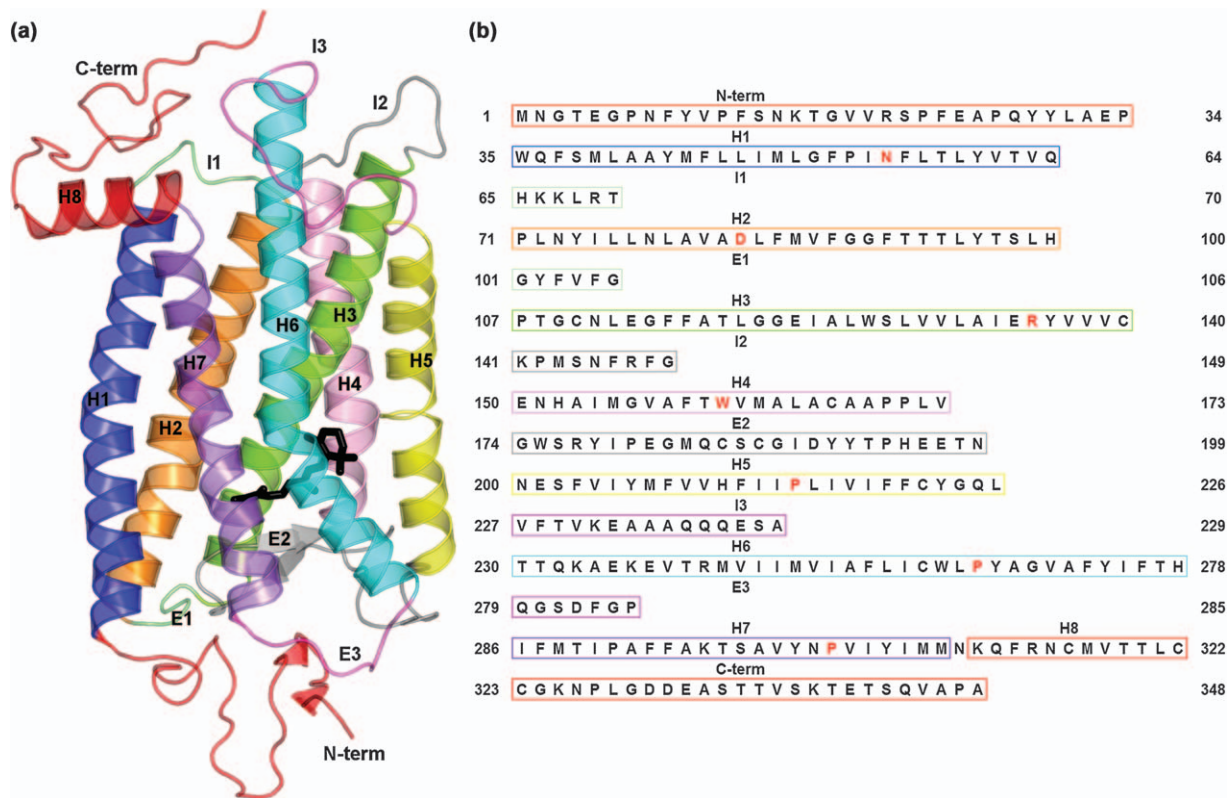


**Figure 8.1** (a) Cartoon representation of the crystal structure of dark bovine rhodopsin (PDB code: 1U19) bound to 11-*cis* retinal. The cytosolic side is at the top. Helices 1, 2, 3, 4, 5, 6, and 7 are blue, orange, green, pink, yellow, aquamarine and violet, respectively. H8, the N- and C-terms are red, E1 and I1 are light green, E2 and I2 are grey, and E3 and I3 are magenta. The 11-*cis* retinal is represented in blue sticks. (b) The primary sequence of bovine rhodopsin is shown. The red label in each helix indicates the amino acid used as a reference in the position-linked numbering by Ballesteros and Weinstein.<sup>62</sup> In such numbering, the first digit indicates the helix whereas the last two digits indicate the position relative to a reference amino acid in the helix numbered as 50.



conservation over the Ras superfamily resides just in the G boxes, also defined as ultra-conserved regions. G1, originally termed as the Walker A motif (motif GxxxxGKS/T), is also called P-loop (phosphate-binding loop). The latter, in fact, contacts the phosphates through main-chain NH groups and lysine side chain. G2 is also called switch I (swI or  $\alpha 1/\beta 2$  loop ( $\alpha F/\beta 2$  loop in the G $\alpha$  proteins)) and G3 is part of the switch II (swII made by the  $\beta 3/\alpha 2$  loop plus the  $\alpha 2$ -helix).

This chapter reports on examples of PSN-MD and PSN-ENM analyses to model the structural communication in GPCRs and G proteins.

## 8.2 Experimental

### 8.2.1 Materials

The freely available Wordom software<sup>39</sup> was employed to perform both PSN-MD and PSN-ENM analyses. Post-processing of the Wordom output relies on a number of in house made programs and scripts. The latter provide numerical output as well as files for 3D output visualisation by the pymol software ([www.pymol.org](http://www.pymol.org)).

The PSN-ENM method has been also implemented in a web server with a friendly graphical user interface.<sup>38</sup> The server allows the user to easily setup the calculation, perform post-processing analyses, and both visualise and download numerical and 3D visualisation outputs.

The GBSW implicit water/membrane model implemented in the CHARMM molecular simulation software<sup>51</sup> was used for GPCR simulations. The GROMACS4 simulation package<sup>52</sup> with the AMBER03 all atoms force field<sup>53,54</sup> was used for equilibrium MD simulations in explicit water of G proteins.

### 8.2.2 Methods

#### 8.2.2.1 Workflow of the PSN-MD and PSN-ENM Approaches

The first step in PSN analysis consists in computing the Protein Structure Graph (PSG), *i.e.* an ensemble of nodes and links.<sup>55</sup> PSG can be computed either on an MD trajectory (PSN-MD approach) or on a single high resolution structural model (PSN-ENM approach). This step provides the basis to search for the shortest paths between pairs of nodes, *i.e.* linked nodes connecting two extremities. In this framework, once the two extremities of interest have been specified, the algorithm first defines all possible shortest communication paths between such extremities and then it filters the results according to the cross-correlation of atomic motions derived from MD trajectories or ENM-NMA. Outcome of this stage is the total pool of paths for the system under study. Paths can be either subjected to clusterisation and selection of the cluster center or to computation of metapaths, coarse representation of a path ensemble (see below).

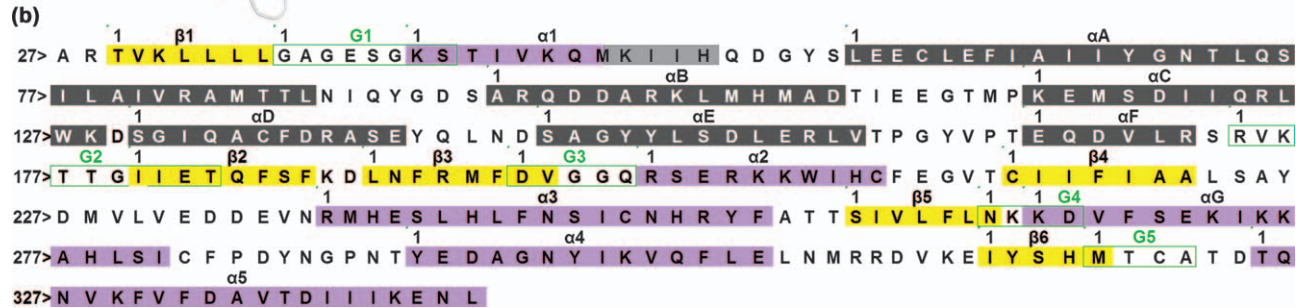


**Figure 8.2** (a) Cartoon representation of the G $\alpha$  protein transducin. The  $\alpha$ -helical domain is grey whereas the Ras-like domain is coloured according to secondary structure (*i.e.* helices, strands, and loops are violet, yellow, and white, respectively). The G boxes are green. The GTP nucleotide is coloured by atom type and represented by sticks. (b) The primary sequence of G $\alpha$  is shown. Helices, strands, and loops are coloured according to the cartoon colours. The G boxes are delimited by green boxes. Black numbers on the left side of the alignment refer to the sequential numbering, whereas black numbers above the sequences indicate the beginning of a secondary structure/G-box motif. An arbitrary numbering of each residue was set, characterised by the label of the secondary structure segment followed by the amino acid position within the segment. In those cases where the G-boxes overlap with the secondary structure segment, positions refer to the G-boxes.

### 8.2.2.2 Building the PSG

Building of the PSG is carried out by means of the PSN module implemented in the Wordom software.<sup>39</sup> PSN analysis is a product of graph theory applied to protein structures.<sup>56</sup> A graph is defined by a set of vertices (nodes) and connections (edges) between them. In a PSG, each amino acid residue is represented as a node and these nodes are connected by edges based on the strength of non-covalent interactions between residues.<sup>19</sup> The strength of interaction between residues  $i$  and  $j$  ( $I_{ij}$ ) is evaluated as a percentage given by the following equation:

$$I_{ij} = \frac{n_{ij}}{\sqrt{N_i N_j}} \times 100 \quad (8.1)$$

where  $I_{ij}$  is the percentage interaction between residues  $i$  and  $j$ ;  $n_{ij}$  is the number of atom–atom pairs between the side chains of residues  $i$  and  $j$  within a distance cut-off (4.5 Å);  $N_i$  and  $N_j$  are normalisation factors for residue types  $i$  and  $j$ , which account for the differences in size of the amino acid side chains and their propensity to make the maximum number of contacts with other amino acids in protein structures. The normalisation factors for the 20 amino acids were derived from the work by Kannan and Vishveshwara.<sup>57</sup> The normalisation index for retinal, 170.13, was computed as the average number of contacts done by the molecule in a dataset of 83 crystallographic structures concerning the different photointermediate states of bacteriorhodopsin, bovine rhodopsin, sensory rhodopsin and squid rhodopsin. The normalisation value for GDP, 220.19, was derived from a database of 55 G proteins whereas that for GTP, 274.78, was derived from a database of 69 G proteins. The normalisation factors for  $\text{Mg}^{2+}$  were 14.65 in the GDP-bound form (based on 41 structures) and 22.01 in the GTP-bound form (based on 68 structures). Finally, the normalisation factor for water, 27, was computed on the crystal structures of rhodopsin and four G $\alpha$  proteins (PDB codes: 1GZM, 1CIP, 1CUL, 1TAG, and 1TND).

Thus,  $I_{ij}$  is calculated for all nodes, excluding  $i \pm n$ , where  $n$  is a given neighbour cut-off, 2 in the case studies reported here. An interaction strength cut-off  $I_{\min}$  is then chosen and any residue pair  $ij$  for which  $I_{ij} \geq I_{\min}$  is considered to be interacting and hence connected in the PSG. Node inter-connectivity is finally used to highlight cluster-forming nodes, where a cluster is a set of connected amino acids in a graph. Node clustering procedure is such that nodes are iteratively assigned to a cluster if they could establish a link with at least one node in such cluster. A node not linkable to existing clusters initiates a novel cluster and so on until the node list is exhausted. Cluster size, defined as the number of nodes, varies as a function of the  $I_{\min}$ , and the size of the largest cluster is used to calculate the  $I_{\text{critic}}$  value. The latter is defined as the  $I_{\min}$  at which the size of the largest cluster is half the size of the largest cluster at  $I_{\min} = 0.0\%$ . At  $I_{\min} = I_{\text{critic}}$  weak node interactions are discarded, emphasizing the effects of stronger interactions

on PSN properties. Generally, the  $I_{\min}$  is set equal to the  $I_{\text{critic}}$  approximated to the second decimal place. Therefore, it is possible to obtain different PSGs for the same protein structure depending on the selected  $I_{\min}$ , and, consequently,  $I_{\min}$  can be varied to obtain graphs with strong or weak interactions forming the edges between the residues. The residues making zero edges are termed as orphans and those that make four or more edges are referred to as hubs at that particular  $I_{\min}$ . The four-link cut-off for hub definition relates to the intrinsic limit in the possible number of non-covalent connections made by an amino acid in protein structures due to steric constraints and it is close to its upper limit. The majority of amino acid hubs indeed make from four to six links.

Stable links can be used to highlight network communities, which are sets of highly interconnected nodes such that nodes belonging to the same community are densely linked to each other and poorly connected to nodes outside the community. Such communities can be considered as fairly independent compartments of a graph. Communities were built by identifying all the  $k=3$ -cliques, *i.e.* sets of three fully interconnected nodes and then merging all those cliques sharing at least one node.

With the PSN-ENM method all edges at the selected  $I_{\min}$  are considered in the PSG, whereas with the PSN-MD method only edges occurring in a given fraction of the trajectory frames, *i.e.* link frequency, enter in the PSG.<sup>21,24,25,37,38,40,58,59</sup> Different states of a molecular system, *e.g.* free or bound, wild type (WT) or mutated, inactive or active, monomeric or oligomeric, *etc.* can be compared in terms of PSGs. PSG differences can be either plotted in histograms or mapped onto the 3D structure.<sup>40,55</sup>

### 8.2.2.3 Search for the Shortest Communication Pathways

The search for the shortest path(s) between pairs of nodes as implemented in the PSN-path module of Wordom relies on the Dijkstra's algorithm.<sup>60</sup> Paths are searched by combining PSN data with cross-correlation of atomic motions calculated by using the linear mutual information (LMI) method, for PSN-MD, or by the covariance matrix inferred from ENM-NMA, for the PSN-ENM method.<sup>40</sup> When dealing with GPCRs, pathways are worth searching between all possible residues in the intracellular and extracellular portions<sup>24,25</sup> or between all residue pairs in the protein except those at sequence distance  $\pm 5$ . The latter setup is employed for the analyses on the G proteins.

Following calculation of the PSG and of correlated motions, the procedure to search for the shortest path(s) consists of (a) searching for the shortest path(s) between each selected amino acid pair based upon the PSN connectivities, and (b) selecting the shortest path(s) that contains at least one residue correlated (*e.g.* with a correlation coefficient  $\geq 0.8$ ) with either one of the two extremities. With the PSN-MD approach, all the shortest paths that pass the filter of correlation of motions are subjected to a further filter based upon path frequency, *i.e.* number of frames containing the selected path divided by the total number of frames in the trajectory.

Collectively, the main differences between PSN-MD and PSN-ENM in terms of path search include the way in which the cross-correlations of atomic motions for path filtering are computed and the application of a frequency-based extra-filter. In detail, whereas with PSN-MD the cross-correlations of atomic motions are computed on the trajectory frames by means of the LMI method, with PSN-ENM they are extracted from the covariance matrix of the deformation modes computed by ENM-NMA. As for the path filtering issue, PSN-MD re-filters those paths that pass the motion correlation filter by finally keeping only those that exceed a recurrence cut-off (*i.e.* presence in a given number of trajectory frames); in contrast, PSN-ENM applies only the motion correlation filter. Thus, whereas with PSN-MD recurrence of network parameters (*i.e.* links and paths) in the trajectory frames dictates the composition of both PSG and path pool, recurrence-based filtering does not apply to PSN-ENM. In spite of these significant differences, the two different approaches tend to produce overlapping outcomes concerning those nodes and links that recur the most in the predicted pathways.<sup>40</sup> The PSN-ENM approach tends to predict a more extended communication, likely due to the less heavy filtering applied to the communication pathways.

Thus, the paths that pass the filtering stage(s) constitute the pool of paths of a system at given  $I_{\min}$  and correlation coefficient cut-offs. The statistical analysis of such pool of paths can lead to the building of global metapaths constituted by the most recurrent nodes and links in the pool. Metapaths made of the most recurrent nodes and links in global pool of paths (global metapaths) or in a cluster of paths (cluster metapath) indeed allow inference of a coarse/global picture of the structural communication in the considered system. Cluster analysis may provide finer information on the predicted pathways. Two path clusterisation methods have been implemented differing both in the clusterisation algorithms and in the score employed to evaluate the similarity between path pairs (*i.e.* similarity score).<sup>40,58</sup> Irrespective of the clusterisation method, for each cluster, following a pairwise comparison of all cluster members, the center is computed, which is the path with the highest average similarity among all the paths in the cluster. The center can be employed as a representative of a given cluster.<sup>40</sup> Path clusters can be analysed through cluster metapaths, cluster centres, as well as computational descriptors of path features. Computational indices describing path features can be used as well to choose representative paths. These indices include as an example the mean square distance fluctuations (MSDF) either computed between the extreme nodes (this case study) or averaged over all node pairs in a path.<sup>40</sup>

## 8.3 Results and Discussion

### 8.3.1 Modelling Allosteric Communication in GPCRs

GPCRs are allosteric proteins whose functioning fundamentals is the communication between the two poles of the helix bundle, *i.e.* the extracellular

side receives and transfers extracellular signals to the intracellular side deputised to recognise and activate intracellular proteins, primarily the G protein transducers, to which these receptors owe their name. Activating or misfolding mutations, functionally different ligands (*e.g.* agonists, biased agonists, inverse agonists, antagonists, allosteric modulators), as well as different homo/heterooligomeric states, are likely to exert differential impacts on such communication.

In the last two decades or so, Fanelli's group paid a lot of effort in setting computational strategies to infer the mechanisms of intramolecular and intermolecular communication in a number of GPCRs of the rhodopsin family in functionally different states.<sup>61</sup> The extensive investigations done so far led to inference that the main effects of activating signals, either from mutations or activating ligands, include perturbations in the interaction pattern of the E/DRY arginine and increases in solvent exposure of selected amino acids in the neighbourhoods of the highly conserved E/DRY receptor motif.<sup>61</sup> The latter effect marks the opening of a solvent accessible crevice between H3 and H6. These effects were suggested to require the integrity of conserved amino acids in H2 and H7 (*e.g.* D2.50, N7.49, and Y7.53 of the NPxxY motif; here the numbering scheme by Ballesteros and Weinstein is used<sup>62</sup>). It was also predicted that such cytosolic crevice would form a receptor docking site for the C-term of the G protein  $\alpha$ -subunit.<sup>61</sup> All together, these predictions found validation in recent advances from structure determinations, showing that the arginine of the E/DRY motif can act both as a structural hallmark of receptor functionality and as a recognition point for the G protein C-term.<sup>63–65</sup>

### 8.3.1.1 PSN-MD Analysis of GPCRs to Infer the Effects of Mutation and Ligand-Protein Binding

The representation of GPCR structures as networks of interacting amino acids can be a meaningful way to decipher the impact of mutation, ligand binding, and/or formation of multi-protein complexes on the structural communication of the protein. The PSN-MD method was employed to infer: (a) the structural effects of rhodopsin mutations linked to autosomal dominant retinitis pigmentosa (adRP);<sup>21,66</sup> (b) the effects of highly conserved amino acids in the structural communication of the luteinising hormone receptor (LHR), both in its inactive and mutation-induced active states;<sup>24</sup> and (c) the effects of ligand binding and dimerisation on the structural communication of the A<sub>2A</sub>-adenosine receptor (A<sub>2A</sub>R).<sup>25</sup>

As for the structural determinants of adRP linked to rhodopsin mutations, RP comprises a group of hereditary human diseases that are the most common cause of inherited blindness in developed countries.<sup>67,68</sup> Almost 89% of the biochemically characterised adRP RHO mutants are likely to be misfolded and mislocalised<sup>69–71</sup> following within the misfolding disease model amenable to treatment with pharmacological



chaperones, small ligands able to correct defects in structure and sub-cellular localisation of the mutated protein. The basic idea of the study was that understanding the structural determinants of the disease implies deciphering the effects of misfolding mutations on the molecular forces related to the folding and stability of the protein. Such issue was, therefore, faced by comparing *in silico* mechanical unfolding of WT rhodopsin with that of selected adRP-linked mutants.<sup>21</sup> The latter share (more or less) marked misfolding effects that, in some cases, may result in impaired binding of the retinal ligand, ER retention, and destabilisation of the dark and/or the signalling states.<sup>70</sup> The variegated distribution of selected mutation sites in almost all seven-helices and the common misfolding effects drove the investigation towards the search for possible intramolecular communications between mutation sites and selected portions related to rhodopsin stability and function. This was done by searching possible effects of misfolding mutations on the native network of intramolecular connections that characterise the folded native state and progressively break during the unfolding process. In this framework, steered MD simulations were instrumental in simulating the unfolding process, whereas PSN analysis was used to infer the native intramolecular connections progressively broken during rhodopsin unfolding, and to investigate the effect of pathogenic mutations on such connection loss.<sup>21</sup> In particular, PSN analysis served to infer the effects of mutations on the native fingerprint of the structurally stable amino acids, *i.e.* the highly connected nodes in the structure network that oppose resistance to connectivity loss in response to an external force. Possible relationships between the structural effects of mutations and the severity of the biochemical defect caused by mutation were found.<sup>21</sup> Thus, the analysis focused on mutational effects on the native stable hub frequency.

The study showed that native stable hubs essentially group in the two poles of the helix bundle and along the main axes of H3 and H6, thus suggesting that they play a role both in protein stability and in signal transfer between extracellular and intracellular sides (Figure 8.3(a)).<sup>21</sup>

Hubs involve highly conserved amino acids and sites of spontaneous misfolding mutation, thus indicating the structural/functional relevance of such nodes. The high concentration of hubs in the retinal binding site is consistent with computational studies highlighting this receptor portion as a part of the stability core and a hinge site in the dynamics of the protein.<sup>21,72,73</sup> Irrespective of their location, misfolding mutations tend to reduce the frequency (compared to wild type) of selected native stable hubs in the retinal binding site (Figure 8.3(a)). The extent of this structural effect turned out to be related to the extent of the biochemical defect associated with the mutation.<sup>21</sup> In a recent study the application of the graph theory to MD trajectories from thermal unfolding simulations was combined with *in vitro* immunofluorescence analysis to infer the effects of 33 adRP rhodopsin mutations on stability and transport of the protein in the absence and presence of the retinal ligand.<sup>66</sup> The discovery of a correlation





between impairment of native structure network and mislocalisation allowed inference of common structural defects held by the mutants in relation to their ER retention and to identify key structural elements of protein responsiveness to small chaperones like the retinal. This represents a step forward towards the discovery of new compounds able to alleviate misfolding and harmful ER retention. The study generated a new classification of adRP mutants based on the multiscale integration of atomic and sub-cellular levels of detail.<sup>66</sup>

As for LHR, PSN-MD analyses allowed the identification of key amino acids that are part of the regulatory network responsible for propagating communication between the extracellular and intracellular poles of the receptor.<sup>24</sup> The number of hubs and link-involving hubs in the WT was found to be higher compared to the two constitutively active mutants (CAMs), consistent with the demonstrated lower stability of the active GPCR states compared to the inactive ones. The study emphasised the role of highly conserved amino acids both in protein stability and in intra-protein allosteric communication as they behaved as stable hubs both in the inactive and active states. Collectively, structural and mutational analyses suggest that the main reason why selected highly conserved amino acids in H2 and H7 are essential for both hormone- and mutation-induced activation of the LHR is that they participate in the structural stability of the protein as they are hubs in both the inactive and active states. Moreover, they behave as the most frequent nodes in the communication paths between the extracellular and intracellular sides in the functionally different states with emphasis on the active ones. In this respect, non-conservative mutations of either D405(2.50), or N619(7.49), or P620(7.50) or Y623(7.53) are expected to impair the most relevant ways of communication between activating mutation sites or the hormone-binding

---

**Figure 8.3** (a) Cartoon representation of the crystal structure of dark bovine rhodopsin (PDB code: 1U19) bound to 11-*cis* retinal. The structure is represented and coloured as explained in the legend to Figure 8.1. The most frequent hubs are shown as spheres centered on the respective  $\alpha$ -atoms. (b) The effects of adRP-linked mutations on the hub frequencies of WT rhodopsin are shown. Colour ranges from green to red and green to blue indicate, respectively, decreasing and increasing frequency values compared to WT. Columns refer to the amino acids that are hubs in more than 20% of the trajectory frames of WT rhodopsin. Hubs include the chromophore, RET. Red labels indicate the amino acids in the extracellular half of the receptor. Rows refer to the molecular systems, thirty-two adRP-linked mutants. Each matrix element was obtained according to the following formula:  $(\text{Mut}_{\text{Freq}} - \text{WT}_{\text{Freq}})/\text{WT}_{\text{Freq}}$ , where  $\text{Mut}_{\text{Freq}}$  and  $\text{WT}_{\text{Freq}}$  are the native hub frequencies of a given mutant and of the WT, respectively. All the relative frequency numbers above 1 and below -1 were changed to 1 and -1, respectively. Remarkably, the PSN analysis and calculation of the hub-frequency matrix shown here is slightly different than the published one.<sup>21</sup> The number of considered mutants differs as well.

domain and G protein recognition regions.<sup>24</sup> Although the two CAMs considered in this study share some common features likely related to their common active state, they show clear differences in the structure network and in the composition of the shortest communication paths, indicative of differences in the information flow between the two poles of the helix bundle.

As for the A<sub>2A</sub>R, the dynamic network of intramolecular interactions characterizing the complex with the ZM241385 antagonist in its monomeric state was compared with that of the same complex in three different dimeric forms, as well as with that of the APO-protomer.<sup>25</sup> The results of the study emphasised the roles of H1 in A<sub>2A</sub>R homo-dimerisation and of highly conserved amino acids in H1, H2, H6, and H7 in maintaining the structure network of the A<sub>2A</sub>R. A<sub>2A</sub>R dimerisation resulted in an effect on the communication networks intrinsic to the receptor fold in a way dependent on the dimer architecture. Certain architectures retained the most recurrent communication paths with respect to the monomeric antagonist-bound form but enhancing path numbers and frequencies, whereas some others impaired ligand-mediated communication. Ligand binding turned out to affect the network as well. Collectively, the study suggested that the communication network that pertains to the functional dynamics of a GPCR is influenced by ligand functionality, oligomeric order, and architecture of the supramolecular assembly.

### 8.3.1.2 PSN-ENM Analysis of GPCRs for Decrypting the Structural Communication Encoded in Crystal Structures

A comparative PSN analysis on all the GPCR crystallographic structures released so far by the PSN-ENM approach has been undertaken (Felline and Fanelli, manuscript in preparation). The goal of the study is to infer functional state-dependent and independent commonalties and differences in the structural communication features at the family and sub-family levels. Results limited to rhodopsin bound to 11-*cis* retinal (in the dark inactive state) and bound to all-*trans* retinal (meta II (MII) active state) have been recently shown.<sup>55,74</sup> It turned out that the structure networks of the dark and MII states of rhodopsin are quite similar in number of nodes, hubs, links, and hub-mediated links; the only difference is that the number of links is slightly higher in the inactive than the active state, consistent with the higher stability of the former (Figure 8.4(a) and (b)).<sup>75</sup> Major differences in the PSG concern the retinal chromophore, which acts as a hub in both forms. However, whereas 11-*cis* retinal is part of a network community, all-*trans* retinal is not.<sup>55,74</sup> Thus, photo-activated retinal is involved in a less dense network compared to the 11-*cis* form. As for hubs, as already found by the PSN-MD approach coupled with mechanical unfolding simulations,<sup>21</sup> rhodopsin hubs crowd in the retinal binding site and involve both highly

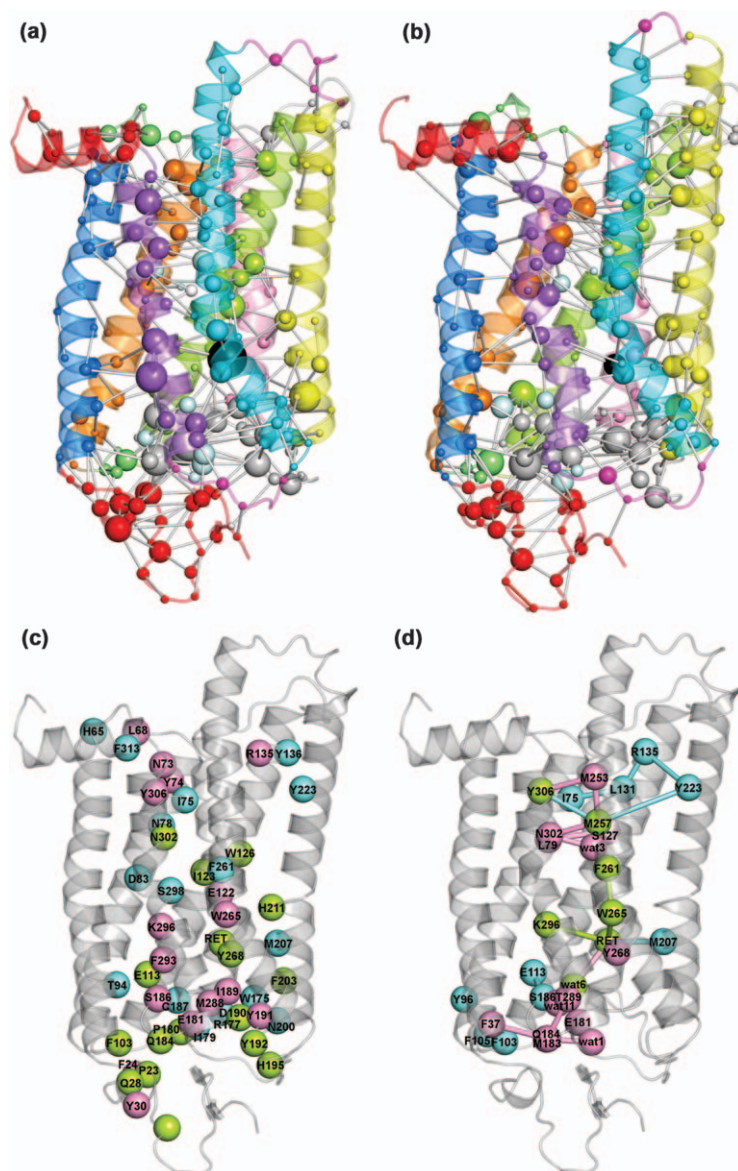


**Figure 8.4** The 3D PSGs concerning dark (PDB code: 1GZM) (a) and MII (PDB code: 3PDX) (b) states are shown. Nodes are coloured according to their location (*i.e.* according to the different receptor regions, see the legend to Figure 8.1). The diameter of the sphere is proportional to the number of links made by the considered node, with the lowest value corresponding to one link. Links are grey. The retinal is represented as a black sphere centered on the C8 atom. (c) Hubs are represented as spheres onto the dark rhodopsin structure. Hubs specific to the dark and MII states are respectively pink and aquamarine, whereas those shared by the two forms are lemon-green. (d) The difference metapath is shown onto the dark rhodopsin structure. Metapath nodes and links specific to the dark and MII states are pink and aquamarine, respectively, whereas those shared by the two different states are lemon-green.

conserved amino acids and adRP mutation sites, the latter being localised essentially in the extracellular half of the protein and in the N-term (Figure 8.4(c)). Remarkably, the number of misfolding mutation sites behaving as hubs is higher in the inactive state than the active one. This is likely linked to 11-*cis* retinal making an extended community of nodes including adRP mutation sites. Another remarkable difference between dark and MII states concerns the hub behavior of highly conserved amino acids in the cytosolic regions. In particular, whereas R135(3.50) of the E/DRY motif, and Y306(7.53) of the NPxxY motif are hubs in the dark state and not in MII, the contrary happens for Y223(5.58) and F312 in H8, likely linked to the structural rearrangements in the cytosolic regions following the photoisomerisation of 11-*cis*-retinal. In general, in the dark state hub specificity essentially locates in the N-term, E2, and H7, whereas in MII it essentially resides in H2, E2, and H5 (Figure 8.4(c)).

As for the shortest communication pathways, rhodopsin's fold is such that, in both functional states, the most likely pathways express a vertical communication between the two poles of the transmembrane helix bundle, in line with the fact that nature made GPCRs competent to transfer signals from the outside to the inside of the cell. Consistent with the denser retinal-involving network in the dark state, the number of pathways that pass through retinal is higher in the dark state compared to MII. On average, paths characterising the inactive state are shorter, stiffer, and tend to involve a major number of hubs than those characterising the MII state (Figure 8.4(d)). Striking different communication modes are expressed by the metapaths of the two functionally different states. In the dark state, the gross of the communication involves E2, retinal, H6, and H7. In contrast, the metapath characterising the MII state describes a vertical communication essentially involving retinal, H6, H5, and the C-terms of H3 and H8. Remarkably, in both states, the W265(6.48)–F261(6.44) aromatic pair in H6 mediates the communication between chromophore binding site and cytosolic regions (Figure 8.4(d)). Thus, metapath comparisons highlight an increase in the contribution of the intracellular regions and a reduction in the contribution of the extracellular regions on going from the dark to the MII state. In the latter, the cytosolic end of the pathways involves Y223(5.58) and the E/DRY arginine. These data suggest that, whereas in the inactive states the structure network serves essentially to ensure structural stability and promote fast signal transfer, in the active form, it is essentially instrumental in maintaining a binding site for the G protein in the cytosolic regions of the receptor.

Collectively, PSN-ENM applied to the dark and MII states of rhodopsin reveals the structure networks as expression of enhanced stability of the inactive state compared to the active one, highlighting also the receptor portions responsible for such stability differences, which strongly involve the retinal-mediated network. Finally, the structural communication appears to hold different functional competences in the inactive and active states.



### 8.3.2 Modelling Allosteric Communication in G Proteins

Approaches like principal component analysis (PCA) and elastic network model-normal mode analysis (ENM-NMA) were recently used to identify the important structural flexibilities that enable proteins in the Ras superfamily to switch between their active and inactive states.<sup>76,77</sup> These analyses, carried out on the crystal structures of representative members led to an hypothesis regarding the evolutionary adaptation of structural deformations by the individual members of the superfamily to fulfil their specialised function. Moreover, atomistic MD simulations were employed to investigate nucleotide effects on the functional dynamics of the G protein<sup>78</sup> and to infer the structural effects of loss-of-function and activating mutations of the G $\alpha$  protein transducin (Gt).<sup>34,79</sup> In this respect, the structural effects of the Nougaret congenital night blindness (NCNB) linked to the G38D loss-of-function mutation of Gt were thoroughly investigated by combining mathematical modelling of visual photo-transduction and atomistic simulations on the major targets of the mutational effect. Mathematical modelling, in line with electrophysiological recordings, indicated reduction of phosphodiesterase 6 (PDE) recognition and activation as the main determinants of the pathological phenotype. Sub-microsecond MD simulations coupled with functional mode analysis improved the resolution of information, showing that such impairment is likely due to disruption of the PDE $\gamma$  binding cavity in transducin. The functional dynamics of wild type and six CAMs of Gt were also investigated by combining atomistic MD simulations with Maxwell–Demod discrete MD (MDdMD) simulations of the receptor-catalysed transition between GDP-bound and nucleotide-free states of the G protein.<sup>79</sup> Compared to the WT, Gt CAMs increase the overall fluctuations of the nucleotide and its binding site. This is accompanied by weakening of native links involving GDP,  $\alpha 1$ , the G boxes,  $\beta 1$ – $\beta 3$ , and  $\alpha 5$ . Collectively, constitutive activation by the considered mutations seems to be associated with weakening of the interfaces between  $\alpha 5$  and the surrounding portions and the interface between G and H domains. These mutational effects are associated with increases in the overall fluctuations of the G and H domains, which reflect on the collective motions of the protein. Gt CAMs, with prominence to G56P, T325A, and F332A prioritise collective motions of the H domain overlapping with the collective motions associated with receptor-catalysed nucleotide release.<sup>79</sup>

MD simulations on a ternary complex between agonist-bound thromboxane A2 receptor (TP) and GDP-bound heterotrimeric Gq were the first attempt in the literature to investigate at the atomic detail GPCR's impact on the G protein dynamics.<sup>80</sup> The study suggested that the formation of a composite receptor-G protein interface, dominated by receptor contacts with the C-term of the  $\alpha$ -subunit, favours concerted motions of selected G protein loops in the nucleotide binding site as well as of the H domain with respect to the G domain, features intrinsic to the G protein fold, but amplified by receptor binding. Such inter-domain uncoupling was found related to increases in

solvent exposure of GDP and hence to nucleotide exit. In spite of the extremely short length of simulations, the study suggested for the first time the displacement of the H domain with respect to the G domain as one of the early events in receptor-catalysed GDP-release. These inferences now find support in the crystallographic complex between the  $\beta$ 2-adrenergic receptor ( $\beta$ 2AR) and nucleotide-free Gs heterotrimer,<sup>65</sup> in the spectroscopic determination of Gi activation by rhodopsin,<sup>81</sup> and in microsecond MD simulations.<sup>82</sup> It was also speculated that the establishment of receptor-G protein contacts is also instrumental in preventing the establishment of intra-subunit or inter-subunit interactions that would occur in receptor-free heterotrimeric G protein. Preventing the formation of such interactions upon receptor binding would contribute to weakening the control exerted by the  $\beta$ -subunit on certain intra-domain and inter-domain motions of the  $\alpha$ -subunit.

### 8.3.2.1 PSN-MD Analysis of Small G Proteins for Investigating the Effects of Mutations and Ligand-Protein Binding

PSN-MD analysis served to investigate the effects of nucleotide on the structural communication in four small G proteins, Arf1, H-Ras, RhoA, and Sec4 in their free, as well as GDP-bound and GTP-bound states.<sup>58</sup>

It was observed that the topology of the Ras-like domain is such that, in the presence of nucleotide, the gross of the network localises in the C-terms of selected  $\beta$ -strands and in the G boxes. As expected, the nucleotide acts as the most linked and persistent hub, by establishing connections with the G boxes. In the APO forms, the absence of nucleotide causes the complete loss of hubs in the G boxes and the reduction of nodes and links compared to the nucleotide-bound states (Figure 8.5(a)–(c)). This would suggest that the nucleotide plays a central role in maintaining the structure network, making selected G box amino acids behave as hubs as well. In contrast, in the APO forms, the absence of nucleotide makes these loops intrinsically dynamic and poorly interconnected. On these bases, it was speculated that the nucleotide-mediated structure network is instrumental in restricting protein dynamics to a few selected regions related to protein function like switches and inter-switch.

In the nucleotide-bound states, the majority of communication pathways involved the nucleotide and was absent in the APO forms (Figure 8.5(d)–(f)). The four considered systems were found to share selected regions deputed to structural communication in both the GDP- and GTP-bound states. These regions include  $\beta$ 4,  $\alpha$ 3, and the G boxes 1 and 4. The G boxes 2, 3, and 5 participated as well in the most relevant paths, though to a lower extent. In almost all systems, the nucleotide was found to participate in  $\geq 75\%$  of the considered paths being involved in structural communication with distal regions such as the effector binding sites (*i.e.* swI and swII) and the N- and C-terminal regions. In the absence of nucleotide, the Rossman fold of Ras GTPases turned out to be significantly deprived of stable communication paths.



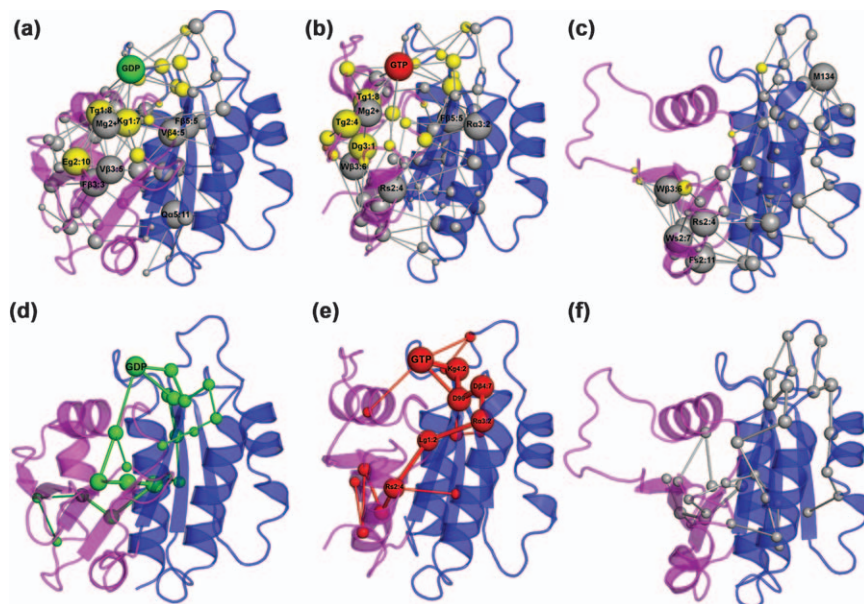


**Figure 8.5** 3D Protein Structure Graphs (PSG) of the Arf1 small G protein in its GDP-bound (a), GTP-bound (b), and APO (c) states are shown. The G protein is shown in cartoons coloured according to the structural subdivision in lobes 1 (magenta) and 2 (blue). Indeed, previous computational experiments suggested that the strands  $\beta 1$  and  $\beta 4$  divide the conserved Ras-like domain into two dynamically distinct lobes, lobe 1 (*i.e.* the N-terminal half of the domain, coloured magenta) and lobe 2 (*i.e.* the C-terminal half of the domain, coloured blue).<sup>78</sup> The spheres centered on the C $\alpha$ -atoms concern those node pairs, which are linked in  $\geq 50\%$  of frames in a 40 ns trajectory. Green and red spheres indicate GDP and GTP, respectively. Labelled nodes behave as hubs in  $\geq 50\%$  of the trajectory frames. The metapaths concerning the GDP-bound (d, green), GTP-bound (e, red), and APO (f, grey) states of Arf1 are shown. The represented nodes, ‘frequent nodes’, are present in  $\geq 10\%$  of the most frequent and non-redundant paths, while the represented links satisfy both conditions of being present in  $\geq 10\%$  of the most frequent and non-redundant paths and of connecting ‘frequent nodes’. The radius of the spheres and the width of links are proportional to the recurrence of nodes and links, respectively. Nodes with frequency  $\geq 50\%$  are labelled.

According to these results, the topology of small GTPases seems to require the presence of the nucleotide to acquire a persistent structural communication network. The bound nucleotide also significantly affects the communication features of small G proteins, where it plays a central and active role in establishing the major structure network and in mediating communication pathways. The structure network indeed undergoes a reshaping depending on the bound nucleotide. In selected families, *i.e.* Arf1 and H-Ras,

this effect is more marked than in others and may be related to the function specialisation-oriented evolutionary flexibility previously inferred from the structural PCA.<sup>76</sup> The central role of the nucleotide in the dynamic structure network of Ras GTPases suggests that such a ligand modifies the intrinsic dynamics of the protein through the establishment of regions of hyperlinked nodes with high occurrence of correlated motions.

PSN-MD analyses were also carried out to infer the structural determinants of the observed slight reduction of the RGS9-catalysed GTPase activity of the G38D NCNB-linked mutant of Gt compared to the WT form. Significant differences between the WT and G38D forms of Gt in complex with GTP and RGS9 could be inferred from the analysis of the shortest communication pathways, which were almost halved in the mutated complex compared to the WT.<sup>34</sup> A clear effect of the Gt mutation was the incapacity to form a stable Q200(G3:5)-mediated communication between the nucleotide and N364 of RGS9. Such communication is instead present in the WT and it is necessary for the action of RGS9, thus suggesting that the observed slight reduction of the RGS9-catalysed GTPase activity is a matter of perturbed inter-protein communication.

## 8.4 Conclusions

In the last decade, important efforts have been devoted to setting and implementing computational approaches relying on the application of graph theory to protein structures to investigate the structural communication in bio-systems. Two PSN-based approaches differing in the way protein dynamics is accounted for can be applied by the Wordom software. The approaches have been extensively used to investigate GPCR and G protein function. It was inferred that conserved amino acids, sites of misfolding mutations, or ligands like the retinal in rhodopsin or the nucleotide in G proteins tend to behave as stable hubs in the structure network. The studies highlighted densely linked regions playing a central role in protein stability and function. They highlighted changes in the communication fingerprints depending on the bound ligand (*i.e.* functional state).

The analyses of structure network coupled to the analyses of sub-cellular localisation could establish a bridge between misfolding and misrouting in rhodopsin mutants linked to inherited blindness. The study identified key structural elements of protein responsiveness to small chaperones with implication for drug discovery. It also generated a new classification of adRP mutants based on the multiscale integration of atomic and sub-cellular levels of detail.

## Acknowledgements

This study was supported by a Telethon-Italy grant [GGP13227] and an Airc-Italy grant [14811] to FF.

## References

1. C. Bode, I. A. Kovacs, M. S. Szalay, R. Palotai, T. Korcsmaros and P. Csermely, *FEBS Lett.*, 2007, **581**, 2776–2782.
2. M. Vendruscolo, E. Paci, C. M. Dobson and M. Karplus, *Nature*, 2001, **409**, 641–645.
3. M. Vendruscolo, N. V. Dokholyan, E. Paci and M. Karplus, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 2002, **65**, 061910.
4. G. Amitai, A. Shemesh, E. Sitbon, M. Shklar, D. Netanel, I. Venger and S. Pietrokovski, *J. Mol. Biol.*, 2004, **344**, 1135–1146.
5. R. Sathyapriya and S. Vishveshwara, *Nucleic Acids Res.*, 2004, **32**, 4109–4118.
6. K. V. Brinda, A. Surolia and S. Vishveshwara, *Biochem. J.*, 2005, **391**, 1–15.
7. K. V. Brinda and S. Vishveshwara, *Biophys. J.*, 2005, **89**, 4159–4170.
8. K. V. Brinda and S. Vishveshwara, *BMC Bioinf.*, 2005, **6**, 296.
9. C. Chennubhotla and I. Bahar, *Mol. Syst. Biol.*, 2006, **2**, 36.
10. A. del Sol, H. Fujihashi, D. Amoros and R. Nussinov, *Mol. Syst. Biol.*, 2006, **2**, 0019.
11. R. Sathyapriya, K. V. Brinda and S. Vishveshwara, *J. Chem. Inf. Model.*, 2006, **46**, 123–129.
12. C. Chennubhotla and I. Bahar, *PLoS Comput. Biol.*, 2007, **3**, 1716–1726.
13. A. Ghosh, K. V. Brinda and S. Vishveshwara, *Biophys. J.*, 2007, **92**, 2523–2535.
14. A. Ghosh and S. Vishveshwara, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 15711–15716.
15. S. Tang, J. C. Liao, A. R. Dunn, R. B. Altman, J. A. Spudich and J. P. Schmidt, *J. Mol. Biol.*, 2007, **373**, 1361–1373.
16. C. Chennubhotla, Z. Yang and I. Bahar, *Mol. Biosyst.*, 2008, **4**, 287–292.
17. A. Ghosh and S. Vishveshwara, *Biochemistry*, 2008, **47**, 11398–11407.
18. R. Sathyapriya, M. S. Vijayabaskar and S. Vishveshwara, *PLoS Comput. Biol.*, 2008, **4**, e1000170.
19. S. Vishveshwara, A. Ghosh and P. Hansia, *Curr. Protein Pept. Sci.*, 2009, **10**, 146–160.
20. M. Bhattacharyya, A. Ghosh, P. Hansia and S. Vishveshwara, *Proteins*, 2010, **78**, 506–517.
21. F. Fanelli and M. Seeber, *FASEB J.*, 2010, **24**, 3196–3209.
22. M. S. Vijayabaskar and S. Vishveshwara, *Biophys. J.*, 2010, **99**, 3704–3715.
23. M. S. Vijayabaskar and S. Vishveshwara, *BMC Bioinf.*, 2010, **11**(Suppl 1), S49.
24. K. Angelova, A. Felling, M. Lee, M. Patel, D. Puett and F. Fanelli, *Cell. Mol. Life Sci.*, 2011, **68**, 1227–1239.
25. F. Fanelli and A. Felling, *Biochim. Biophys. Acta.*, 2011, **1808**, 1256–1266.
26. R. Sathyapriya and S. Vishveshwara, *Proteins*, 2007, **68**, 541–550.
27. M. Bhattacharyya and S. Vishveshwara, *BMC Struct. Biol.*, 2009, **9**, 8.
28. K. V. Brinda and S. Vishveshwara, *Mol. Biosyst.*, 2010, **6**, 391–398.
29. M. Bhattacharyya and S. Vishveshwara, *BMC Struct. Biol.*, 2010, **10**, 27.

30. M. Bhattacharyya and S. Vishveshwara, *Biochemistry*, 2011, **50**, 6225–6236.
31. A. Pandini, A. Fornili, F. Fraternali and J. Kleinjung, *FASEB J.*, 2012, **26**, 868–881.
32. E. Papaleo, K. Lindorff-Larsen and L. De Gioia, *Phys. Chem. Chem. Phys.*, 2012, **14**, 12515–12525.
33. M. Pasi, M. Tiberti, A. Arrigoni and E. Papaleo, *J. Chem. Inf. Model.*, 2012, **52**, 1865–1874.
34. S. Mariani, D. Dell'Orco, A. Felling, F. Raimondi and F. Fanelli, *PLoS Comput. Biol.*, 2013, **9**, e1003207.
35. A. J. Venkatakrishnan, X. Deupi, G. Lebon, C. G. Tate, G. F. Schertler and M. M. Babu, *Nature*, 2013, **494**, 185–194.
36. T. Flock, C. N. Ravarani, D. Sun, A. J. Venkatakrishnan, M. Kayikci, C. G. Tate, D. B. Veprintsev and M. M. Babu, *Nature*, 2015, **524**, 173–179.
37. F. Raimondi, A. Felling and F. Fanelli, *J. Chem. Inf. Model.*, 2015, **55**, 1878–1893.
38. M. Seeber, A. Felling, F. Raimondi, S. Mariani and F. Fanelli, *Bioinformatics*, 2015, **31**, 779–781.
39. M. Seeber, A. Felling, F. Raimondi, S. Muff, R. Friedman, F. Rao, A. Caflisch and F. Fanelli, *J. Comput. Chem.*, 2011, **32**, 1183–1194.
40. F. Raimondi, A. Felling, M. Seeber, S. Mariani and F. Fanelli, *J. Chem. Theory Comput.*, 2013, **9**, 2504–2518.
41. F. Fanelli and P. G. De Benedetti, *Chem. Rev.*, 2005, **105**, 3297–3351.
42. R. J. Lefkowitz, *Nat. Cell Biol.*, 2000, **2**, E133–E136.
43. K. L. Pierce, R. T. Premont and R. J. Lefkowitz, *Nat. Rev. Mol. Cell Biol.*, 2002, **3**, 639–650.
44. A. E. Brady and L. E. Limbird, *Cell. Signalling*, 2002, **14**, 297–309.
45. I. R. Vetter and A. Wittinghofer, *Science*, 2001, **294**, 1299–1304.
46. J. Colicelli, *Sci. STKE*, 2004, **2004**, RE13.
47. A. Wittinghofer and I. R. Vetter, *Annu. Rev. Biochem.*, 2011, **80**, 943–971.
48. A. G. Gilman, *Annu. Rev. Biochem.*, 1987, **56**, 615–649.
49. D. E. Clapham, *Nature*, 1996, **379**, 297–299.
50. W. M. Oldham and H. E. Hamm, *Q. Rev. Biophys.*, 2006, **39**, 117–166.
51. B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus, *J. Comput. Chem.*, 1983, **4**, 187–217.
52. B. Hess, C. Kutzner, D. Van Der Spoel and E. Lindahl, *J. Chem. Theory Comput.*, 2008, **4**, 435–447.
53. 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.
54. E. J. Sorin and V. S. Pande, *Biophys. J.*, 2005, **88**, 2472–2493.
55. F. Fanelli, A. Felling and F. Raimondi, *Methods Cell Biol.*, 2013, **117**, 43–61.
56. S. Vishveshwara, K. V. Brinda and N. Kannan, *J. Theor. Comput. Chem.*, 2002, **1**, 187–211.

57. N. Kannan and S. Vishveshwara, *J. Mol. Biol.*, 1999, **292**, 441–464.
58. F. Raimondi, A. Fellingine, G. Portella, M. Orozco and F. Fanelli, *J. Biomol. Struct. Dyn.*, 2013, **31**, 142–157.
59. S. Mariani, D. dell'Orco, A. Fellingine, F. Raimondi and F. Fanelli, *PLoS Comput. Biol.*, 2013, **9**, e1003207.
60. E. W. Dijkstra, *Numer. Math.*, 1959, **1**, 269–271.
61. F. Fanelli and P. G. De Benedetti, *Chem. Rev.*, 2011, **111**, PR438–PR535.
62. J. A. Ballesteros and H. Weinstein, *Methods Neurosci.*, 1995, **25**, 366–428.
63. J. H. Park, P. Scheerer, K. P. Hofmann, H. W. Choe and O. P. Ernst, *Nature*, 2008, **454**, 183–187.
64. P. Scheerer, J. H. Park, P. W. Hildebrand, Y. J. Kim, N. Krauss, H. W. Choe, K. P. Hofmann and O. P. Ernst, *Nature*, 2008, **455**, 497–502.
65. S. G. Rasmussen, B. T. Devree, Y. Zou, A. C. Kruse, K. Y. Chung, T. S. Kobilka, F. S. Thian, P. S. Chae, E. Pardon, D. Calinski, J. M. Mathiesen, S. T. Shah, J. A. Lyons, M. Caffrey, S. H. Gellman, J. Steyaert, G. Skiniotis, W. I. Weis, R. K. Sunahara and B. K. Kobilka, *Nature*, 2011, **469**, 175–181.
66. J. P. Behnen, A. Fellingine, A. Comitato, M. T. Di Salvo, F. Raimondi, V. Marigo and F. Fanelli, 2017, submitted.
67. A. Kennan, A. Aherne and P. Humphries, *Trends Genet.*, 2005, **21**, 103–110.
68. H. F. Mendes, J. van der Spuy, J. P. Chapple and M. E. Cheetham, *Trends Mol. Med.*, 2005, **11**, 177–185.
69. M. P. Krebs, D. C. Holden, P. Joshi, C. L. Clark, 3rd, A. H. Lee and S. Kaushal, *J. Mol. Biol.*, 2010, **395**, 1063–1078.
70. A. D. Briscoe, C. Gaur and S. Kumar, *Gene*, 2004, **332**, 107–118.
71. E. P. Rakoczy, C. Kiel, R. McKeone, F. Stricher and L. Serrano, *J. Mol. Biol.*, 2011, **405**, 584–606.
72. B. Isin, A. J. Rader, H. K. Dhiman, J. Klein-Seetharaman and I. Bahar, *Proteins*, 2006, **65**, 970–983.
73. O. Tastan, E. Yu, M. Ganapathiraju, A. Aref, A. J. Rader and J. Klein-Seetharaman, *Photochem. Photobiol.*, 2007, **83**, 351–362.
74. F. Fanelli, A. Fellingine, F. Raimondi and M. Seeber, *Biochem. Soc. Trans.*, 2016, **44**, 613–618.
75. S. M. A. Khan, W. Bole, P. A. Hargrave, M. M. Santoro and J. H. McDowell, *Eur. J. Biochem.*, 1991, **200**, 53–59.
76. F. Raimondi, M. Orozco and F. Fanelli, *Structure*, 2010, **18**, 402–414.
77. F. Fanelli and F. Raimondi, *Curr. Pharm. Des.*, 2013, **19**, 4214–4225.
78. F. Raimondi, G. Portella, M. Orozco and F. Fanelli, *PLoS Comput. Biol.*, 2011, **7**, e1001098.
79. A. Fellingine, S. Mariani, F. Raimondi, L. Bellucci and F. Fanelli, *J. Chem. Theory Comput.*, 2017, **13**, 886–899.
80. F. Raimondi, M. Seeber, P. G. De Benedetti and F. Fanelli, *J. Am. Chem. Soc.*, 2008, **130**, 4310–4325.

81. N. Van Eps, A. M. Preininger, N. Alexander, A. I. Kaya, S. Meier, J. Meiler, H. E. Hamm and W. L. Hubbell, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**, 9420–9424.
82. R. O. Dror, T. J. Mildorf, D. Hilger, A. Manglik, D. W. Borhani, D. H. Arlow, A. Philippsen, N. Villanueva, Z. Yang, M. T. Lerch, W. L. Hubbell, B. K. Kobilka, R. K. Sunahara and D. E. Shaw, *Science*, 2015, **348**, 1361–1365.

# ***Current Challenges in the Computational Modelling of Molecular Recognition Processes***

LUCÍA PÉREZ-REGIDOR, JOAN GUZMÁN-CALDENTEY,  
CARLOS F. RODRÍGUEZ, JEAN-MARC BILLOD, JUAN NOGALES  
AND SONSOLES MARTÍN-SANTAMARÍA\*

Center for Biological Research, CIB-CSIC. Ramiro de Maeztu, 9.  
28040-Madrid, Spain

\*Email: smsantamaria@cib.csic.es

## **9.1 Modelling the Dynamics of the Proteins**

Molecular dynamics (MD) simulations represent a possibility to simultaneously study the structure and dynamics of macromolecules at atomistic resolution. Although MD simulations cannot model the underlying physics exactly, their applications can provide a sufficiently close approximation to capture a wide range of protein conformations that can be associated with particular biological effects. Some representative examples are here provided from recently published works in which different types of MD simulation protocols have been used to study protein folding.

Sborgi *et al.* used a combination of nuclear magnetic resonance (NMR) and long time-scale MD simulations to study the single-domain protein gpW.<sup>1</sup> This 62-residue protein folds into an  $\alpha + \beta$  (two  $\alpha$ -helices and one



$\beta$ -hairpin) topology in microseconds, making it a suitable target to be studied by MD simulations. Using Anton, a supercomputer specifically designed for MD simulations, the authors performed a 250  $\mu$ s simulation at 340 K and four independent simulated tempering simulations of 200  $\mu$ s each.<sup>2</sup> In all cases, starting from an extended state, transitions from non-folded to folded state or native state (RMSD below 1 Å from the NMR structure) were observed. Furthermore, the RMSD analysis of the 340 K simulation showed several metastable states, being consistent with the kinetic-clustering analysis showing, on a microsecond time-scale, the interconversion of these metastable states. They also used a Markov model to calculate the folding free energy surface resulting in the identification of two different intermediate states and two misfolded kinetic traps. Combining these data, they were able to elucidate the folding pathway: it starts with the folding of helix 1, forming intermediate 1, and the folding of helix 2, forming intermediate 2. Finally, folding of the  $\beta$ -hairpin led to the formation of the native state. Intermediate 2 is in fast equilibrium with both native state and intermediate 1, which is in agreement with the different structures that were elucidated by NMR experiments. The very large time scale (200  $\mu$ s) needed for this work was made possible by the use of the Anton supercomputer. However, other methods can be considered to achieve similar results in shorter time scales.

For example, Miao *et al.* used accelerated molecular dynamics (*aMD*) simulations to study the dynamics of chignolin,<sup>3</sup> a 10 amino acid fast folding peptide designed by segment statistics that adopts a  $\beta$ -hairpin conformation.<sup>4</sup> Unlike conventional MD, *aMD* allows to capture events happening in the millisecond-timescale in only hundreds of nanoseconds by adding a non-negative boost potential to decrease energy barriers and thus accelerate transitions between different low-energy states, while conserving the shape of the potential (such that minima are maintained as minima and barriers are preserved as barriers). To study chignolin folding, the authors performed a minimisation of the system (*i.e.* extended protein) followed by 50 ns of conventional MD simulation to calculate the acceleration parameters that needed to be modified in the force field equation. Three independent 300 ns *aMD* simulations were run starting from the final structure of the previous simulation. In all three cases, they observed folding of the protein and a decreasing RMSD value in relation to the structure resolved by NMR. Furthermore, the protein became more compact during folding. The simulations led to two low energy conformations, the folded and the unfolded one, the former being the global minimum energy conformational state. The information at atomic detail inferred from these techniques also permitted to identify that the turn formation is led by the C-terminal region and it propagates through the rest of the protein, which is in agreement with previous conventional MD simulations performed at a microsecond-timescale. Therefore, this *aMD* simulation protocol was shown to be an effective tool to study folding events that occur in the microsecond time-scale.

Another interesting application of MD simulation to the study of the protein folding is the replica-exchange molecular dynamics (REMD), initially developed by Sugita and Okamoto in 1999.<sup>5</sup> REMD is a generalized ensemble simulation method applied to sample a wide range of the conformational space. A number of simulations are performed at different temperatures in parallel (replicas). Exchange of the replicas at different temperatures is performed periodically, so a free random walk can be performed in the potential-energy space, and the simulation can escape from local-minimum free-energy states. Since the development of this method, several teams have applied it to study protein dynamics and folding. In a recent example, Jiang *et al.* have been able to fold fourteen small proteins using REMD in combination with a residue-specific force field.<sup>6</sup> After minimisation of all systems, they performed a 2  $\mu$ s REMD simulation for each protein. In all cases, they were able to reproduce the X-ray crystallographic structures with high accuracy and determine the folding pathways for the 14 proteins. Zerze *et al.* also applied the REMD approach to study the folding stability of  $\beta$ -hairpins, including tryptophan zippers that contain two pairs of stabilising tryptophan residues.<sup>7</sup> They were able to fold all the proteins in a relatively short period of time, successfully reproducing the differences found in experimental studies with the mutated peptides.

Adaptive steered molecular dynamics (ASMD) have also been applied to the understanding of folding pathways and energetics. This approach is based on steered MD but it is performed in stages to obtain more efficient convergence.<sup>8</sup> In a recent work by Hernandez *et al.*, the authors applied this method to determine the energetics of two similar  $\beta$ -hairpin peptides, chignolin and trpzip1.<sup>9</sup> These two peptides shared very similar secondary structures and hydrogen bonding profiles, but they presented different stabilities and melting temperature. The problem was approached from a reverse perspective: the ASMD was started from the completely folded peptide which was forced to unfold. Using this strategy, the authors could easily calculate the energetics of the unfolding, *i.e.* the energy that requires the peptide to unfold. It was observed that the calculated energetic cost to stretch the trpzip1 peptide from an initially folded state to an unstructured state was higher than that for the chignolin. Also, the atomistic resolution of ASMD allowed the identification of pairwise interactions present in trpzip1 that confer more stability to this peptide.

The study of the dynamics, folding and conformational ensembles is of particular relevance when studying the structure and function of intrinsically disordered proteins (IDPs).<sup>10</sup> All-atom and coarse-grained MD simulations have proven to be particularly relevant for studying these proteins, and for complementing the experiments. Advances in algorithm development and computational potency are also helping in the elucidation of the connection between the intrinsic disorder and the biological properties. Out of the scope of this chapter, but wanting to mention the rapid advance in this field, we here refer to some recent publications.<sup>11–14</sup>

## 9.2 Three-dimensional Structure Prediction and Homology Modelling

Another important and challenging issue when studying molecular recognition processes is the prediction of the 3D structure of the proteins involved when the experimental 3D structure is not available. Enormous efforts have been devoted to developing user-friendly servers, and secondary and tertiary structure prediction platforms have risen in the recent years to make structure prediction more accessible to a wide range of investigators, with the additional attractiveness of being usually freely available. All these servers are based on the Anfinsen's dogma relating the structure of a given protein, and thus its function, to its amino acid sequence, and on the principle that the native structure of the protein is the one that has the lowest energy.<sup>15</sup> In particular, homology modelling methods use a homologous protein of known structure as a template to determine the structure of the problem protein. The principal steps in a homology modelling protocol are template selection, alignment, backbone and side chain prediction and structure optimisation. The target protein is aligned to the template sequence and the backbone atoms are mapped onto the three-dimensional template structure. The non-conserved side chains orientations are predicted. Finally, a force field is used to optimise the structure, which basically includes the removal of steric clashes and the optimisation of hydrogen bond network.<sup>16</sup>

Here we briefly present some examples from a wide variety of servers for protein 3D structure prediction. Rosetta<sup>17</sup> is a fragment-based method that combines native-like structures of unrelated proteins to search the most probable 3D structure of a given amino acid sequence. The scoring function includes biological information, like solvation and residue-pair interactions, and the 3D structures are generated by combining fragments of known structures based on a large number of sequences with known structure present in the protein data bank (PDB). Another example is the I-TASSER server,<sup>18</sup> an online resource for automated protein structure prediction that uses the information stored in the PDB to perform the alignment of related sequences. Then, the protein of study is divided into aligned and unaligned regions, these regions are connected, and a Monte-Carlo (MC) simulation is performed for the assembling and refinement. I-TASSER combines different methods to add backbone atoms, side chains, and to refine the final structure. SWISS-MODEL<sup>19</sup> is also a popular homology modelling web server that presents three different modelling protocols, depending on the degree of similarity between the input sequence and the template, allowing to predict structures with a very low percentage of similarity. Another homology modelling software that uses comparative methods is MODELLER.<sup>20,21</sup>

The combination of 2D/3D structure prediction servers and MD simulations has also proved to yield successful results. As a representative example, it is here presented the interesting work performed by Reichen *et al.* for the design of Armadillo (ARM)-repeat proteins using both Rosetta

software and MD simulations to modulate their peptide recognition properties.<sup>22</sup> ARM-repeat proteins are involved in cell signalling, regulation, and protein trafficking. They are characterised by containing a repeating ~42 amino acid motif composed of three  $\alpha$ -helices. Tandem ARM-repeat units fold together as a superhelix, forming a platform for interactions with many protein partners. The binding selectivity depends on the different curvatures of the loops more than on the protein sequence, so the 3D protein structure acquires here more relevance and, consequently, the designed ARM-repeat proteins should present a uniform curvature through all the structure. With this aim, Rosetta software was implemented to find an optimal amino acid sequence to achieve the desired curvature using an ARM-repeat protein as a template. The newly designed protein was expressed in *Escherichia coli*, but the crystallisation failed due to the presence of high flexibility regions. To stabilise these regions, multiple MD simulations were carried out to identify the particular amino acids that could induce high flexibility. Site mutations were performed and the resulting protein was successfully crystallised, presenting high affinity and selectivity for the desired peptide.

### 9.3 Modelling of Protein–Protein Interactions

Protein–protein interactions (PPIs) are on the basis of molecular recognition processes. The availability of the 3D structure of many protein complexes, and their biophysical and structural investigations, for example through X-ray crystallography analysis, NMR and electron microscopy (EM), together with the assessment of binding affinity, are providing insights on PPIs at the atomic level and helping to elucidate the interplay of structures.<sup>23,24</sup> This approach can be useful to understand the principles of the biological pathways and, consequently, to facilitate new roads toward drug development for the treatment of many diseases.<sup>25</sup> Computational techniques have complemented the missing information from experimental techniques by studying protein surfaces, site flexibility, and protein–protein complementarity (docking).<sup>26</sup> Currently, the modulation of protein–protein interactions is one of the most promising but also challenging approaches in chemical biology and drug discovery.<sup>27</sup> The total number of PPIs in an organism, the so-called ‘interactome’, has been predicted to be enormous in humans.<sup>28–30</sup> Consequently, finding successful strategies to modulate the PPIs with small drug-like molecules would expand the opportunities for pharmacological intervention.<sup>31</sup> In a review by Ottmann *et al.*,<sup>32</sup> a number of biochemical techniques for the identification of hits as PPI modulators are reported, together with the chemical methods for accessing designed small molecules as modulators of PPIs. Among them, the following are highlighted as the most promising ones: peptide binding epitopes as lead structures, oligomeric structures such as foldamers, and natural products as a constant source of inspiration. Also, biological and chemical approaches to screen compound libraries (including experimental and virtual protocols) are pointed out as current strategies for the identification of active PPI

modulators. Interestingly, this review also brings to light the enormous potential of PPI stabilisation as a drug discovery strategy, a concept which remains unexplored so far.<sup>32</sup>

The identification of all the interactions inside a proteome is a key milestone to greatly deepen the understanding of the functioning of a cell. Although experimental methods have considerably contributed to reach this aim, nevertheless these methods have their limitations and computational prediction of protein interactions allows the validation and completion of the missing interactions. As a result, vast amounts of experimental and computational PPI data have been accumulated in diverse sets of databases. An integrative combination of both experimental and computational approaches may help to build the proteome-scale interactome. The majority of these computational approaches are simulation-based and statistical/machine-learning-based methods. The former methods, that include MD simulations and docking, aim to model the forces governing interactions of proteins, usually at atomistic level, and to compute the strength of the interactions. The latter, statistical and machine-learning based methods, can be used at large scale to infer PPIs from the information of known interacting proteins.<sup>33</sup> The information from the large and diverse data (sequence, structure, gene expression, binding affinity, *etc.*) is extracted and classified in order to derive a predictive model that can be applied to a new pair of proteins.<sup>33–35</sup> At a large scale, other computational predictions of pairwise PPIs and their analysis can also include interolog mapping, gene/domain fusion-based methods, learning-based prediction using sequence information, domain co-occurrence, and gene co-expression. These prediction methods are reviewed in the literature with excellent discussions on their performance and availability.<sup>36</sup>

The available information about PPIs is catalogued in various publicly/freely available databases. These databases are primarily elaborated by collecting data from individual scientific publications. Protein–protein interactions are determined with a large array of experimental methods, which produce results presented in a non-homogenous nomenclature, rendering the classification process difficult. In order to homogenise the nomenclature and create a proteomics standard for PPI, the International Molecular Exchange consortium (IMEx)<sup>37</sup> was formed. The main databases are BioGRID,<sup>38</sup> MINT,<sup>39</sup> BIND,<sup>40</sup> DIP,<sup>41</sup> IntAct,<sup>42</sup> and HPRD.<sup>43</sup> Some of these databases have a pairwise overlap reaching up to 75%, thus it is necessary to combine data from several datasets to collect a broad set of interactions.<sup>44</sup>

## 9.4 Prediction of Protein–Protein Interactions: Docking

For the understanding of protein function, prediction of the 3D structure of protein–protein complexes is a crucial and critical step, also with implications for structure-based drug design. A number of protein–protein docking programs are available,<sup>45</sup> and several of them are freely available

web servers. Some leading servers include ClusPro,<sup>46–49</sup> HADDOCK,<sup>45,50</sup> ZDOCK,<sup>45,51</sup> and PatchDock.<sup>52–54</sup> Some algorithms (fast Fourier transform approaches, FFT) treat the proteins as rigid bodies, which permits an effective sample of the intermolecular space of the complex. Some other approaches (that make use of MC search, MD simulation or energy minimisation), treat the proteins as flexible but limit the intermolecular space search only to the areas of the protein surface that may be interacting patches. These should be predicted by bioinformatics methods or should be derived from experimental data.<sup>55</sup> It has been developed the idea that the limitations of the two sampling approximations would be overcome by the combined use of docking methods belonging to these two different philosophies.<sup>56</sup> Effective combination of independently developed docking protocols has also indicated that using diverse search and scoring functions can improve protein–protein docking results.<sup>57</sup>

Here we mention some of the most popular programs for protein–protein docking. ClusPro<sup>58</sup> currently offers three different scoring schemes, the first for enzyme–inhibitor complexes, the second for antigen–antibody pairs, and the third for other, primarily signal transduction complexes. An algorithm to predict molecular assemblies of homo-oligomers has also been implemented. This approach to protein–protein docking includes three main steps. First, it runs PIPER, a rigid body docking program based on the FFT correlation approach, extended to use pairwise interactions potentials. Second, the 1000 best energy conformations are clustered, and the 30 largest clusters are retained for refinement. Third, the stability of the clusters is analysed by short MC simulations, and the structures are refined by the medium-range optimisation method SDU (semidefinite programming-based underestimation). The first two steps of this approach are implemented in the ClusPro 2.0 protein–protein docking server.<sup>46</sup> Despite being fully automated, the last step is computationally too expensive to be included in the server. ClusPro-DC adds dimer classification (DC) as a new option to the protein–protein docking server ClusPro,<sup>59</sup> leading to the discrimination between crystallographic and biological dimers by docking the two subunits to exhaustively sample the interaction energy landscape.

High ambiguity driven biomolecular docking (HADDOCK)<sup>60</sup> uses an energy minimisation approach. It makes use of biochemical and/or biophysical interaction data such as chemical shift perturbation data resulting from NMR titration experiments, mutagenesis data or bioinformatic predictions. This information is introduced as ambiguous interaction restraints (AIRs) to drive the docking process. An AIR is defined as an ambiguous distance between all residues shown to be involved in the interaction. Active residues correspond to residues experimentally identified or predicted to be at the interface. Passive residues are surface neighbours of active residues. AIRs are defined for each active residue with the effective distance being calculated from the sum of all individual distances between any atom of an active residue and any atom of all active and passive residues on the partner molecule. HADDOCK has been extended to deal with a large variety of data



and complexes: protein–protein docking, protein–nucleic acid, protein–DNA, protein–RNA, protein–oligosaccharides and protein–ligand complexes with high accuracy.<sup>61</sup>

ZDOCK performs a full rigid-body search of docking orientations between two proteins.<sup>45,62</sup> The current version includes performance optimisation and a novel pairwise statistical energy potential. M-ZDOCK is a modification of ZDOCK to predict symmetric assemblies using the structure of a subunit, and ZRANK is a docking refinement program developed to provide fast and accurate rescoring of models from initial-stage docking (*e.g.* from ZDOCK), as well as refined docking models (*e.g.* from RosettaDock). There is also an older docking refinement protocol that utilises structural minimisation and rescoring of initial-stage docking models.

The PatchDock method performs structure prediction of protein–protein and protein–small molecule complexes. PatchDock is a geometry-based molecular docking algorithm inspired by object recognition and image segmentation techniques used in computer vision. It is aimed at finding docking transformations that yield good molecular shape complementarity. Such transformations, when applied, induce both wide interface areas and small amounts of steric clashes. A wide interface is ensured to include several matched local features of the docked molecules that have complementary characteristics. The PatchDock algorithm divides the Connolly dot surface representation of the molecules into concave, convex and flat patches.<sup>63</sup> Then, complementary patches are matched in order to generate candidate transformations. Each candidate transformation is further evaluated by a scoring function that considers both geometric fit and atomic desolvation energy.<sup>64</sup> Finally, an RMSD clustering is applied to the candidate complexes to discard redundant solutions. The main reason behind PatchDock's high efficiency is its fast-transformational search, which is driven by local feature matching rather than brute force searching of the six-dimensional transformation space. It further speeds up the computational processing time by utilizing advanced data structures and spatial pattern detection techniques, such as geometric hashing and pose clustering that were originally developed in the field of computer vision. The run time of PatchDock for two input proteins of average size (about 300 amino acids) is <10 min on a single 1.0 GHz PC processor under the Linux operating system.<sup>65</sup>

An interesting integrative application of these protocols for 3D structure prediction and protein–protein docking is represented by the CASP/CAPRI experiment.<sup>45</sup> The CASP/CAPRI experiment is a project started in 2014 as a junction between critical assessment of protein structure prediction (CASP) and critical assessment of predicted interactions (CAPRI) communities with the aim of bringing together experts from the structure prediction and protein–protein docking fields. The project consisted on the elucidation of secondary, tertiary and quaternary structure of 25 target proteins, comprising homodimers, homotetramers and heterodimers. 23 CAPRI and 7 CASP groups uploaded docking and homology models for each target, and



12 CAPRI groups performed the scoring experiment, being in total more than 9500 submitted models. The results showed that targets with large subunit interfaces were more suitable to be predicted by homology modelling and docking techniques than the ones with ambiguous interfaces. The analysis also showed that docking protocols perform better than homology modelling algorithms when trying to elucidate the structure of protein complexes.

## 9.5 Computational Studies of Complex Protein Systems

Complex molecular systems or sophisticated molecular mechanisms can be entirely studied by computational techniques or these techniques can be used as a complement to experimental studies. Here, we briefly mention the computational studies of some challenging systems carefully chosen from a vast literature.

The ribosome is the protein synthetic machinery of a cell.<sup>66,67</sup> Its role is to translate the genetic information stored in the messenger ribonucleic acid to synthesise the corresponding protein. During the translation process, the transfer RNA (tRNA), a molecule typically composed of 76 to 90 nucleotides, carries an amino acid and brings it to the ribosome which, in turn, will add it to the growing peptide chain to form a protein. MD simulations techniques permit to unveil functional details at atomic level of this complex molecular machine, helping to understand the functioning. Since the publication of the X-ray crystal structure of the complete *Thermus thermophilus* 70S ribosome (PDB-ID<sup>68</sup> 4V42, superseded and combined from entries 1GIX and 1GIY) by Yusupov *et al.*,<sup>69</sup> experts in the field of computational science displayed a growing interest in using computational approaches to study the ribosome. Interestingly, in 2002, Karplus and McCammon commented: “Another complex that is waiting to be studied by molecular and normal mode dynamics is the ribosome, whose structure was determined recently. With the pace of advance in both computer hardware and algorithms, simulation of such large systems for the time required to obtain meaningful results is expected to be possible in the very near future”.<sup>70</sup> In fact, in 2005, the developers of the NAMD software acknowledged the advances in scaling MD simulations to study a system as big and complex as the ribosome: “Striking progress has been achieved in characterizing the immense machines of the cell, such as the ribosome, at the atomic level”.<sup>71</sup> In 2004, Tung and Sanbonmatsu<sup>72</sup> reported a model of the ribosome constructed by assembling experimentally resolved structures of several constituents of the ribosome. They later used a refined version of this model to perform simulations of the movement of tRNA into the ribosome during decoding.<sup>73</sup> Another interesting study was the investigation of the interaction between tRNA and the L1 stalk, a mobile structure implicated in directing tRNA movement during translocation through the ribosome.<sup>74</sup> Combination of data from cryo-EM, X-ray crystallography, and

MD simulations led to interesting insights into the L1 stalk dynamics, and the interactions with tRNA, with potential application for designing new labelling schemes to probe different conformational transitions of the ribosome. Finally, a simulation of the entire ribosome revealed the conformational changes involving the movement of the L1 stalk. These conformations were correlated with tRNA and intersubunit rotation states during translocation: open stalk (vacant E-site, non-rotated ribosome), half-closed stalk (E/E tRNA, post-translocation non-rotated ribosome), and fully closed stalk (P/E tRNA, pre-translocation rotated ribosome). The MD simulations supported the hypothesis that L1 stalk opening facilitates release the E-site tRNA from the ribosome (Figure 9.1).

Enhanced sampling methods, such as coarse grained (CG)<sup>75</sup> using elastic network models (ENMs) and normal mode analysis (NMA) studies,<sup>76</sup> have also been applied to the ribosome.<sup>77</sup> These techniques are of great help when applied to big systems, since they permit overcoming of the limitations in size and in times of the standard MD all-atoms simulation methods.

In an interesting perspective article, Šponer *et al.* reviewed the history of MD simulations of nucleic acids in explicit solvent from a tetranucleotide system to the ribosome, explaining the advantages and the limitations of the simulation techniques.<sup>78</sup> The authors enlightened the fast growing potential of such simulations and future challenges, arguing that aiming to reach accurate simulations, thus computationally costly, is vain when using poorly defined starting structures. However, they stated that atomistic simulations provide unique insights that can deepen the understanding gained from CG and experimental studies.<sup>78</sup>

The CRISPR/Cas9 system (CRISPR, clustered regularly interspaced short palindromic repeats; Cas, CRISPR-associated system) is another interesting molecular machinery to be studied by computational techniques. The first time that the use of CRISPR/Cas9 was suggested for genome editing applications can be traced back to a paper published in *Science* in an article titled “A Programmable Dual-RNA Guided DNA Endonuclease in Adaptive Bacterial Immunity”.<sup>79</sup> It is only very recently that the CRISPR-Cas9 system was first extensively studied by MD simulations. Palermo *et al.* recently published a work reporting a multi-microsecond time scale MD simulations of the Cas9 system to study the conformational plasticity of Cas9.<sup>80</sup> Extensive MD simulations have permitted to reveal the critical role played by non-target DNA to initiate major conformational changes necessary for the nucleic acid association. More specifically, the authors unveiled that the non-target DNA greatly interacts with the L2 loop, suggesting that more efficient Cas9 could be developed by engineering this loop. Overall the authors reported an atomic-level model of the dynamics of the CRISPR-Cas9 system that can serve as a reference for future studies.

Another interesting example is illustrated by the computational studies performed on the immunity Toll-like receptors (TLRs). Research around TLRs is currently experiencing a lively activity given that they are promising therapeutic targets for the development of novel drugs for the treatment of

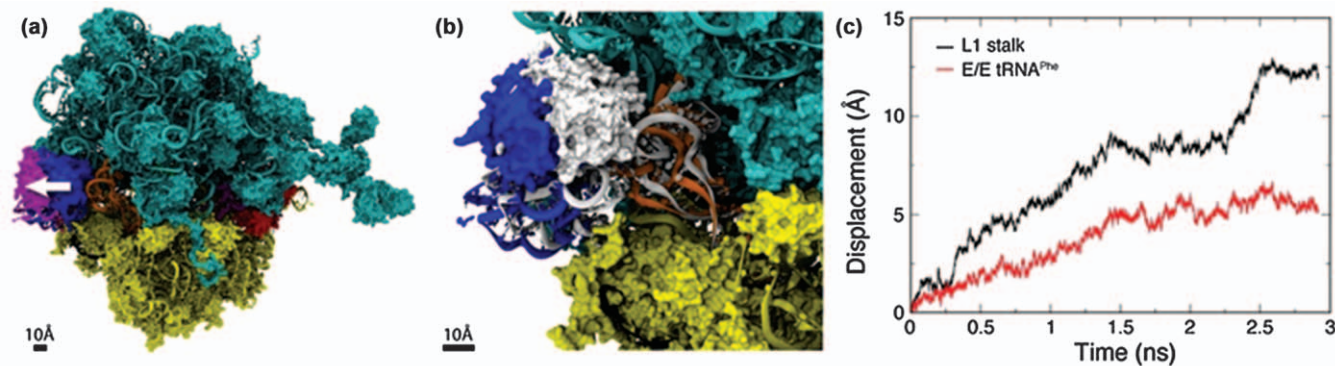


**Figure 9.1** MD simulation of the L1 stalk opening. (a) 3D model of the ribosome corresponding to the release from the ribosome of the E tRNA. The initial half-closed L1 stalk is shown in blue; the open conformation is shown in magenta. The L1 stalk opening is illustrated by the white arrow. (b) Detail of the 3D model showing the initial structure of half-closed E/E tRNA and L1 stalk in white, and the final (open) structure in blue and the same colours as in (a). The E/E tRNA follows the movement of the L1 stalk. (c) Displacement of the centres of mass of the L1 stalk and the E/E tRNA elbow during the L1 stalk opening simulation. Reprinted from *Journal of Molecular Biology*, Volume 402, Schulten *et al.*, The Role of L1 Stalk-tRNA Interaction in the Ribosome Elongation Cycle, 741–760, Copyright 2010, with permission from Elsevier.<sup>74</sup>

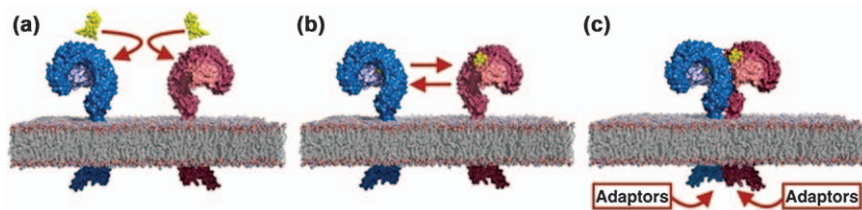


**Figure 9.2** Representation of the LPS-induced dimerisation of the TLR4/MD-2 complex leading to immune system activation. Red arrows indicate motion and mutual recognition. (a) Two LPS molecules (yellow) are engaged by two distinct TLR4/MD-2 extracellular domains (TLR4 in blue and red, MD-2 in pale blue and pale red). (b) Two TLR4/MD-2/LPS complexes dimerise by protein-protein interactions at the dimerisation interface. (c) Dimerisation brings together the two intracellular TLR4 domains providing a suitable molecular surface for recruiting downstream adaptors.

Reprinted with permission from Martín-Santamaría *et al.*<sup>81</sup> Published in *Molecules* (<https://creativecommons.org/licenses/by/4.0/>).

sepsis, inflammatory diseases, and rheumatoid arthritis, and novel adjuvant therapies in cancer treatments. In particular, Toll-like receptor 4 (TLR4), together with its accessory protein myeloid differentiation factor 2 (MD-2), builds a heterodimeric complex able to specifically recognise lipopolysaccharides (LPS), present on the cell wall of Gram-negative bacteria. Through the dimerisation of the ectodomains, the intracellular domains are brought together, allowing the recruitment of the downstream adaptors leading to the final activation of the innate immune response (Figure 9.2).

The relatively recent elucidation of the X-ray crystallographic structure of the TLRs has open new opportunities for the study of the ligand recognition process and the mechanism of functioning from a computational perspective. These advances have been recently reviewed by several authors and range from extensive MD simulations<sup>81,82</sup> to homology modelling<sup>83</sup> and virtual screening strategies for the identification of novel chemical scaffolds avoiding toxicity and physical-chemical problems associated to glycolipid-based TLR4 modulators.<sup>84</sup>

## 9.6 Computational Modelling of Nanostructures

Nanostructures (nanoparticles, nanotubes, nanodevices) are used in biological systems to study biological processes and biomolecules interactions, and are nowadays in the mind of scientists from different fields as a very promising area to explore. Originally under the scope of nanostructure scientists, the interesting chemical and physical properties of those materials have attracted the attention of chemists and biologists. Some examples in the potential application of nanomaterials in chemical and biological sciences include: graphene as a thin molecular support; fullerenes proposed as drug carriers against cancer or bacterial infection; nanoparticles like gold

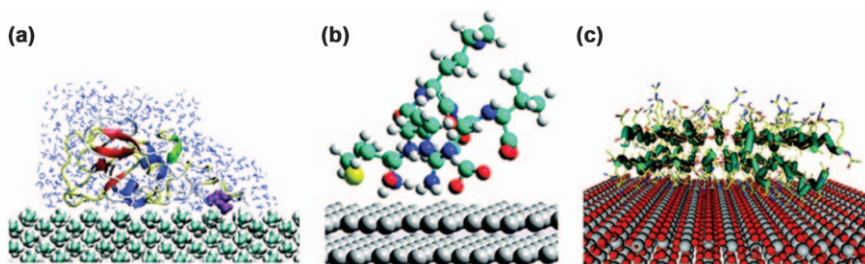


**Figure 9.3** Representation of different proteins interacting with different types of surfaces from computational simulations. (a) Lysozyme on a polyethylene surface with some solvating water molecules. Reprinted with permission from ref. 94. Copyright 2011 American Chemical Society. (b) The MRKDV peptide interacting on a bare silver surface. Reprinted with permission from ref. 91. Copyright 2011 American Chemical Society. (c) RAD16II  $\beta$ -sheet filaments onto a rutile surface. Reprinted with permission from ref. 93. Copyright 2007 American Chemical Society.

nanomaterials lacking toxic effects and nanowires allowing miniaturisation of biosensors and eventually nanocomputers.<sup>85–89</sup> This field experiences a sparkling activity and is also benefited from the application of computational tools to unveil molecular recognition events with atomic detail.

In an interesting review, Wade *et al.* report the computational studies of different types of inorganic surfaces and their interactions with proteins.<sup>90</sup> The understanding of the protein-surface binding mechanisms is of paramount importance for the design of novel materials with many applications in biomaterial sciences, nanomedicine and nanobiotechnology. Some examples of the different types of surfaces that can be modelled include noble metals, such as gold and silver,<sup>91</sup> widely used as biosensors and biomaterials;<sup>92</sup> oxide and mineral surfaces interacting with proteins or peptides;<sup>93</sup> self-assembled monolayers or polymers surfaces,<sup>91</sup> and different carbon allotrope surfaces (Figure 9.3).<sup>90</sup>

### 9.6.1 Modelling of Gold Nanoparticles

Gold nanoparticles (AuNPs) are important participants in the development of new methodologies in biomedicine: drug delivery, magnetic therapies, biosensing, and imaging.<sup>95,96</sup> The interaction between proteins and nanoparticles has been a hot topic in nanoscience. Computational studies are important to understand the physicochemical interactions between these materials and biological molecules, aiming to rationally design AuNPs applications.<sup>97,98</sup> Here, we comment some of the most recent successful computational studies involving gold nanoparticles.

Wade *et al.* have reported atomistic simulations on protein-gold-surface interactions.<sup>98</sup> Since the specific mechanisms for proteins recognising gold

nanoparticles are poorly understood, ubiquitin was chosen as target, given that it is a well characterised small protein present in all eukaryote cells, and free of cysteine (to avoid the binding of the gold atom to the thiol group). Moreover, experimental data from dynamic light scattering and NMR were available. Therefore, it was a good model to study the molecular basis of the protein/AuNP interactions. Considering the binding of naked AuNPs (with no citrate surfactants) to ubiquitin, Brownian dynamics (BD) simulations were carried out, in which the internal structure of the protein was kept rigid. Measuring and classifying the interaction energies using ProMetCS (protein-metal continuum solvent model),<sup>99</sup> five clusters of interactions were identified. The interactions were described by three main terms: van der Waals energy, adsorbate-metal electrostatic interaction energy (dominated by charged side chains) and desolvation energy of both protein and metal surface. Five different orientations of ubiquitin attached to gold surface described over 98% of all the complexes obtained in the BD docking. In three of them, binding was mostly driven by the van der Waals energy, whereas in the other two complexes the binding was mainly due to electrostatic interactions between protein and gold, with arginine and specially lysine residues in contact to the gold surface. These five resulting protein/AuNP complexes were submitted to a protocol of combined annealing and MD simulations, using the GoIP force-field and the GROMACS software.<sup>100</sup> After the simulations, the five starting binding poses finally converged to three stable complexes, thus providing a model for the more likely interactions between the protein and the gold surface. Furthermore, since citrate is one of the most common surfactants for gold nanoparticles during the synthesis, it may also be involved in protein binding. For this case, citrate-AuNP complex was used as docking target, revealing that citrate was mostly affecting the relative population between previously described ubiquitin-AuNP complexes. In fact, DFT calculations indicated that the direct interaction of H and N backbone atoms with the AuNP and the citrate are the origin of the large changes in the NMR chemical shifts of such atoms.

### 9.6.2 Modelling of Nanowires

Computational methods can also be used to describe the formation of interesting biological polymers for nanotechnology, the DNA-protein nanowires being an interesting example. They are composed of double strand DNA and an *in silico* optimised peptide containing the homodimerisation domain and the DNA-binding domain. In one recent example,<sup>101</sup> the scaffold chosen for homodimerisation was the *Drosophila* engrailed homeodomain (ENH), a transcription factor whose X-ray crystallographic structure is available (PDB-ID 1ENH). Using fast Fourier transform symmetrical docking for dimerisation<sup>102</sup> the authors selected one homodimer for further optimisation.<sup>102</sup> To improve the binding energy of the dimer, the authors used a home-built protein design software, allowing mutations of the interfacial residues of the selected homodimer. The Rosetta



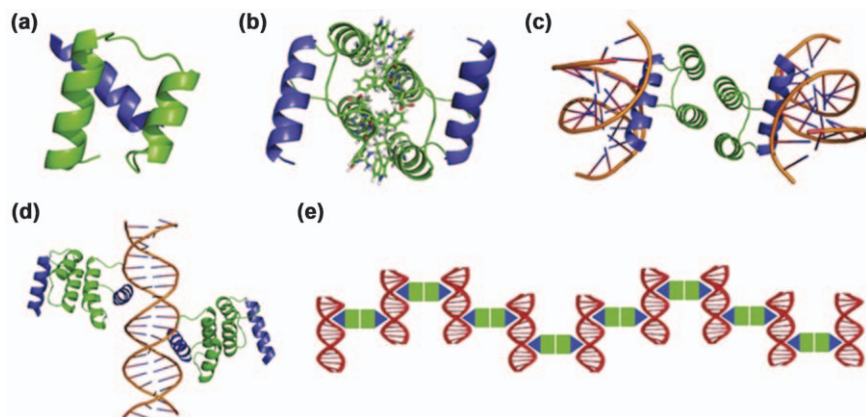


**Figure 9.4** Computationally designed nanowire. (a) Helix 1 and helix 2 (both in green) of ENH as the target for the homodimerisation domain; helix 3 (in blue) is the DNA binding domain. (b) Docked model of dual-ENG for engineering a stable dimerisation interface. (c) Model of the designed dual-ENG binding two dsDNA fragments. (d) Assemble of two dual-ENG onto a dsDNA fragment. (e) Schematic representation of the designed nanowire.

Reprinted by permission from Macmillan Publishers Ltd: *Nature* (ref. 101), copyright 2015.

force field was used for energy evaluation,<sup>103</sup> and FASTER was used for rotamer optimisation.<sup>104</sup> The optimisation of the dimer ended up with 17 rational mutations predicted *in silico* in the dimerisation interface of this new dual-ENG system, keeping unchanged the third loop, responsible of the DNA binding. The newly designed dual-ENG was synthesised and characterised using atomic forced microscopy, fluorescence microscopy and X-ray crystallography. It was shown that this new system was able to bind two dsDNA molecules on the sides, and that the assembly of the dual-ENG/DNA nanowire was driven by non-covalent interactions (Figure 9.4). Therefore, self-assembly of this nanomaterial can be controlled by changing the reaction conditions, such as the molar ratio, the temperature, the ionic strength, and the pH. This inspiring work lays the foundation for the development of new classes of protein–DNA hybrid materials.

### 9.6.3 Modelling of Nanotubes

Carbon nanotubes have the potential use for sensitive detection of chemical and biological molecules. Specifically, single-wall carbon nanotube field effect transistors (SWNT-FETs) have been described as low selective but high sensitive sensors. Therefore, the development of ligands and coats to increase the sensitivity is leading to interesting research in the nanotechnology field.<sup>105</sup>

The use of biomolecules, such as DNA and peptides, as highly specific ligands in combination with SWNT-FETs opens multiple possibilities in the



nanosensor technology. However, it is necessary to understand how biomolecules interact with these devices, since the loss of activity or disrupting structural changes of proteins and peptides, when binding to carbon nanotubes, has been described.<sup>106,107</sup> In order to understand the events in SWNT-biomolecule interaction, computational approaches may be useful to predict and analyse the properties required for a proper binding at the nanoscale. As a representative example, the case of the chimeric peptide P1ASP1C is here presented.<sup>108</sup> P1ASP1C is a peptide recognition element present in a SWNT-biomolecule device designed trinitrotoluene (TNT) detection, composed by ASP1 (bee odour binding protein) binding peptide for chemical agents, linked to an SWNT-binding domain from a phage display library. The 3D coordinates of P1ASP1C in solution were predicted using Rosetta,<sup>103</sup> and REMD simulations, performed with AMBER,<sup>109</sup> were used to explore the possible peptide conformations at different temperatures ranging from 270 K to 700 K. Moreover, to validate the equilibrated conformation, they set up MD simulations starting from independent initial conformations such as fully extended  $\beta$ -strand and  $\alpha$ -helical structures. In all the cases, a final structure with hairpin conformation was obtained with hydrophobic residues from N-terminal and C-terminal facing each other as double  $\beta$ -sheet, and having a kink in the middle due to the presence of large chain amino acids. P1ASP1C was placed in the SWNT system in two different conformations: one perpendicular disposition as a SWNT 'wrapper', and another disposition parallel to the axis of SWNT. After MD simulations, an equilibrated conformation was obtained for each parallel and perpendicular conformation of the peptide-SWNT system.

In order to validate the performance of the computationally designed SWNT-P1ASP1C nanosensor, the docking of different compounds (2,4,6-trinitrotoluene, TNT; cyclotrimethylenetrinitramine, RDX; and 2-heptanone, HPT) into the equilibrated structures was performed with AutoDock.<sup>110</sup> The studies revealed the specificity of this SWNT-P1ASP1C for TNT, by a difference in the predicted binding energy of approximately 9 kcal mol<sup>-1</sup> compared to a bare SWNT. Additionally, as negative control the binding energy of RDX and HPT to the carbon nanotube was not affected by the presence of P1ASP1C in the system.

#### 9.6.4 Modelling of Nanomachines

In the field of molecular machines and molecular switches, Leigh's group has extensively explored the use of rotaxanes to design artificial nanomachines, such as the sequence-specific peptide synthesis machine.<sup>111</sup> Understanding the motion of rotaxanes in redox reactions, light absorption and other conditions are key in the design of new nanodevices. Fumaric and maleic amides are small molecules widely known as photo-controllable rotaxanes *via cis-trans* isomerisation that can be studied using very accurate high-level *ab initio* methods. The isomerisation energy landscape of those compounds was calculated using quantum mechanics

calculations on dimethyl-fumaramide and tetramethyl-fumaramide.<sup>112</sup> These computational investigations have provided a rational for fumaramide photochemistry.

## 9.7 Models of Signalling Networks

Any phenotype from a given biological system (network) arises from the complex interaction of its simple components. For practical reasons, such components are often categorised based on their general functional characteristics as belonging to metabolic, regulatory or to signalling networks. While metabolic networks account for those catabolic and anabolic chemical transformations sustaining life, regulatory networks focus on tuning the genetic program outputs. Signalling networks are the biological layer responsible to compute and connect environmental perturbations with genotype making proper biological decisions.<sup>113</sup> The highly conserved structure of such networks, *e.g.* inputs (sensors), integration (transducers), and outputs (actuators), facilitates cross talking, feedforward and feedback control resulting in optimal decision making (Figure 9.5(a)).<sup>114</sup> Therefore, the understanding of biological systems under perturbations requires the full understanding of signalling networks.

Inherent properties from signalling networks, in contrast with metabolic networks, such as the high crosslinking behavior and the lack of conservation along the life-tree, make that several challenges remains in reconstructing and analysing signalling networks remains. Despite the large contribution of *omic* technologies to the identification and characterisation of biological parts, it is still extremely hard describing the full array of biological components from any signalling network, and even harder deciphering their emergent properties. In spite of that, it is evident that not only has been a substantial progress in modelling signalling network during last years but also in their applications. For instance, they are now being used to identify new drug targets and drug action mechanisms.<sup>116,117</sup>

In the simplest description of a signalling pathway the most upstream component interacts with an external environmental input transferring this information to an effector which is able to elicit a biological response (Figure 9.5(a)). Along with this simplified system the relative concentration of a large array of signalling components, including receptors, transducers, enzymes, and diffusible second messengers determine the network properties.<sup>118</sup> Such complexity, even in the simplest scenario, makes it difficult to understand the organisation and intuitively predict, in advance, the performance of signalling networks. Therefore, analysis of signalling networks requires a combination of experimental and theoretical approaches including the development and analysis of quantitative models which have become essential to understand cellular signalling and information processing. Two main types of computational approaches have become in the current gold standard for modelling signalling networks: (i) network models, and (ii) dynamical models. In addition, constraints-based approach is

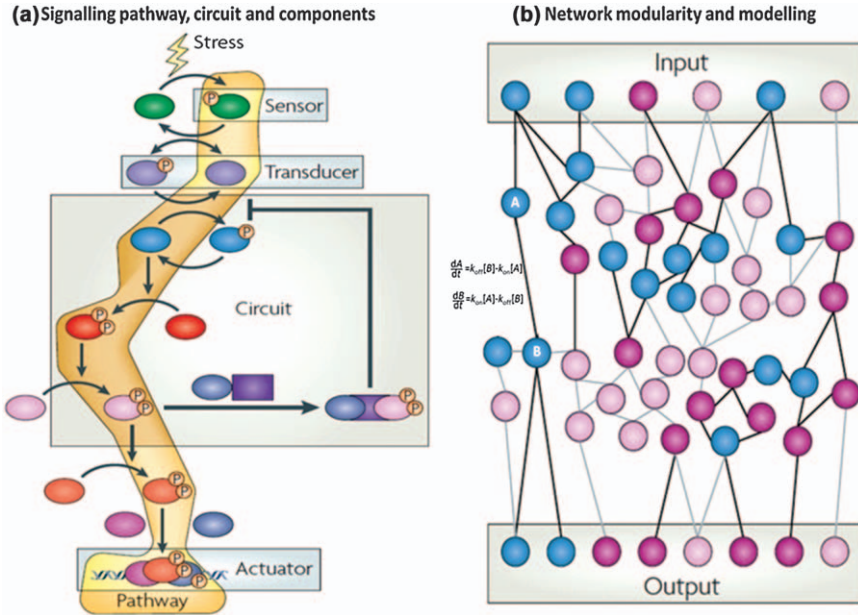


**Figure 9.5** Signalling networks modified from Hyduke *et al.*<sup>115</sup> (a) Signalling pathway, circuit and components. A basic signalling network is comprised of sensors (sensing environmental perturbations), transducers (propagating signal) and actuators (eliciting biological response). Complex signalling networks ensemble circuits which integrate/compute (un)related signals and modulate the flow of information through the network. (b) Network modularity and modelling. Complex signalling networks are subject to crosstalking and crosslinking increasing robustness and integration between different signals. Modularity is also present based on correlated expression between different components. In this example, there are two expressed modules (blue circles and purple circles) and one unexpressed module (pink circles). Although there is a route from one, or more, input to each output, not all of the outputs are currently accessible from the set of active inputs. Modelling formalisms include, network and dynamic approaches. Network analysis considers every single network component as a node. Connectivity between two components such as A and B in the example are denoted as edges. Dynamic approach accounts in addition for detailed information about reaction rates describing network interactions and changes in the concentration of each component by using mainly differential equations. Reprinted by permission from Macmillan Publishers Ltd: *Nature Reviews Genetics* (ref. 115), copyright 2010.

emerging as promising new approach for modelling signalling networks, at genome-scale.

Network or topological models have taken advantage of graph theory. For analysing the properties of a biological system, every single network component is considered as a node. In addition, these models include additional level of complexity such as the connectivity observed between the different

components within the network, which are denoted as edges (Figure 9.5(b)).<sup>119</sup> This connectivity can be direct, (when two components directly interact chemically and/or structurally each other) or indirect (wherein changes in a component drives changes in other network component without direct interaction). The ability of topological models predicting information-processing relies on the identification of large connected nodes (hubs) and network motifs, this is small organising units such as feedback and feedforward loops. Therefore, network models provide insights about network organisation while identifying underlying relationships between nodes. A well-known limitation of network models is that they only provide a static view of the system being unable to predict network changes over time. Despite these important shortcomings, topological models allow accurate predictions about the effects of perturbations and mutations at the cellular level. Thus, they have shown to be powerful computational tools identifying new drug targets and biomarkers in complex diseases such as cancer, diabetes, and Alzheimer's disease.<sup>120–123</sup> It is therefore not surprising that a network-based analysis approach is widely used for system analysis of multivariable diseases. An excellent and more authorised overview about this approach applied to cancer research can be found in ref. 124.

Dynamical or kinetic models, in contrast to graph-theory-based models, provide spatio-temporal understanding of signal propagation. In addition to topology, kinetic models account for detailed information about reaction rates between components in the network, which allow the accurate description about how a system changes over time (Figure 9.5(b)). Kinetic models are categorised in two classes, deterministic and stochastics. The fundamental differences between deterministic and stochastic approaches in chemical kinetic modelling are that, while deterministic simulation uses reaction kinetics and initial concentrations to describe changes in concentration of each species over time, stochastic simulations depends upon the probability of each reaction occurring (reaction propensities). Deterministic models are based on differential equations which describe network interactions. This is, by using reaction stoichiometry and reaction rates, the signal flow propagation can be represented by formulating a system of differential equations. This equations system can be then solved as a function of time to compute the state of each individual component over time. Deterministic formulation can be further improved by using partial differential equations in which each component includes a diffusion coefficient as new attribute in addition to specific reaction rate, allowing the prediction of the final state over time but also as a function of the space.<sup>125</sup> Stochastics models consider the probability of the interaction as a function of time, volume, and number of molecules and uses Gillespie's algorithm for computation.<sup>126</sup> This large number of variables makes stochastics model computationally expensive and therefore a deterministic approach is more commonly used in signalling network modelling. As in the case of network modelling, kinetic models have proven to be ideal computational tools in

biomedicine. For instance and without wishing to be exhaustive, the mammalian signalling networks modelling has provided numerous examples of model-driven discovery targeted on (i) the identification feedback loops governing p53 oscillations, (ii) activation of the nuclear factor- $\kappa$ B (NF- $\kappa$ B) and (iii) apoptotic pathways (thoroughly revised by Hughey *et al.*).<sup>127</sup> In another example, the dynamic modelling of oestrogen signalling provided insights into the molecular foundations of endocrine responsiveness in breast cancer.<sup>128</sup>

Because of the large and hard-to-assess parameters required for kinetic modelling, such as reactions rates and component concentrations, they are more suitable for small pathways and the large-scale signalling modelling using this approach is still challenging. Constraint-based approaches have been suggested for exploring the properties of cellular signalling networks (Figure 9.5).<sup>115</sup> Constraint-based modelling imposes a set of suitable constraints on the space of possible solutions and allows filtering out behaviours that are not biologically feasible in a large-scale manner.<sup>129,130</sup> Although this approach has shown to be useful for reconstructing individual signalling pathways such as the TLR signalling,<sup>131</sup> genome-scale reconstruction of signalling networks still is in its infancy and requires the complete list of parts and organising principles.

In summary, signalling network models are now being successfully used in driving new biological knowledge, identifying new drug targets, drug regimens and mechanisms of drug action. Therefore, it is expected that the introduction of new modelling formalisms, as well as the identification of the complete list of components, would lead models to be featured more prominently in the future for these purposes.

## Acknowledgements

This work was funded by grants from the Spanish Ministry for Economy and Competitiveness (MINECO CTQ2014-57141-R), and the European Commission (TOLLerant H2020-MSCA-ETN-642157).

## References

1. L. Sborgi, A. Verma, S. Piana, K. Lindorff-Larsen, M. Cerminara, C. M. Santiveri, D. E. Shaw, E. de Alba and V. Munoz, *J. Am. Chem. Soc.*, 2015, **137**, 6506–6516.
2. E. Marinari and G. Parisi, *EPL*, 1992, **19**, 451.
3. Y. Miao, F. Feixas, C. Eun and J. A. McCammon, *J. Comput. Chem.*, 2015, **36**, 1536–1549.
4. S. Honda, K. Yamasaki, Y. Sawada and H. Morii, *Structure*, 2004, **12**, 1507–1518.
5. Y. Sugita and Y. Okamoto, *Chem. Phys. Lett.*, 1999, **314**, 141–151.
6. F. Jiang and Y. D. Wu, *J. Am. Chem. Soc.*, 2014, **136**, 9536–9539.
7. G. H. Zerze, B. Uz and J. Mittal, *Proteins*, 2015, **83**, 1307–1315.

8. G. Ozer, E. F. Valeev, S. Quirk and R. Hernandez, *J. Chem. Theory Comput.*, 2010, **6**, 3026–3038.
9. H. R. Bureau, E. Hershkovits, S. Quirk and R. Hernandez, *J. Chem. Theory Comput.*, 2016, **12**, 2028–2037.
10. R. van der Lee, M. Buljan, B. Lang, R. J. Weatheritt, G. W. Daughdrill, A. K. Dunker, M. Fuxreiter, J. Gough, J. Gsponer, D. T. Jones, P. M. Kim, R. W. Kriwacki, C. J. Oldfield, R. V. Pappu, P. Tompa, V. N. Uversky, P. E. Wright and M. M. Babu, *Chem. Rev.*, 2014, **114**, 6589–6631.
11. R. B. Best, *Curr. Opin. Struct. Biol.*, 2017, **42**, 147–154.
12. M. Varadi, W. Vranken, M. Guharoy and P. Tompa, *Front. Mol. Biosci.*, 2015, **2**, 45.
13. L. C. DaSilva, T. Gurry and C. M. Stultz, *Intrinsically Disord. Proteins*, 2014, **2**, e970902.
14. Z. Ziegler, M. Schmidt, T. Gurry, V. Burger and C. M. Stultz, *Bioinformatics*, 2016, **32**, 2545–2547.
15. C. B. Anfinsen, *Science*, 1973, **181**, 223–230.
16. A. Kryshtafovych and K. Fidelis, *Drug Discovery Today*, 2009, **14**, 386–393.
17. C. A. Rohl, C. E. Strauss, K. M. Misura and D. Baker, *Methods Enzymol.*, 2004, **383**, 66–93.
18. A. Roy, A. Kucukural and Y. Zhang, *Nat. Protoc.*, 2010, **5**, 725–738.
19. K. Arnold, L. Bordoli, J. Kopp and T. Schwede, *Bioinformatics*, 2006, **22**, 195–201.
20. A. Sali and T. L. Blundell, *J. Mol. Biol.*, 1993, **234**, 779–815.
21. B. Webb and A. Sali, *Curr. Protoc. Bioinformatics*, 2016, **54**, 5.6.1–5.6.37.
22. C. Reichen, S. Hansen, C. Forzani, A. Honegger, S. J. Fleishman, T. Zhou, F. Parmeggiani, P. Ernst, C. Madhurantakam, C. Ewald, P. R. Mittl, O. Zerbe, D. Baker, A. Caflisch and A. Pluckthun, *J. Mol. Biol.*, 2016, **428**, 4467–4489.
23. O. Byron and B. Vestergaard, *Curr. Opin. Struct. Biol.*, 2015, **35**, 76–86.
24. M. Zhou, Q. Li and R. Wang, *ChemMedChem*, 2016, **11**, 738–756.
25. D. K. Johnson and J. Karanicolas, *Curr. Top. Med. Chem.*, 2017, DOI: 10.2174/1568026617666170508153904.
26. Y. Murakami, L. P. Tripathi, P. Prathipati and K. Mizuguchi, *Curr. Opin. Struct. Biol.*, 2017, **44**, 134–142.
27. L. Bonetta, *Nature*, 2010, **468**, 851–854.
28. M. P. Stumpf, T. Thorne, E. de Silva, R. Stewart, H. J. An, M. Lappe and C. Wiuf, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 6959–6964.
29. K. Venkatesan, J. F. Rual, A. Vazquez, U. Stelzl, I. Lemmens, T. Hirozane-Kishikawa, T. Hao, M. Zenkner, X. Xin, K. I. Goh, M. A. Yildirim, N. Simonis, K. Heinzmann, F. Gebreab, J. M. Sahalie, S. Cevik, C. Simon, A. S. de Smet, E. Dann, A. Smolyar, A. Vinayagam, H. Yu, D. Szeto, H. Borick, A. Dricot, N. Klitgord, R. R. Murray, C. Lin, M. Lalowski, J. Timm, K. Rau, C. Boone, P. Braun, M. E. Cusick, F. P. Roth, D. E. Hill, J. Tavernier, E. E. Wanker, A. L. Barabasi and M. Vidal, *Nat. Methods*, 2009, **6**, 83–90.



30. K. Luck, G. M. Sheynkman, I. Zhang and M. Vidal, *Trends Biochem. Sci.*, 2017, **42**, 342–354.
31. L. G. Milroy, L. Brunsveld and C. Ottmann, *ACS Chem. Biol.*, 2013, **8**, 27–35.
32. L.-G. Milroy, T. N. Grossmann, S. Hennig, L. Brunsveld and C. Ottmann, *Chem. Rev.*, 2014, **114**, 4695–4748.
33. J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li and H. Jiang, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 4337–4341.
34. C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.
35. Y. Ofra and B. Rost, *Bioinformatics*, 2007, **23**, e13–e16.
36. O. Keskin, N. Tuncbag and A. Gursoy, *Chem. Rev.*, 2016, **116**, 4884–4909.
37. <http://www.imexconsortium.org/>, (accessed on 24th May 2017).
38. <https://thebiogrid.org/>, (accessed on 24th May 2017).
39. <http://mint.bio.uniroma2.it/>, (accessed on 24th May 2017).
40. <http://binddb.org>, (accessed on 24th May 2017).
41. <http://dip.doe-mbi.ucla.edu/dip/Main.cgi>, (accessed on 24th May 2017).
42. <http://www.ebi.ac.uk/intact/>, (accessed on 24th May 2017).
43. <http://www.hprd.org/>, (accessed on 24th May 2017).
44. B. Lehne and T. Schlitt, *Hum. Genomics*, 2009, **3**, 291–297.
45. M. F. Lensink, S. Velankar, A. Kryshchuk, S. Y. Huang, D. Schneidman-Duhovny, A. Sali, J. Segura, N. Fernandez-Fuentes, S. Viswanath, R. Elber, S. Grudinin, P. Popov, E. Neveu, H. Lee, M. Baek, S. Park, L. Heo, G. Rie Lee, C. Seok, S. Qin, H. X. Zhou, D. W. Ritchie, B. Maigret, M. D. Devignes, A. Ghosh, M. Torchala, R. A. Chaleil, P. A. Bates, E. Ben-Zeev, M. Eisenstein, S. S. Negi, Z. Weng, T. Vreven, B. G. Pierce, T. M. Borrmann, J. Yu, F. Ochsenbein, R. Guerois, A. Vangone, J. P. Rodrigues, G. van Zundert, M. Nellen, L. Xue, E. Karaca, A. S. Melquiond, K. Visscher, P. L. Kastoris and A. M. Bonvin, *Proteins*, 2016, **84**(Suppl 1), 323–348.
46. D. Kozakov, D. Beglov, T. Bohnuud, S. E. Mottarella, B. Xia, D. R. Hall and S. Vajda, *Proteins*, 2013, **81**, 2159–2166.
47. D. Kozakov, R. Brenke, S. R. Comeau and S. Vajda, *Proteins*, 2006, **65**, 392–406.
48. <https://cluspro.bu.edu/login.php>, (accessed on 24th May 2017).
49. S. R. Comeau, D. W. Gatchell, S. Vajda and C. J. Camacho, *Bioinformatics*, 2004, **20**, 45–50.
50. <http://www.nmr.chem.uu.nl/haddock/>, (accessed on 24th May 2017).
51. <http://zdock.umassmed.edu/>, (accessed on 24th May 2017).
52. D. Duhovny, R. Nussinov and H. J. Wolfson, in *Algorithms in Bioinformatics: Second International Workshop, WABI 2002 Rome, Italy, September 17–21, 2002 Proceedings*, ed. R. Guigó and D. Gusfield, Springer Berlin Heidelberg, Berlin, Heidelberg, 2002, pp. 185–200.
53. D. Schneidman-Duhovny, Y. Inbar, R. Nussinov and H. J. Wolfson, *Nucleic Acids Res.*, 2005, **33**, W363–W367.



54. <http://bioinfo3d.cs.tau.ac.il/PatchDock/index.html>, (accessed on 24th May 2017).
55. A. D. J. van Dijk, R. Boelens and A. M. J. J. Bonvin, *FEBS J.*, 2005, **272**, 293–312.
56. S. Vajda and D. Kozakov, *Curr. Opin. Struct. Biol.*, 2009, **19**, 164–170.
57. J. J. Gray, S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C. A. Rohl and D. Baker, *J. Mol. Biol.*, 2003, **331**, 281–299.
58. D. Kozakov, D. R. Hall, D. Beglov, R. Brenke, S. R. Comeau, Y. Shen, K. Li, J. Zheng, P. Vakili, I. Paschalidis and S. Vajda, *Proteins*, 2010, **78**, 3124–3130.
59. C. Yueh, D. R. Hall, B. Xia, D. Padhorny, D. Kozakov and S. Vajda, *J. Mol. Biol.*, 2017, **429**, 372–381.
60. G. C. P. van Zundert, J. P. G. L. M. Rodrigues, M. Trellet, C. Schmitz, P. L. Kastritis, E. Karaca, A. S. J. Melquiond, M. van Dijk, S. J. de Vries and A. M. J. J. Bonvin, *J. Mol. Biol.*, 2016, **428**, 720–725.
61. M. S. Titushin, Y. Feng, G. A. Stepanyuk, Y. Li, S. V. Markova, S. Golz, B. C. Wang, J. Lee, J. Wang, E. S. Vysotski and Z. J. Liu, *J. Biol. Chem.*, 2010, **285**, 40891–40900.
62. B. G. Pierce, Y. Hourai and Z. Weng, *PLoS One*, 2011, **6**, e24657.
63. M. L. Connolly, *Science*, 1983, **221**, 709–713.
64. C. Zhang, G. Vasmatzis, J. L. Cornette and C. DeLisi, *J. Mol. Biol.*, 1997, **267**, 707–726.
65. E. Mashlach, D. Schneidman-Duhovny, A. Peri, Y. Shavit, R. Nussinov and H. J. Wolfson, *Proteins*, 2010, **78**, 3197–3204.
66. T. M. Schmeing and V. Ramakrishnan, *Nature*, 2009, **461**, 1234–1242.
67. T. A. Steitz, *Nat. Rev. Mol. Cell Biol.*, 2008, **9**, 242–253.
68. G. Ozer, S. Quirk and R. Hernandez, *J. Chem. Theory Comput.*, 2012, **8**, 4837–4844.
69. M. M. Yusupov, G. Z. Yusupova, A. Baucom, K. Lieberman, T. N. Earnest, J. Cate and H. F. Noller, *Science*, 2001, **292**, 883–896.
70. M. Karplus and J. A. McCammon, *Nat. Struct. Mol. Biol.*, 2002, **9**, 646–652.
71. J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale and K. Schulten, *J. Comput. Chem.*, 2005, **26**, 1781–1802.
72. C.-S. Tung and K. Y. Sanbonmatsu, *Biophys. J.*, 2004, **87**, 2714–2722.
73. K. Y. Sanbonmatsu, S. Joseph and C.-S. Tung, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 15854–15859.
74. L. G. Trabuco, E. Schreiner, J. Eargle, P. Cornish, T. Ha, Z. Luthey-Schulten and K. Schulten, *J. Mol. Biol.*, 2010, **402**, 741–760.
75. F. Tama, M. Valle, J. Frank and C. L. Brooks, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 9319–9323.
76. P. Chacón, F. Tama and W. Wriggers, *J. Mol. Biol.*, 2003, **326**, 485–492.
77. J. Trylska, *J. Phys.: Condens. Matter*, 2010, **22**, 453101.
78. J. Šponer, P. Banáš, P. Jurecka, M. Zgarbová, P. Kührová, M. Havrila, M. Krepl, P. Stadlbauer and M. Otyepka, *J. Phys. Chem. Lett.*, 2014, **5**, 1771–1782.

79. M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna and E. Charpentier, *Science*, 2012, **337**, 816–821.
80. G. Palermo, Y. Miao, R. C. Walker, M. Jinek and J. A. McCammon, *ACS Cent. Sci.*, 2016, **2**, 756–763.
81. J. M. Billod, A. Lacetera, J. Guzman-Caldentey and S. Martin-Santamaria, *Molecules*, 2016, **21**, 994.
82. N. A. Berglund, V. Kargas, M. L. Ortiz-Suarez and P. J. Bond, *Prog. Biophys. Mol. Biol.*, 2015, **119**, 72–83.
83. M. S. Murgueitio, C. Rakers, A. Frank and G. Wolber, *Trends Pharmacol. Sci.*, 2017, **38**, 155–168.
84. L. Perez-Regidor, M. Zarioh, L. Ortega and S. Martin-Santamaria, *Int. J. Mol. Sci.*, 2016, **17**, 1508.
85. Y. Amir, E. Ben-Ishay, D. Levner, S. Ittah, A. Abu-Horowitz and I. Bachelet, *Nat. Nanotechnol.*, 2014, **9**, 353–357.
86. Z. G. Wang, C. Song and B. Ding, *Small*, 2013, **9**, 2210–2222.
87. P. Sharma, N. K. Mehra, K. Jain and N. K. Jain, *Curr. Drug Delivery*, 2016, **13**, 796–817.
88. F. M. Tonelli, V. A. Goulart, K. N. Gomes, M. S. Ladeira, A. K. Santos, E. Lorencon, L. O. Ladeira and R. R. Resende, *Nanomedicine*, 2015, **10**, 2423–2450.
89. M. Shah, V. D. Badwaik and R. Dakshinamurthy, *J. Nanosci. Nanotechnol.*, 2014, **14**, 344–362.
90. M. Ozboyaci, D. B. Kokh, S. Corni and R. C. Wade, *Q. Rev. Biophys.*, 2016, **49**, e4.
91. A. E. Aliaga, H. Ahumada, K. Sepúlveda, J. S. Gomez-Jeria, C. Garrido, B. E. Weiss-López and M. M. Campos-Vallette, *J. Phys. Chem. C*, 2011, **115**, 3982–3989.
92. X. Liu, P. K. Chu and C. Ding, *Mater. Sci. Eng., R*, 2004, **47**, 49–121.
93. S. Monti, *J. Phys. Chem. C*, 2007, **111**, 16962–16973.
94. T. Wei, M. A. Carignano and I. Szleifer, *Langmuir*, 2011, **27**, 12074–12081.
95. V. Biju, *Chem. Soc. Rev.*, 2014, **43**, 744–764.
96. S. Jiang, K. Y. Win, S. Liu, C. P. Teng, Y. Zheng and M. Y. Han, *Nanoscale*, 2013, **5**, 3127–3148.
97. P. Charchar, A. J. Christofferson, N. Todorova and I. Yarovsky, *Small*, 2016, **12**, 2395–2418.
98. G. Brancolini, D. B. Kokh, L. Calzolari, R. C. Wade and S. Corni, *ACS Nano*, 2012, **6**, 9863–9878.
99. D. B. Kokh, S. Corni, P. J. Winn, M. Hoefling, K. E. Gottschalk and R. C. Wade, *J. Chem. Theory Comput.*, 2010, **6**, 1753–1768.
100. D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark and H. J. Berendsen, *J. Comput. Chem.*, 2005, **26**, 1701–1718.
101. Y. Mou, J. Y. Yu, T. M. Wannier, C. L. Guo and S. L. Mayo, *Nature*, 2015, **525**, 230–233.
102. P. S. Huang, J. J. Love and S. L. Mayo, *J. Comput. Chem.*, 2005, **26**, 1222–1232.

103. R. Das and D. Baker, *Annu. Rev. Biochem.*, 2008, **77**, 363–382.
104. B. D. Allen and S. L. Mayo, *J. Comput. Chem.*, 2006, **27**, 1071–1075.
105. C. Yang, M. E. Denno, P. Pyakurel and B. J. Venton, *Anal. Chim. Acta*, 2015, **887**, 17–37.
106. S. S. Karajanagi, A. A. Vertegel, R. S. Kane and J. S. Dordick, *Langmuir*, 2004, **20**, 11594–11599.
107. G. R. Dieckmann, A. B. Dalton, P. A. Johnson, J. Razal, J. Chen, G. M. Giordano, E. Munoz, I. H. Musselman, R. H. Baughman and R. K. Draper, *J. Am. Chem. Soc.*, 2003, **125**, 1770–1777.
108. Z. Kuang, S. N. Kim, W. J. Crookes-Goodson, B. L. Farmer and R. R. Naik, *ACS Nano*, 2010, **4**, 452–458.
109. D. A. Case, T. A. Darden, T. E. Cheatham, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, M. Crowley, R. C. Walker, W. Zhang, K. M. Merz, B. Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossváry, K. F. Wong, F. Paesani, J. Vanicek, X. Wu, S. R. Brozell, T. Steinbrecher, H. Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, D. H. Mathews, M. G. Seetin, C. Sagui, V. Babin and P. A. Kollman, *AMBER 10*, University of California, San Francisco, 2010.
110. 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.
111. B. Lewandowski, G. De Bo, J. W. Ward, M. Papmeyer, S. Kuschel, M. J. Aldegunde, P. M. Gramlich, D. Heckmann, S. M. Goldup, D. M. D'Souza, A. E. Fernandes and D. A. Leigh, *Science*, 2013, **339**, 189–193.
112. P. Altoe, N. Haraszkiwicz, F. G. Gatti, P. G. Wiering, C. Frochot, A. M. Brouwer, G. Balkowski, D. Shaw, S. Woutersen, W. J. Buma, F. Zerbetto, G. Orlandi, D. A. Leigh and M. Garavelli, *J. Am. Chem. Soc.*, 2009, **131**, 104–117.
113. J. D. Jordan, E. M. Landau and R. Iyengar, *Cell*, 2000, **103**, 193–200.
114. N. J. Eungdamrong and R. Iyengar, *Biol. Cell*, 2004, **96**, 355–362.
115. D. R. Hyde and B. O. Palsson, *Nat. Rev. Genet.*, 2010, **11**, 297–307.
116. M. J. Lee, A. S. Ye, A. K. Gardino, A. M. Heijink, P. K. Sorger, G. MacBeath and M. B. Yaffe, *Cell*, 2012, **149**, 780–794.
117. L. B. Kleiman, T. Maiwald, H. Conzelmann, D. A. Lauffenburger and P. K. Sorger, *Mol. Cell*, 2011, **43**, 723–737.
118. G. Weng, U. S. Bhalla and R. Iyengar, *Science*, 1999, **284**, 92–96.
119. G. A. Pavlopoulos, M. Secrier, C. N. Moschopoulos, T. G. Soldatos, S. Kossida, J. Aerts, R. Schneider and P. G. Bagos, *BioData Min.*, 2011, **4**, 10.
120. J. T. Erler and R. Linding, *J. Pathol.*, 2010, **220**, 290–296.
121. J. Wang, Y. Zuo, Y.-G. Man, I. Avital, A. Stojadinovic, M. Liu, X. Yang, R. S. Varghese, M. G. Tadesse and H. W. Ransom, *Cancer*, 2015, **6**, 54–65.
122. T. Jing and A. Tero, *Curr. Pharm. Des.*, 2014, **20**, 23–36.
123. J. D. Deco, A. M. Gonzalez, J. Diaz, V. Mato, D. Garcia-Frank, J. Alvarez-Linera, A. Frank and J. A. Hernandez-Tamames, *Curr. Top. Med. Chem.*, 2013, **13**, 652–662.

124. The Mutation Consequences and Pathway Analysis working group of the International Cancer Genome Consortium, P. Creixell, J. Reimand, S. Haider, W. Guanming, S. Tatsuhiko, M. Vazquez, V. Mustonen, A. Gonzalez-Perez, J. Pearson, C. Sander, B. J. Raphael, D. S. Marks, B. F. Ouellette, A. Valencia, G. D. Bader, P. C. Boutros, J. M. Stuart, R. Linding, N. Lopez-Bigas and L. D. Stein, *Nat. Methods*, 2015, **12**, 615–621.
125. A. MacNamara, D. Henriques and J. Saez-Rodriguez, in *In Silico Systems Biology*, ed. M. V. Schneider, Humana Press, Totowa, NJ, 2013, pp. 89–105.
126. D. T. Gillespie, *Annu. Rev. Phys. Chem.*, 2007, **58**, 35–55.
127. J. J. Hughey, T. K. Lee and M. W. Covert, *Wiley Interdiscip. Rev.: Syst. Biol. Med.*, 2010, **2**, 194–209.
128. J. J. Tyson, W. T. Baumann, C. Chen, A. Verdugo, I. Tavassoly, Y. Wang, L. M. Weiner and R. Clarke, *Nat. Rev. Cancer*, 2011, **11**, 523–532.
129. J. D. Orth, I. Thiele and B. O. Palsson, *Nat. Biotechnol.*, 2010, **28**, 245–248.
130. E. J. O'Brien, J. M. Monk and B. O. Palsson, *Cell*, 2015, **161**, 971–987.
131. F. Li, I. Thiele, N. Jamshidi and B. Ø. Palsson, *PLoS Comput. Biol.*, 2009, **5**, e1000292.

# *Novel Insights into Membrane Transport from Computational Methodologies*

VICTORIA OAKES<sup>a</sup> AND CARMEN DOMENE<sup>\*a,b</sup>

<sup>a</sup> Department of Chemistry, Britannia House, 7 Trinity Street, King's College London, London SE1 1DB, UK; <sup>b</sup> Chemistry Research Laboratory, Mansfield Road, University of Oxford, Oxford OX1 3TA, UK  
\*Email: mcdn20@bath.ac.uk

## 10.1 Introduction

The plasma membrane is an integral constituent of both prokaryotic and eukaryotic cells, enclosing the cytoplasm and other cell components. In the latter, the biological membrane is the primary instrument governing the passage of solutes to the cell interior. A phospholipid bilayer forms the basic structural unit establishing a hydrophobic barrier in the membrane core, thus whilst hydrophobic molecules may be admitted passage, additional biological assemblies are required for the transmission of many molecules. Membrane proteins, as these assemblies are known, also play a role in cell signalling, invoking internal processes on recognition of external stimuli. A schematic of membrane transport phenomenon can be found in Figure 10.1.

The availability of atomic resolution structural information is pivotal to discern the underlying principles of membrane transport and communication phenomenon. Vast developments in X-ray crystallographic techniques, NMR and innovative structural determination methods, such as

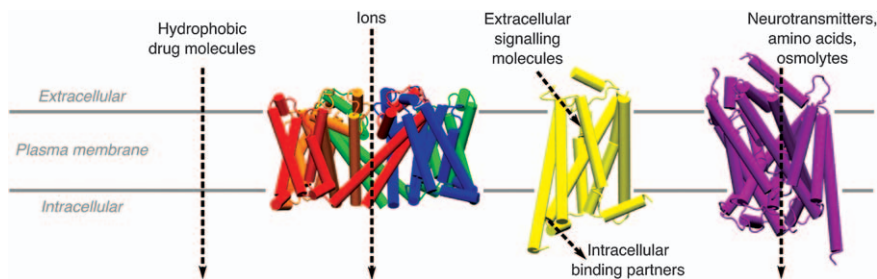


**Figure 10.1** Schematic of transport phenomenon across the plasma membrane: hydrophobic drug molecules transit directly through the lipid membrane; ions passively diffuse down the electrochemical gradient of the membrane *via* ion channels, represented as the  $\text{Na}_v\text{Ab}$  sodium voltage-gated channel (individual subunits coloured blue, red, yellow and green); binding of extracellular molecules to G-protein coupled receptors initiates coupling with intracellular partners and subsequent signalling, represented as the  $\beta_2$ -adrenergic receptor (yellow); molecules can also be actively transported across the membrane, such as amino acids through  $\text{Na}^+$ -coupled secondary symporters, represented as the leucine transporter (purple).

electron cryo-microscopy, have revolutionised our understanding of the three-dimensional structure of membrane proteins.<sup>1</sup> In combination with the current availability of state-of-the-art computational algorithms and high-performance computing facilities, molecular dynamics (MD) simulations of membrane proteins in a model environment can now be performed routinely to explore a wide range of biological phenomena providing insight for which no experimental methods are applicable. Exploration of the dynamic behaviour of such entities in atomic detail has become particularly important in the realm of medicinal chemistry, where familiarity of ligand transport and binding processes, and subsequent activation or deactivation mechanisms, can lead to the rational design of drugs with improved potency and efficacy, reduced toxicity and subtype specificity. In the following text, we aim to review the considerable insights gained from computational methodologies in recent years.

## 10.2 Computational Methods

A wide variety of computational methods are currently applied in the field of computational chemistry. MD is a common method used to obtain time-dependent behaviour of a chemical system. In this method, the motion of interacting particles is calculated by the integration of Newton's equations of motion (EOM). The potential energy of the system and the forces, derived from the negative gradient of the potential with respect to displacement in a specified direction, can be used to calculate the acceleration, and hence forecast the time evolution of the system, in the form of a trajectory.

In systems with a couple of atoms, solutions of the EOM can be gained analytically, resulting in a continuous trajectory over time. However, in larger systems, the subsistence of a continuous potential instigates a many body problem for force evaluations, rendering analytic solutions unattainable. In these circumstances, finite difference methods can be used; forces are assessed at discrete intervals, and considered constant in the hiatus. Positions and velocities at the next time step, as these intervals are known, are computed using force evaluations for each atom combined with current positions and velocities. Forces are then recalculated and this procedure is repeated, propagating a trajectory describing the flux of the atomic coordinates over time in a given equilibrium state, which can then be analysed for the properties of interest.

The potential energy can be obtained by quantum or molecular mechanical methods. In the former, only the nuclear motion is evaluated using classical physics; atoms and bonds are considered as balls and springs, respectively, and an analytical expression for the energy of a system, known as a force field, can be derived. This is composed entirely of inter- and intramolecular energetic contributions, including bond stretching, angle bending, bond rotations and non-bonded terms. This architecture neglects electronic properties, such as dipole moments, polarisabilities and vibrational frequencies, but allows for the evaluation of molecular motion for (biological) systems with a large number of degrees of freedom due to reduced computational expense. A wide variety of force fields for biological molecules are available including, but not limited to CHARMM (Chemistry at Harvard Molecular Mechanics),<sup>2</sup> AMBER (Assisted Model Building with Energy Requirement),<sup>3</sup> and OPLS (Optimised Potentials for Liquid Simulations).<sup>4</sup> Each varies in their functional form, parameterisation protocol and parameters therein, which are generally obtained to provide a suitable reproduction of experimental and/or quantum mechanical data.

Most commonly, individual terms for bond lengths and angles utilise simple harmonic potentials on the basis of an energetic penalty associated with a deviation from the equilibrium value.<sup>2</sup> A torsional angle potential function is also typical to model steric barriers associated with the rotation of atoms. With respect to non-bonded interactions, Coulomb and Lennard-Jones potentials are used to express electrostatic and van der Waals forces respectively. In order to ascertain electrostatic forces *via* the Coulomb formula, fixed-point charges are assigned at the nuclei. However, in reality, molecules are subject to polarisation effects, *i.e.* the charge fluctuates response to an external electric field, induced by the presence of additional molecules, thus highlighting a limitation of additive force fields.<sup>5</sup>

Significant efforts to develop polarisable FF have been undertaken in recent years. Several schemes have been proposed, among which are the fluctuating charge model, the induced dipole model and the Drude oscillator approach. In the fluctuating charge model, adopted in the CHEQ force field<sup>6,7</sup> for example, molecular charges remain constant throughout the simulation with individual point charges readjusted in consonance with



the electronegativity. AMOEBA is the most noteworthy force field utilising an induced dipole model, where atomic multipoles are used explicitly to represent electrostatics; multipoles are calculated *via* a self-consistent field procedure, heightening the computational expense of this protocol.<sup>8</sup> Finally, in the Drude oscillator approach, a subsidiary charged particle is attached to the nucleus by a harmonic spring, and treated as an extra degree of freedom.<sup>9</sup> A comprehensive review of the field can be found in ref. 10.

Despite the universal availability of MD algorithms and force fields applicable to biological macromolecules, the size of model systems of intrinsic membrane proteins, and the computing resources this requires, poses inherent limitations. The time step suitable for stable dynamics is dictated by the highest vibrational frequency of the molecule, usually the C–H bond; 1 fs is commonplace, with 2 fs permitted when supplementary algorithms such as SETTLE,<sup>11</sup> an analytical version of the SHAKE function, are employed to the movement of these bonds. In combination with system size up to 1 000 000 atoms, dynamics can be extremely intensive.<sup>12</sup> Recent expansions in computer hardware and high-performance computing facilities, means MD simulations on a nanosecond timescale are now routine, with microsecond simulations attainable in recent years. Such methodologies are therefore relevant to study a wide range of biological phenomena. Many biological phenomena that occur on extended timescales, such as protein folding, complex association and conformational changes associated with gating, are generally unattainable by atomistic equilibrium MD without the use of tailor-made software. The Shaw Group has pioneered the production of millisecond long unbiased simulations by the development of the Supercomputer Anton, optimised for use with MD software.<sup>13</sup> However, this technology is not widely available, leading to the development of alternative approaches to accelerate sampling.

Using a reduced representation is one such approach, utilising classical MD simulations, to increase the speed and hence timescales obtainable, as well as simulate larger systems with increased complexity. Coarse-grained (CG) molecular dynamics, as this is known, reduces the number of degrees of freedom in a simulation system by treating a group of atoms as a single entity, significantly curtailing the computational expense of each step. To convert an all-atom structure to a coarse-grained model, hydrogen atoms are not considered, with a number of heavy atoms (typically three or four) grouped into a single interaction site, known as a ‘bead’. Interaction potentials are then characterised dependent on the CG model, with required parameters generally developed to reproduce microscopic properties recorded in atomistic simulations and thermodynamic data derived by experimental means. Originally developed to capture the extended time and large-scale behaviour of membranes, a number of CG force fields have emerged from the research groups of Klein,<sup>14,15</sup> Marrink<sup>16,17</sup> and others.<sup>18</sup> The implementation of these schemes reduces the number of required calculations, as well as increasing the time step of each iteration (20–40 fs) providing a powerful tool to accelerate molecular simulations. Methods to

maintain a level of chemical specificity are also embedded in each scheme; in the Martini force field, from Marrink *et al.*, sites are classified by the chemical nature of the region, *i.e.* polar, non-polar, apolar (a mixture of polar and non-polar groups) and charged, for example. Additional techniques are often required to sample phenomena where both atomistic and coarse-grain methodologies are not appropriate, such as those that occur on an extended timescale or involve an energetic barrier.

From MD trajectories of the system in time structural and dynamical quantities as well as kinetic and thermodynamic properties can be calculated using the principles of statistical mechanics. Among the thermodynamic quantities, it is the free energy that provides a direct link between statistical mechanics and thermodynamics, and through which other thermodynamic quantities can be obtained. A long-standing method to calculate the underlying free energy of a system is known as free energy perturbation (FEP). In this method, alchemical transformations are performed to overcome energetic barriers, and the relative free energy differences are calculated by a thermodynamic cycle.<sup>19,20</sup> Several algorithms also exist to accelerate sampling along a pre-defined set of reaction coordinates and estimate the potential of mean force (PMF), such as umbrella sampling, metadynamics, adaptive biasing force or steered MD. Such reaction coordinates, known as collective variables (CV), are chosen to elucidate a specific conformational transition. A free energy estimate as a function of the collective variables, as well as the equilibrium properties, can be obtained providing a wealth of information about the simulation system at a fraction of the expense of traditional all-atom MD.

Umbrella sampling (US)<sup>21</sup> is perhaps one of the most popular enhanced sampling method in this field, where a bias potential along user-defined CVs provokes conversion between stable thermodynamic states. Independent MD simulations are performed at intermediary steps, known as windows, which can be combined using the weight histogram analysis method (WHAM)<sup>22,23</sup> or umbrella integration.<sup>24</sup> Each window represents equilibrium sampling of energetically distinct locales; accordingly such evaluation estimates the consolidated equilibrium free-energy surface.

Another method in this group is steered molecular dynamics (SMD) simulations.<sup>25</sup> SMD are akin to well-established experimental techniques, such as atomic force microscopy or optical tweezers, where an external force is applied to an atom, or group of atoms, to overcome barriers and sample a specific process.<sup>25</sup> Relative free energies can then be obtained by the Jarzynski equality.<sup>26</sup>

In metadynamics, a superficial bias potential is utilised to advance sampling along suitable CVs, diverting from configurational space previously inhabited. The biasing potential is adjusted by the addition of a Gaussian function, augmenting the energy of the system and departing from local free energy minima, allowing the exploration of alternative thermodynamic states separated by energetic barriers. Once convergence has been achieved, effectively when the entire free energy profile has been flattened, it can be

easily reconstructed to provide an unbiased estimate of the landscape as a function of the CVs.

The adaptive biasing force method (ABF)<sup>27</sup> is largely based on thermodynamic integration, whereby the instantaneous force along the reaction coordinate is evaluated directly and counteracted by an external biasing force of equal and opposite magnitude. This effectively provides a smooth energy landscape, and uniform sampling irrespective of energetic barriers allowing accelerated dynamics.

In the following sections, a wide range of literature will be discussed to illustrate how conventional and accelerated MD methods have been used to provide crucial insights into the functioning of the plasma membrane and its constituents on an atomistic level.

### 10.3 Unassisted Diffusion Across Lipid Bilayers

In the first instance, MD simulations have become an established tool to characterise unassisted transport across lipid bilayers, as a model for the biological membrane. The properties of lipid bilayers are generally characterised by a hydrophilic exterior and hydrophobic interior, favourably interacting with aqueous intra- and extracellular compartments (Figure 10.2(a)). In reality, of course, the situation is significantly more complex; membranes are highly heterologous systems that are capable of transporting small solutes and impermeable to many others. Even though overall permeability coefficients can be obtained by experimental and computational means, the latter is required to explicitly represent structural and dynamical fluctuations across lipid bilayers and, hence, gain atomistic mechanistic details of transport phenomenon. Although embedded proteins are instrumental in a manifold of transport events as described in later sections, unassisted diffusion is the prevailing manner by which small molecules, including drugs, gain access to the cell. In the following section, we aim to review a selection of recent publications to illustrate the use of computational simulations to

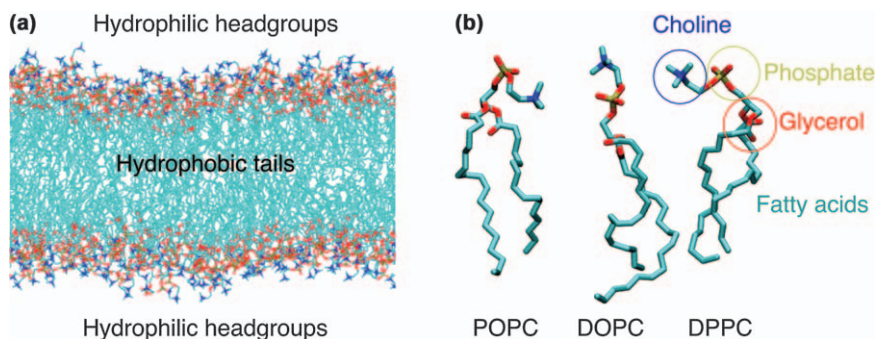


**Figure 10.2** (a) Representation of a single-component lipid bilayer typically used as a model plasma membrane. (b) Structures of common phospholipid molecules used in MD simulations.

explore membrane permeability properties. Model membranes comprised of a single species of phospholipid, such as 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC), 1,2-dioleoyl-sn-glycero-3-phosphocholine (DOPC) and 1,2-dipalmitoyl-sn-glycero-3-phosphocholine (DPPC) are typically used in this context (Figure 10.2(b)).

Passive diffusion across lipid bilayers is historically characterised by a solubility–diffusion mechanism; the solute diffuses from the extracellular solution and negotiates the membrane prior to accessing the cytoplasm. If partitioning at the membrane–water interfaces is assumed to be at equilibrium, and the membrane is considered as a homogenous oil slab in line with Overton’s correlation between membrane permeability and oil/water partitioning coefficient of a solute<sup>28</sup> a simple relationship between the oil/water-partitioning coefficient ( $K$ ), the diffusion coefficient of the solute in solvent ( $D$ ) and the membrane thickness can be used to yield the membrane permeability coefficient ( $P$ ):<sup>29</sup>

$$P = \frac{KD}{h} \quad (10.1)$$

The inhomogeneous solubility–diffusion model provides notable improvements by taking into account the heterogeneous behaviour of lipid bilayers, procuring permeability coefficients as a function of membrane depth.<sup>30</sup> Comprehensive sampling of the entire diffusion process is currently unattainable by all-atom MD simulations, due to the computational expense and the stochastic nature of the technique. Therefore, the applicability of this model to MD simulations<sup>31</sup> is dependent on enhanced sampling techniques; the original implementation utilised the z-constraint method<sup>31</sup> whereby the centre of mass of the molecule in question is constrained at defined positions along the z-axis, obtaining free-energies and depth-dependent diffusion coefficients as a function of the constraining force required at each position. Methods outlined in computational methods such as umbrella sampling<sup>21</sup> can also be applied. Such practices remain commonplace today, with alternative techniques to calculate such parameters emerging in recent years.<sup>32–34</sup>

The earliest studies utilising MD simulations to probe passive diffusion primarily focused on small molecules, such as water, oxygen and carbon dioxide.<sup>31</sup> The predictions obtained qualitatively agreed with known transport phenomenon of such molecules: oxygen and carbon dioxide are able to permeate relatively unhindered, whereas water diffusion is impeded and therefore, requires aquaporins for fast and efficient transport. This analysis was extended to the most common functional groups. However, the inaccuracy of observed permeability coefficients was notable in comparison to experimental determinations. Fortunately, advancements in computer hardware and empirical force fields have enabled the characterisation of transport properties to be undertaken at increasing timescales and levels of accuracy.

In a recent study by Riahi and Rowley,<sup>35</sup> the permeability of water and hydrogen sulphide was evaluated using the Drude polarisable model, where experimental diffusion coefficients have been used to optimise parameters for both substances<sup>36–38</sup> and the lipid DPPC.<sup>39</sup> Incorporation of induced polarisation effects enabled agreement between calculated and experimental diffusion coefficients at a value of  $(2.6 \pm 0.5) \times 10^{-5} \text{ cm s}^{-1}$  for water,<sup>40</sup> previously calculated as two orders of magnitude greater by additive force fields.<sup>41</sup> In comparison, a coefficient of  $11.9 \pm 0.7$  was obtained, suggesting hydrogen sulphide permeates approximately 400 000 times faster, and reinforcing experimental suggestions that hydrogen sulfide permeability is at least four orders of magnitude greater than water<sup>35</sup> rendering facilitated transport unnecessary.<sup>42</sup> This is attributed to distinct energetic barriers arising from electrostatic contributions ( $27 \text{ kJ mol}^{-1}$  for water *vs.*  $2.6 \text{ kJ mol}^{-1}$  for hydrogen sulfide); the dominant hydrogen bonding character of water cannot be obtained in the hydrophobic tail region disfavours the permeation of individual molecules; in contrast, hydrogen sulfide is relatively hydrophobic, thus hydrogen bonding capabilities with water are limited, and occupation of the inner membrane is encouraged. Therefore, hydrogen sulfide demonstrates permeability on a similar scale to non-polar solutes,<sup>43</sup> undeterred by its innate polarity. This phenomenon likely contributes to the remarkable potency of hydrogen sulfide in biological systems.<sup>44</sup>

The inherent relationship between membrane permeability and biological activity is also important in a drug discovery context. Penetration of cell membranes is imperative to gain access to target sites; the evaluation of permeability coefficients and free-energy profiles of drug molecules evaluation, in a similar manner to small molecules, can be used to assess the likelihood of reaching such sites and hence elicit a functional response. Experimental determination of drug permeabilities requires intensive experimental techniques, thus a plethora of computational studies to characterise the behaviour of clinically relevant molecules in model membranes have emerged to complement experimental work. A chemically diverse range of molecules has been investigated, establishing the partitioning properties of a broad spectrum of drug classes. Orsi and Essex, for example, demonstrated steroid hormones are centralized in the glycerol region, whereas,  $\beta$ -blockers tended towards the headgroup region, due to favourable interactions between the central oxygen's and polar moieties in the membrane.<sup>45</sup> Overall, a general consensus has emerged that large drug molecules primarily accumulate at the extended polar–apolar interface to accommodate their amphiphilic nature.<sup>46</sup> Such molecules are found to have a lower energy than in bulk water, with small barriers to overcome, providing an energetically permissible route for drug transit. The calculated  $\Delta G$  values can sometimes diverge from those determined experimentally, which could be attributed to the use of bilayers containing a single species of lipid and/or consideration of a single drug molecule in the majority of models, even though simultaneous diffusion of a number of molecules may actually occur.<sup>47</sup>

The interrelationship between solute concentration and membrane permeabilities has been examined *via* MD simulations. Comparison of ABF-generated free energy profiles of the anti-cancer drug, paclitaxel permeation through a pure POPC bilayer and POPC bilayers containing 12 mol% paclitaxel suggested incorporation of paclitaxel endorsed the transportability of the drug by increasing the partitioning from water to the bilayer ( $\sim 9$  vs.  $\sim 21$  kcal mol<sup>-1</sup>) and decreasing the barrier to transfer between leaflets in the bilayer ( $\sim 7$  vs.  $\sim 4$  kcal mol<sup>-1</sup>).<sup>48</sup> In the paclitaxel-rich bilayers, aggregates are formed from both random and lattice-based starting configurations, which decrease the lipid tail order parameter and promote translocation of water into the membrane, confirming earlier experimental predictions of pore formation in a concentration-dependent manner.<sup>49</sup>

The ability of small membrane bound molecules to cause significant perturbations in membrane structure and accelerate drug permeation has been exploited particularly where transdermal drug delivery is concerned.<sup>50</sup> Molecules such as dimethyl sulfoxide,<sup>51</sup> ethanol,<sup>52</sup> acetone<sup>53</sup> and oleic acid<sup>54</sup> have been proposed as chemical penetration enhancers, to breach the skin barrier by improving permeability of the lipid bilayers of the stratum corneum. The molecular mechanisms by which such molecules act have been investigated in a number of computational studies to date.<sup>55-61</sup>

Dimethyl sulfoxide molecules, for example, were found to freely partition into bilayer and accumulate beneath the lipid headgroups, swelling the distance between adjacent headgroups and dispersing the tails, increasing membrane flexibility and decreasing thickness.<sup>55</sup> Above a certain concentration threshold structural defects generate transient water pores, in a similar manner to paclitaxel, with a further increase expelling individual lipids from the membrane and destroying the bilayer structure.<sup>56</sup> The concentration dependence of dimethyl sulfoxide permeability enhancement experimentally<sup>52</sup> and for pore formation computationally suggests the two phenomena could be related, providing a feasible explanation for the amplified permeabilities of both hydrophobic and hydrophilic compounds through the skin.<sup>52</sup> Comparison of DPPC and DOPC showed that the latter is less susceptible to these effects, demonstrating enhanced stability and diminished dimethyl sulfoxide diffusion.<sup>57</sup>

These examples demonstrate how computational methodologies have provided a mechanistic understanding of experimentally established phenomenon, such as the exclusion of charged and hydrophilic molecules, the partitioning properties of hydrophobic drug molecules and membrane perturbations as a result of molecule entry. The alternative entry routes to membrane passage, *via* embedded proteins, will be explored in the following sections.

## 10.4 Passive Transport by Ion Channels

Ion channels facilitate the passive diffusion of ionic species down their electrochemical gradient from the extracellular medium into the cell



cytoplasm. Permeation is controlled by the onset of various external stimuli, such as transmembrane voltage, heat, ligand binding, and mechanical stretch, and is responsible for regulating electrical signals across the cellular membrane.

The voltage-gated ion channel (VGIC) family that specifically conduct  $\text{Na}^+$  ( $\text{Na}_v$  channels) and  $\text{K}^+$  ions ( $\text{K}_v$  channels), for example, are responsible for the generation of action potentials in excitable cells in various tissues in the heart, brain and nervous system, and thus play a crucial physiological role. VGIC's are a common target for antiarrhythmic agents, local anaesthetics, anticonvulsants and pain therapeutics.<sup>62,63</sup> Understanding how such proteins assemble and function is therefore of great pharmacological importance.

Elucidation of the crystal structure of the KcsA channel from *Streptomyces lividans* in 1998, provided the first atomistic description of the pore structure of an ion channel, a tetrameric arrangement with each monomer containing two transmembrane  $\alpha$ -helices and an intermittent pore loop (Figure 10.3).<sup>64</sup>

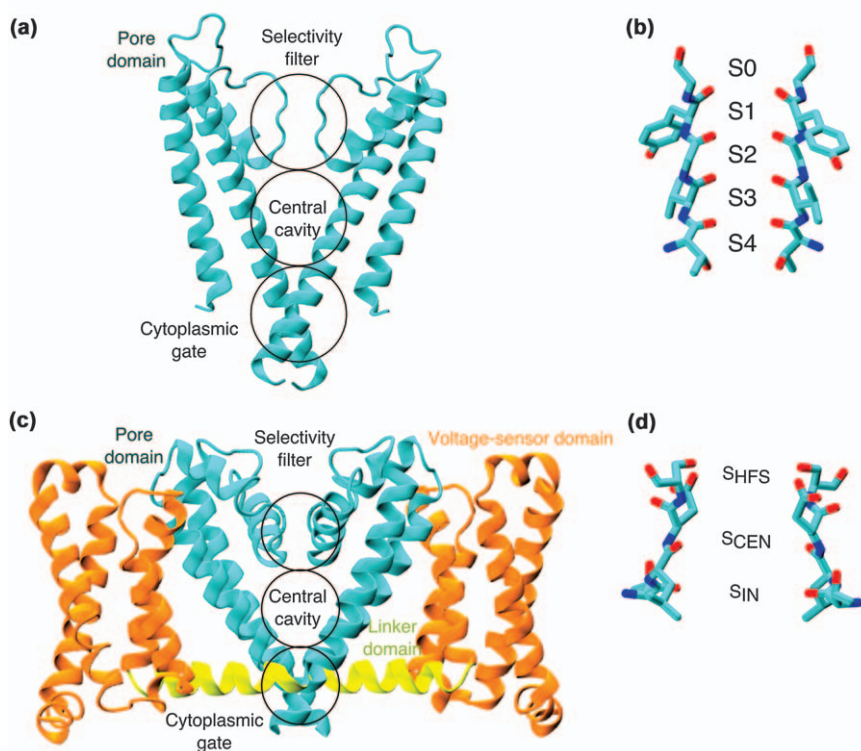


**Figure 10.3** (a) Structure of KcsA potassium channel. (b) Proposed ion binding sites in the KcsA  $\text{K}^+$ -channel selectivity filter. (c) Structure of voltage-gated sodium channel  $\text{Na}_v\text{Ab}$ . (d) Proposed ion binding sites in  $\text{Na}_v\text{Ab}$   $\text{Na}^+$ -channel selectivity filter. Only two domains are shown for clarity throughout.



The latter was found to contain a pore helix and the signature selectivity sequence and thus has become known as the selectivity filter.<sup>65,66</sup> Between the selectivity filter and the cytoplasm, a water-filled cavity is present to provide an ideal environment for ion transfer.<sup>67</sup> MD simulations, using KcsA, have extensively analysed the mechanism by which  $K^+$  channels selectively and efficiently conduct  $K^+$  ions.<sup>68–72</sup> The abundance of  $K^+$  channel structures now available has enabled various aspects of  $K^+$  channel function, such as voltage-sensing, cytoplasmic gating and drug blockage to be examined by computational analyses.<sup>73–75</sup> Several comprehensive reviews are available on this subject.<sup>76,77</sup> Similarly, since the publication of the first crystal structure of the Na<sub>v</sub>Ab from *Arcobacter butzleri*,<sup>78</sup> and subsequent structures from bacterial sources,<sup>79–82</sup> significant efforts have been focused towards elucidating analogous functional properties in NaV channels. The pore structure of NaVAb showed similar characteristics to  $K^+$  channels, displaying a voltage-domain attached *via* a linker domain, as shown in Figure 10.3. In this section, recent insights gained from the Na<sub>v</sub>Ab channel will be used to illustrate the application of computational methodologies to complex questions regarding ion channels.

The resolved selectivity filter structure identified three possible ion-binding sites which were apparently capable of coordinating partially hydrated ions; at the extracellular entrance, the S<sub>HFS</sub> site formed of a ring of anionic glutamate residues Glu177, was proposed to attract extracellular cations; deeper sites proposed (S<sub>IN</sub> and S<sub>CEN</sub>) were composed of carbonyl backbones Leu76 and Thr175, respectively, (Figure 10.3(d)). The observed ion dynamics in early MD simulations was consistent with predictions from structural data.<sup>83</sup> The binding sites were also confirmed as energy minima in single and multi-ion PMF conduction profiles<sup>84,85</sup> and later MD simulations on a microsecond timescale. The conformation of the residues forming the binding sites was largely unchanged irrespective of the presence of ions,<sup>86</sup> in contrast with  $K^+$  channels whose sites depend on the presence of  $K^+$  ions for stability.<sup>87–89</sup>

The precise sequence of events inducing Na<sup>+</sup> conduction was examined *via* umbrella sampling calculations, sampling conduction along the permeation axis. As expected, a deep energy well between  $-5 \text{ kcal mol}^{-1}$ <sup>90</sup> and  $-8 \text{ kcal mol}^{-1}$ <sup>85</sup> was observed at S<sub>HFS</sub>. Entrance to further sites, however, was subject to a barrier of  $4 \text{ kcal mol}^{-1}$ ,<sup>85,90</sup> suggesting movement of isolated ions was not the primary mechanism of conduction, as proposed from flux measurements.<sup>91</sup> Accompanying multi-ion profiles displayed a reduced energy barrier below  $3 \text{ kcal mol}^{-1}$  for penetration of the selectivity filter, advocating translocation of Na<sup>+</sup> ions likely occurs *via* sequential movement through doubly occupied extracellular/S<sub>HFS</sub>, S<sub>HFS</sub>/S<sub>HFS</sub>, S<sub>HFS</sub>/S<sub>CEN</sub>, S<sub>HFS</sub>/S<sub>IN</sub>, S<sub>HFS</sub>/intracellular conformations, in a so-called loosely coupled ‘knock-on’ mechanism. Additional efforts to characterise the minimum energy pathway utilising metadynamics, taking into account binding as a function of the radial distribution around the pore axis, identified an additional ‘drive-by’ mechanism of conduction.<sup>92</sup> Extensive MD simulations of  $\sim 22 \mu\text{s}$  in the

absence of a transmembrane voltage, provided direct observations of knock-on/off transitions, multi-ion configurations constituted over 90% of the simulation trajectory.<sup>93</sup> Interestingly, more than 20% were triply occupied states of the selectivity filter, which have also been observed in bias-exchange metadynamics simulations.<sup>94</sup>

The presence of such states has been associated with conformational states of the selectivity filter, divergent from the crystal structure. Multi-microsecond simulations revealed Glu177 could occupy an additional conformational state directed towards the selectivity filter, as opposed to the extracellular medium, which is thought to catalyse  $\text{Na}^+$  permeation.<sup>93</sup> Boiteux *et al.* disclosed PMF profiles with a double energy well at this site, resulting in reduced energy barriers, supporting this proposal.<sup>95</sup> An increased occupancy of the selectivity filter is observed in this case, providing further evidence of a conduction mechanism involving three ions. Studies have since proposed distinct multi-ion conduction mechanisms may be in operation during inward and outward conduction.<sup>96</sup>

The protonation state of Glu177 has also been under intense scrutiny throughout MD simulations. Multiple studies utilizing both equilibrium<sup>95</sup> and free-energy MD.<sup>84,97</sup> methodologies conclude conduction is favoured in the wholly deprotonated state, and is unfeasible when multiple residues are protonated.

The mechanism by which  $\text{Na}^+$  channels exclude other monovalent and divalent ions has been explored by comparison of single and multi-ion PMF profiles.<sup>84,85</sup> These profiles demonstrated that even though  $\text{K}^+$  ions were capable of penetrating the channel with a favourable network of coordinating ligands, a heightened barrier was identified in the plane of the Glu177 side-chains.<sup>84</sup> This observation was rationalised using geometric arguments, as  $\text{K}^+$  is unable to permeate this region unperturbed in an optimum geometry, and supported by a systematic comparison between selectivity and pore radius in the same study. The overall free-energy difference for  $\text{K}^+$  relative to  $\text{Na}^+$  was calculated as  $3 \text{ kcal mol}^{-1}$  in line with experimental permeability ratios. PMF profiles of  $\text{Ca}^{2+}$  conduction displayed distinct energy landscapes from both  $\text{Na}^+$  and  $\text{K}^+$ , suggesting further mechanisms of selectivity in the  $\text{Na}_v$  filter. Corry *et al.* proposed the desolvation energies of  $\text{Ca}^{2+}$  likely results in hindered permeation;<sup>90</sup> in addition, Ke *et al.* advocated that transfer from  $S_{\text{CEN}}$  to the central pore was energetically unfavourable, thus sustained inhabitancy of this site blocks inward conduction of  $\text{Ca}^{2+}$ .<sup>98</sup>

As well as exploring mechanistic aspects of  $\text{Na}_v$  channel conduction and selectivity, aspects of channel modulation have also been explored computationally.  $\text{Na}_v$  channels represent a putative target for local and general anaesthetics, yet the functional binding sites of such molecules, and the pathway by which they can access them are not widely understood. A multitude of experimental studies have proposed the existence of a 'hydrophobic pathway' enabling the entrance of hydrophobic molecules, when entrance is obstructed from both extracellular and intracellular vestibules.<sup>99</sup> The  $\text{Na}_v\text{Ab}$  structure revealed the presence of hydrophobic

side-portals, termed fenestrations, thought to represent such a route. Thus, the dynamics of these sites and their accessibility to drug molecules have been the focus of several studies in recent years.

Using a structural model of NaChBac, Raju *et al.* identified three binding sites of the general anaesthetic isoflurane, in the extracellular, linker and pore domains using flooding simulations, with subsequent FEP calculations to estimate the free energy of binding of each site.<sup>100</sup> Entry/exit routes were observed on the fly, including traversal of the fenestrations to the central pore site, confirming the feasibility of the portals as access pathways. The evolution of the size and dynamics of the fenestrations and hence, the size restrictions imposed on incoming drug molecules were subsequently explored by Kaczmarek and Corry.<sup>101</sup> A range of bacterial NaV channels displayed an average bottleneck radii between  $\sim 1.6$  and  $2.2$  Å, reaching a maximum bottleneck radii between  $\sim 2.6$  and  $2.8$  Å. Thus, in the maximally extended state, the portals are capable of accommodating phenyl rings, a key constituent of most sodium-channel blocking drugs.<sup>102</sup> Entrance of larger drugs would, therefore, require considerable perturbations of the fenestration or drug conformation. In Na<sub>v</sub>Ab, F203 acted as the central gate to fenestration size, illustrating that fluctuations in bottleneck radii are primarily determined by rapid side-chain rotations. The entrance of lipid molecules was also found to modulate fenestration size and dynamics. Further studies investigated access of the local anaesthetic benzocaine and the anti-epileptic drug phenytoin to the central pore; in agreement with the predicted physical constrictions, free energy maps constructed from umbrella sampling simulations,<sup>103</sup> and extensive unbiased MD simulations indicated a larger barrier for the entrance for the larger phenytoin, although both exhibited a minimum energy pathway through the lateral fenestrations.<sup>104</sup> The possible existence of lateral fenestrations in K<sup>+</sup> channels has subsequently been investigated using MD simulations, identifying tuneable openings in the two-pore domain K<sup>+</sup> channel family.<sup>105</sup>

In conclusion, MD simulations have revealed key principles of ion conduction in bacterial Na<sub>v</sub> channels. It is now understood that permeation at high-throughput requires two ions at a minimum, which are loosely coupled with each other and water molecules in the selectivity filter, in stark contrast to K<sup>+</sup> channels. Furthermore, lateral fenestrations have been confirmed as viable entry routes for small hydrophobic molecules to reach high-affinity binding sites in the central pore. The molecular determinants of conduction and selectivity of Ca<sup>2+</sup> channels are yet to be distinguished, and will likely be the subject of computational studies when high-resolution structural information is available.

## 10.5 Facilitated Diffusion by Transporters

Membrane transporters are a highly specialised class of membrane proteins, which couple substrate translocation to a variety of cellular energy sources. A wide range of chemical species are admitted passage *via* transporters,

often against the electrochemical gradient of the plasma membrane. This process is thought to occur by means of an *alternating access* mechanism, whereby the transporter interior is sequentially exposed to the intracellular and extracellular frontiers of the membrane. Primary transporters utilise energy input directly from chemical reactions, whilst secondary transporters are driven by an electrochemical gradient. MD simulations have provided considerable insights into the interrelationship between transporters and chemical driving forces, and how this is coupled to the large-scale conformational changes underlying the alternating access model. In this section, we have limited our discussion to the Leucine transporter (LeuT), which has become a prototype for structural and dynamical analyses of neurotransmitter sodium symporter (NSS) family, to illustrate the progression of our understanding in this field.

NSS's are responsible for the selective re-uptake of substrates to terminate neurotransmission at synapses.<sup>106</sup> These transporters function in a  $\text{Na}^+$  dependent manner, coupling the passive diffusion of  $\text{Na}^+$  down the electrochemical gradient, to the active transport of substrate molecules, including monoamine neurotransmitters (serotonin, dopamine, norepinephrine), amino acids and osmolytes (betaine, taorine, creatine). Eukaryotic NSS are pharmacological targets for important drugs, such as anti-depressants,<sup>107</sup> as well as psychoactive substances, such as cocaine.<sup>108</sup> Determination of the structure of the bacterial amino acid transporter LeuT elucidated the basic transmembrane architecture of NSS's,<sup>109</sup> denoted the 'LeuT fold', comprised of two five-helical bundles in an anti-parallel arrangement (Figure 10.4(a)).<sup>110</sup> A wealth of atomic resolution structural information of LeuT from *Aquifex aeolicus* has since emerged,<sup>111–114</sup> revealing various novel aspects of NSS assembly. The available structural information advocates a mechanism of alternating access in which  $\text{Na}^+$  ions and the substrate bind to an outward-facing open ( $\text{OF}_\text{O}$ ) state, and released by a series of concerted transitions between outward-facing closed ( $\text{OF}_\text{C}$ ), inward-facing closed ( $\text{IF}_\text{C}$ ) and inward-facing open ( $\text{IC}_\text{O}$ ) states, where the transporter can loop back to the initial conformation. The availability of LeuT structures has allowed in-depth investigation of its behaviour by MD simulations, as a paradigm for the NSS transporter function.

Crystallographic information of LeuT has elucidated the presence of two  $\text{Na}^+$  binding sites ( $\text{Na1}$ ,  $\text{Na2}$ ), in close proximity to the substrate-binding site ( $\text{S1}$ ) in the membrane core (Figure 10.4(b)).<sup>110</sup>  $\text{Na1}$  explicitly coordinates the substrate, alongside residues Ala22, Asn27, Thr254 and Asn286 in an octahedral arrangement. In contrast,  $\text{Na2}$  is entirely composed by LeuT residues Gly20, Val23, Ala351, Thr354 and Ser355 coordinating ions in a trigonal bipyramidal manner. In the initial computational studies, the dynamics of such sites was investigated.

In the first MD study, Celik *et al.* performed a series of unbiased MD simulations on the initial crystal structure of an  $\text{OF}_\text{C}$ , which contained the bound substrate and two  $\text{Na}^+$  ions, to explore dynamics of this LeuT complex.<sup>115</sup> The most notable interactions in the observed substrate binding

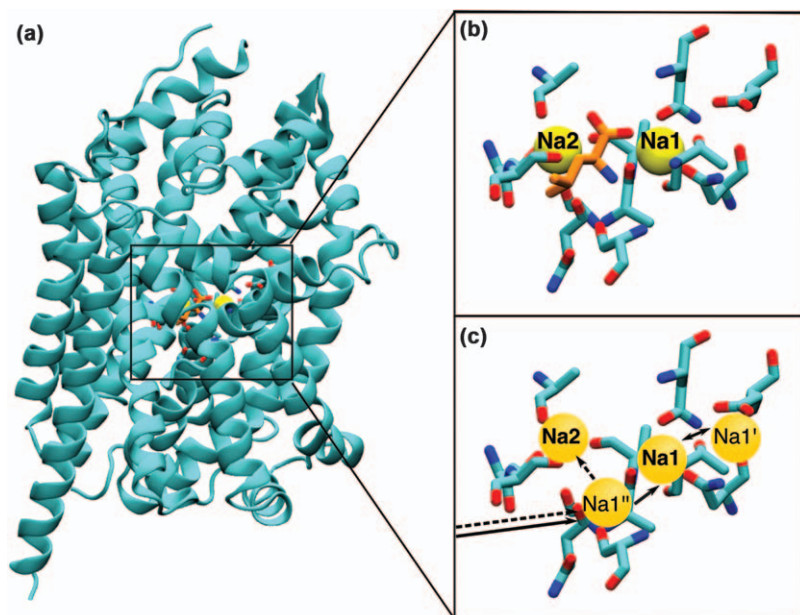


**Figure 10.4** (a) Crystal structure of LeuT receptor in OF<sub>C</sub> conformation with two Na<sup>+</sup> ions and leucine bound. (b) Close-up of substrate and ion binding sites. (c) Proposed sequence of ion binding; straight and dashed lines represent the movement of the first and second ions respectively. Throughout, protein residues are shown in liquorice representation, with blue, red, cyan and orange parts representing nitrogen, oxygen, carbon in LeuT and carbon in leucine. Sodium ions are shown as yellow spheres.

process of the zwitterionic leucine involved a salt bridge in the extracellular vestibule (Arg30 and Asp404) and an ion occupying Na1 on occupation of the canonical binding site. FEP/MD simulations were undertaken by Noskov and Roux to discern the mechanism by which other cations are excluded from Na<sup>+</sup> binding sites, and hence the specificity of LeuT to Na<sup>+</sup> ions.<sup>116</sup> Interestingly, distinct mechanisms of selectivity were observed; in Na1, coordination to a negatively charged residue results in preferred binding of Na<sup>+</sup>, whereas geometric constraints were proposed to operate in Na2. The influence of ion occupation on substrate binding was also assessed using FEP simulations, revealing optimum substrate coupling to Na<sup>+</sup> when both sites are occupied, and enhanced structural stability of the substrate and Na<sup>+</sup> site selectivity when Na2 is inhabited.<sup>117</sup>

Celik *et al.* also undertook a series of steered MD experiments to probe the formation of IF<sub>C</sub> complex. Entrance to the site required an open state of the aromatic lid (Phe353), which closed upon stable binding, thus illustrating as a key transition between OF<sub>O</sub> and OF<sub>C</sub> states.<sup>115</sup> Later unbiased simulations of the OF<sub>C</sub> state, with the substrate removed, displayed spontaneous opening to a conformation poised for substrate binding, similar to the OF<sub>O</sub> crystal

structure, demonstrating  $\text{Na}^+$  binding in isolation to biases the transporter towards this state and further elucidating the conformational changes involved in this transition.<sup>118</sup> The authors noted that the affinity for Na1 is inversely proportional with the progression of the transition of the  $\text{IF}_\text{O}$  state, thus it is likely occupation of Na2 contributes to the stability of this conformation. This has recently been confirmed by mutagenesis of the Na2 site (Thr354 and Ser355), which was found to stabilize the  $\text{IF}_\text{O}$  state by direct interactions with helices 1 and 8.<sup>119</sup>

Computational analyses have also been able to predict additional  $\text{Na}^+$  sites, not identified during crystallographic data, in order to delineate the precise series of events constituting  $\text{Na}^+$  and substrate entrance (Figure 10.4(c)). Zhao *et al.* identified an additional binding site in close proximity to Na1 located on Glu290, referred to as Na1' from this point forward.<sup>118</sup> The evolution of the protonation state of Glu290 is known to be an integral part of the transport cycle, corroborating this prediction.<sup>120</sup> Using extensive all-atom MD simulations ( $\sim 20 \mu\text{s}$ ), Zomot *et al.* identified an additional binding site, namely Na1'', constituted of Ser256 and Ser355 side-chains and the backbone carbonyl of Asn21, and have provided an exhaustive account of  $\text{Na}^+$  translocation events, and associated conformational changes.<sup>121</sup> Site Na1'' constitutes the first point of contact for  $\text{Na}^+$  ions, attracting anions from the extracellular medium throughout; within hundreds of nanoseconds, this ion shifts to Na1 where it maintains residence, or transiently occupies Na1'. The Na1'' site may remain vacant whilst Na1 is occupied, or hold a further incoming ion, which can subsequently enter Na2. Simultaneous population of both sites is consistently correlated with expansion of the extracellular entrance, permitting access of the substrate to S1. Following binding, the entrance is concealed by local rearrangements, notably the side-chain isomerization of Phe253. Subsequently, the surrounding helices undergo global rearrangements to evolve conformations capable of releasing the bound species.

In the  $\text{OF}_\text{O}$  state, Zhao and Noskov observed the formation of water wires from the cytoplasm to the S1 and Na2 sites in the  $\text{IFO}$  state, which may facilitate ion release and prompt inter-helical arrangement and flooding of the intracellular opening.<sup>122</sup> Using free energy calculations, Thr345 was shown to occupy single rotameric state when Na2 is occupied, as opposed to two degenerate states when it is vacant, and thus may act as a switch to vacate Na2 and advance the transport cycle.

Significant efforts were directed towards prediction of further intermediates in the transport cycle, such as the  $\text{IF}_\text{O}$  structure.<sup>123–125</sup> Tajkhorshid *et al.* produced an  $\text{IF}_\text{O}$  model by homology modelling of the inward-facing state,<sup>126</sup> based on the structures of outward-facing LeuT<sup>110</sup> and inward facing vSGLT.<sup>127</sup> Simulations of this state consistently displayed intracellular release of Na2. Quick *et al.* proposed a similar structure, in addition to other unknown states, by varying the presence of leucine in the binding site and  $\text{Na}^+$  ions in known crystal structures and performing accelerated MD simulations in combination with principal component analysis.<sup>123</sup> All



inward-facing conformations were found to contain a vacant Na2 site. The eventual crystallization of a LeuT inward-facing state<sup>111</sup> further corroborated earlier predictions that Na2 initiates intracellular release; Na<sup>+</sup> ions were weakly coupled with the Na2 site, persistently exiting to the intracellular medium and initiating release of the bound substrate and ion in Na1 in the same direction.<sup>128</sup> The protonated/neutral state of Glu290 (Na1' site) in the inward-open state is proposed to enhance dissociation of ions to the intracellular solution.<sup>129</sup> These observations provide a mechanistic understanding of how negative charges in close proximity to the characteristic Na<sup>+</sup> binding sites, such as Glu290 in the LeuT transporter or chloride ions in eukaryotic NSS, may regulate ion binding and release.<sup>120</sup>

The overall mechanism of release has been a source of controversy throughout the literature. Early SMD simulations by Shi *et al.* revealed a second substrate-binding site (S2), leading to proposals of an allosteric mechanism of transport whereby occupation of the secondary binding site, triggers release of the inhabitants of S1 and Na1.<sup>130</sup> Furthermore, the observed overlap of the S2 site with the binding site of tricyclic antidepressants has led to suggestions that the S2 site could exert an activator or inhibitory effect dependent on the manner of binding.<sup>112,113</sup>

Championed by Javitch and Weinstein,<sup>130–132</sup> the presence of the S2 binding site has been widely disputed throughout the literature, with Gauaux *et al.*, for example, endorsing the functional significance of the S1 site only.<sup>110,133–136</sup> Differing reports have also emerged in computational studies. Extensive unbiased simulations by Zomot *et al.*, do not observe alanine binding, and only partial leucine binding in the S2 site.<sup>121</sup> However, using a complex protocol combining of accelerated, targeted and conventional MD, Cheng *et al.* have identified an increase in the substrate-binding affinity of the S2 site whilst progressing towards an inward-facing state and subsequent displacements in the putative S1 site.<sup>137</sup> Furthermore, unbiased simulations with Ala and Leu and inhibitors known inhibitors bound at the S2 site revealed key structural alterations in the extracellular portion of TM6 which may be propagated throughout the transporter and influence its functional state.<sup>138</sup> FEP/MD simulations yielded favourable absolute binding free energies for tricyclic antidepressants in the range of  $-12$  to  $-14$  kcal mol<sup>-1</sup>.<sup>139</sup> These energies were dissimilar on removal of the substrate, supporting a thermodynamic coupling mechanism between the two sites.<sup>112</sup>

The final question remaining of LeuT transport involves the inward-to-outward transition to restart the translocation mechanism. The recent crystallisation of WT LeuT in a Na<sup>+</sup> and substrate-free state by Malinauskaite *et al.* provided the first insights into the structure of an intermediate involved in this transformation.<sup>114</sup> MD simulations confirmed both Na<sup>+</sup> sites are inaccessible to extracellular ions in the structure, with a distinct Na1 conformation observed stable throughout. Highly conserved residue, Leu25, is found to consistently occupy the S1 site, and occlude Glu290 from the extracellular environment, reportedly acting as gatekeeper for Na<sup>+</sup> binding



and playing an intimate role in  $H^+$ -counter-transport during the return transition. The authors propose release of the counter-ion stimulates re-orientation of Leu25, allowing entry to the sites and initiation of the forward transport cycle.

Remarkably, using the available crystal structures of LeuT, MD simulations have provided a full atomistic description of the entire transport cycle. Key questions concerning the locality and behaviour of  $Na^+$  binding sites, the interrelationship between such sites and substrate binding at different stages of the transport cycle, as well as a possible allosteric mechanism involving a second substrate-binding site have been addressed.

In a wider context, these studies have examined the key principles of membrane active transport: molecular determinants of substrate binding, coupling to external energy sources and conformational changes constituting the alternative access mechanism. These fundamental principles of active transport have been examined in a number of transporter families, and it is likely that a detailed description of active transport in these assemblies will emerge in the future.

## 10.6 Signalling *via* Receptors

Cell surface receptors are intrinsic membrane proteins forming the primary communication mechanism between the cell exterior and interior. The functionality of receptors is dependent on the transformation of the protein transmembrane between conformational states, in response to the binding of extracellular molecules. Ligand-gated ion channels are a family of membrane receptors that allow ion influx/efflux in response to bound neurotransmitters. G-protein coupled receptors (GPCR's) constitute a significant class of membrane receptors, which couple to heterotrimeric G-proteins on activation to initiate intracellular signalling cascades. GPCR's are susceptible to hormones, neurotransmitters, and sensory stimuli, thus they are critical for basic physiological function of eukaryotic organisms.<sup>140</sup> As a consequence, GPCR's have been implicated in neurological disorders, cardiac failure, cancer and diabetes, and are the target of a significant proportion of pharmaceuticals available on the market today.<sup>141,142</sup>

As GPCR activation is a profoundly dynamic process reliant on large-scale conformational changes, elucidation of high-resolution three-dimensional structures has proven difficult.<sup>143</sup> The ground breaking crystal structure of light sensitive pigment, rhodopsin, revealed the conserved transmembrane arrangement of seven  $\alpha$ -helices of class A GPCR's and provided high-resolution data appropriate for MD simulations. In recent years, computational analyses of GPCR's has been significantly aided by the increase in X-ray crystallographic information and state-of-art homology modelling tools, exploring numerous facets of GPCR behaviour, such as ligand binding, G-protein coupling, lipid modulation the highly concerted conformational changes associated with activation/inactivation processes. The  $\beta_2$ -adrenergic receptor has been the focus of seminal studies exploring the

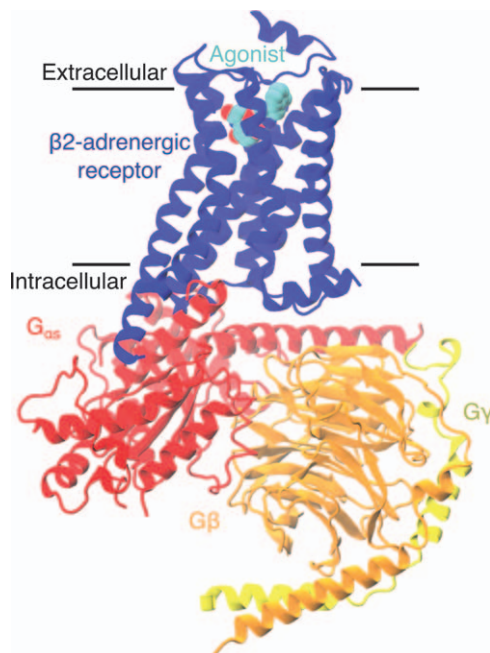


**Figure 10.5** Crystal structure of  $\beta 2$ -adrenergic receptor (PDB ID 3SN6) in fully activated state with bound agonist and heterotrimeric G-protein. The position of the plasma membrane is indicated by a solid black line.

latter, due to the availability of structural information in multiple functional states, including a fully activated state in complex with extracellular agonist and intracellular G-protein complex (Figure 10.5). Thus, it will be used as an archetypal example of how computational methodologies have advanced our understanding of the relationship between GPCR structure and function.

Multiple studies have advocated that an ensemble of conformational states exist for individual GPCR functional states.<sup>144,145</sup> The initial  $\beta 2$ AR structures, in complex with inverse agonist carazolol and timolol,<sup>146–149</sup> allowed theoreticians to characterise the ensemble of conformational states representing the inactive state of the receptor.<sup>150</sup> Unexpectedly, a salt bridge between the intracellular ends of helices III and VI, dubbed the ‘ionic lock’, was broken in the inactive state structures, in defiance of biochemical evidence arguing this event represents a crucial activation step.<sup>151</sup> Microsecond simulations of these structures were able to demonstrate that inactive  $\beta 2$ AR actually exists in conformational equilibrium between states with a broken and intact ionic lock.<sup>150</sup> These observations suggest the receptor is likely biased towards the broken state upon receptor activation, reconciling with the previous experimental work mentioned.<sup>151</sup>

The elucidation of the agonist bound  $\beta 2$ A receptor lead to further investigation of GPCR conformational states and the transitions between them. Rosenbaum *et al.* explored the dynamics of this complex on an

extended timescale using the Anton Supercomputer.<sup>152</sup> This seminal study demonstrated the feasibility of simulations up to 30  $\mu$ s in length for the study of integral proteins, and provided unprecedented insights into the behaviour of this complex. After approximately 11  $\mu$ s, the active state receptor spontaneously transitioned to the inactive structure that remained stable for the duration of the simulations, advocating that binding to an intracellular partner was absolutely required for stabilisation of a fully activated conformation. Later experimental and computational works supported this hypothesis.<sup>153</sup>

Dror *et al.* proposed an atomically detailed activation mechanism based on the transition between functionally active and inactive states, *via* a number of previously unreported intermediates, observed reproducibly in over 30 independent simulations.<sup>154</sup> Three functionally important regions were identified, the intracellular G-protein-binding site, the extracellular ligand-binding site and the junction in between, the so-called 'connector' region. The regions are loosely coupled, although the connector region is largely responsible for communicating small-scale ligand related movements and large-scale helical movements surrounding the intracellular binding pocket. Interestingly, the simulations suggest the activation process originates from the latter; outward movements of helix VI initiate conversion to an intermediate state, which incorporates an expanded G-protein site, and an equilibrium between active and inactive states in the ligand binding site and connector region. Subsequent agonist binding biases this equilibrium towards active conformations. Finally, an intracellular binding partner may interact and trigger the final stage of the activation process. Detailed understanding of specific receptor conformations, as described here, may contribute to the design of drugs targeting specific functional states of GPCR's. Knowledge of the interrelationship between cellular entities, such as G-proteins and arrestins, and the multitude of conformational states of the  $\beta$ 2AR will refine this scheme further.

A key question that has not been addressed throughout these mechanistic studies is the assembly of GPCR dimers and higher-order oligomers, and their functional role of in the intracellular signalling pathway. Hydrophobic mismatch, where the length of the membrane spanning segments conflicts with the length of the hydrophobic core of the membrane, has been proposed to stimulate GPCR dimerisation, and has consequently been explored using MD simulations.<sup>155</sup> Using a novel multi-scale approach known as continuum-molecular dynamics, Mondal *et al.* quantified energetic penalties emerging from this phenomenon in individual transmembrane helices of GPCR's, enabling the prediction of energetically favourable contact interfaces by comparison of such energetic costs in monomeric and oligomeric states.<sup>156</sup> In the  $\beta$ 2A receptor, the hydrophobic mismatch was considerably reduced in the transmembrane helices I, IV and V consistent with the typical interfacial regions elucidated from unbiased CGMD (helices I-I, V-V and IV-V),<sup>157</sup> with similar results obtained using a bilayer containing 10% cholesterol. Duplicate analysis of the  $\beta$ 1A receptor displayed distinct

oligomerisation patterns, with a single predicted contact interface localized on helix I. Overall, these results were akin to experimental evidence showing the  $\beta$ 1A preferentially forms dimers,<sup>158</sup> whereas the  $\beta$ 2A receptor can form more extensive dimers as well as higher-order oligomers.<sup>159</sup>

Prasanna *et al.* performed extensive CGMD of the  $\beta$ 2A receptor in bilayers containing increasing concentrations of cholesterol to examine the effect of cholesterol on the dimerisation process.<sup>160</sup> In 0% cholesterol, the receptors form a homo-interface involving helices IV and V; in 50% cholesterol, a homo-interface between helices II and I is formed; whilst in the concentrations between (9 and 30%) a hetero-interface is observed comprising a combination of the two. Increased cholesterol occupancy at helix V is observed throughout, advocating specific cholesterol interactions bias the oligomer towards distinct oligomeric states.

It is well established throughout the literature that functionality of integral membrane proteins is dependent on the composition of the plasma membrane, thus it is possible the influence of membrane organisation extends further than GPCR dimerisation. Many GPCR's have demonstrated a functional dependence on membrane cholesterol in particular, which is considered to act either by direct interactions or indirect effects involving the biophysical properties of the membrane.

Unbiased MD simulations, either CG or atomistic, of GPCR's embedded in lipid bilayers enriched with cholesterol molecules have become a popular tool to identify cholesterol binding sites.<sup>161–164</sup> Utilising this protocol, Cang *et al.* revealed the presence of three extracellular and four intracellular high occupancy sites. The accuracy of such predictions was supported by the likeness between three sites and those observed in crystal structures of GPCR's with bound cholesterol molecules. Of note, two cholesterol molecules occupy the surface of helix I and VIII, consistent with that observed in the dimeric structure of the  $\beta$ 2AR (PDB 2RH1), contrasting with claims that the observed organization was an artefact caused by crystal packing.<sup>147</sup> Furthermore, a cleft between the extracellular ends of helices I, II and VII, accommodated a cholesterol molecule in a stable manner for the duration of the trajectories. The authors suggest habitation of this site, and the resulting stabilisation of residue Trp313, may facilitate ligand-receptor binding.

Prasanna *et al.* identified a specific POPC binding site between helices I and VII, in a similar position in the extracellular leaflet.<sup>165</sup> This site had been previously suggested to accommodate lipids from X-ray crystallographic information of the  $\alpha$ 2-adenosine receptor. Neale *et al.* explored the relationship between specific phospholipid interactions and receptor activation state.<sup>161</sup> Using MD simulations with a cumulative time of 0.25 ms, the authors observed individual phospholipid molecules entering the receptor by an opening in the cytoplasmic leaflet between helices VI and VII and forming a salt-bridge that directly impedes ionic lock formation, thus stabilising the active state. This phenomenon was enhanced by the presence of anionic lipids, elucidating a possible mechanism by which such lipids perpetuate receptor activation.

In conclusion, MD has provided crucial insights into how GPCR's convert between functional states to convey signals, as a result of highly synchronised structural transitions emanating from atomic fluctuations. The driving forces of receptor dimerisation have been studied, and the functional significance of higher order oligomers considered. The close association between membrane lipids and receptor state has been explored, with the identification of multiple functional binding sites of both cholesterol and phospholipid molecules.

## 10.7 Conclusions

Molecular dynamics simulations have become an essential tool to study the dynamics of biological systems in atomic resolution, and elucidate the molecular mechanisms of numerous phenomena that cannot be gained by experimental means. A detailed understanding of how many small molecules interact with lipid bilayers, and consequential diffusion or exclusion processes has been gained, which can be applied in a medicinal chemistry context to assess the likelihood of drug molecules reaching intracellular targets. Furthermore, computational methodologies in combination with high-resolution structural information of integral membrane proteins have provided crucial insights into how such assemblies function as highly efficient transport machinery. For ion channels, multiple conduction mechanisms for selective transport through  $K^+$  and  $Na^+$  channels have emerged, as well as characterisation of a novel pore access pathway through lateral fenestrations in the latter. In the case of GPCR's, conformational ensembles of active and inactive functional states have been described, and the inter-conversion transitions delineated, providing a general mechanism of GPCR activation. The influence of the membrane environment on GPCR dynamics by direct and indirect effects has been explored intensively. Finally, significant advances have been made in understanding the alternating access mechanism underlying transporter function, and how such changes are driven by cellular energy sources. Overall, such insights have considerably contributed to our understanding of membrane transport, and will significantly advance the rational design of drugs in the future. The increasing availability of high-resolution structural information, growth in computer capabilities and development of state-of-the art MD algorithms and accompanying force fields will markedly amplify the use of computational simulations for the study of intrinsic membrane proteins in the coming years.

## Acknowledgements

V.O. acknowledges a BBSRC-CASE studentship, in association with Pfizer Neusentis. C.D. acknowledges use of ARCHER, the UK National Supercomputing Service, the National Service for Computational Chemistry Software (NSCCS), the 'Red Española de Supercomputación' (RES), and the

Hartree Center. Research in the Domene group is supported by the Engineer and Physical Sciences Research Council (EPSRC) and the Biotechnology and Biological Sciences Research Council (BBSRC).

## References

1. M. Luckey, *Membrane Structural Biology: With Biochemical and Biophysical Foundations*, Cambridge University Press, 2014.
2. B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus, CHARMM: A program for macromolecular energy, minimization, and dynamics calculations, *J. Comput. Chem.*, 1983, **4**(2), 187–217.
3. W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell and P. A. Kollman, A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules, *J. Am. Chem. Soc.*, 1995, **117**(19), 5179–5197.
4. W. L. Jorgensen and J. Tirado-Rives, The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin, *J. Am. Chem. Soc.*, 1988, **110**(6), 1657–1666.
5. I. V. Vorobyov, V. M. Anisimov and A. D. MacKerell, Polarizable Empirical Force Field for Alkanes Based on the Classical Drude Oscillator Model, *J. Phys. Chem. B*, 2005, **109**(40), 18988–18999.
6. S. Patel and C. L. Brooks, CHARMM fluctuating charge force field for proteins: I parameterization and application to bulk organic liquid simulations, *J. Comput. Chem.*, 2004, **25**(1), 1–16.
7. S. Patel, A. D. Mackerell Jr. and C. L. Brooks, 3rd, CHARMM fluctuating charge force field for proteins: II protein/solvent properties from molecular dynamics simulations using a nonadditive electrostatic model, *J. Comput. Chem.*, 2004, **25**(12), 1504–1514.
8. Y. Shi, Z. Xia, J. Zhang, R. Best, C. Wu, J. W. Ponder and P. Ren, Polarizable Atomic Multipole-Based AMOEBA Force Field for Proteins, *J. Chem. Theory Comput.*, 2013, **9**(9), 4046–4063.
9. G. Lamoureux, A. D. MacKerell and B. Roux, A simple polarizable model of water based on classical Drude oscillators, *J. Chem. Phys.*, 2003, **119**(10), 5185–5197.
10. J. A. Lemkul, J. Huang, B. Roux and A. D. MacKerell, An Empirical Polarizable Force Field Based on the Classical Drude Oscillator Model: Development History and Recent Applications, *Chem. Rev.*, 2016, **116**(9), 4983–5013.
11. S. Miyamoto and P. A. Kollman, SETTLE: an analytical version of the SHAKE and RATTLE algorithm for rigid water models, *J. Comput. Chem.*, 1992, **13**(8), 952–962.
12. J. R. Perilla, B. C. Goh, C. K. Cassidy, B. Liu, R. C. Bernardi, T. Rudack, H. Yu, Z. Wu and K. Schulten, Molecular dynamics simulations of large macromolecular complexes, *Curr. Opin. Struct. Biol.*, 2015, **31**, 64–74.



13. D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan and W. Wriggers, Atomic-Level Characterization of the Structural Dynamics of Proteins, *Science*, 2010, **330**(6002), 341–346.
14. J. C. Shelley, M. Y. Shelley, R. C. Reeder, S. Bandyopadhyay and M. L. Klein, A Coarse Grain Model for Phospholipid Simulations, *J. Phys. Chem. B*, 2001, **105**(19), 4464–4470.
15. J. C. Shelley, M. Y. Shelley, R. C. Reeder, S. Bandyopadhyay, P. B. Moore and M. L. Klein, Simulations of Phospholipids Using a Coarse Grain Model, *J. Phys. Chem. B*, 2001, **105**(40), 9785–9792.
16. S. J. Marrink, A. H. de Vries and A. E. Mark, Coarse Grained Model for Semiquantitative Lipid Simulations, *J. Phys. Chem. B*, 2004, **108**(2), 750–760.
17. C. A. Lopez, Z. Sovova, F. J. van Eerden, A. H. de Vries and S. J. Marrink, Martini Force Field Parameters for Glycolipids, *J. Chem. Theory Comput.*, 2013, **9**(3), 1694–1708.
18. M. G. Saunders and G. A. Voth, Coarse-graining methods for computational biology, *Annu. Rev. Biophys.*, 2013, **42**, 73–93.
19. R. W. Zwanzig, High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases, *J. Chem. Phys.*, 1954, **22**, 1420.
20. D. L. Beveridge and F. M. DiCapua, Free Energy Via Molecular Simulation: Applications to Chemical and Biomolecular Systems, *Annu. Rev. Biophys. Biophys. Chem.*, 1989, **18**(1), 431–492.
21. G. M. Torrie and J. P. Valleau, Monte Carlo free energy estimates using non-Boltzmann sampling: Application to the sub-critical Lennard-Jones fluid, *Chem. Phys. Lett.*, 1974, **28**(4), 578–581.
22. S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen and P. A. Kollman, The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method, *J. Comput. Chem.*, 1992, **13**(8), 1011–1021.
23. M. Souaille and B. T. Roux, Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations, *Comput. Phys. Commun.*, 2001, **135**(1), 40–57.
24. J. Kästner and W. Thiel, Bridging the gap between thermodynamic integration and umbrella sampling provides a novel analysis method: “Umbrella integration”, *J. Chem. Phys.*, 2005, **123**(14), 144104.
25. S. Park and K. Schulten, Calculating potentials of mean force from steered molecular dynamics simulations, *J. Chem. Phys.*, 2004, **120**(13), 5946–5961.
26. S. Park, F. Khalili-Araghi, E. Tajkhorshid and K. Schulten, Free energy calculation from steered molecular dynamics simulations using Jarzynski’s equality, *J. Chem. Phys.*, 2003, **119**(6), 3559–3566.
27. E. Darve and A. Pohorille, Calculating free energies using average force, *J. Chem. Phys.*, 2001, **115**(20), 9169–9183.
28. E. Overton, *Vierteljahrsschr. Naturforsch. Ges. Zurich*, 1896, **41**, 383.



29. A. Finkelstein, Water and nonelectrolyte permeability of lipid bilayer membranes, *J. Gen. Physiol.*, 1976, **68**(2), 127–135.
30. J. M. Diamond; Y. Katz, Interpretation of nonelectrolyte partition coefficients between dimyristoyl lecithin and water, *J. Membr. Biol.*, 1974, **17**(1), 121–154.
31. S. J. Marrink and H. J. C. Berendsen, Simulation of water transport through a lipid membrane, *J. Phys. Chem.*, 1994, **98**(15), 4155–4168.
32. G. Parisio, M. Stocchero and A. Ferrarini, Passive Membrane Permeability: Beyond the Standard Solubility-Diffusion Model, *J. Chem. Theory Comput.*, 2013, **9**(12), 5236–5246.
33. J. Comer, K. Schulten and C. Chipot, Calculation of Lipid-Bilayer Permeabilities Using an Average Force, *J. Chem. Theory Comput.*, 2014, **10**(2), 554–564.
34. J. Comer, K. Schulten and C. Chipot, Diffusive Models of Membrane Permeation with Explicit Orientational Freedom, *J. Chem. Theory Comput.*, 2014, **10**(7), 2710–2718.
35. S. Riahi and C. N. Rowley, Why Can Hydrogen Sulfide Permeate Cell Membranes?, *J. Am. Chem. Soc.*, 2014, **136**(43), 15111–15113.
36. G. Lamoureux, E. Harder, I. V. Vorobyov, B. Roux, MacKerell Jr. and A. D. , A polarizable model of water for molecular dynamics simulations of biomolecules, *Chem. Phys. Lett.*, 2006, **418**(1–3), 245–249.
37. S. Riahi, C. N. Rowley and A. Drude, Polarizable Model for Liquid Hydrogen Sulfide, *J. Phys. Chem. B*, 2013, **117**(17), 5222–5229.
38. S. Riahi and C. N. Rowley, Solvation of Hydrogen Sulfide in Liquid Water and at the Water–Vapor Interface Using a Polarizable Force Field, *J. Phys. Chem.*, 2014, **118**(5), 1373–1380.
39. J. Chowdhary, E. Harder, P. E. M. Lopes, L. Huang, A. D. MacKerell and B. Roux, A Polarizable Force Field of Dipalmitoylphosphatidylcholine Based on the Classical Drude Model for Molecular Dynamics Simulations of Lipids, *J. Phys. Chem. B*, 2013, **117**(31), 9142–9160.
40. E. Terreno, A. Sanino, C. Carrera, D. D. Castelli, G. B. Giovenzana, A. Lombardi, R. Mazzon, L. Milone, M. Visigalli and S. Aime, Determination of water permeability of paramagnetic liposomes of interest in MRI field, *J. Inorg. Biochem.*, 2008, **102**(5–6), 1112–1119.
41. D. Bemporad, J. W. Essex and C. Luttmann, Permeation of Small Molecules through a Lipid Bilayer: A Computer Simulation Study, *J. Phys. Chem. B*, 2004, **108**(15), 4875–4884.
42. J. C. Mathai, A. Missner, P. Kügler, S. M. Saparov, M. L. Zeidel, J. K. Lee and P. Pohl, No facilitator required for membrane transport of hydrogen sulfide, *Proc. Natl. Acad. Sci.*, 2009, **106**(39), 16633–16638.
43. W. K. Subczynski, J. S. Hyde and A. Kusumi, Oxygen permeability of phosphatidylcholine–cholesterol membranes, *Proc. Natl. Acad. Sci. U. S. A.*, 1989, **86**(12), 4474–4478.
44. R. O. Beauchamp Jr., J. S. Bus, J. A. Popp, C. J. Boreiko and D. A. Andjelkovich, A critical review of the literature on hydrogen sulfide toxicity, *Crit. Rev. Toxicol.*, 1984, **13**(1), 25–97.

45. M. Orsi and J. W. Essex, Permeability of drugs and hormones through a lipid bilayer: insights from dual-resolution molecular dynamics, *Soft Matter*, 2010, **6**(16), 3797–3808.
46. M. Palonc'ová, R. DeVane, B. Murch, K. Berka and M. Otyepka, Amphiphilic Drug-Like Molecules Accumulate in a Membrane below the Head Group Region, *J. Phys. Chem. B*, 2014, **118**(4), 1030–1039.
47. H. Wang, X. Ren and F. Meng, Molecular dynamics simulation of six  $\beta$ -blocker drugs passing across POPC bilayer, *Mol. Simul.*, 2016, **42**(1), 56–63.
48. M. Kang and S. M. Loverde, Molecular Simulation of the Concentration-Dependent Interaction of Hydrophobic Drugs with Model Cellular Membranes, *J. Phys. Chem. B*, 2014, **118**(41), 11965–11972.
49. M. Ashrafuzzaman, C. Y. Tseng, M. Duszyk and J. A. Tuszynski, Chemotherapy drugs form ion pores in membranes due to physical interactions with lipids, *Chem. Biol. Drug Des.*, 2012, **80**(6), 992–1002.
50. R. Notman and J. Anwar, Breaching the skin barrier — Insights from molecular simulation of model membranes, *Adv. Drug Delivery Rev.*, 2013, **65**(2), 237–250.
51. A. N. C. Anigbogu, A. C. Williams, B. W. Barry and H. G. M. Edwards, Fourier transform raman spectroscopy of interactions between the penetration enhancer dimethyl sulfoxide and human stratum corneum, *Int. J. Pharm.*, 1995, **125**(2), 265–282.
52. A. C. Williams and B. W. Barry, Penetration enhancers, *Adv. Drug Delivery Rev.*, 2004, **56**(5), 603–618.
53. P. Kumar, S. K. Singh, D. N. Mishra and P. Girotra, Enhancement of ketorolac tromethamine permeability through rat skin using penetration enhancers: An ex-vivo study, *Int. J. Pharm. Invest.*, 2015, **5**(3), 142–146.
54. B. W. Barry, Mode of action of penetration enhancers in human skin, *J. Controlled Release*, 1987, **6**(1), 85–97.
55. R. Notman, M. Noro, B. O'Malley and J. Anwar, Molecular Basis for Dimethylsulfoxide (DMSO) Action on Lipid Membranes, *J. Am. Chem. Soc.*, 2006, **128**(43), 13982–13983.
56. A. A. Gurtovenko and J. Anwar, Modulating the structure and properties of cell membranes: the molecular mechanism of action of dimethyl sulfoxide, *J. Phys. Chem. B*, 2007, **111**(35), 10453–10460.
57. Z. E. Hughes, A. E. Mark and R. L. Mancera, Molecular Dynamics Simulations of the Interactions of DMSO with DPPC and DOPC Phospholipid Membranes, *J. Phys. Chem. B*, 2012, **116**(39), 11911–11923.
58. A. P. Dabkowska, F. Foglia, M. J. Lawrence, C. D. Lorenz and S. E. McLain, On the solvation structure of dimethylsulfoxide/water around the phosphatidylcholine head group in solution, *J. Chem. Phys.*, 2011, **135**(22), 225105.
59. A. A. Gurtovenko and J. Anwar, Interaction of Ethanol with Biological Membranes: The Formation of Non-bilayer Structures within the

- Membrane Interior and their Significance, *J. Phys. Chem. B*, 2009, **113**(7), 1983–1992.
60. Y. O. Posokhov and A. Kyrychenko, Effect of acetone accumulation on structure and dynamics of lipid membranes studied by molecular dynamics simulations, *Comput. Biol. Chem.*, 2013, **46**, 23–31.
  61. R. Notman, M. G. Noro and J. Anwar, Interaction of Oleic Acid with Dipalmitoylphosphatidylcholine (DPPC) Bilayers Simulated by Molecular Dynamics, *J. Phys. Chem. B*, 2007, **111**(44), 12748–12755.
  62. N. W. John and B. James, Voltage-Gated Sodium Channel Blockers; Target Validation and Therapeutic Potential, *Curr. Top. Med. Chem.*, 2005, **5**(6), 529–537.
  63. S. K. Bagal, M. L. Chapman, B. E. Marron, R. Prime, R. I. Storer and N. A. Swain, Recent progress in sodium channel modulators for pain, *Bioorg. Med. Chem. Lett.*, 2014, **24**(16), 3690–3699.
  64. D. A. Doyle, J. Morais Cabral, R. A. Pfuetzner, A. Kuo, J. M. Gulbis, S. L. Cohen, B. T. Chait and R. MacKinnon, The structure of the potassium channel: molecular basis of  $K^+$  conduction and selectivity, *Science*, 1998, **280**(5360), 69–77.
  65. O. S. Smart, J. M. Goodfellow and B. A. Wallace, The pore dimensions of gramicidin A, *Biophys. J.*, 1993, **65**(6), 2455–2460.
  66. L. Heginbotham, Z. Lu, T. Abramson and R. MacKinnon, Mutations in the  $K^+$  channel signature sequence, *Biophys. J.*, 1994, **66**(4), 1061–1067.
  67. A. Parsegian, Energy of an Ion crossing a Low Dielectric Membrane: Solutions to Four Relevant Electrostatic Problems, *Nature*, 1969, **221**(5183), 844–846.
  68. S. Berneche and B. Roux, Energetics of ion conduction through the  $K^+$  channel, *Nature*, 2001, **414**(6859), 73–77.
  69. S. Y. Noskov, S. Berneche and B. Roux, Control of ion selectivity in potassium channels by electrostatic and dynamic properties of carbonyl ligands, *Nature*, 2004, **431**(7010), 830–834.
  70. J. Aqvist and V. Luzhkov, Ion permeation mechanism of the potassium channel, *Nature*, 2000, **404**(6780), 881–884.
  71. S. Furini and C. Domene, Atypical mechanism of conduction in potassium channels, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**(38), 16074–16077.
  72. D. A. Köpfer, C. Song, T. Gruene, G. M. Sheldrick, U. Zachariae and B. L. de Groot, Ion permeation in  $K^+$  channels occurs by direct Coulomb knock-on, *Science*, 2014, **346**(6207), 352–355.
  73. C. Jorgensen, L. Darré, K. Vanommeslaeghe, K. Omoto, D. Pryde and C. Domene, In Silico Identification of PAP-1 Binding Sites in the Kv1.2 Potassium Channel, *Mol. Pharm.*, 2015, **12**(4), 1299–1307.
  74. Y. Li, F. Barbault, M. Delamar, R. Zhang and R. Hu, Targeted molecular dynamics (TMD) of the full-length KcsA potassium channel: on the role of the cytoplasmic domain in the opening process, *J. Mol. Model.*, 2013, **19**(4), 1651–1666.

75. L. Monticelli, K. M. Robertson, J. L. MacCallum and D. P. Tieleman, Computer simulation of the KvAP voltage-gated potassium channel: steered molecular dynamics of the voltage sensor, *FEBS Lett.*, 2004, **564**(3), 325–332.
76. B. Roux, Ion Conduction and Selectivity in  $K^+$  Channels, *Annu. Rev. Biophys. Biomol. Struct.*, 2005, **34**(1), 153–171.
77. S. Furini and C. Domene,  $K^+$  and  $Na^+$  Conduction in Selective and Nonselective Ion Channels Via Molecular Dynamics Simulations, *Biophys. J.*, 2013, **105**(8), 1737–1745.
78. J. Payandeh, T. Scheuer, N. Zheng and W. A. Catterall, The crystal structure of a voltage-gated sodium channel, *Nature*, 2011, **475**(7356), 353–358.
79. J. Payandeh, T. M. Gamal El-Din, T. Scheuer, N. Zheng and W. A. Catterall, Crystal structure of a voltage-gated sodium channel in two potentially inactivated states, *Nature*, 2012, **486**(7401), 135–139.
80. E. C. McCusker, C. Bagneris, C. E. Naylor, A. R. Cole, N. D'Avanzo, C. G. Nichols and B. A. Wallace, Structure of a bacterial voltage-gated sodium channel pore reveals mechanisms of opening and closing, *Nat. Commun.*, 2012, **3**, 1102.
81. C. Bagn  ris, P. G. DeCaen, B. A. Hall, C. E. Naylor, D. E. Clapham, C. W. M. Kay and B. A. Wallace, Role of the C-terminal domain in the structure and function of tetrameric sodium channels, *Nat. Commun.*, 2013, **4**, 2465.
82. D. Shaya, F. Findeisen, F. Abderemane-Ali, C. Arrigoni, S. Wong, S. R. Nurva, G. Loussouarn, Minor Jr. and D. L. , Structure of a Prokaryotic Sodium Channel Pore Reveals Essential Gating Elements and an Outer Ion Binding Site Common to Eukaryotic Channels, *J. Mol. Biol.*, 2014, **426**(2), 467–483.
83. V. Carnevale, W. Treptow and M. L. Klein, Sodium Ion Binding Sites and Hydration in the Lumen of a Bacterial Ion Channel from Molecular Dynamics Simulations, *J. Phys. Chem. Lett.*, 2011, **2**(19), 2504–2508.
84. B. Corry and M. Thomas, Mechanism of Ion Permeation and Selectivity in a Voltage Gated Sodium Channel, *J. Am. Chem. Soc.*, 2012, **134**(3), 1840–1846.
85. S. Furini and C. Domene, On conduction in a bacterial sodium channel, *PLoS Comput. Biol.*, 2012, **8**(4), e1002476.
86. H. Qiu, R. Shen and W. Guo, Ion solvation and structural stability in a sodium channel investigated by molecular dynamics calculations, *Biochim. Biophys. Acta*, 2012, **1818**(11), 2529–2535.
87. C. Domene, M. L. Klein, D. Branduardi, F. L. Gervasio and M. Parrinello, Conformational changes and gating at the selectivity filter of potassium channels, *J. Am. Chem. Soc.*, 2008, **130**(29), 9474–9480.
88. Y. Zhou, J. H. Morais-Cabral, A. Kaufman and R. MacKinnon, Chemistry of ion coordination and hydration revealed by a  $K^+$  channel-Fab complex at 2.0   resolution, *Nature*, 2001, **414**, 43–48.
89. C. Domene and S. Furini, Dynamics, energetics, and selectivity of the low- $K^+$  KcsA channel structure, *J. Mol. Biol.*, 2009, **389**(3), 637–645.

90. B. Corry,  $\text{Na}^+/\text{Ca}^{2+}$  selectivity in the bacterial voltage-gated sodium channel NavAb, *PeerJ*, 2013, **1**, e16.
91. D. J. Benos, B. A. Hyde and R. Latorre, Sodium flux ratio through the amiloride-sensitive entry pathway in frog skin, *J. Gen. Physiol.*, 1983, **81**(5), 667–685.
92. L. Stock, L. Delemotte, V. Carnevale, W. Treptow and M. L. Klein, Conduction in a Biological Sodium Selective Channel, *J. Phys. Chem. B*, 2013, **117**(14), 3782–3789.
93. N. Chakrabarti, C. Ing, J. Payandeh, N. Zheng, W. A. Catterall and R. Pomès, Catalysis of  $\text{Na}^+$  permeation in the bacterial sodium channel NavAb, *Proc. Natl. Acad. Sci.*, 2013, **110**(28), 11331–11336.
94. C. Domene, P. Barbini and S. Furini, Bias-Exchange Metadynamics Simulations: An Efficient Strategy for the Analysis of Conduction and Selectivity in Ion Channels, *J. Chem. Theory Comput.*, 2015, **11**(4), 1896–1906.
95. C. Boiteux, I. Vorobyov and T. W. Allen, Ion conduction and conformational flexibility of a bacterial voltage-gated sodium channel, *Proc. Natl. Acad. Sci.*, 2014, **111**(9), 3454–3459.
96. S. Ke, E. N. Timin and A. Strydom, Different inward and outward conduction mechanisms in NaVMs suggested by molecular dynamics simulations, *PLoS Comput. Biol.*, 2014, **10**(7), e1003746.
97. S. Furini, P. Barbini and C. Domene, Effects of the Protonation State of the EEEE Motif of a Bacterial  $\text{Na}^+$ -channel on Conduction and Pore Structure, *Biophys. J.*, 2014, **106**(10), 2175–2183.
98. S. Ke, E.-M. Zangerl and A. Strydom, Distinct interactions of  $\text{Na}^+$  and  $\text{Ca}^{2+}$  ions with the selectivity filter of the bacterial sodium channel NavAb, *Biochem. Biophys. Res. Commun.*, 2013, **430**(4), 1272–1276.
99. B. Hille, Local anesthetics: hydrophilic and hydrophobic pathways for the drug-receptor reaction, *J. Gen. Physiol.*, 1977, **69**(4), 497–515.
100. S. G. Raju, A. F. Barber, D. N. LeBard, M. L. Klein and V. Carnevale, Exploring Volatile General Anesthetic Binding to a Closed Membrane-Bound Bacterial Voltage-Gated Sodium Channel via Computation, *PLoS Comput. Biol.*, 2013, **9**(6), e1003090.
101. J. A. Kaczmarek and B. Corry, Investigating the size and dynamics of voltage-gated sodium channel fenestrations: A molecular dynamics study, *Channels*, 2014, **8**(3), 264–277.
102. A. Nardi, N. Damann, T. Hertrampf and A. Kless, Advances in targeting voltage-gated sodium channels with small molecules, *ChemMedChem*, 2012, **7**(10), 1712–1740.
103. L. J. Martin and B. Corry, Locating the Route of Entry and Binding Sites of Benzocaine and Phenytoin in a Bacterial Voltage Gated Sodium Channel, *PLoS Comput. Biol.*, 2014, **10**(7), e1003688.
104. C. Boiteux, I. Vorobyov, R. J. French, C. French, V. Yarov-Yarovoy and T. W. Allen, Local anesthetic and antiepileptic drug access and binding to a bacterial voltage-gated sodium channel, *Proc. Natl. Acad. Sci.*, 2014, **111**(36), 13057–13062.

105. C. Jorgensen, L. Darré, V. Oakes, R. Torella, D. Pryde and C. Domene, Lateral Fenestrations in  $K^+$ -Channels Explored Using Molecular Dynamics Simulations, *Mol. Pharm.*, 2016, **13**(7), 2263–2273.
106. G. Rudnick, Mechanisms of Biogenic Amine Neurotransmitter Transporters, in *Neurotransmitter Transporters: Structure, Function, and Regulation*, ed. M. E. A. Reith, Humana Press, Totowa, NJ, 1997, pp. 73–100.
107. L. Iversen, Neurotransmitter transporters and their impact on the development of psychopharmacology, *Br. J. Pharmacol.*, 2006, **147**(Suppl 1), S82–S88.
108. S. G. Amara and M. S. Sonders, Neurotransmitter transporters as molecular targets for addictive drugs, *Drug Alcohol Depend.*, 1998, **51**(1–2), 87–96.
109. I. Bisha and A. Magistrato, The molecular mechanism of secondary sodium symporters elucidated through the lens of the computational microscope, *RSC Adv.*, 2016, **6**(12), 9522–9540.
110. A. Yamashita, S. K. Singh, T. Kawate, Y. Jin and E. Gouaux, Crystal structure of a bacterial homologue of  $Na^+/Cl^-$  dependent neurotransmitter transporters, *Nature*, 2005, **437**(7056), 215–223.
111. H. Krishnamurthy and E. Gouaux, X-ray structures of LeuT in substrate-free outward-open and apo inward-open states, *Nature*, 2012, **481**(7382), 469–474.
112. S. K. Singh, A. Yamashita and E. Gouaux, Antidepressant binding site in a bacterial homologue of neurotransmitter transporters, *Nature*, 2007, **448**(7156), 952–956.
113. Z. Zhou, J. Zhen, N. K. Karpowich, R. M. Goetz, C. J. Law, M. E. Reith and D. N. Wang, LeuT-desipramine structure reveals how antidepressants block neurotransmitter reuptake, *Science*, 2007, **317**(5843), 1390–1393.
114. L. Malinauskaite, S. Said, C. Sahin, J. Grouleff, A. Shahsavar, H. Bjerregaard, P. Noer, K. Severinsen, T. Boesen, B. Schiott, S. Sinning and P. Nissen, A conserved leucine occupies the empty substrate site of LeuT in the  $Na^+$ -free return state, *Nat. Commun.*, 2016, **7**, 11673.
115. L. Celik, B. Schiøtt and E. Tajkhorshid, Substrate Binding and Formation of an Occluded State in the Leucine Transporter, *Biophys. J.*, 2008, **94**(5), 1600–1612.
116. S. Y. Noskov and B. Roux, Control of ion selectivity in LeuT: two  $Na^+$  binding sites with two different mechanisms, *J. Mol. Biol.*, 2008, **377**(3), 804–818.
117. D. A. Caplan, J. O. Subbotina and S. Y. Noskov, Molecular mechanism of ion-ion and ion-substrate coupling in the  $Na^+$ -dependent leucine transporter LeuT, *Biophys. J.*, 2008, **95**(10), 4613–4621.
118. C. Zhao, S. Stolzenberg, L. Gracia, H. Weinstein, S. Noskov and L. Shi, Ion-controlled conformational dynamics in the outward-open transition from an occluded state of LeuT, *Biophys. J.*, 2012, **103**(5), 878–888.
119. S. Tavoulari, E. Margheritis, A. Nagarajan, D. C. DeWitt, Y.-W. Zhang, E. Rosado, S. Ravera, E. Rhoades, L. R. Forrest and G. Rudnick, Two



- Na<sup>+</sup> Sites Control Conformational Change in a Neurotransmitter Transporter Homolog, *J. Biol. Chem.*, 2015.
120. E. Zomot, A. Bendahan, M. Quick, Y. Zhao, J. A. Javitch and B. I. Kanner, Mechanism of chloride interaction with neurotransmitter:sodium symporters, *Nature*, 2007, **449**(7163), 726–730.
121. E. Zomot, M. Gur and I. Bahar, Microseconds simulations reveal a new sodium-binding site and the mechanism of sodium-coupled substrate uptake by LeuT, *J. Biol. Chem.*, 2015, **290**(1), 544–555.
122. C. Zhao and S. Y. Noskov, The role of local hydration and hydrogen-bonding dynamics in ion and solute release from ion-coupled secondary transporters, *Biochemistry*, 2011, **50**(11), 1848–1856.
123. J. R. Thomas, P. C. Gedeon, B. J. Grant and J. D. Madura, LeuT conformational sampling utilizing accelerated molecular dynamics and principal component analysis, *Biophys. J.*, 2012, **103**(1), L1–L3.
124. M. H. Cheng and I. Bahar, Complete mapping of substrate translocation highlights the role of LeuT N-terminal segment in regulating transport cycle, *PLoS Comput. Biol.*, 2014, **10**(10), e1003879.
125. M. Gur, E. Zomot, M. H. Cheng and I. Bahar, Energy landscape of LeuT from molecular simulations, *J. Chem. Phys.*, 2015, **143**(24), 243134.
126. S. A. Shaikh and E. Tajkhorshid, Modeling and dynamics of the inward-facing state of a Na<sup>+</sup>/Cl<sup>−</sup> dependent neurotransmitter transporter homologue, *PLoS Comput. Biol.*, 2010, **6**(8), e1000905.
127. S. Faham, A. Watanabe, G. M. Besserer, D. Cascio, A. Specht, B. A. Hirayama, E. M. Wright and J. Abramson, The crystal structure of a sodium galactose transporter reveals mechanistic insights into Na<sup>+</sup>/sugar symport, *Science*, 2008, **321**(5890), 810–814.
128. J. Grouleff, S. Sondergaard, H. Koldso and B. Schiott, Properties of an inward-facing state of LeuT: conformational stability and substrate release, *Biophys. J.*, 2015, **108**(6), 1390–1399.
129. R. Chen and S.-H. Chung, Molecular dynamics simulations of Na<sup>+</sup> and leucine transport by LeuT, *Biochem. Biophys. Res. Commun.*, 2015, **464**(1), 281–285.
130. L. Shi, M. Quick, Y. Zhao, H. Weinstein and J. A. Javitch, The mechanism of a neurotransmitter:sodium symporter–inward release of Na<sup>+</sup> and substrate is triggered by substrate in a second binding site, *Mol. Cell*, 2008, **30**(6), 667–677.
131. M. Quick, L. Shi, B. Zehnpfennig, H. Weinstein and J. A. Javitch, Experimental conditions can obscure the second high-affinity site in LeuT, *Nat. Struct. Mol. Biol.*, 2012, **19**(2), 207–211.
132. Y. Zhao, D. S. Terry, L. Shi, M. Quick, H. Weinstein, S. C. Blanchard and J. A. Javitch, Substrate-modulated gating dynamics in a Na<sup>+</sup>-coupled neurotransmitter transporter homologue, *Nature*, 2011, **474**(7349), 109–113.
133. S. K. Singh, C. L. Piscitelli, A. Yamashita and E. Gouaux, A competitive inhibitor traps LeuT in an open-to-out conformation, *Science*, 2008, **322**(5908), 1655–1661.



134. C. L. Piscitelli, H. Krishnamurthy and E. Gouaux, Neurotransmitter/sodium symporter orthologue LeuT has a single high-affinity substrate site, *Nature*, 2010, **468**(7327), 1129–1132.
135. H. Wang and E. Gouaux, Substrate binds in the S1 site of the F253A mutant of LeuT, a neurotransmitter sodium symporter homologue, *EMBO Rep.*, 2012, **13**(9), 861–866.
136. H. Wang, J. Elferich and E. Gouaux, Structures of LeuT in bicelles define conformation and substrate binding in a membrane-like context, *Nat. Struct. Mol. Biol.*, 2012, **19**(2), 212–219.
137. M. H. Cheng and I. Bahar, Coupled global and local changes direct substrate translocation by neurotransmitter-sodium symporter orthologue LeuT, *Biophys. J.*, 2013, **105**(3), 630–639.
138. M. Quick, A. M. Winther, L. Shi, P. Nissen, H. Weinstein and J. A. Javitch, Binding of an octylglucoside detergent molecule in the second substrate (S2) site of LeuT establishes an inhibitor-bound conformation, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**(14), 5563–5568.
139. C. Zhao, D. A. Caplan and S. Y. Noskov, Evaluations of the Absolute and Relative Free Energies for Antidepressant Binding to the Amino Acid Membrane Transporter LeuT with Free Energy Simulations, *J. Chem. Theory Comput.*, 2010, **6**(6), 1900–1914.
140. N. King, C. T. Hittinger and S. B. Carroll, Evolution of Key Cell Signaling and Adhesion Protein Families Predates Animal Origins, *Science*, 2003, **301**(5631), 361–363.
141. K. A. Jacobson, New paradigms in GPCR drug discovery, *Biochem. Pharmacol.*, 2015, **98**(4), 541–555.
142. R. Lappano and M. Maggiolini, G protein-coupled receptors: novel targets for drug discovery in cancer, *Nat. Rev. Drug Discovery*, 2011, **10**(1), 47–60.
143. R. C. Stevens, V. Cherezov, V. Katritch, R. Abagyan, P. Kuhn, H. Rosen and K. Wuthrich, The GPCR Network: a large-scale collaboration to determine human GPCR structure and function, *Nat. Rev. Drug Discovery*, 2013, **12**(1), 25–34.
144. S. Galandrin and M. Bouvier, Distinct signaling profiles of beta1 and beta2 adrenergic receptor ligands toward adenylyl cyclase and mitogen-activated protein kinase reveals the pluridimensionality of efficacy, *Mol. Pharmacol.*, 2006, **70**(5), 1575–1584.
145. B. K. Kobilka and X. Deupi, Conformational complexity of G-protein-coupled receptors, *Trends Pharmacol. Sci.*, 2007, **28**(8), 397–406.
146. S. G. F. Rasmussen, H.-J. Choi, D. M. Rosenbaum, T. S. Kobilka, F. S. Thian, P. C. Edwards, M. Burghammer, V. R. P. Ratnala, R. Sanishvili, R. F. Fischetti, G. F. X. Schertler, W. I. Weis and B. K. Kobilka, Crystal structure of the human [bgr]2 adrenergic G-protein-coupled receptor, *Nature*, 2007, **450**(7168), 383–387.
147. V. Cherezov, D. M. Rosenbaum, M. A. Hanson, S. G. F. Rasmussen, F. S. Thian, T. S. Kobilka, H.-J. Choi, P. Kuhn, W. I. Weis, B. K. Kobilka and R. C. Stevens, High Resolution Crystal Structure of an Engineered

- Human  $\beta(2)$ -Adrenergic G protein-Coupled Receptor, *Science*, 2007, **318**(5854), 1258–1265.
148. D. M. Rosenbaum, V. Cherezov, M. A. Hanson, S. G. F. Rasmussen, F. S. Thian, T. S. Kobilka, H.-J. Choi, X.-J. Yao, W. I. Weis, R. C. Stevens and B. K. Kobilka, GPCR Engineering Yields High-Resolution Structural Insights into  $\beta(2)$ -Adrenergic Receptor Function, *Science*, 2007, **318**(5854), 1266–1273.
149. M. A. Hanson, V. Cherezov, M. T. Griffith, C. B. Roth, V.-P. Jaakola, E. Y. T. Chien, J. Velasquez, P. Kuhn and R. C. Stevens, A Specific Cholesterol Binding Site Is Established by the 2.8 Å Structure of the Human  $\beta(2)$ -Adrenergic Receptor, *Structure*, 2008, **16**(6), 897–905.
150. R. O. Dror, D. H. Arlow, D. W. Borhani, M. Ø. Jensen, S. Piana and D. E. Shaw, Identification of two distinct inactive conformations of the  $\beta(2)$ -adrenergic receptor reconciles structural and biochemical observations, *Proc. Natl. Acad. Sci.*, 2009, **106**(12), 4689–4694.
151. J. A. Ballesteros, A. D. Jensen, G. Liapakis, S. G. Rasmussen, L. Shi, U. Gether and J. A. Javitch, Activation of the beta 2-adrenergic receptor involves disruption of an ionic lock between the cytoplasmic ends of transmembrane segments 3 and 6, *J. Biol. Chem.*, 2001, **276**(31), 29171–29177.
152. D. M. Rosenbaum, C. Zhang, J. A. Lyons, R. Holl, D. Aragao, D. H. Arlow, S. G. F. Rasmussen, H.-J. Choi, B. T. DeVree, R. K. Sunahara, P. S. Chae, S. H. Gellman, R. O. Dror, D. E. Shaw, W. I. Weis, M. Caffrey, P. Gmeiner and B. K. Kobilka, Structure and function of an irreversible agonist-[bgr] $\beta(2)$  adrenoceptor complex, *Nature*, 2011, **469**(7329), 236–240.
153. R. Nygaard, Y. Zou, R. O. Dror, T. J. Mildorf, D. H. Arlow, A. Manglik, A. C. Pan, C. W. Liu, J. J. Fung, M. P. Bokoch, F. S. Thian, T. S. Kobilka, D. E. Shaw, L. Mueller, R. S. Prosser and B. K. Kobilka, The dynamic process of beta(2)-adrenergic receptor activation, *Cell*, 2013, **152**(3), 532–542.
154. R. O. Dror, D. H. Arlow, P. Maragakis, T. J. Mildorf, A. C. Pan, H. Xu, D. W. Borhani and D. E. Shaw, Activation mechanism of the beta2-adrenergic receptor, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**(46), 18684–18689.
155. A. V. Botelho, T. Huber, T. P. Sakmar and M. F. Brown, Curvature and hydrophobic forces drive oligomerization and modulate activity of rhodopsin in membranes, *Biophys. J.*, 2006, **91**(12), 4464–4477.
156. S. Mondal, G. Khelashvili, J. Shan, O. S. Andersen and H. Weinstein, Quantitative modeling of membrane deformations by multihelical membrane proteins: application to G-protein coupled receptors, *Biophys. J.*, 2011, **101**(9), 2092–2101.
157. S. Mondal, J. M. Johnston, H. Wang, G. Khelashvili, M. Filizola and H. Weinstein, Membrane Driven Spatial Organization of GPCRs, *Sci. Rep.*, 2013, **3**, 2909.

158. D. Calebiro, F. Rieken, J. Wagner, T. Sungkaworn, U. Zabel, A. Borzi, E. Cocucci, A. Zürn and M. J. Lohse, Single-molecule analysis of fluorescently labeled G-protein-coupled receptors reveals complexes with distinct dynamics and organization, *Proc. Natl. Acad. Sci.*, 2013, **110**(2), 743–748.
159. J. J. Fung, X. Deupi, L. Pardo, X. J. Yao, G. A. Velez-Ruiz, B. T. DeVree, R. K. Sunahara and B. K. Kobilka, Ligand-regulated oligomerization of  $\beta(2)$ -adrenoceptors in a model lipid bilayer, *EMBO J.*, 2009, **28**(21), 3315–3328.
160. X. Prasanna, A. Chattopadhyay and D. Sengupta, Cholesterol Modulates the Dimer Interface of the  $\beta_2$ -Adrenergic Receptor via Cholesterol Occupancy Sites, *Biophys. J.*, 2014, **106** (6), 1290–1300.
161. C. Neale, H. D. Herce, R. Pomès and A. E. García, Can Specific Protein-Lipid Interactions Stabilize an Active State of the Beta 2 Adrenergic Receptor? *Biophys. J.*, 2015, **109** (8), 1652–1662.
162. A. C. Kalli, M. S. P. Sansom and R. A. F. Reithmeier, Molecular Dynamics Simulations of the Bacterial UraA  $H^+$ -Uracil Symporter in Lipid Bilayers Reveal a Closed State and a Selective Interaction with Cardiolipin, *PLoS Comput. Biol.*, 2015, **11**(3), e1004123.
163. J. N. Horn, T.-C. Kao and A. Grossfield, Coarse-grained Molecular Dynamics Provides Insight into the Interactions of Lipids and Cholesterol with Rhodopsin, *Adv. Exp. Med. Biol.*, 2014, **796**, 75–94.
164. D. Sengupta and A. Chattopadhyay, Molecular dynamics simulations of GPCR-cholesterol interaction: An emerging paradigm, *Biochim. Biophys. Acta.*, 2015, **1848**(9), 1775–1782.
165. X. Prasanna, A. Chattopadhyay and D. Sengupta, Role of Lipid-Mediated Effects in  $\beta_2$ -Adrenergic Receptor Dimerization, in *Biochemical Roles of Eukaryotic Cell Surface Macromolecules*, ed. Chakrabarti, A., Surolia, A., Springer International Publishing, Cham, 2015, pp. 247–261.

## CHAPTER 11

# *Application of Molecular Modelling to Speed-up the Lead Discovery Process*

IUNI M. L. TRIST,<sup>a</sup> MAURIZIO BOTTA<sup>\*a,b</sup> AND  
ANNA LUCIA FALLACARA<sup>a</sup>

<sup>a</sup> Università degli Studi di Siena, Dipartimento di Biotecnologie Chimica e Farmacia, Via A. Moro 2, Siena 53100, Italy; <sup>b</sup> Sbarro Institute for Cancer Research and Molecular Medicine, Center of Biotechnology, College of Science and Technology, Temple University, BioLife, Science Bldg., Suite 333, 1900 N 12th Street, Philadelphia, PA 19122, USA

\*Email: botta.maurizio@gmail.com

## 11.1 Introduction

### 11.1.1 The ‘Pharmaceutical Crisis’

Drug discovery has improved life expectancy by an average of two months each year and our quality of life in general by transforming many life threatening diseases to almost negligible problems.<sup>1–3</sup> Behind each new molecular entity (NME) that reaches the market there is a particular ‘story’, but the overall process is generally long (around 14 years), costly (over 2.5 billion dollars) and most of the time unpredictable.<sup>1,4</sup> For this reason, the potential return of a drug should be enough to cover the costs for its development and invest in future research projects (R&D), in particular in pharmaceutical industries.<sup>1</sup>

Since 1940, 1679 NMEs were approved for commercialisation, and most of them are small molecules.<sup>3,5</sup> The trend of NMEs reaching the market each year is approximately flat, with an average of 24 NMEs per year and a few peaks registered between 1995 and 1999, in 2012 (39 NMEs), in 2014 (41 NMEs) and in 2015 (45 NMEs).<sup>5</sup> At the same time, a 12.3% per year increase in the needed expenditure in drug R&D has been registered, pointing towards a ‘pharmaceutical crisis’.<sup>3,4,6</sup> This has been linked to the fact that candidates are dropped late in the drug discovery process mainly due to limited efficacy and problems in the absorption, distribution, metabolism, excretion and toxicity (ADMET) profiles, thus leading to a big loss of investments.<sup>7</sup>

‘Time is money’ and several strategies have been employed to speed up the R&D process and improve this scenario as, for example, parallel synthesis and high throughput screening (HTS) assays. However, these did not produce the expected improvements.<sup>1,8,9</sup> Based on information technology, an important aid in sensibly reducing time and costs of R&D is given by molecular modelling. It was recognised already in the 1950s that “*the only difficulty which exists in the evaluation of the energy and wave function of any molecule ... is the amount of computing necessary*” (S. Francis Boys)<sup>10</sup> and in the 1970s, molecules were described with pseudo 3D models for the first time. Nevertheless, it is only in more recent years that the popularity of this scientific field has increased.<sup>11,12</sup> Indeed, tremendous efforts in algorithms, computer hardware and software development have been necessary for pharmaceutical R&D to be able to benefit of modern molecular modelling, in several stages of the drug discovery process, as described later in paragraph 11.1.3.<sup>12</sup>

### 11.1.2 The Drug Discovery Process

The process of drug design and development can follow many different routes, but the general process is summarised in Figure 11.1. Briefly, after target selection, several strategies (*e.g.* from natural products, serendipity, HTS campaigns on large chemical databases or rational drug design) can be applied to identify a molecule, namely the *hit compound*, that is able to produce an appreciable effect. This molecule is then improved to a *lead compound*, which has a selective therapeutic effect, but that still needs to be further optimised for the improvement of both its pharmacodynamics and pharmacokinetics properties. In this optimisation process, many analogue molecules are synthesised and evaluated *in vitro* not only for activity assessments, but also for known toxic mechanisms that need to be avoided. The most promising compounds proceed to the pre-clinical phase, where they are also tested *in vivo* with animal models in order to have a first proof-of-concept (POC) of the therapeutic strategy pursued and to verify the ADMET profile. If the compound(s) prove to be safe enough, the process continues with three clinical trials in humans. In the first one (Phase I), ascending doses are administered to healthy volunteers in order to assess the safety, tolerability pharmacokinetics and bioavailability properties of the potential drug. In Phase II, the efficacy of the compounds is tested for the

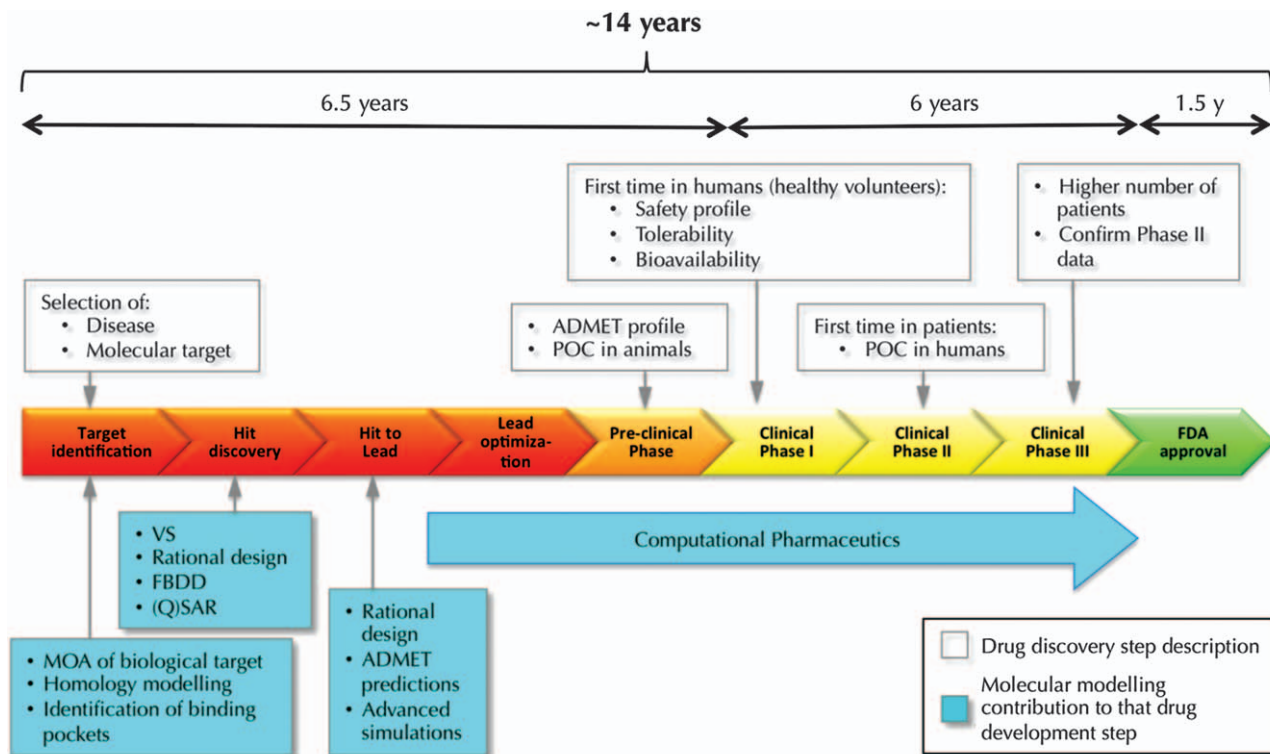


**Figure 11.1** The drug discovery process: the principle steps and molecular modelling contributions.

first time in humans and compared to placebo or, where possible, to the gold standard marketed therapy. Phase II, indeed, provides the POC in humans. Here, usually between 100 and 300 patients are recruited. Finally, in Phase III larger patient groups are involved, with the aim of confirming what has been observed in phase II, but with stronger statistics. These trials also look for unwanted effects that are only apparent when large patient populations are exposed to the drug. However, since the compound has never been tested on a large scale as the World population, its effects are monitored also after approval, through pharmacosurveillance.<sup>1</sup>

### 11.1.3 The Contribution of Molecular Modelling to Improve Drug Discovery

The challenges that are affecting the drug discovery process are pushing towards reducing time and costs by exploiting cutting-edge techniques, such as molecular modelling, that use information technology. It is in the 1980s that drug discovery has started to use computers not only for data management, but for actual computer-aided drug design (CADD).<sup>13–15</sup> Now, CADD is well established both in academia and in pharmaceutical industry R&D.<sup>16,17</sup> The continued technological and algorithm improvements themselves that are contributing notably to the reliability of the results obtained and the fact that virtual experiments are intrinsically cheaper, faster and safer than real experiments are pushing towards the increased use of molecular modelling in research projects.<sup>2,12,18</sup>

Computational techniques are exploited to explain complex chemical systems in terms of models. Models are heuristic and simplified descriptions of a system that allow understanding and predicting macroscopic properties, starting from knowledge on an atomic scale. They exploit experimental data or use a theoretical model or a combination of the two.<sup>19</sup>

The involvement of molecular modelling in drug discovery is very wide, contributing to several aspects of the process, from target identification to the pre-clinical stage, as exemplified in Figure 11.1.<sup>2,12,17,18</sup> At the beginning of a drug design project it is fundamental to assess the potential to generate a molecule with the desired pharmaceutical activity. Molecular modelling can be pivotal in this evaluation, bridging biology and medicinal chemistry. For example, pockets within 3D structures can be assessed according to their putative ability to bind ligands or inhibitors; homology models can be built; molecular dynamics (MD) simulations can be performed in order to understand the mechanism of action (MOA) of a protein or a molecular system; pharmacophores can be built based on available ligands. After target selection, molecular modelling can speed up the hit discovery process in several ways, as for example by virtual screening (VS) of large libraries of diverse or similar compounds; rationally designing a molecule on the base of a co-crystallised ligand in the target protein structure; growing and/or merging fragments (fragment-based drug design, FBDD); using (quantitative) structure–activity relationship, or (Q)SAR, assessments from known



ligands. Then, hit and lead optimisation can be aided by modellers for example by rational modifications to the molecular structures; structure–activity assessments; ADMET predictions; advanced simulations for a more accurate binding evaluation.<sup>12,20</sup>

We now also have the ability to study *in silico* a wide range of systems, from chemical reactions up to nanoparticles and even powder. In recent years, the emerging of a new branch called ‘computational pharmaceuticals’ that can provide knowledge for the design and optimisation of drug delivery and formulation systems, has enhanced the impact of computational techniques in the drug development process even more.<sup>21</sup> Despite its importance, computational pharmaceuticals is beyond the scope of this chapter and will not be further discussed.

#### 11.1.4 Quantum and Molecular Mechanics in Drug Design

Computational chemistry can study a system at two different levels: Quantum Mechanics (QM) and Molecular Mechanics (MM). QM studies molecules on the base of the molecular orbital theory and Schrödinger’s equation. Ideally, QM is the most precise approach available, but the complexity of the calculations described above, make these methods too computational expensive and applicable only to small molecular systems.<sup>18,22,23</sup> Indeed, QM is usually applied to problems that depend on electronic distribution, like electron density studies and chemical reactions in which bonds are broken or formed.<sup>18,22</sup>

On the other side, MM applies classical mechanics to the modelling of molecular systems, simplifying the physical and mathematical description of molecules and lowering the complexity of the problem, reducing computational costs. MM is indeed more often applied in drug discovery studies and, for this reason, the molecular techniques described in this chapter are mainly MM-based. However, it is worth mentioning that hybrid methods, like QM/MM, treat different portions of the molecular systems with different precision and are sometimes applied in the understanding of the behaviour of the target–ligand complex.<sup>19,23</sup>

#### 11.1.5 An Introduction to Structure- and Ligand-based Molecular Modelling

Molecular recognition, the steering of binding to a biological target and thus drug activity, relies on specific attractive interactions between molecules, as for example hydrogen bonds, water mediated bridges or hydrophobic interactions. This makes the understanding of the characteristics that an ideal compound should have essential for efficient drug design.<sup>24,25</sup> Molecular modelling can achieve this knowledge mainly through two approaches: *structure-based* and *ligand-based* drug design (SBDD and LBDD, respectively).<sup>26</sup> The first one is based on the availability or the possibility to

build 3D structures of the biological target and applies molecular modelling techniques to identify and optimise drug target interactions.<sup>24</sup> A valid alternative is *ligand-based* drug design, which exploits experimental data on ligands of a biological target to predict and optimise biological properties of compounds. Descriptors obtained from the chemical structure of compounds can also be used for the prediction and optimisation of ADMET properties in order to enhance the opportunity for the molecule to interact with the molecular target.<sup>26</sup> The molecular modelling techniques used in these two approaches will be further discussed later in this chapter, in paragraphs 11.2 and 11.3.

## 11.2 Structure-based Molecular Modelling

Since one of the principal objectives of drug design is the exploitation of specific drug-molecular target interactions to achieve a good activity with low side effects, the knowledge of the target structure is of key interest. The SBDD process can be an iterative procedure in which, after the identification of a hit, there is a compound optimisation through drug–target interactions improvements based on CADD, synthesis and biological evaluation.<sup>27,28</sup> SBDD approaches are several and can be very different according to the objective of the study. For example, new hits can be identified from compound libraries with known chemistry through VS campaigns, based on 3D pharmacophores or ligand docking, or they can be developed *de novo* within the target binding site. Furthermore, known hits and leads can be optimised through the introduction of new interactions and the MOA of compounds can be explored.

### 11.2.1 Sources of 3D Structures

The first SBDD projects appeared in the drug discovery process in the late 1970s and became well established in the late 1980s–early 1990s.<sup>27,29</sup> Of the earliest examples, it is worth mentioning the development of Zanamivir (influenza virus neuroaminidase inhibitor) and Nelfinavir (non-peptide HIV protease inhibitor), known with the brand names Relenza and Viracept, respectively.<sup>30,31</sup> Since then, there has been an exponential increase in the number of determined three-dimensional structures of potential molecular targets (proteins, nucleic acids, *etc.*). Analysing the protein data bank, the worldwide repository for the 3D structures of proteins, nucleic acids, and complex assemblies since 1971, it is clear how structural biology is becoming an important player in the identification of a target at the beginning of the drug discovery process and in compound optimisation. Each year, the number of deposited structures is increasing dramatically, with an exponential yearly growth. Figures are impressive: while only 142 structures were deposited in 1990, bringing the total number of structures to 507, 10 897 structures were deposited just in 2016 (for a total number of 125 454

structures at the end of that year) and over 1300 structures in the first month and a half in 2017.<sup>32</sup>

The major source for 3D structures is still X-ray crystallography (with over 100 000 entries), followed by nuclear magnetic resonance (NMR) and electron microscopy. In addition to the amount of 3D information nowadays available, the improvement in quality (*e.g.* most of the deposited structures now have resolution between 1.5 and 2.0 Å) and the possibility of obtaining structures of targets alone (apo form) or in complexes with ligands, are providing to molecular modelling the right tools to understand the MOA of potential leads and speed up lead discovery and optimisation.<sup>29,32,33</sup>

Despite this very positive scenario, not all structures are complete or in the right conformation and there is still a considerable gap between known structures and sequenced genome portions. In these cases, molecular modelling becomes a further source of structures. In the specific, one of the most used techniques is homology modelling, which 'builds' of macromolecular structures exploiting the existence of structural data of similar molecules (templates). Given that structure is more conserved than sequence, this approach has proven to be sufficiently reliable when proteins are evolutionary related and they share at least 30% or the primary sequence.<sup>28,34–36</sup> After template selection, the query and the template sequences are aligned following one of the several sequence-alignment methods and programs that have been developed, as for example the Needleman–Wunsch or the Smith–Waterman algorithms, the heuristic method adopted by BLAST and the progressive approach adopted by T-Coffee.<sup>37–40</sup> This is a crucial step on which the quality of the model depends greatly. It aims to pair amino acids in such way to obtain a match in corresponding structural or functional features.<sup>36</sup> Often, more than one template might be required for model building and in this case, the alignment must take in account multiple sequences, making sure that structural features are correctly aligned. Following alignment, conserved regions are built by transferring the corresponding coordinates of the template backbone and side chains directly into the model, while for structurally variable regions only backbone coordinates are usually kept from the template and side chains are optimised (*e.g.* exploring the rotamers). Some portions (likely placed in flexible regions or 'loops') of the query sequence might be missing in the template (insertion) or *vice versa* (deletion). In the case of deletions, usually a minimisation is sufficient, while in the case of insertions the loop must be built, using either knowledge-based (*e.g.* by similarity with other loops in the PDB) or by energy-based methods (*e.g.* MD or Monte Carlo simulations).<sup>36</sup> The most used programs for building an homology model are for example: MODELLER,<sup>41</sup> Prime,<sup>42</sup> Rosetta,<sup>43</sup> Swiss-Model,<sup>44</sup> Yasara.<sup>45</sup> Some of them are also helpful for model evaluation and validation through stereochemistry check, molecular mechanics energy based functions, statistical potentials or machine-learning approaches. In addition, web servers, as for example: ProCHECK,<sup>46</sup> ProSA,<sup>47</sup> ModFOLD4,<sup>48</sup> Verify 3D,<sup>49</sup> are often used for model evaluation. If the model results not adequate, it can be



proteins are generally at an energy-minimum state. Methods are classified as knowledge-based, if they reconstruct proteins following known folds, or as physical methods, if they use atom interaction principles and free-energy landscapes, such as MD. The increase of knowledge and of computational power are making *ab initio* target structure determination more feasible. However, the computational cost of proper conformation sampling, in particular for large proteins with multi-basin energy landscapes, and the analysis of sampled conformations are still important challenges.<sup>23,34,58</sup>

### 11.2.2 Docking

Given the main scope of understanding the principles by which drugs recognise and interact with their targets, docking is one of the most used and one of the most essential tools in SBDD. The term was ‘coined’ in the late 1970s, when it was restricted to the refinement or optimisation of a model of a complex. Soon later, in the 1980s, the relative orientation was allowed to vary, making docking the election tool for the prediction of the binding mode of a small molecule in respect to its molecular target. This allowed to predict the binding affinity, which often translates into activity.<sup>59,60</sup> The most popular docking programs currently in use are for example: DOCK,<sup>61</sup> FlexX,<sup>62</sup> GOLD,<sup>63</sup> AutoDock,<sup>64</sup> AutoDock Vina,<sup>65</sup> Glide,<sup>66,67</sup> RosettaLigand,<sup>68</sup> SwissDock.<sup>69</sup> Regardless of the program used, the process can be generalised in three moments (poses generation, poses scoring and final pose selection), which are often repeated in cycles until the best solution is identified.<sup>26,59,70</sup>

Poses generation aims the correct positioning of the ligand within the binding site of the target through search algorithms. In the first developed and simplest model, namely rigid docking, both the small molecule and the protein conformations are maintained rigid and the ligand can translate and rotate within the binding site.<sup>26,28,60,70</sup> In reality, flexibility often plays a very important role in ligand binding and is now accounted in several ways, enhancing in the accuracy of the results.<sup>26,71</sup> Ligand flexibility alone was historically introduced first. Given that a rigorous conformational search of the ligand is too computational expensive, docking programs apply either one of the two main sampling methods: systematic and stochastic methods. In systematic search, possible conformations are achieved through the incremental (*e.g.* every 30°) rotation of torsional angles, probing the energy landscape and converging to the (local) minimum energy solution. Conversely, stochastic methods randomly change the conformation of the ligand and then evaluate the poses with a pre-defined probability function.<sup>59,70,71</sup> Genetic algorithms (GA) are a particular application of stochastic methods that apply the theory of evolution and natural selection.<sup>59</sup> Protein flexibility is also often important in ligand binding and can be considered in a docking simulation up to a certain extent. In the case of ‘soft docking’ techniques, some overlap is allowed between the flexible ligand and the rigid target during pose generation, resolving unfavourable contacts due to overlap at a

post-processing stage. This introduces some localised flexibility with no dramatic increase in the computational cost of the calculation.<sup>72,73</sup> Also induced fit docking does not affect greatly computational time as it limits protein flexibility to residues side chains and/or selected backbone bonds within or near the binding site.<sup>74</sup> Other methods account full receptor flexibility, as in the cases of ensemble docking. In this scenario, the small molecules are docked on multiple low-energy conformations of the target binding site that are obtained prior to the binding simulation through sampling, for example by means of MD.<sup>73</sup>

The evaluation of the binding poses is performed through the estimation of binding free energies. In this respect, the non-feasibility of the exhaustive calculation of target–protein interactions have brought to the development of scoring functions. These take in consideration selected physical–chemical phenomena involved in binding, like intermolecular interactions, desolvation and entropic effects. The accuracy of the calculation ameliorates with the number of parameters evaluated in the scoring function, but this increases the computational time required. Consequently, scoring functions must be carefully selected according to the scope of the docking simulation.<sup>59</sup> Scoring functions are classified as force field-, empirical- or knowledge-based.<sup>26,70,71,75</sup> The first estimate binding energies by calculating through MM and classical mechanics and summing the receptor–ligand interactions and the internal energy of the ligand. Differently, in empirical-based scoring functions the physical–chemical phenomena involved in binding are calibrated on the base of experimental data on binding energies or affinities. Finally, knowledge-based scoring functions use energy potentials extracted from experimental structures, with the assumption that crystallographic (or NMR) data resembles energetically favoured conformations.<sup>59,70,75</sup> Given their approximate nature, each scoring function carries advantages and imperfections. A common procedure to overcome this issue is to combine different scoring methodologies into consensus scores, improving the probability to identify favourable ligands.<sup>22,70</sup> Despite the fact that a consensus score increases the procedure's performance, not all docking programs are able to successfully predict binding poses for all molecular systems, therefore the intervention of the medicinal chemist through visual inspection of the actual ligand–protein interactions is still pivotal for successful docking.<sup>28,70,75</sup>

The concept of docking is not limited to the correct placement of a small molecule within a target binding site anymore. Some compounds exert their therapeutic effect through inactivating their biological targets by covalently binding to them. The docking algorithms described above are not able to explore the energy landscape available to the ligand when it is linked to the receptor or to evaluate the binding energetics of the interaction because the formation of the bond is more appropriately handled by QM methods. Probably in relation to covalent inhibitors common drawbacks (toxicity, high reactivity and lack of specificity), molecular modelling methods that address docking of these compounds are not as developed as the ones for



non-covalent binding and, thus, these techniques will not be further discussed in this chapter.<sup>59</sup>

An increase of interest in protein–protein interfaces was recorded in the last years, because of the growing importance of protein–protein interactions (PPI) in several biological and pathological pathways.<sup>76–82</sup> However, 3D structures of protein assemblies are not always available and, thus, must be built through protein–protein docking. In this case, the recreation of the interaction between the counterparts is more complex than in small-molecule docking. In particular, taking in account of flexibility becomes more complex due to the dimensions of the counterparts.<sup>60</sup> Despite, challenges are identified, programs have benefited from experiments, such as the critical assessment of prediction of interactions (CAPRI),<sup>83</sup> making protein–protein docking simulations a promising tool in SBDD.<sup>59,60</sup> For example, the correct construction DENV capsid (C) tetramers was fundamental in the determination of the MOA of ST-148, a promising antiviral compound that inhibits DENV replication through the stabilisation of C self-interaction.<sup>84</sup>

### 11.2.3 *De Novo* Drug Design

*De novo* drug design consists in suggesting either compound modifications for lead structure optimisation or novel chemotypes that do not belong to already explored chemical space. Molecules with known chemistry might not always be suitable for all drug targets and to generate novel molecular entities ‘from scratch’ has the great advantage to produce structures not present in literature and free from intellectual property rights.<sup>26,85,86</sup> The most popular programs currently used for *de novo* drug design are for example: GenStar,<sup>87</sup> LigBuilder,<sup>88</sup> SYNOPSIS,<sup>89</sup> SPROUT,<sup>90</sup> ReCore,<sup>91</sup> CombiGlide.<sup>92</sup>

Based on the assumption that portions, or fragments, of the molecule itself are collectively responsible for binding, *de novo* hit identification undergoes the so-called fragment-based drug design (FBDD) approach. FBDD popularity is increasing in both pharma industry and academia and has many successful examples, like the approval of Vemurafenib in 2011.<sup>85,93–96</sup> This method is based on the identification of small chemical fragments that bind to the target, followed by their elaboration in order to improve affinity and produce a lead. Differently from compounds used in HTS or VS campaigns that follow the Lipinski’s ‘rule of five’ (a set of rules used for defining a compound like ‘druggable’), fragments usually follow the ‘rule of three’: they have molecular weight (MW) lower than 300 Da, a log *P* lower than three, less than three hydrogen bond donors (HBD) and less than three hydrogen bond acceptor (HBA) groups.<sup>93,94,97–99</sup> Reducing the number of atoms decreases the available chemical space and enhances the capability of exploration with reasonable sized chemical libraries. A positive outcome of the study relies on the library quality and thus, unwanted scaffolds (e.g. toxic, unstable or reactive) are removed from the collection. Fragments are usually screened *in vitro* against the target and the positive



ones are then elaborated *in silico*.<sup>93,94,99</sup> A limitation to this screen is that being smaller molecules, fragments have a low affinity for the target (around 0.1–10 mM). To avoid possible false positives or negatives, the common approach is to use several biochemical and biophysical techniques and searching for consensus. The most commonly used assays are for example thermal shift assays, surface plasmon resonance, differential scanning fluorimetry, isothermal titration calorimetry, weak affinity chromatography, mass spectroscopy and microscale thermophoresis.<sup>100,101</sup> These, are always coupled with structural techniques (X-ray crystallography or NMR) to reveal the binding modes of the scaffolds and proceed with the *in silico* elaboration.<sup>29,93,94,99</sup> Alternatively, fragments can also be placed within the target binding site by means of docking simulations.<sup>70</sup> As schematised in Figure 11.3, the *in silico* elaboration of the fragments aims the enhancement of binding affinity and involves fragment growing, merging and linking. In all three the cases, ‘growing sites’ are identified on the starting fragments. In fragment growing, functional groups are added in these sites, increasing the number of interactions with the target. If multiple fragments are used as a starting point, these can be either linked or merged according to their distance in space, with the objective of obtaining a compound with good affinity, deriving from the cumulative contribution of the starting points. In all the cases, an important evaluation that must be done for the molecules designed by FBDD is their synthetic feasibility because they are not commercially available and must be synthesised from scratch. For this reason, molecular modelling programs are now taking in consideration is the synthetic feasibility of the final molecules.<sup>85,102</sup>

As already mentioned, *de novo* drug design is most commonly applied also for compound optimisation, usually exploiting molecule growing (see above) or the replacement of the molecule central portion through scaffold hopping, maintaining synthetic feasibility and activity. Scaffold replacement is a common practice in medicinal chemistry for improving pharmacological

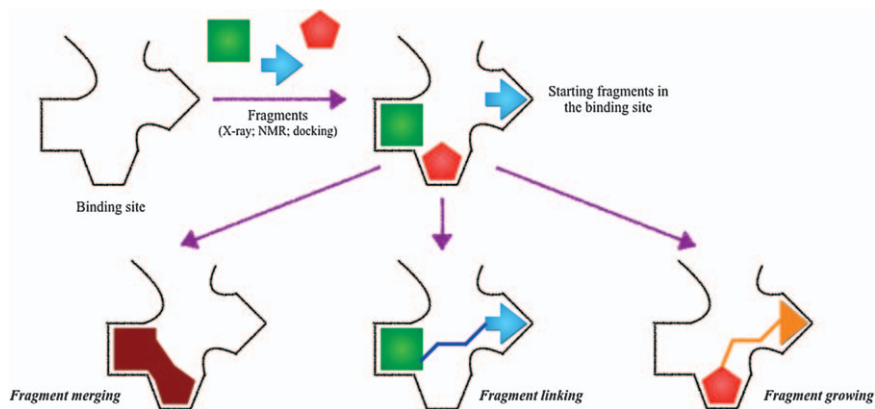


**Figure 11.3** *In silico* elaboration of fragments in a FBDD approach.

properties of compounds, circumventing an intellectual property by identifying new active chemical entities or making more synthetically accessible a chemically complex natural product. It is based on the assumption that two chemically different structures can exhibit similar biological activity if they are able to interact in the same way with the target. In this respect, chemical modifications have historically involved bioisosteric substitutions, heterocycle replacements, ring opening/closure, *etc.* In addition to this, in computer-based scaffold hopping the algorithms can also replace the molecular portion defined by the user with chemically different structure but similar in the meaning that they own, for example, the same shape, fingerprints, autocorrelation vectors or pharmacophore features in order to reproduce activity.<sup>103–106</sup>

### 11.2.4 Introducing Dynamics

Among the SBDD approaches, docking is surely the most used method for the identification of new hits for its rapidity and facility of use. However, the simulated binding interactions are not always predicted with sufficient accuracy mainly because, despite the efforts described above, target flexibility, including effects induced by ligand binding, is not fully taken in account.<sup>107</sup> An example of this is a recent study that aimed the development of novel compounds that stopped influenza A through the inhibition of the polymerase acidic protein (PA)–basic protein 1 (PB1) protein–protein interaction.<sup>81</sup> After the identification of a hit scaffold,<sup>80</sup> the analysis of 15 synthesised analogues allowed to gain knowledge on the SAR for this class of molecules. However, docking results did not always correlate well with biological data. In particular, docking poses highlighted a nitrile as important for compound binding for its interaction of PA Lys643. Nevertheless, two compounds lacking this group and therefore not interacting with PA Lys643, retained low micromolar biological activity. Only MD analysis of PA in complexes with an inhibitor possessing the nitrile (compound **1** in ref. 81) or with inhibitors lacking this group (compounds **11** and **15** in ref. 81) showed that the interaction with Lys643 could be either direct or through a water bridge (compounds **1** and **15**, respectively) and that the lack of this interaction (as for compound **11**) was compensated by a higher presence over time of other fundamental PA–inhibitor interactions, as shown in Figure 11.4.<sup>81</sup>

Among the techniques commonly used for conformational sampling, MD is indeed one of the most preferred as it also provides information of time scales and kinetics. The most popular programs currently used to perform MD simulations are for example: GROMACS,<sup>108</sup> CHARMM,<sup>109</sup> AMBER,<sup>110</sup> NAMD,<sup>111</sup> ACEMD.<sup>112</sup> Classical MD (cMD) simulations are based on MM functions for the calculation of the properties of the *N* interacting atoms of the system and for moving them collectively, following Newton's equation of motion.<sup>26</sup> The simulation returns a series of successive conformations at equilibrium conditions close to the starting conformation of the molecular

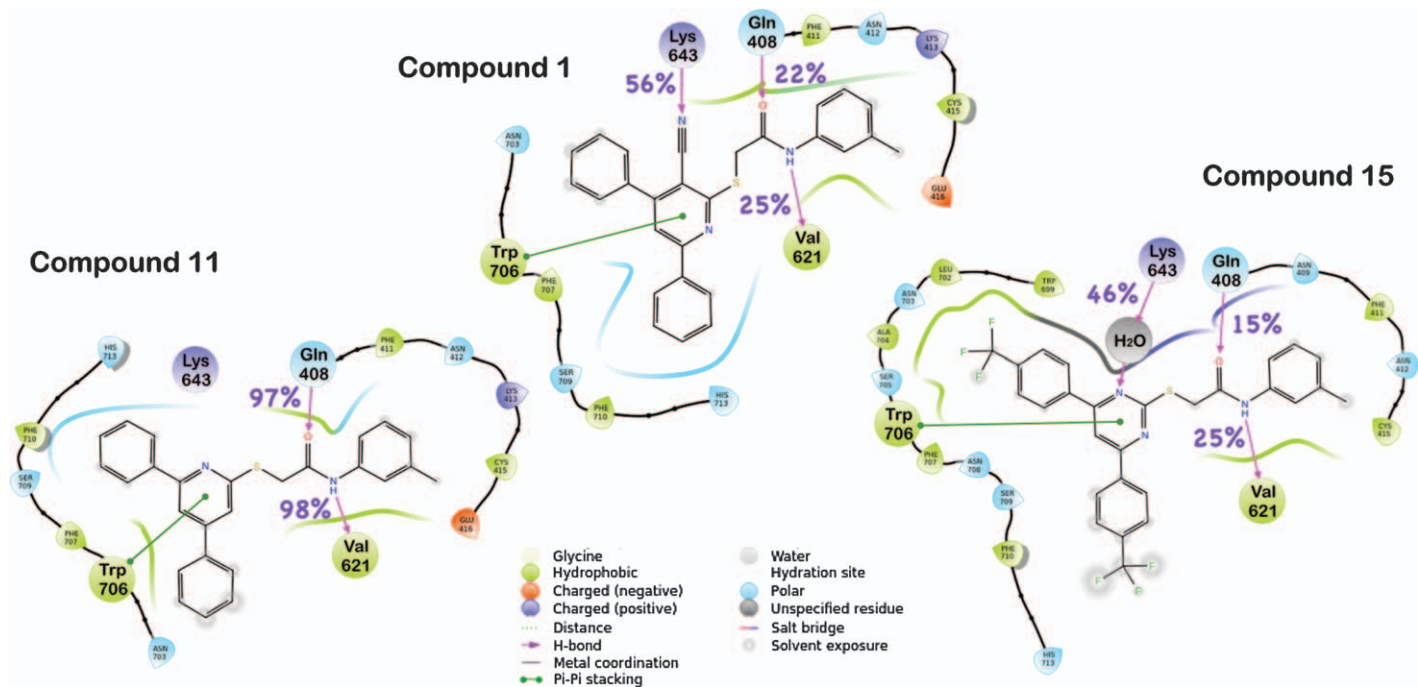


**Figure 11.4** Different interaction occupancies observed in MD simulations of Influenza A pa protein in complex with compounds 1, 11 and 15.

system. As for the PA-PB1 inhibitors example reported above, or for its application within VS workflows, cMD might be enough for some drug discovery studies as it returns useful and atomistic detailed insights on biomolecular dynamics involving individual and correlated motions between macromolecular units of a complex.<sup>23,113,114</sup> Starting from cMD trajectories it is possible to estimate the energy differences between the ligand-bound and the ligand-free states of the targets and to study the affinity of a ligand towards its biological counterpart. These binding free energies can be calculated by means of several methods, from thermodynamic integration (TI) and linear interaction energy (LIE), to the most commonly used approaches: free energy perturbation (FEP) and MM-Poisson Boltzmann surface area (MMPBSA) and MM-Generalised Born surface area (MMGBSA) methods.<sup>114</sup> In FEP, free energy differences between two states are calculated from MD or Monte Carlo (MC) simulations, which use statistical mechanics.<sup>115</sup> FEP is often used in SBDD to assess the change in affinity for the target when small chemical changes are applied to the compounds. Small chemical changes might be, for example, halogen (chlorine, bromine, and fluorine) and hydroxyl scans on a particular small molecule position as we have performed in the optimisation of Pyrazolo[3,4d]pyrimidines c-Src inhibitors active against neuroblastoma. In this study, the MC/FEP scans of the C4 anilino hydrogens showed that the *ortho* and *meta* positions for substitution are not equivalent and lead our hit-to-lead optimisation process that produced a *meta*-hydroxyl substituted Pyrazolo[3,4d]pyrimidine with enhanced activity and water solubility.<sup>116</sup> In MMPBSA and MMGBSA methods, the binding free energy is calculated using MD trajectories and continuum solvent models.<sup>117</sup> Similarly to what we have done in a study for the design of novel Fyn kinase inhibitors, these free energies estimations are often calculated on similar ligands and compared to understand the biological activity rationale. In our specific study, we discovered that the most active Fyn kinase inhibitors were also the ones that exhibited lower MMGBSA binding free energy values, and that this correlation was good (with Pearson's correlation coefficient,  $r^2$  of 0.83).<sup>118</sup>

The advancements in software and in computational power, in particular with the introduction of graphics-processing-units (GPU), have allowed to perform milliseconds MD simulations for small systems and hundreds of nanoseconds MD simulations for systems of over millions of atoms.<sup>119,120</sup> However, in the context of fundamental processes, including molecular recognition or allosteric regulation, these simulation times still represent only snapshots of real systems. Indeed, most biologically relevant events take place on timescales of the order of milliseconds, or sometimes even longer times and the gap between MD simulation timescales and those of biological processes is still relevant.<sup>119–121</sup> Furthermore, rare events or events occurring in regions of the energy landscape of the molecular system that are accessible through high energetic barriers, cannot be explored by means of cMD in computationally practical time periods. In all of these cases, enhanced algorithms are necessary.<sup>23,107</sup> This is particularly

important to study in details the MOA of compounds, in particular if it is based on effects as induced fit or conformational selection or, above all, if the compound activity is due to an allosteric effect on the biological target.<sup>122</sup> Extensive reviews of enhanced sampling MD algorithms have been reported and details of these methods can be found elsewhere.<sup>123–127</sup> One possibility to expand the simulation time for complex systems (as for example cell membranes with their proteins, entire virions, *etc.*)<sup>128,129</sup> is to simplify their representation. For instance, in coarse-grained MD, ‘pseudoatoms’ group together amino acids or amino acids portions, reducing in this way the number of the degrees of freedom and, consequently, the computational costs needed to study large conformational changes.<sup>130</sup> Sampling enhancement can also be achieved by applying a bias to the system. A first group of methods applies biases to collective variables (CV). In this case, the selection of the correct CV to bias is crucial to avoid the risk of not identifying the correct representation of the free energy landscape and therefore a good forehand knowledge of the system is required. CV-biased enhanced sampling methods include: umbrella sampling, biased or steered MD, conformational flooding, targeted MD and many others.<sup>23,131</sup> Differently, the unconstrained enhanced sampling methods explore conformational space without the need of the same amount of *a priori* knowledge about the system because they do not require predefined CVs. These methods include: replica exchange MD, parallel tempering, Langevin dynamics, accelerated MD, and many more.<sup>23,131</sup>

MD has brought huge advancements in the investigation of biomolecular systems and in SBDD, as it was also recognised by the 2013 Nobel Prize in Chemistry.<sup>132</sup> However, each MD-based method (either classical or with enhanced sampling) is characterised by advantages and limitations. It is therefore important to have a good knowledge of the real applicability of each technique in order to employ the most appropriate one according to the research question that the MD simulation has to answer to.<sup>119</sup>

### 11.3 Ligand-based Molecular Modelling

One of the key steps in drug development is to identify of a set of compounds, active against a target protein or disease, and then optimise them till the obtainment of a lead candidate. Although the number of proteins’ solved structures continues to increase every year,<sup>32</sup> there are many important drug targets without a known 3D structure, such as many G-proteins coupled receptor. In other cases, an active compound can be identified from a big cell-based HTS campaign, but without knowing the molecular target of such molecule. Therefore, whenever (i) a target receptor structure is not available, (ii) the protein involved in the development of the disease is completely unknown or (iii) the chemical dataset to be screened is very large, the most advantageous approach to apply is LBDD.<sup>133</sup>

LBDD approaches appeared earlier in the history of drug discovery than SBDD. Already between the 1950s and 1960s, many drugs that are still in use today, as for example propranolol or cimetidine, were discovered through

the synthesis and biological evaluation of several derivatives to deduce QSAR and pharmacophore models.<sup>27,134,135</sup>

LBDD methods perform a comparison between a known active compound (or a small set of active compounds) and molecules in a dataset, without considering any interaction with their receptor. Several approaches are used for representing compounds structure: 1D representation (SMILES,<sup>136</sup> SMARTS<sup>137</sup>), 2D representation and 3D representation. 1D and 2D representations have been widely used for identifying mainly close chemical analogues of known active compounds.<sup>138</sup> 1D and 2D representations are limited and often fail to predict activity differences because they just consider structural similarities and functional groups, while protein–ligand binding is governed by atomic interactions that occur in the 3D space. Thus, it is common that molecules having similar 3D shape and properties share biological activities even if their 2D and 1D representations are not similar.

However, the use of 3D methods requires a high storage space and is computational expensive, in particular taking in account the importance to treat the conformational flexibility of each ligand and the fact that the biologically active conformation of a compound often does not match the conformation with the lowest energy, adding complexity to the research.

In this part of the chapter, we will briefly review the main LBDD methods used for the identification of active compounds: the similarity search method, the pharmacophore model approach and the QSAR analysis. Then, an example of a combination of these methods applied to a virtual screening campaign will be provided.

### 11.3.1 Similarity Searching: Same Shape, Same Activity

#### 11.3.1.1 Molecular Descriptors

In similarity search, the query is an entire molecule which is compared to all entries contained in a database for the calculation of a similarity coefficient. In a typical scenario, the query molecule possesses biological activity and the objective is to identify compounds that will hopefully show the same potency. Therefore, similarity must be calculated on the base of a series of ‘characteristics’ owned by the query molecule. These useful ‘characteristics’ are known as molecular descriptors.

Molecular descriptors are usually easily calculated from molecular formula or computed from 3D conformations. The inclusion of experimentally determined descriptors is also possible, but it is often not feasible because of their unavailability for most of the screened databases commonly constituted by molecules not yet synthesised. The optimal descriptors’ set to use depends on the biological function predicted as well as the LBDD technique used. Molecular descriptors can include properties such as MW, surface area, ring contents, rotatable bonds, geometry and many others.<sup>139–142</sup> They can be generated by many different algorithms that rely on knowledge-based, graph-theoretical, MM or QM methods and are classified according to



the 'dimensionality' of the chemical representation from which they are computed.<sup>143</sup>

Descriptors like MW or the number of HBD are very simple to calculate, while physicochemical properties are more complex. Given the high level of information provided, the latter are however considered critical for developing effective models.<sup>144</sup> An example is the prediction of the logarithmic of partition coefficient between *n*-octanol and water ( $\log P$ ), which is widely used in the early stage of a drug development process. Traditionally,  $\log P$  can be experimentally calculated and has been shown to be important for solubility, oral availability transport, penetration through the blood-brain barrier and toxicity.<sup>145</sup> Not by chance, also Lipinski's 'rule of five' includes the measurement of  $\log P$ .<sup>97,98</sup> In the last decades, several methods have been developed, for *in silico* calculation of  $\log P$ , all based on molecular constitution. Among the first, the most popular was the Rekker and Manhold<sup>146</sup> method, which summed molecule basic fragments' contributions to  $\log P$ . Subsequently, more efficient methods were enhanced with the incorporation of molecular properties.<sup>147</sup> In 1997 Wang *et al.* developed the  $x\log P$  method,<sup>148</sup> based on the definition of atom types (80 at first) for carbon, nitrogen, oxygen, sulfur, phosphorous, halogens and on the application of correction factors to take in account specific intramolecular interactions.  $x\log P$  was improved in 2000 by adding a new set of atom types and correction factors. In 2002, Xing and Glen elaborated a statistical model based on the evidence that  $\log P$  was largely influenced from molecular size and hydrogen-bonding ability.<sup>149</sup> Their model showed how polarisability was more significant than atomic charges for  $\log P$  prediction, and that an increase in polarisability was correlated with an increase of  $\log P$ . Polarisability itself is considered as a good molecular descriptor, contributing to the dispersion forces and influencing intermolecular interactions.<sup>150-152</sup>

The number of descriptors and possible combinations of them that can be used for a medicinal chemistry task is huge.<sup>142</sup> A group of descriptors might work well for a target but be a poor choice for another target, thus it is important to keep in mind that the objectives for their final selection.

### 11.3.1.2 Calculating Similarity

Once the descriptors are defined, the next step is computing a quantitative measure of the similarity. The most common molecular coefficients used for similarity calculation are: Tanimoto coefficient, Dice coefficient and Cosine coefficient. Of them, Tanimoto coefficient (Tc) is the most used similarity metric in library design, although many reports suggest it is size-dependent favouring large molecules for similarity and small molecules for diversity selection.<sup>153,154</sup>

All molecular coefficients can be divided into three classes: association coefficients that assign a value of 1 for identity and 0 for diversity; correlation coefficients that calculate a degree of correlation between two descriptors vectors and assign a value of 1 for identical molecules and -1 for dissimilar;



distance coefficients that assign values in a range from 0 to  $\infty$ , according to similarity.

### 11.3.1.3 Molecular Fingerprints

One of the most important problems when trying to compare molecules for computing similarity is the complexity of the task, which is correlated to the complexity of the molecular representation used. The molecular fingerprint method aims to represent molecules in a way that allows rapid structural comparison. Fingerprints are bit string representations of molecular structures or properties.<sup>155,156</sup> They encode various molecular descriptors as predefined bit settings, where 1 means the presence and 0 the absence of a feature.<sup>157</sup> This method is less computationally expensive or hypothesis-driven than others because it simply relies on chemical structures not considering biological activity and evaluating the whole molecule, without focusing only on the portion that is thought to be the most important.

Comparison between molecules must then be quantifiable. There are many ways to assess the similarity between two vectors, but the overall standard for molecular fingerprints is Tc (see paragraph 11.3.1.2).

Different types of fingerprints methods can be used depending on how molecular representations are converted into a bit string. For example, in substructure key-based fingerprints the bits of the string indicate the presence or absence of a substructure in a molecule. The topological or path-based fingerprints analyse all the possible fragments of a given molecule following a path. They produce a fingerprint that is useful for a fast searching, but hashed because a single bit cannot be traced back to a specific feature. Another useful method handles circular fingerprints that are not suitable for substructure queries but for full structure searching.<sup>158</sup>

The current trend in similarity searching by using fingerprints seems to be the combination of different methods through data fusion. The advantage of this approach is that by capturing different chemical information the highest ranked hits will be those highly ranked by several approaches, reducing the possible selection of false positive.

## 11.3.2 Pharmacophore Modelling

The term 'pharmacophore' was defined for the first time by Ehrlich as "*a molecular framework that carries the essential features for a drug's biological activity*".<sup>159</sup> After a century from that simple but effective definition, the basic pharmacophore concept still remains unchanged, but its application range has been expanded considerably. Nowadays, according to the recent definition by the International Union of Pure and Applied Chemistry (IUPAC),<sup>160</sup> a pharmacophore model is "*an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response*". This simple definition introduced a very complex concept: the biological activity

of a compound is not only dependent on functional groups but also, and mostly, from their 3D arrangement. Following this definition, any others have been formulated but the basic concept remained the same.<sup>161–163</sup>

Pharmacophore models can be generated using two different approaches depending on the available information. The structure-based approach is used when an experimentally determined ligand–target complex is available, so that the interaction pattern of a ligand and its targets can be directly extracted. The structure-based approach also gives the possibility to include exclusion volumes as features, in order to mimic the geometry of the binding site avoiding an overlapping of the ligand with protein portions. In the ligand-based approach, the 3D structures of a training set of molecules (one or two active compounds might be enough) are aligned and the common pharmacophore features identified. In this case, all of the common chemical features from the pharmacophore have to be presumed as essential because the interaction pattern is completely unknown and excluding features could prejudice the predictivity of the model; whereas, in a structure-based approach, chemical features can be assessed as important for binding or not. Datasets containing molecules with known activity (test sets) are usually employed to assess the quality of the developed models. Training and test sets need to be designed carefully, because they largely influence the quality and the success of the study. For the development of a good pharmacophore model datasets should only include molecules for which there is experimental evidence of direct interaction with the receptor or of interference with the target enzyme activity.<sup>164–167</sup> Results from cell-based assays should be avoided because the compounds' activity that they measure depends also on other factors, as for example off-target effects or pharmacokinetic properties. Finally, whenever possible, the dataset should contain molecules with a certain degree of structural diversity.<sup>166</sup> In datasets, experimentally confirmed inactive compounds are as important as the active ones. Several public compound repositories such as ChEMBL,<sup>168</sup> Drugbank,<sup>169</sup> or OpenPHACTS<sup>170</sup> contain target-based activity data. However, if no confirmed inactive molecules are available, decoys, a set of compounds with unknown biological activity but assumed to be inactive, might be used instead. These decoy-datasets need to be adapted for each protocol because they should contain compounds with similar 1D properties<sup>171–173</sup> but different topologies compared to the known active molecules. It is possible to generate a set of optimised decoys on the DUD-E (Database of Useful Decoys: Enhanced) database, starting from the uploaded active molecules. In general, a ratio of about 1 : 50 for the number of active molecules and decoys is recommended.<sup>174</sup>

### 11.3.3 QSAR

Considering the rapid development of structure-based design approaches, the classical chemometric method QSAR might be considered today out of fashion. The number of deposited 3D structures has increased tremendously in the last two decades, so structure-based methods have had a big rise in

their use (see paragraph 11.2). In addition, a mathematical equation is much more difficult to understand and much less impressive than a coloured 3D structure of a protein. However, there is no denying that QSAR analysis represents a milestone in the field of drug design and still plays an important role.

The QSAR approach can be described as the application of both mathematical and statistical analysis for developing methods used to predict biological activity, or other properties, on the basis of chemical structures of compounds. Any QSAR model implies a simple concept: compounds with similar structure are expected to have similar biological activity. The main workflow for developing a rational QSAR model is reported in Figure 11.5.

Curating the starting database is the first crucial step in a QSAR protocol. An interesting study<sup>175</sup> showed that on average there are at least two structural errors per each medicinal chemistry publication. These errors could translate into inability to calculate descriptors for erroneous chemical records. In a recent review, Tropsha *et al.* described the importance of addressing chemical data accuracy in QSAR models and reported some integrated protocols that should be followed at the onset of any molecular modelling investigation.<sup>176</sup>

After preparation of the starting structures dataset, the biological data to include in our model need to be selected. Also in this case, it is imperative for biological data to be both accurate and precise to develop a meaningful model. For use in QSAR, standard biological equilibrium constants such as  $K_i$  and  $K_m$  should be preferred. Percentage of activity (e.g. % of inhibition of an enzyme at a given concentrations) is not considered appropriate because

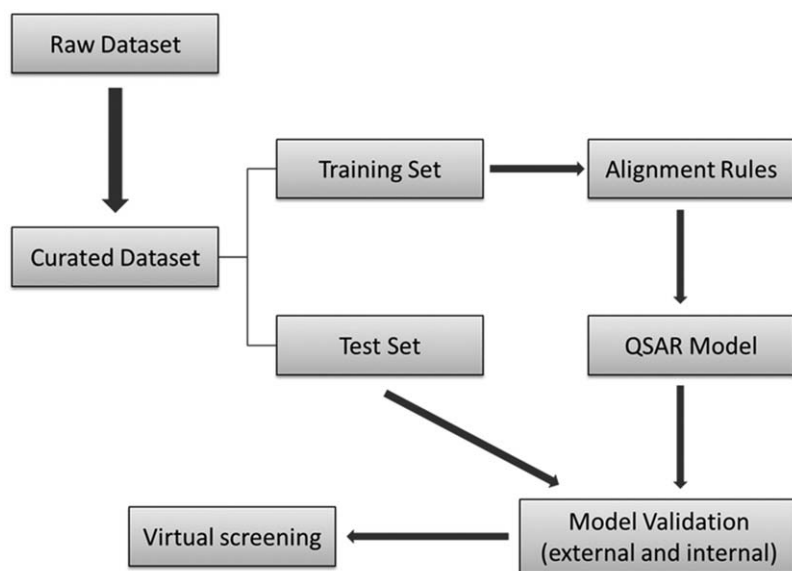


Figure 11.5 QSAR modelling (simplified workflow).

of non-linear characteristic of the dose–response relationships. Despite the understandable reluctance of medicinal chemists to synthesise inactive molecules, the inclusion of these compounds into a dataset is important. This widens the range of activities available and increases the predictive quality of the model.

Following biological parameters and structure database selection, a fraction of compounds is randomly designated as an external set, while the remaining part is divided into a training set (used for the QSAR model building) and a test set (used for the validation of the model). Many types of statistical analysis methods have been developed for the biological effect–chemical structure correlation, but the most widely used is the multiple regression analysis (MRA).<sup>177</sup>

In over a century of QSAR model development, the methods have evolved from Hansch<sup>145,178,179</sup> and Free–Wilson's<sup>180</sup> one or two dimensional linear free-energy relationship, *via* Crammer's<sup>181</sup> three-dimensional QSAR to Hopfinger's<sup>182</sup> fourth and Vedani's fifth<sup>183</sup> and sixth-dimensions.<sup>184</sup> Among all, 3D-QSAR methods are worth of note because of their popularity and use as successful applications in drug design. The detailed description of the theoretical basis of QSAR can be found in literature and is beyond the aim of this chapter, so it will not be discussed here.

However, in a paragraph about QSAR, it is impossible not mention the CoMFA (Comparative Molecular Field Analysis) methodology. In 1987, Cramer developed the so called Dynamic Lattice-Oriented Molecular Modelling System (DYLOMMS) that used principal component analysis (PCA) to extract vectors from molecular interaction fields, which were then correlated with biological activity.<sup>181</sup> Soon after he modified the method by combining two existing techniques, GRID and PLS, to develop a powerful 3D-QSAR methodology CoMFA.<sup>139</sup> Today CoMFA has become a prototype of 3D-QSAR methods.<sup>185</sup> QSAR has matured over the last decades in terms of descriptors, models, and methods of analysis. However, the most exciting challenge it is facing today regards the development of models to predict pharmacokinetic (PK) properties of potential drug candidates. The extension of the QSAR approach to PK parameters is referred to as quantitative structure–pharmacokinetic relationships (QSPkR) modelling.

### 11.3.3.1 QSAR Models for Predicting Clinical PK Parameters

The ability to predict PK properties is considered crucial for the drug development process. The average rate for potential drug candidates entering a Phase I clinical trial to become a commercial drug is about 11% with the 39% of them failing for poor PK and ADMET properties.<sup>186</sup> Indeed, to be competitive on market a new drug must have optimal PK properties, acceptable safety and good efficacy.

Despite QSAR models have been widely described<sup>187,188</sup> and used for early prediction of ADMET properties from pharmaceutical companies, the use of QSAR for predicting PK parameters is almost uncommon.

The oral administration of drugs is a cost-effective and desired route that is associated with high patient compliance. This makes human intestinal absorption one of the most influential properties to be assessed early<sup>189</sup> in the development of oral drugs because poor aqueous solubility or slow dissolution rate can negatively influence oral absorption and bioavailability.<sup>190</sup> QSAR models have been widely used in intestinal absorption prediction. Differently from qualitative models, which only classify compounds into high, low or medium absorption classes, quantitative models rank compounds according to their relative values of intestinal absorption. Given the difficulty in analysing active transport, only the passive mechanism is usually considered for *in silico* prediction of intestinal absorption. Among the available models, the one built by Thomas *et al.* uses Caco-2 permeability in combination with kinetic solubility data. This model has demonstrated to be superior to several previously reported models based only on permeability or solubility.<sup>191</sup>

Drugs can bind to plasma proteins at constant rates, decreasing bioavailability and causing undesirable drug-drug interactions.<sup>192</sup> Many ligand- and structure based *in silico* models have been proposed and take into consideration two fundamental aspects: (1) binding rate and affinity, directly used to evaluate how tightly a drug binds to human serum albumin (HSA); (2) binding sites and poses, which may provide useful information for structure modification. Li *et al.* developed a multiple linear regression (MLR) model, in which both intramolecular descriptors (ligand properties) and intermolecular interaction descriptors (from docking results) are considered.<sup>193</sup>

The blood-brain barrier (BBB) plays a pivotal role in separating the brain from the blood. High BBB penetration is needed for most of the drugs targeting the central nervous system (CNS), whereas it should be minimised for non-CNS drugs to avoid undesired side-effects. Compounds may cross the BBB by passive diffusion or *via* a variety of catalysed transport systems (active efflux).<sup>194</sup> Different methodologies have been developed to measure the potential for novel compounds to penetrate the BBB and among these, *in vivo* brain uptake experiments provide the most reliable evaluation of BBB penetration. As in the case of absorption, most *in silico* models of BBB penetration are based on the assumption that compounds are transported by passive diffusion. To account for the contribution of transporters, Garg and Verma developed an ANN model based of the molecular structural parameters, and the probability of compounds to be P-glycoprotein (P-gp) substrates.<sup>195</sup> The result showed improved prediction performance, indicating that P-gp substrate probability plays an important role in BBB permeability. Lanevskij *et al.* developed a simple QSAR model with good predictive power of the brain/blood distribution ratio (log BB) and based on log *P*, p*K*<sub>a</sub>, and considering the influence of brain tissue binding by estimating the negative logarithm of the fraction that is unbound in the brain.<sup>196</sup> Recently, Carpenter *et al.* predicted the log BB and permeability surface-area product (log PS) of 12 small molecules by a using MD and binding free energy calculations in a simple BBB virtual model.<sup>197</sup>

Sudden death induced by hERG K<sup>+</sup> channels blockage is widely recognised as the predominant cause of drug-induced QT interval prolongation. Since a diverse range of drug structures can cause hERG toxicity, the early regulatory detection of compounds with this undesirable side effect has become another important objective in the pharmaceutical industry.<sup>198</sup> Seierstad and Agrafiotis developed a QSAR model for hERG toxicity prediction.<sup>199</sup> This model can quantitatively assess the cardiotoxicity of newly designed compounds and provide alerts for them. Good hERG toxicity prediction is also performed by SciQSAR, a comprehensive QSAR modelling system that enables researchers to establish reliable QSARs and QSPRs (quantitative structure–property models), create new calculators for *in silico* screening and generate new compound libraries.<sup>200,201</sup>

In 2002 Wajima *et al.* developed a MLR model, considering seven descriptors, to predict human clearance (CL) for a set of 68 drugs.<sup>202</sup> More recently, in 2013, Vijay *et al.* developed QSAR models for prediction of systemic CL and steady-state volume of distribution ( $V_{dss}$ ) from intravenous (i.v.) dosing in humans.

*In silico* ADMET modelling has made significant advances in the past decade. Many models have been proposed to address different aspects of pharmacokinetic and safety evaluations of drug-like molecules. However, this situation is still far from the prediction paradise that was expected by van de Waterbeemd and Gifford 13 years ago, in which *in silico* methods could support automated decision making in drug discovery.<sup>203</sup> Challenges such as the insufficient prediction reliability, the disconnection between experiments and models, and the lack of systematic perspective on intracorporal processes may be the main reasons why *in silico* ADMET modelling has not met this expectation yet. To resolve this problem, the communication between synthetic chemists and computational researchers must be strengthened, and the quality of models must be improved.

#### 11.3.4 Use of In Silico Ligand-based Approaches: A Practical Case Study on Antitubercular Agents

A recent example of how LBDD can be used for the discovery of new active hits is the identification and development of compounds against tuberculosis (TB).<sup>204</sup> In this study, we first identified a hit molecule (compound **24** in ref. 204) through pharmacophore modelling and screening. The pharmacophore was generated with Discovery Studio<sup>205</sup> using literature data and it was refined using a training set of 71 in-house molecules synthesised and already described as TB inhibitors *in vitro*. This model was then used to virtually screen the whole ASINEX database collection and an in-house library of azole compounds. The number of positively matching molecules was then reduced by keeping only those that showed at least 50% shape similarity with the most active compound of the training set. Of the final

27 small molecules, compound **24** inhibited *in vitro* TB with minimum inhibitory concentration (MIC) of  $8\ \mu\text{g mL}^{-1}$ . Thus, compound **24** was used as a starting query for the selection commercially available analogue compounds from the ZINC<sup>206</sup> database. After the application of a customised drug-like filter implemented in Omega2<sup>207</sup> the compounds were searched for shape matching with vROCS.<sup>208</sup> The procedure allowed the improvement of TB inhibition with the identification of a molecule with MIC of  $2\ \mu\text{g mL}^{-1}$ .

## 11.4 Conclusions

In an era of 'pharmaceutical crisis', R&D has turned to cutting-edge technology to reduce needed time and costs for the development of new drugs. Dramatic improvements in both software and hardware solutions have allowed molecular modelling to increasingly contribute in the last three decades at different levels in the drug discovery process, from target identification to the design and optimisation of drug delivery systems. Molecular modelling is able to support such a diversity of tasks because it collects a big variety of techniques that are able to study a system at different levels of theory: QM and MM. In this chapter, we have given a broad overview of the most important cutting-edge tools that are used in CADD to speed up the drug discovery process. Following the classical classification, we have divided the methods into structure- and ligand-based according to the type of the available source information. In structure-based molecular modelling we have discussed approaches like homology modelling, *de novo* protein structure building, docking, FBDD, core hopping, *de novo* lead optimisation practices and MD; whereas in ligand-based molecular modelling we have examined molecular descriptors, similarity assessments, molecular fingerprints, pharmacophores, QSAR models, CoMFA and QSPkR modelling. A fundamentally superior or universal technique does not exist and, indeed, real CADD studies select and combine *ad hoc* structure- and ligand-based tools according to situation. For example, a VS campaign might be ligand- or structure-based only or use a mixture of hierarchical filters, comprising for example 3D-pharmacophores, substructure and ADMET filters, docking simulations and MD. Going through the various methods, we have also examined their constant development towards always more reliable techniques. The always better level of accuracy of the results provided by programs and the broad variety of possible tool combinations are the reasons for which molecular modelling is able to deal with the most important issues related to the failure of drug development projects, like unsuitable ADMET properties. Given the positive trend in improvement, it is foreseeable that computational chemistry will not only continue to contribute significantly drug discovery, but that it will have an increasingly important role in reducing time and costs of the process. We are not there yet, but maybe one day computational chemistry will even override some experiments.



## Abbreviations

|              |                                                                          |
|--------------|--------------------------------------------------------------------------|
| NME          | new molecular entity                                                     |
| R&D          | research and development                                                 |
| ADMET        | absorption, distribution, metabolism, excretion and toxicity             |
| HTS          | high throughput screening                                                |
| POC          | proof of concept                                                         |
| CADD         | computer aided drug design                                               |
| MD           | molecular dynamics                                                       |
| MOA          | mechanism of action                                                      |
| VS           | virtual screening                                                        |
| FBDD         | fragment-based drug design                                               |
| (Q)SAR       | (quantitative) structure–activity relationship                           |
| QM           | quantum mechanics                                                        |
| MM           | molecular mechanics                                                      |
| SBDD         | structure-based drug design                                              |
| LBDD         | ligand-based drug design                                                 |
| NMR          | nuclear magnetic resonance                                               |
| GA           | genetic algorithm                                                        |
| PPI          | protein–protein interaction                                              |
| CAPRI        | critical assessment of prediction of interaction                         |
| C            | capsid                                                                   |
| MW           | molecular weight                                                         |
| HBD          | hydrogen bond donor                                                      |
| HBA          | hydrogen bond acceptor                                                   |
| PA           | polymerase acidic protein                                                |
| PB1          | polymerase basic protein 1                                               |
| cMD          | classical molecular dynamics                                             |
| TI           | thermodynamic integration                                                |
| LIE          | linear interaction energy                                                |
| FEP          | free energy perturbation                                                 |
| MMPBSA       | MM-Poisson Boltzmann surface area                                        |
| MMGBSA       | MM-Generalised Born surface area                                         |
| MC           | Monte Carlo                                                              |
| GPU          | graphical processing unit                                                |
| CV           | collective variable                                                      |
| log <i>P</i> | logarithmic of partition coefficient between <i>n</i> -octanol and water |
| Tc,          | Tanimoto coefficient                                                     |
| IUPAC        | International Union of Pure and Applied Chemistry                        |
| MRA          | multiple regression analysis                                             |
| CoMFA        | Comparative Molecular Field Analysis                                     |
| DYLOMMS      | Dynamic Lattice-Oriented Molecular Modeling System                       |
| PCA          | principal component analysis                                             |
| PK           | pharmacokinetic                                                          |
| QSPkR        | quantitative structure–pharmacokinetic relationships                     |

|        |                                   |
|--------|-----------------------------------|
| HSA    | human serum albumin               |
| MLR    | multiple linear regression        |
| BBB    | blood–brain barrier               |
| CNS    | central nervous system            |
| P-gp   | P-glycoprotein                    |
| log BB | brain/blood distribution ratio    |
| log PS | permeability surface-area product |
| QSPR   | quantitative structure–property   |
| CL     | clearance                         |
| TB     | tuberculosis                      |
| MIC    | minimum inhibitory concentration  |

## Acknowledgements

The authors are grateful to past and present members of the research team supervised by Prof. Maurizio Botta at the University of Siena for the scientific work conducted in the projects mentioned in this chapter.

## References

1. T. L. Lemke and D. A. Williams, *Foye's Principles of Medicinal Chemistry*, Lippincott Williams & Wilkins, Philadelphia, 5th edn, 2002, Design and development of drugs: general aspects, p. 12.
2. D. C. Young, *Computational Drug Design - A Guide for Computational and Medicinal Chemists*, John Wiley & Sons, Inc., New Jersey, 2009, p. 1.
3. B. Munos, *Nat. Rev. Drug Discovery*, 2009, **8**(12), 959.
4. R. M. Plenge, *Sci. Transl. Med.*, 2016, **8**(349), 345ps15.
5. <http://www.fda.gov> (last accessed March 2017).
6. K. I. Kaitin, *Clin. Pharmacol. Ther.*, 2010, **87**(3), 356.
7. H. van de Waterbeemd and E. Gifford, *Nat. Rev. Drug Discovery*, 2003, **2**(3), 192.
8. [http://www.sciencemag.org/site/products/ddbt\\_0207\\_Final.xhtml](http://www.sciencemag.org/site/products/ddbt_0207_Final.xhtml) (Accessed March 2017).
9. R. Macarron, M. N. Banks, D. Bojanic, D. J. Burns, D. A. Cirovic, T. Garyantes, D. V. S. Green, R. P. Herzberg, W. P. Janzen, J. W. Paslay, U. Schopfer and G. S. Sittampalam, *Nat. Rev. Drug Discovery*, 2011, **10**(3), 188.
10. S. F. Boys, *Proc. R. Soc. London, Ser. A.*, 1950, **200**, 542.
11. H.-D. Höltje and G. Folkers in *Molecular Modeling Basic Principles and Applications*, ed. R. Mannhold, H. Kubinyi and H. Timmerman, Wiley-VCH, Weinheim, New York, 1996, ch. 1, Introduction, p. 1.
12. A. Hillisch, N. Heinrich and H. Wild, *ChemMedChem*, 2015, **10**(12), 1958.
13. P. N. Kaul, G. Edwards, A. H. Weston, M. Rohmer, R. W. Rockhold, T. D. Johnson, J. M. Colacino and K. A. Staschke, *Progress in Drug*

- Research*, Birkhäuser Basel, Basel, ed. E. Jucker, 1998, vol. 50, Drug Discovery: past present and future, p. 9.
14. C. M. Song, S. J. Lim and J. C. Tong, *Briefings Bioinf.*, 2009, **10**(5), 579.
  15. M. H. Baig, K. Ahmad, S. Roy, J. M. Ashraf, M. Adil, M. H. Siddiqui, S. Khan, M. A. Kamal, I. Provazník and I. Choi, *Curr. Pharm. Des.*, 2016, **22**(5), 572.
  16. B. Kuhn, W. Cuba, J. Hert, D. Banner, C. Bissantz, S. Ceccarelli, W. Haap, M. Körner, A. Kuglstatter, C. Lerner, P. Mattei, W. Neidhart, E. Pinard, M. G. Rudolph, T. Schulz-Gasch, T. Woltering and M. Stahl, *J. Med. Chem.*, 2016, **59**(9), 4087.
  17. Loughney, B. L. Claus and S. R. Johnson, *Drug Discovery Today*, 2011, **16**(13–14), 548.
  18. T. L. Lemke and D. A. Williams, *Foye's Principles of Medicinal Chemistry*, Lippincott Williams & Wilkins, Philadelphia, 5th edn, 2002, Molecular Modeling and In Silico Drug Design, p. 68.
  19. A. R. Leach, *Molecular Modelling Principles and Applications*, Pearson Education Limited, Harlow, 2nd edn, 2001, Useful Concepts in Molecular Modelling, p. 1.
  20. W. Guba Stahl and M. Kansy, *Drug Discovery Today*, 2006, **11**(7–8), 326.
  21. Ouyang and S. C. Smith in *Computational Pharmaceutics Application of Molecular Modeling in Drug Delivery*, ed. D. Douroumis, A. Fahr, J. Siepmann, M. Snowden and V. Torchilin, Wiley and Sons, Ltd, Chichester, West Sussex, 2015, Preface, p. xv.
  22. A. R. Leach, *Molecular Modelling Principles and Applications*, Pearson Education Limited, Harlow, 2nd edn, 2001, An Introduction to Computational Quantum Mechanics, p. 26.
  23. T. Maximova, R. Moffatt, B. Ma, R. Nussinov and A. Shehu, *PLoS Comput. Biol.*, 2016, **12**(4), e1004619.
  24. C. Bissantz, B. Kuhn and M. Stahl, *J. Med. Chem.*, 2010, **53**(14), 5061.
  25. E. Persch, O. Dumele and F. Diederich, *Angew. Rev.*, 2015, **54**(11), 3290.
  26. A. R. Leach, *Molecular Modelling Principles and Applications*, Pearson Education Limited, Harlow, 2nd edn, 2001, The Use of Molecular Modelling and Chemoinformatics to Discover and Design New-Molecules, p. 640.
  27. M. Kuhnert and W. E. Diederich, *Synlett*, 2016, **27**(5), 641.
  28. A. C. Anderson, *Chem. Biol.*, 2003, **10**(9), 787.
  29. D. G. Brown and E. J. Shotton, *Philos. Trans. R. Soc., A*, 2015, **373**(2036), 20140468.
  30. M. Von Itzstein, W.-Y. Wu, G. B. Kok, M. S. Pegg, J. C. Dyason, B. Jin, T. Van Phan, M. L. Smythe, H. F. White, S. W. Oliver, P. M. Colman, J. N. Varghese, D. M. Ryan, J. M. Woods, R. C. Bethell, V. J. Hotham, J. M. Cameron and C. R. Penn, *Nature*, 1993, **363**(6428), 418.
  31. S. Wang, G. W. Milne, X. Yan, I. J. Posey, M. C. Nicklaus, L. Graham and W. G. Rice, *J. Med. Chem.*, 1996, **39**(10), 2047.
  32. <http://www.rcsb.org/pdb/home/home.do> (accessed March 2017).
  33. S. G. Dahl and I. Sylte, *Basic Clin. Pharmacol. Toxicol.*, 2005, **96**(3), 151.

34. Y. Zhang, *Curr. Opin. Struct. Biol.*, 2009, **19**(2), 145.
35. C. Sander and R. Schneider, *Proteins: Struct., Funct., Bioinf.*, 1991, **9**(1), 56.
36. A. R. Leach, *Molecular Modelling Principles and Applications*, Pearson Education Limited, Harlow, 2nd edn, 2001, Protein Structure Prediction, Sequence Analysis and Protein Folding, p. 509.
37. S. B. Needleman and C. D. Wunsch, *J. Mol. Biol.*, 1970, **48**(3), 443.
38. T. F. Smit and M. S. Waterman, *J. Mol. Biol.*, 1981, **147**(1), 195.
39. S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, *J. Mol. Biol.*, 1990, **215**(3), 403.
40. C. Notredame, D. G. Higgins and J. Heringa, *J. Mol. Biol.*, 2000, **302**(1), 205.
41. M. A. Marti-Renom, A. Stuart, A. Fiser, R. Sánchez, F. Melo and A. Sali, *Annu. Rev. Biophys. Biomol. Struct.*, 2000, **29**, 291.
42. M. P. Jacobson, D. L. Pincus, C. S. Rapp, T. J. F. Day, B. Honig, D. E. Shaw and R. A. Friesner, *Proteins: Struct., Funct., Bioinf.*, 2004, **55**(2), 351.
43. Y. Song, F. DiMaio, R. Y. Wang, D. Kim, C. Miles, T. Brunette, J. Thompson and D. Baker, *Structure*, 2013, **21**(10), 1735.
44. M. Biasini, S. Bienert, A. Waterhouse, K. Arnold, G. Studer, T. Schmidt, F. Kiefer, T. G. Cassarino, M. Bertoni, L. Bordoli and T. Schwede, *Nucleic Acids Res.*, 2014, **42**, W252.
45. E. Krieger, K. Joo, J. Lee, J. Lee, S. Raman, J. Thompson, M. Tyka, D. Baker and K. Karplus, *Proteins: Struct., Funct., Bioinf.*, 2009, **77**, 114.
46. A. L. Morris, M. W. MacArthur, E. G. Hutchinson and J. M. Thornton, *Proteins: Struct., Funct., Bioinf.*, 1992, **12**(4), 345.
47. M. Wiederstein and M. Sippl, *Nucleic Acid Res.*, 2007, **35**, W407.
48. L. J. McGuffin, M. T. Buenavista and D. B. Roche, *Nucleic Acid Res.*, 2013, **41**, W368.
49. D. Eisenberg, R. Lüthy and J. U. Bowie, *Methods Enzymol.*, 1997, **277**, 396.
50. R. Fazi, C. Tintori, A. Brai, L. Botta, M. Selvaraj, A. Garbelli, G. Maga and M. Botta, *J. Chem. Inf. Model.*, 2015, **55**(11), 2443.
51. V. S. Yedavalli, C. Neuvent, Y. H. Chi, L. Kleiman and K. T. Jeang, *Cell*, 2004, **119**(3), 381.
52. Y. Ariumi, M. Kuroki, K. Abe, H. Dansako, M. Ikeda, T. Wakita and N. Kato, *J. Virol.*, 2007, **81**(24), 13922.
53. C. G. Noble, Y. L. Chen, H. Dong, F. Gu, S. P. Lim, W. Schul, Q. Y. Wang and P. Y. Shi, *Antiviral Res.*, 2010, **85**(3), 450.
54. H. S. Chahar, S. Chen and N. Manjunath, *Virology*, 2013, **436**(1), 1.
55. M. Högbom, R. Collins, S. van der Berg, R. M. Jenvert, T. Karlberg, T. Kotenyova, A. Flores, G. B. Karlsson Hadestam and L. H. Schiavone, *J. Mol. Biol.*, 2007, **372**(1), 150.
56. M. Radi, F. Falchi, A. Garbelli, A. Samuele, V. Bernardo, S. Paolucci, F. Baldanti, S. Schenone, F. Manetti, G. Maga and M. Botta, *Bioorg. Med. Chem. Lett.*, 2012, **22**(5), 2094.

57. A. Brai, R. Fazi, C. Tintori, C. Zamperini, F. Bugli, M. Sanguinetti, E. Stigliano, J. Esté, R. Badia, S. Franco, M. A. Martinez, J. P. Martinez, A. Meyerhans, F. Saladini, M. Zazzi, A. Garbelli, G. Maga and M. Botta, *Proc. Natl. Acad. Sci. U. S. A.*, 2016, **113**(19), 5388.
58. S. Kalyaanamoorthy and Y.-P. P. Chen, *Prog. Biophys. Mol. Biol.*, 2014, **114**(3), 123.
59. L. G. Ferreira, R. N. Dos Santos, G. Oliva and A. D. Andricopulo, *Molecules*, 2015, **20**(7), 13384.
60. J. de Ruyck, G. Brysbaert, R. Blossey and M. F. Lensink, *Adv. Appl. Bioinf. Chem.*, 2016, **9**, 1.
61. T. J. Ewing, S. Makino, A. G. Skillman and I. D. Kuntz, *J. Comput.-Aided Mol. Des.*, 2001, **15**(5), 411.
62. B. Kramer, M. Rarey and T. Lengauer, *Proteins: Struct., Funct., Bioinf.*, 1999, **37**(2), 228.
63. M. L. Verdonk, J. C. Cole, M. J. Hartshorn, C. W. Murray and R. D. Taylor, *Proteins: Struct., Funct., Bioinf.*, 2003, **52**(4), 609.
64. D. S. Goodsell, G. M. Morris and A. J. Olson, *J. Mol. Recognit.*, 1996, **9**(1), 1.
65. O. Trott and A. J. Olson, *J. Comput. Chem.*, 2010, **31**(2), 455.
66. R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, D. Shaw, M. Shelley, J. K. Perry, P. Francis and P. S. Shenkin, *J. Med. Chem.*, 2004, **47**(7), 1739.
67. 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**(7), 1750.
68. I. W. Davis and D. Baker, *J. Mol. Biol.*, 2009, **385**(2), 381.
69. A. Grosdidier, V. Zoete and O. Michielin, *Nucleic Acid Res.*, 2011, **39**, W270.
70. D. B. Kitchen, H. Decornez, J. R. Furr and J. Bajorath, *Nat. Rev. Drug Discovery*, 2004, **3**(11), 935.
71. P. J. Gane and P. M. Dean, *Curr. Opin. Struct. Biol.*, 2000, **10**(4), 401.
72. A. M. Ferrari, B. Q. Wei, L. Costantino and B. K. Shoichet, *J. Med. Chem.*, 2004, **47**(21), 5076.
73. D. B. Kokh, R. C. Wade and W. Wenzel, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2011, **1**(2), 298.
74. A. R. Leach, *J. Mol. Biol.*, 1994, **235**(1), 345.
75. J. Liu and R. Wang, *J. Chem. Inf. Model.*, 2015, **55**(3), 475.
76. C. Tintori, A. Brai, A. L. Fallacara, R. Fazi, S. Schenone and M. Botta, *Curr. Opin. Pharmacol.*, 2014, **18**, 1.
77. A. Loregian and G. Palù, *J. Cell. Physiol.*, 2005, **204**(3), 750.
78. U. Kessler, D. Castagnolo, M. Pagano, D. Deodato, M. Bernardini, B. Pilger, C. Ranadheera and M. Botta, *Bioorg. Med. Chem. Lett.*, 2013, **23**(20), 5575.
79. M. Pagano, D. Castagnolo, M. Bernardini, A. L. Fallacara, I. Laurenzana, D. Deodato, U. Kessler, B. Pilger, L. Stergiou, S. Strunze, C. Tintori and M. Botta, *ChemMedChem*, 2014, **9**(1), 129.
80. C. Tintori, I. Laurenzana, A. L. Fallacara, U. Kessler, B. Pilger, L. Stergiou and M. Botta, *Bioorg. Med. Chem. Lett.*, 2014, **24**(1), 280.

81. I. M. L. Trist, G. Nannetti, C. Tintori, A. L. Fallacara, D. Deodato, B. Mercorelli, G. Palù, M. Wijtmans, T. Gospodova, E. Edink, M. Verheij, I. de Esch, L. Viteva, A. Loregian and M. Botta, *J. Med. Chem.*, 2016, **59**(6), 2688.
82. M. Mori, G. Vignaroli and M. Botta, *Drug Discovery Today Technol.*, 2013, **10**(4), e541.
83. J. Janin, K. Henrick, J. Moult, L. T. Eyck, M. J. Sternberg, S. Vajda, I. Vakser and S. J. Wodak, *Proteins: Struct., Funct., Bioinf.*, 2003, **52**(1), 2.
84. P. Scaturro, I. M. Trist, D. Paul, A. Kumar, E. G. Acosta, C. M. Byrd, R. Jordan, A. Brancale and Bartenschlager, *J. Virol.*, 2014, **88**(19), 11540.
85. G. Schneider, *Drug Discovery Today: Technol.*, 2013, **10**(4), e453.
86. A. C. Anderson, *Chem. Biol.*, 2003, **10**(9), 787.
87. S. H. Rotstein and M. A. Murcko, *J. Comput.-Aided Mol. Des.*, 1993, **7**(1), 23.
88. R. X. Wang, Y. Gao and L. Lai, *J. Mol. Model.*, 2000, **6**(7), 498.
89. H. M. Vinkers, M. R. de Jonge, F. F. Daeyaert, J. Heeres, L. M. Koymans, J. H. van Lenthe, P. J. Lewi, H. Timmerman, K. Van Aken and P. A. Janssen, *J. Med. Chem.*, 2003, **46**(13), 2765.
90. V. Gillet, A. P. Johnson, P. Mata, S. Sike and P. Williams, *J. Comput.-Aided Mol. Des.*, 1993, **7**(2), 127.
91. P. Maass, T. Schulz-Gasch, M. Stahl and M. Rarey, *J. Chem. Inf. Model.*, 2007, **47**(2), 390.
92. CombiGlide, Schrödinger, LLC, New York, NY, 2016.
93. P. J. Hajduk and J. Greer, *Nat. Rev. Drug Discovery*, 2007, **6**(3), 211.
94. M. Congreve, G. Chessari, D. Tisi and A. J. Woodhead, *J. Med. Chem.*, 2008, **51**(13), 3661.
95. G. E. de Kloe, D. Bailey, R. Leurs and I. J. P. de Esch, *Drug Discovery Today*, 2009, **14**(13–14), 630.
96. F. Benmansour, I. Trist, B. Coutard, E. Decroly, G. Querat, A. Brancale and K. Barral, *Eur. J. Med. Chem.*, 2017, **125**, 865.
97. C. A. Lipinski, F. Lombardo, B. V. Dominy and P. J. Feeney, *Adv. Drug Delivery Rev.*, 1997, **23**, 3.
98. C. A. Lipinski, *Drug Discov. Today Technol.*, 2004, **1**, 337.
99. D. E. Scott, A. G. Coyne, S. A. Hudson and C. Abell, *Biochemistry*, 2012, **51**(25), 4990.
100. D. E. Scott, C. Spry and C. Abell, in *Fragment-based Drug Discovery*, ed. D. A. Erlanson and W. Jahnke, Wiley-VCH Verlag GmbH & Co., Weinheim, 2016, Differential Scanning Fluorimetry as Part of a Biophysical Screening Cascade, p. 139.
101. S. Ohlson, M.-D. Duong-Thi, *Fragment-based Drug Discovery*, ed. D. A. Erlanson and W. Jahnke, Wiley-VCH Verlag GmbH & Co., Weinheim, 2016, Emerging Technologies for Fragment Screening, p. 173.
102. R. E. Hubbard, *Fragment-based Drug Discovery*, ed. D. A. Erlanson and W. Jahnke, Wiley-VCH Verlag GmbH & Co., Weinheim, 2016, The role of Fragment-based Discovery in Lead Finding, p. 173.

103. H.-J. Böhm, A. Flohr and M. Stahl, *Drug Discovery Today Technol.*, 2004, 1(3), 217.
104. H. Sun, G. Tawa and A. Wallqvist, *Drug Discovery Today*, 2012, 17(7–8), 310.
105. A. Shuffenhauer, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2012, 2(6), 842.
106. Y. Hu, D. Stumpfe and J. Bajorath, *J. Med. Chem.*, 2017, 60(4), 1238.
107. I. G. Rodriguez-Bussey, U. Doshi and D. Hamelberg, *Biopolymers*, 2015, 105(1), 35.
108. M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess and E. Lindahl, *SoftwareX*, 2015, 1–2, 19.
109. B. R. Brooks, C. L. Brooks, A. D. Mackerell Jr, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York and M. Karplus, *J. Comput. Chem.*, 2009, 30(10), 1545.
110. R. Salomon-Ferrer, D. A. Case and R. C. Walker, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2013, 3(2), 198.
111. J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale and K. Schulten, *J. Comput. Chem.*, 2005, 26(16), 1781.
112. H. Harvey, G. Giupponi and G. De Fabritis, *J. Chem. Theory and Comput.*, 2009, 5(6), 1632.
113. H. Zhao and A. Caflisch, *Eur. J. Med. Chem.*, 2015, 91, 4.
114. S. Kalyaanamoorthy and Y.-P. P. Chen, *Prog. Biophys. Mol. Biol.*, 2014, 114(3), 123.
115. W. L. Jorgensen and L. T. Thomas, *J. Chem. Theory Comput.*, 2008, 4(6), 869.
116. C. Tintori, A. L. Fallacara, M. Radi, C. Zamperini, E. Dreassi, E. Crespan, G. Maga, S. Schenone, F. Musumeci, C. Brullo, A. Richters, F. Gasparrini, A. Angelucci, C. Festuccia, S. Delle Monache, D. Rauh and M. Botta, *J. Med. Chem.*, 2015, 58(1), 347.
117. S. Genheden and U. Ryde, *Expert Opin. Drug. Discovery*, 2015, 10(5), 449–461.
118. C. Tintori, G. La Sala, G. Vignaroli, L. Botta, A. L. Fallacara, F. Falchi, M. Radi, C. Zamperini, E. Dreassi, L. Dello Iacono, D. Orioli, G. Biamonti, M. Garbelli, A. Lossani, F. Gasparrini, T. Tuccinardi, I. Laurenzana, A. Angelucci, G. Maga, S. Schenone, C. Brullo, F. Musumeci, A. Desogus, E. Crespan and M. Botta, *J. Med. Chem.*, 2015, 58(11), 4590.
119. F. Feixas, S. Lindert, W. Sinko and J. A. McCammon, *Biophys. Chem.*, 2014, 186, 31.
120. R. E. Amaro and M. Bansal, *Curr. Opin. Struct. Biol.*, 2014, 25, iv.
121. B. K. Ho and D. A. Agard, *PLoS Comput. Biol.*, 2009, 5(4), e1000343.



122. J. R. Wagner, C. T. Lee, J. D. Durrant, R. D. Malmstrom, V. A. Feher and R. E. Amaro, *Chem. Rev.*, 2016, **116**(11), 6370.
123. A. Liwo, C. Czaplewski, S. Ołdziej and H. A. Scheraga, *Curr. Opin. Struct. Biol.*, 2008, **18**(2), 134.
124. M. Christen and W. F. van Gunsteren, *J. Comput. Chem.*, 2008, **29**(2), 157.
125. A. Abrams and G. Bussi, *Entropy*, 2014, **16**(1), 163.
126. V. Spiwok, Z. Sucur and P. Hosek, *Biotechnol. Adv.*, 2015, **33**(6), 1130.
127. R. C. Bernardia, M. C. R. Melob and K. Schulten, *Biochim. Biophys. Acta, Gen. Subj.*, 2015, **1850**(5), 872.
128. M. Chavent, A. L. Duncan and M. S. P. Sansom, *Curr. Opin. Struct. Biol.*, 2016, **40**, 8.
129. T. Reddy and M. S. P. Sansom, *Biochim. Biophys. Acta, Biomembr.*, 2016, **1858**(7), 1610.
130. S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A. E. Dawid and A. Kolinski, *Chem. Rev.*, 2016, **116**(14), 7898.
131. Y. Miao and J. A. McCammon, *Mol. Simul.*, 2016, **42**(13), 1046.
132. W. Thiel and G. Hummer, *Nature*, 2013, **504**, 96.
133. C. Lemmen and T. Lengauer, *J. Comput. Aided Mol. Des.*, 2000, **14**, 215.
134. J. W. Black, A. F. Crowther, R. G. Shanks, L. H. Smith and A. C. Dornhorst, *Lancet*, 1964, **1**(7342), 1080.
135. R. W. Brimblecombe, W. A. Duncan, G. J. Durant, C. R. Ganellin, M. E. Parsons and J. W. Black, *Br. J. Pharmacol.*, 1975, **53**(3), 435P.
136. J. Schwartz, M. Awale and J. L. Reymond, *J. Chem. Inf. Model.*, 2013, **53**, 1979.
137. J. M. Durant, B. A. Leland, D. R. Henry and J. G. Nourse, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 1273.
138. L. Mavridis, B. D. Hudson and D. W. Ritchie, *J. Chem. Inf. Model.*, 2007, **47**, 1787.
139. R. D. Cramer, D. E. Patterson and J. D. Bunce, *J. Am. Chem. Soc.*, 1988, **110**, 5959.
140. M. Randic, *J. Chem. Inf. Comput. Sci.*, 1995, **35**, 373.
141. J. H. Shuur, P. Selzer and J. Gasteiger, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 334.
142. R. Todeschini, V. Consonnini, R. Mannhold, *Molecular Descriptors for Chemoinformatics*, ed. H. Kubinyi and G. Folkers, Wiley-VCH Verlag GmbH & Co, Weinheim, 2nd edn, 2009, vol. 41, p. 1257.
143. S. Ekins, J. Mestres and B. Testa, *Br. J. Pharmacol.*, 2007, **152**, 9.
144. T. Zhou, D. Huang and A. Caflisch, *Curr. Top. Med. Chem.*, 2010, **10**, 33–45.
145. C. Hansch, P. P. Maloney, T. Fujita, R. M. Muir, *Nature*, 1962, **194**, 178.
146. R. F. Rekker, R. Mannhold, *Calculation of Drug Lipophilicity: The Hydrophobic Fragment Constant Approach*, VCH, Weinheim, New York, 1992.
147. G. E. Kellog, S. F. Semus and D. J. Abraham, *J. Comput.-Aided Mol. Des.*, 1991, **5**, 545.

148. R. X. Wang, Y. Fu and L. Lai, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 615.
149. L. Xing and R. C. Glen, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 796.
150. J. I. Brauman and L. K. Blair, *J. Am. Chem. Soc.*, 1968, **90**, 6561.
151. J. Gasteiger and M. G. Hutchings, *J. Am. Chem. Soc.*, 1984, **106**, 6489.
152. R. C. Glen, *J. Comput.-Aided Mol. Des.*, 1994, **8**, 457.
153. S. L. Dixon and R. T. Koehler, *J. Med. Chem.*, 1999, **42**, 2887.
154. M. A. Fligner, J. S. Verducci and P. E. Blower, *Technometrics*, 2002, **44**, 110.
155. J. Bajorath, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 233.
156. J. Bajorath, *Nat. Rev. Drug Discovery Today*, 2002, **1**, 882.
157. J. Auer and J. Bajorath, *Methods Mol. Boil.*, 2008, **453**, 327.
158. A. Cereto-Massaguè, M. J. Ojeda, C. Vallas, M. Mulero, S. Garcia-Vallvé and G. Pujadas, *Methods*, 2014, **71**, 58.
159. P. Ehrlich, *Ber. Dtsch. Chem. Ges.*, 1909, **42**, 17.
160. C. G. Wermuth, C. R. Ganellin, P. Lindberg and L. A. Mitscher, Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1997).
161. *Pharmacophore Perception, Development, and Use in Drug Design*, ed. P. Günd and O. F. Güner, International University Line (IUL), La Jolla CA, 2000, vol 1, pp. 171–189.
162. C. G. Wermuth, *Pharmacophores: historical perspective and viewpoint from a medicinal chemist*, ed. T. Langer, R. D. Hoffmann, Wiley-VCH, Weinheim, 2006, **32**, 3.
163. O. Dror, A. Shulman-Peleg, R. Nussinov and H. J. Wolfson, *Front. Med. Chem.*, 2006, **3**, 551–584.
164. M. Akram, T. Kaserer and D. Schuster, *Pharmacophore modeling and screening. In silico Drug Discovery and Design: Theory, Methods, Challenges and Applications*, ed. C. Cavasotto, CRC Press, Boca Raton, FL, USA, 2015, 123.
165. A. Vuorinen and D. Schuster, *Methods*, 2015, **71**, 113.
166. N. Lagarde, J. F. Zagury and M. Montes, *J. Chem. Inf. Model.*, 2015, **55**, 1297.
167. K. Heikamp and J. Bajorath, *J. Chem. Inf. Model.*, 2013, **53**, 1595.
168. A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani and J. P. Overington, *Nucleic Acids Res.*, 2011, **40**, D1100.
169. D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang and J. Woolsey, *Nucleic Acids Res.*, 2006, **34**, D668.
170. A. J. Williams, L. Harland, P. Groth, S. Pettifer, C. Chichester, E. L. Willighagen, C. T. Evelo, N. Blomberg, G. Ecker, C. Goble and B. Mons, *Drug Discovery Today*, 2012, **17**, 1188.
171. M. L. Verdonk, V. Berdini, M. J. Hartshorn, W. T. M. Mooij, C. W. Murray, R. D. Taylor and P. Watson, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 793.
172. J. Kirchmair, P. Markt, S. Distinto, G. Wolber and T. Langer, *J. Comput. Aided Mol. Des.*, 2008, **22**, 213.
173. N. Huang, B. K. Shoichet and J. J. Irwin, *J. Med. Chem.*, 2006, **49**, 6789.

174. M. M. Mysinger, M. Carchia, J. J. Irwin and B. K. Shoichet, *J. Med. Chem.*, 2012, **55**, 6582.
175. M. Olah, M. Mracec, L. Ostopovici, R. Rad, A. Bora, N. Hadaruga, I. Olah, M. Banda, Z. Simon, M. Mracec and T. I. Oprea, *Chemical biology: from small molecules to systems biology and drug design*, ed. S. L. Schreiber, T. M. Kapoor and G. Wess, Wiley-VCH, Weinheim, 2007, vols. 1–3, issue 13, p. 760.
176. A. Tropsha, *Mol. Inf.*, 2010, **29**, 476.
177. N. R. Draper and H. Smith, *Applied Regression Analysis*, 2nd edn, 1981, Jhon Wiley and Sons, New York.
178. C. Hansch and T. Fujita, *J. Am. Chem. Soc.*, 1964, **861**, 616.
179. C. Hansch, *Acc. Chem. Res.*, 1969, **2**, 232.
180. S.M. Free and J.W. Wilson, *J. Med. Chem.*, 1964, **7**, 395.
181. M. Wise, R. D. Cramer and D. Smith, Exman I progress in 3D drug design: the use of real time colour graphics and computer postulationof bioactive ,olecules in DYLOMMSIn quantitative approaches in drug design Elsevier Amsterdam, uk 1983 pp. 145–146.
182. A. Hopfinger, S. Wang, J. Tokarski, B. Jin, M. Albuquerque, P. Madhav and C. Duraiswami, *J. Am. Chem. Soc.*, 1997, **119**, 10509.
183. A. Vedani and M. Dolber, *J. Med. Chem.*, 2002, **45**, 2139.
184. A. Vedani, M. Dolber and M. A. Lill, *J. Med. Chem.*, 2005, **48**, 3700.
185. B. L. Podlogar and D. M. Ferguson, *Drug Des. Discovery*, 2000, **17**, 4.
186. J. Wang and L. Urban, *Drug Discovery. World*, 2004, **5**, 73.
187. U. Norinder and C. A. S. Bergstroem, *Chem. Biol.*, 2007, **3**, 1003.
188. V. K. Gombar, I. S. Silver and Z. Zhao, *Curr. Topics Med. Chem.*, 2000, **3**, 1205.
189. P. Artusson and L. Karlsson, *Biochem. Biophys. Res. Commun.*, 1991, **175**, 880.
190. M. D. Wessel, P. C. Jurs, J. W. Tolan and S. M. Muskal, *J. Comp. Inf. Comput. Sci.*, 1998, **38**, 726.
191. S. Thomas, F. Brightman, H. Gill, S. Lee and B. Pufong, *J. Pharma. Sci.*, 2008, **97**, 4557.
192. G. L. Trainor, *Expert Opin. Drug Discov.*, 2007, **2**, 51.
193. H. Y. Li, Z. X. Chen, X. J. Xu, X. F. Sui, T. Guo, W. Liu and J. W. Zhang, *Biopharm. Drug Disp.*, 2011, **32**, 333.
194. D. E. Clark, *Drug Discovery Today*, 2003, **8**, 927.
195. P. Garg and J. Verma, *J. Chem. Inf. Mod.*, 2006, **46**, 289.
196. K. Lanevskij, J. Dapkunas, L. Juska, P. Japertas and R. Didziapetris, *J. Pharma. Sci.*, 2011, **100**, 2147.
197. T. S. Carpenter, D. A. Kirshner, E. Y. Lau, S. E. Wong, J. P. Nilmeier and F. C. Lightstone, *Biophys. J.*, 2014, **107**, 630.
198. A. M. Aronov, *Drug Discovery Today*, 2005, **10**, 149.
199. M. Seierstad and D. K. Agrafiotis, *Chem. Biol. Drug Des.*, 2006, **67**, 284.
200. J. F. Contrera, E. J. Matthews and R. D. Benz, *Reg. Tox. Pharm.*, 2003, **38**, 243.
201. W. G. MUuster, A. Breidenbach, H. Fischer, S. Kirchner, L. Muller and A. Pahler, *Drug Discovery Today*, 2008, **13**, 303.

202. T. Wajima, K. Fukumura, Y. Yano and T. Oguma, *J. Pharm. Sci.*, 2002, **91**, 2489.
203. H. Van De Waterbeemd and E. Gifford, *Nat. Rev. Drug Discovery*, 2003, **2**, 192.
204. D. De Vita, F. Pandolfi, R. Cirilli, L. Scipione, R. Di Santo, L. Friggeri, M. Mori, D. Fiorucci, G. Maccari, R. S. Arul Christophel, C. Zamperini, V. Pau, A. De Logu, S. Tortorella and M. Botta, *Eur. J. Med. Chem.*, 2016, **121**, 169.
205. Inc Accelrys, Catalyst 4.6 Program, Accelrys, Inc., San Diego, CA, 2000.
206. J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad and R. G. Coleman, *J. Chem. Inf. Model.*, 2012, **52**, 1757.
207. OpenEye, OMEGA 2.5.1.4: OpenEye Scientific Software, Santa Fe, NM.
208. OpenEye, ROCS 3.2.0.4 OpenEye Scientific Software, Santa Fe, NM.

# *Molecular Modelling and Simulations Applied to Challenging Drug Discovery Targets*

MARCO DE VIVO,<sup>\*a,b</sup> MATTEO MASETTI<sup>\*c</sup> AND  
GIULIA ROSSETTI<sup>\*b,d,e</sup>

<sup>a</sup> Laboratory of Molecular Modeling and Drug Discovery, Istituto Italiano di Tecnologia, Via Morego 30, 16163 Genoa, Italy; <sup>b</sup> IAS-S/INM-9 Computational Biomedicine Forschungszentrum Jülich, Wilhelm-Johnen-Straße 52428 Jülich, Germany; <sup>c</sup> Department of Pharmacy & Biotechnology, Alma Mater Studiorum, University of Bologna, Via Belmeloro 6, 40126 Bologna, Italy; <sup>d</sup> Department of Oncology, Hematology and Stem Cell Transplantation, University Hospital Aachen, RWTH Aachen University, Pauwelsstraße 30, 52074 Aachen, Germany; <sup>e</sup> Jülich Supercomputing Centre (JSC) – Division Computational Science – Simulation Laboratory Biology, Forschungszentrum Jülich, 52425 Jülich, Germany  
\*Email: marco.devivo@iit.it; g.rossetti@fz-juelich.de; matteo.masetti4@unibo.it

## **12.1 Introduction**

Molecular modelling and simulations are computational approaches that are nowadays routinely used to speed up the challenging and costly drug discovery process.<sup>1–3</sup> Computations have had a positive impact on all aspects

of designing and optimising promising small molecules.<sup>4</sup> In structure-based drug design, for example, where the target protein's structural data are available, standard computational methods such as docking and virtual screening have found broad applications and effectiveness in the hit identification phase, in which chemical compounds with a promising activity against the target are identified.<sup>5</sup> However, one of docking's major limitations is that it does not account for full target flexibility. A more accurate description of the ligand–target interaction is needed in the lead generation and optimisation phase, in which hit compounds are improved to produce drug-like molecules that can exert their beneficial pharmacological effect in patients. Here, other computational methods such as molecular dynamics (MD) and quantum mechanics (QM) can be more appropriate and informative.<sup>1,6,7</sup>

MD simulations, for example, are currently used to investigate and describe the time evolution of drug targets such as proteins and receptors. MD simulation timescales can reach up to micro and milliseconds, with simulations accounting for explicit protein solvation in water and (when needed) in the presence of the membrane.<sup>8–10</sup> MD is particularly appropriate for drug discovery applications when coupled to methods for calculating the free energy of binding.<sup>1,11</sup> MD can thus depict the overall process of the ligand's binding/unbinding to/from its target, fully accounting for the system's flexibility. This is MD's major advantage for drug discovery. Furthermore, MD is currently used and tested as a practical computational tool for studying the kinetics of a ligand binding to its target.<sup>12</sup> This is important for lead optimisation and represents a new frontier in computationally driven drug discovery. Indeed, compounds have recently been evaluated not only for their ability to bind tightly to the target but also for their capacity to remain bound to the target for a long time (*i.e.* by considering  $k_{\text{on}}$  and  $k_{\text{off}}$  of binding), which increases the chances of efficacy *in vivo*.<sup>8–10</sup>

Similarly, QM-based methods, including the hybrid quantum mechanics/molecular mechanics (QM/MM) approach,<sup>13,14</sup> are also on the verge of finding practical utility in drug discovery.<sup>7</sup> So far, QM-based drug design methods have mainly involved QM-corrected scoring functions for docking calculations and binding affinity estimations.<sup>15</sup> Other examples include QM-based electrostatic potential maps used to characterise the binding site of receptors, with calculations that investigate the protonation states of crucial residues within the binding pocket. Finally, QM is often used to decipher the reaction mechanisms of pharmaceutically relevant enzymes.<sup>15–18</sup>

In this chapter, we will demonstrate how molecular modelling and simulations can nowadays return informative mechanistic insights into difficult drug discovery targets.<sup>19–21</sup> Our major focus will be MD simulations and QM-based methods. Using examples drawn mostly from our own research, we will report on computational investigations of complex structural molecules that are important drug discovery targets. These molecules include G-protein-coupled receptors (GPCRs), transmembrane proteins and

ion channels, and metalloproteins for nucleic acid processing. While not exhaustive, these selected examples will show how computational methods can be used effectively to facilitate de novo enzyme design, molecular medicine, and, ultimately, drug discovery.

## 12.2 Deciphering Metalloenzyme Catalysis *via* Computations

We will first consider some representative studies, mainly from our own research, of pharmaceutically relevant metalloenzymes. Metalloenzymes are ubiquitous in all life kingdoms, being crucial for a variety of vital processes.<sup>22</sup> As a result, they are often targeted to treat serious diseases like cancer and neurodegeneration,<sup>23–27</sup> but they are challenging drug design targets. Metalloenzymes use metal ions to efficiently perform catalysis. When considering their function, the central challenge is thus to understand how metal ions help maintain structural stability during catalysis, while also allowing the conformational rearrangements necessary for an efficient catalysis.

### 12.2.1 Ribonuclease H

Here, we will consider the QM/MM studies of the ribonuclease H (RNase H).<sup>28</sup> This enzyme catalyses nucleotidyl transfer reactions, with two  $\text{Mg}^{2+}$  ions contained in the catalytic site. Importantly, RNase H is a promising target for antiviral drugs.<sup>29–31</sup>

RNase H hydrolyses the phosphodiester linkages in the RNA strand of RNA–DNA hybrids.<sup>32,33</sup> To perform this reaction, RNase H uses two metal ions in the catalytic site. These two metal ions are needed to accommodate the large negative charge generated by the backbone of substrate RNA and DNA strands.<sup>29</sup> This divalent bimetal architecture of the RNase H catalytic site is critical in maintaining the structural integrity of the protein–substrate complex, promoting nucleophilic attack on the scissile phosphate, and stabilising the transition state (TS) and leaving group exit during catalysis.<sup>32–34</sup> However, extensive experimental studies of RNase H's enzymatic activity have revealed that variations in the catalytic metal ion and/or its concentration impact the enzyme's catalytic efficiency.

To understand the mechanistic details of the bimetal-aided nucleotidyl transfer reaction, hybrid QM/MM Car–Parrinello molecular dynamics (CPMD) were run to study the enzymatic reaction with two different reagent states (Figure 12.1). Namely, the nucleophilic species was either a water molecule ( $P_{\text{WAT}}$ ) or a hydroxide ion ( $P_{\text{OH}^-}$ ).

It was found that  $P_{\text{OH}^-}$  had the lowest free-energy barrier of  $\sim 10.5 \text{ kcal mol}^{-1}$ . However, the reaction with  $P_{\text{WAT}}$  was competitive, with a free-energy barrier of  $\sim 16 \text{ kcal mol}^{-1}$ , if  $\sim 3 \text{ kcal mol}^{-1}$  of dehydration energy was also taken into consideration. Both  $P_{\text{OH}^-}$  and  $P_{\text{WAT}}$  showed an in-line  $\text{SN}_2$ -like nucleophilic



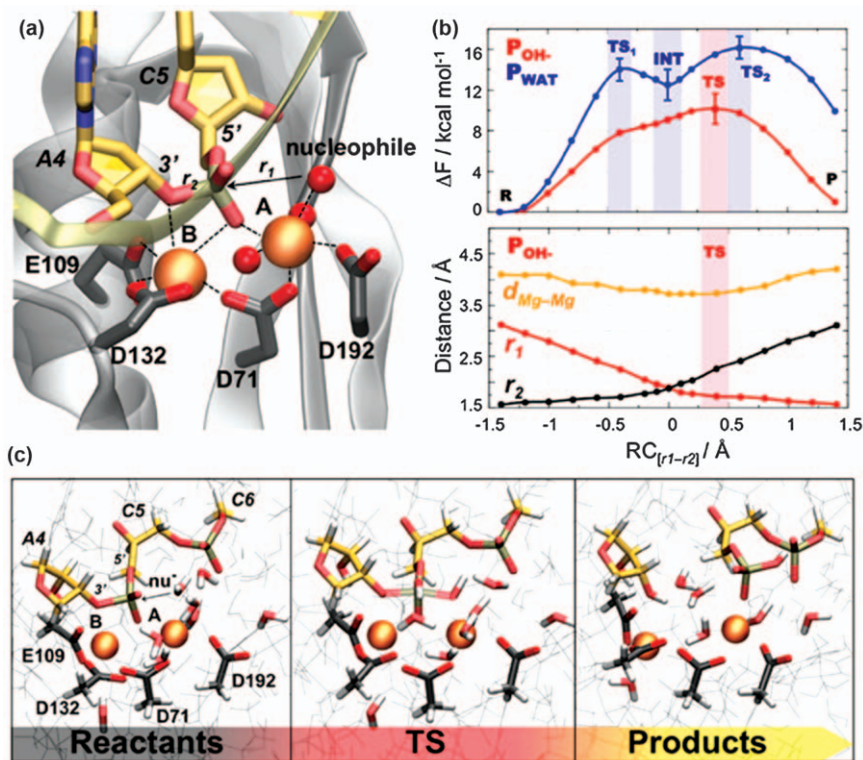


**Figure 12.1** (a) Catalytic site of the RNase H in complex with an RNA–DNA hybrid,<sup>32</sup> showing typical two-metal-aided architecture. (b) Energetic (upper graph) and structural (lower graph) properties for the RNA phosphodiester cleavage catalysed by RNase H *via* the two-metal mechanism, as from Car-Parrinello QM/MM MD.<sup>28</sup> The free-energy profiles for the two investigated pathways ( $P_{\text{WAT}}$ , or one hydroxide ion,  $P_{\text{OH}^-}$ ) are shown. Transition state (TS) and the metastable intermediate (INT) are shown. Average bond distances [ $r_1$ ,  $r_2$ , and the metal–metal distance ( $d_{\text{Mg-Mg}}$ )] along the lower free-energy path ( $P_{\text{OH}^-}$ ) are reported in the lower graph. (c) Representative snapshots from the QM/MM MD show the reactants, TS, and product states of RNase H catalysis. QM atoms are shown explicitly; MM atoms are in thin lines. Reproduced from G. Palermo *et al.*, *Acc. Chem. Res.*, 2015, **48**(2), 220–228. Copyright (2015) American Chemical Society.

attack on the scissile phosphorus (Figure 12.1), suggesting an associative mechanism with formation of phosphorane-like transition states.<sup>35</sup> Interestingly, this phosphorane intermediate has, to date, only been observed in the debated  $\beta$ -phosphoglucomutase crystal.<sup>36</sup> However, the presence of this intermediate was also later suggested by the QM/MM studies of Elsässer *et al.*<sup>37</sup> and Rosta *et al.*<sup>38,39</sup> Importantly, there was qualitative agreement with the kinetic data for substrate analogs for HIV-1 RNase H activity, which corroborated these mechanistic details.<sup>40</sup>

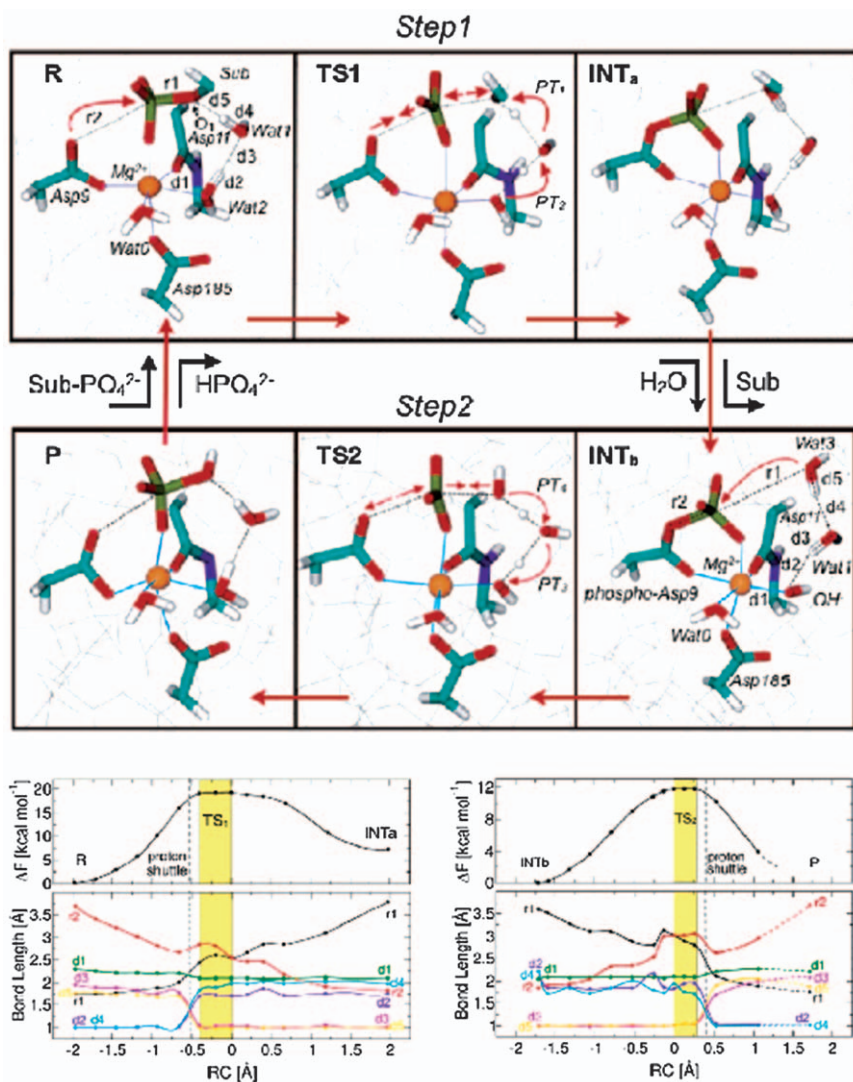
Importantly, these QM/MM studies demonstrated that the two  $\text{Mg}^{2+}$  ions act in a co-operative fashion to help the reaction efficiency. Indeed, the study revealed that the two  $\text{Mg}^{2+}$  ions operate simultaneously to catalyse both nucleophile formation and leaving group stabilisation. In particular, the two ions draw closer to each other in the TS geometry to aid the formation of a phosphorane-like geometry.<sup>28</sup>

### 12.2.2 Epoxide hydrolase

The same computational approach was used to study novel metal ( $\text{Mg}^{2+}$ )-dependent phosphatase activity in the N-terminal domain of soluble epoxide hydrolase (sEH).<sup>41,42</sup> Importantly, this ( $\text{Mg}^{2+}$ )-dependent phosphatase activity opened a new frontier for drug discovery efforts centered on fatty acid metabolism.<sup>43–45</sup> Here too, crystal structures<sup>46–48</sup> were used to set up and run CP QM/MM simulations, which returned a principles-based interpretation of the experimental findings. Again, the most interesting detail revealed by these studies was how the  $\text{Mg}^{2+}$  ion helps the reaction efficiency (Step 1 and Step 2, Figure 12.2). The simulations demonstrated the crucial role of metal–substrate-connecting water bridges (WBs) in transferring the protons that are essential for nucleophile formation (water deprotonation) and leaving group exit. The electrostatic effects of the metal ion and the metal–substrate-connecting WBs were found to be crucial during the two phosphoryl transfers of the catalytic cycle. These computational results were in fair agreement with the experimental findings, pointing to Step 1 ( $\sim 19 \text{ kcal mol}^{-1}$ ) as the catalytic cycle's rate-determining step (Figure 12.2). The most important contribution to enhancing catalytic efficiency came from the nucleophile and leaving group stabilisation *via* WB-mediated proton shuttles. This was mostly induced by the metal ion's electrostatic effects.

It is instructive to compare the QM/MM results for RNase H (two  $\text{Mg}^{2+}$  cations) and those for sEH (one  $\text{Mg}^{2+}$  cation). The different phosphoryl transfer mechanisms (associative *vs.* dissociative) seem to be dictated by the geometry and stoichiometry of the metal(s) during catalysis. In the RNase H, the attacking and leaving groups are stabilised by the two  $\text{Mg}^{2+}$  cations, while the metaphosphate TS is located between the two metals, with formation of a phosphorane-like TS geometry. In the sEH, the metaphosphate group at the TS is stabilised during catalysis by its apical co-ordination to the only  $\text{Mg}^{2+}$  ion in the catalytic site.<sup>49</sup> This supports the mechanistic hypothesis (proposed for the first time)<sup>28</sup> that two ions can more easily facilitate the formation of a meta-stable intermediate, as in the case of RNase H. An additional key aspect is the role of water molecules in solvating the metal centre, which can catalyse the migration of protons involved in the phosphoryl transfer reaction in both sEH and RNase H. These mechanistic aspects, retrieved using QM/MM simulations, help us to understand and appreciate how the enzymes create better attacking and leaving groups during phosphoryl transfers. This is reinforced

by a recent investigation of the metalloenzyme topoisomerase II (topoII),<sup>50,51</sup> where two metal ions aid the phosphodiester bond cleavage of the DNA substrate. Here, hybrid Born–Oppenheimer QM/MM MD simulations were applied to reconstruct a catalytically competent state with a two-metal-ion architecture such as that in RNase H.<sup>28,52</sup> The two Mg ions might therefore act similarly here, as in several other two-metal-ion phosphodiesterases, which suggests that these metalloenzymes likely cut the substrate DNA with a similar mechanism as that observed for RNase H.<sup>23,53,54</sup>



## 12.3 Simulating Membrane Proteins

### 12.3.1 Membrane Enzymes: The Case of FAAH

One challenging class of drug discovery targets is those that are anchored to the membrane, forming a heterogeneous system comprising the enzyme, the membrane, and the solvent (water). The fatty acid amide hydrolase (FAAH), for example, is a promising drug discovery target. This key enzyme is involved in endocannabinoid metabolism, which is fundamental for human health and crucial to regulating pathophysiological processes such as pain and inflammation.<sup>55,56</sup> FAAH is an intracellular serine hydrolase that degrades endocannabinoids. Thus, inhibiting FAAH increases the level of endogenous cannabinoids, which is a promising strategy to treat many pathologies ranging from pain to inflammatory-related diseases to cancer.<sup>57</sup>

Remarkably, a wealth of structural data on FAAH has revealed the structural features of the enzyme's catalytic site.<sup>58,59</sup> The FAAH binding site comprises a catalytic triad (Ser241–Ser217–Lys142) for the hydrolysis of the endocannabinoids, while an oxyanion hole (Gly239–Gly240–Ser241) stabilises the substrate for catalysis. Additional structural data have clarified the mechanism of FAAH inhibition by potent enzyme inhibitors. In particular, the co-crystallisation of some potent covalent inhibitors in FAAH has clarified how they block FAAH activity through a covalent bond with the nucleophilic Ser241.<sup>60</sup>

---

**Figure 12.2** Upper scheme: selected snapshots taken from our computer simulations of the two investigated phosphoryl transfers comprising the catalytic cycle. (Top) Nucleophilic attack of Asp9 at the  $\text{Mg}^{2+}$ -coordinated phosphoryl group, with substrate cleavage and phosphoenzyme intermediate formation **INTa**. In the middle, the transition state structure **TS1** shows the concomitant proton shuttle (labelled PT1 and PT2) from a  $\text{Mg}^{2+}$  coordinated water molecule to the leaving group oxygen *via* a bridging solvent water. (Bottom) The second phosphoryl transfer from the phospho-Asp9 to one attacking solvent water, leading to the product state, with a second proton shuttle (labelled PT3 and PT4 in **TS2**) now traveling in the reverse direction to create the nucleophile  $\text{OH}^-$ . Lower left: Free-energy profile (top) and selected average bond distances (bottom) along the first catalytic step of phosphoenzyme formation (**INTa**). Bond distance labels as in Figure 2; notably,  $r1$  and  $r2$  are the breaking and forming P–O bond lengths, respectively. The proton shuttle occurs at RC # –0.5A (vertical dashed line), just before the system reaches the TS plateau (orange region). Note the shortening of the  $\text{Mg}^{2+}$ -ligand distance,  $d1$ , upon proton donation and transfer along the H-bond wire ( $d2/d3$  and  $d4/d5$  crossing). Lower right: free-energy profile (top) and selected average bond distances (bottom) along the second catalytic step. Bond distance labels as in the **INTb** panel in the upper scheme. Here, the proton shuttle (dashed vertical line) occurs in the reverse direction (note the  $d2/d3$  and  $d4/d5$  crossing) after the TS plateau (orange region). Reproduced from M. De Vivo, B. Ensing, M. Dal Peraro, G. A. Gomez, D. W. Christianson and M. L. Klein, *J. Am. Chem. Soc.*, 2007, **129**, 387–394. Copyright (2007) American Chemical Society.

Most of the covalent FAAH inhibitors are potent electrophilic molecules characterised by an activated carbonyl group, such as trifluoromethyl ketones,  $\alpha$ -keto amides,  $\alpha$ -keto esters, and  $\alpha$ -ketoheterocycles, such as OL-135 and carbamates.<sup>61,62</sup>

The ability of FAAH to hydrolyse amides and esters at similar rates had suggested that both carbamates and ureas could act as good carbamoylating compounds. Indeed, researchers had discovered potent FAAH inhibitors based on cyclic piperidine and piperazine aryl ureas that formed a covalent enzyme-inhibitor adduct with Ser241. Nevertheless, the exact mechanism for the potency of piperidine and piperazine inhibitors of FAAH was unclear.

Comparative MD and QM/MM calculations<sup>63</sup> were used to test the functional hypothesis based on the partial flexibility of piperidine- and piperazine-based inhibitors and their conformational distortion for forming a covalent bond between the inhibitor and Ser241. The piperidine-based PF750 and piperazine-based JNJ1661010 inhibitors are two lead molecules used to generate clinical candidates (Figure 12.3). The extensive computational analysis of these two molecules highlighted differences in their conformational flexibility in water and in complex with FAAH. Within FAAH's binding site, the piperidine and piperazine inhibitors adopt a specific twisted conformation of the amide bond and incomplete pyramidalisation at the amide bond nitrogen (Figure 12.3), which facilitate nucleophilic attack for amide hydrolysis, compared to their planar analogues. Additional QM/MM calculations returned a lower  $\Delta E_{\text{LUMO\_HOMO}}$  for distorted amides with respect to the planar analogues, further corroborating the crucial role of the protein/ligand complex's flexibility.

The relevance of structural flexibility in FAAH catalysis was also demonstrated by microsecond-long MD simulations of wild-type and double mutant models (Phe432Ala and Trp531Ala) of FAAH, embedded in a realistic membrane/water environment. These calculations, integrated with mutagenesis and kinetic experiments, were used to propose a novel structural framework for a lipid selection mechanism. This mechanism is indeed mediated by structural flexibility to selectively accommodate the flexible substrate into a multi-pocket binding site and facilitate pre-reactive conformations of the substrate for efficient hydrolysis.<sup>64,65</sup>

Taken together, these computational studies integrate classical MD and QM/MM simulations, demonstrating that the enzyme-induced twist of the amide bond is crucial in catalysing amide bond hydrolysis.<sup>66</sup> They are thus a representative and noteworthy example of how MD and QM/MM calculations, once considered prohibitive for everyday drug design, can now be developed into routinely used tools for drug design.

### 12.3.2 Ion Channels: The Case of the Kv11.1 Channel

Ion channels are another important and large class of pharmaceutical targets. Ion channels are transmembrane proteins responsible for the passive and selective transport of ions through biological membranes. By helping to



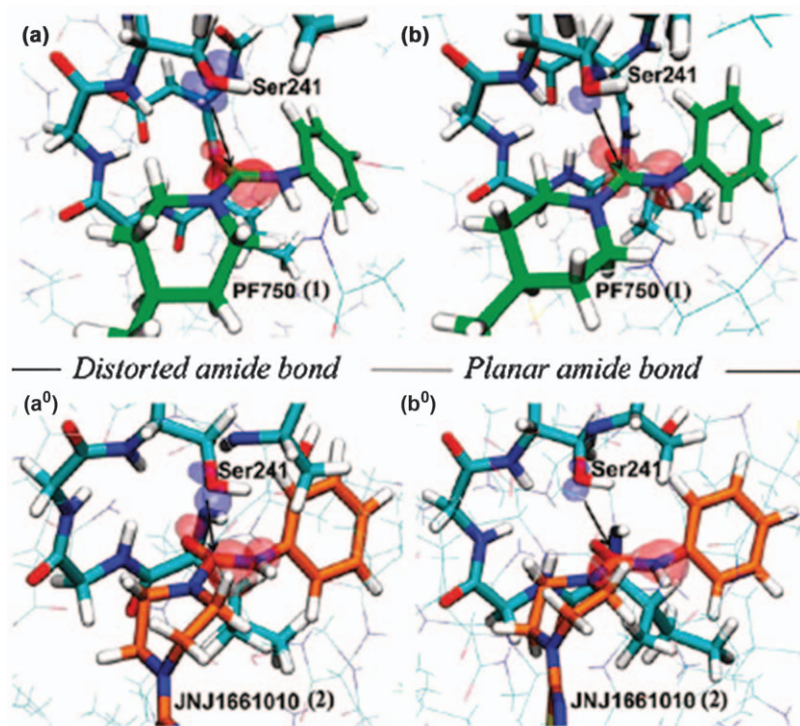


**Figure 12.3** Shape of the frontier orbitals for the FAAH 3 1 (top) and FAAH 3 2 (bottom) complexes. Two representative snapshots characterised by distorted (a and a0) and planar (b and b0) amide bonds are shown for 1 and 2, respectively. FAAH residues Ile238, Gly239, Gly240, Ser241, and Ser217 are in cyan; 1 is in green, and 2 is in orange. The HOMO orbital of the Ser241 nucleophile is in blue, while the LUMO orbital of the electrophile (i.e. the 1 and 2 carbonyl) is in red.<sup>17</sup> Reproduced from G. Palermo, D. Branduardi, M. Masetti, A. Lodola, M. Mor, D. Piomelli, A. Cavalli and M. De Vivo, *J. Med. Chem.*, 2011, 54, 6612–6623. Copyright (2011) American Chemical Society.

regulate the excitability and osmotic equilibrium of cells, ion channels are involved in many essential functions, including transmission of nerve impulses and contraction/relaxation of muscles.<sup>67</sup> Given the relevance of these processes for human physiology, it is not surprising that ion channel dysfunctions are usually associated with serious pathologies, thus drawing the attention of drug discovery programs.<sup>68</sup>

One major example is the hERG (Kv11.1) channel, which is a voltage-gated potassium channel that is mainly expressed in heart muscle tissue.<sup>69</sup> The hERG channel is responsible for the rapidly activating delayed rectifying current ( $I_{Kr}$ ), the most critical current involved in the repolarisation (phase III) of the action potential.<sup>70</sup> hERG has been a target of interest ever since the discovery that impairing its activity, whether through inherited

mutations or through pharmaceutical channel blocking, could contribute to long QT syndrome (LQTS). This cardiac disorder predisposes individuals to a potentially lethal form of arrhythmia (*torsades des pointes*, TdP).<sup>71</sup> Although TdP rarely manifests, several drugs were withdrawn from the market due to the severe adverse effects related to unintentional channel blocking.<sup>72</sup> As a result of these drug safety concerns, hERG is now one of the most popular (or infamous) anti-targets in pharmaceutical research.

In this context, it is most desirable to assess blocking activity in the early stages of drug development. Computational methods are optimal tools for understanding the molecular features responsible for channel blocking. However, for several reasons, it is extremely challenging to model hERG–drug interactions. First and foremost, the channel’s structure is still unknown. This means that ligand-based approaches are the most appropriate techniques for rationalising structure–activity relationships (SAR). In this respect, a number of quantitative SAR (QSAR) and classification models have been developed over the years, relying on either 2D<sup>73–77</sup> or 3D<sup>78–84</sup> descriptors. However, despite their predictive power, ligand-based models lack a clear physicochemical interpretation, which is an advantage of structure-based approaches. 3D-QSAR models are also strongly dependent on the conformations chosen for the molecules. This information requires prior knowledge of the drug binding site’s structural features.

In general, hERG shares low sequence conservation with the available templates. However, it is nowadays widely accepted that the binding site for the blockers is located inside the channel’s cavity, in particular on the sixth transmembrane helix (Figure 12.4), where sequence identity significantly

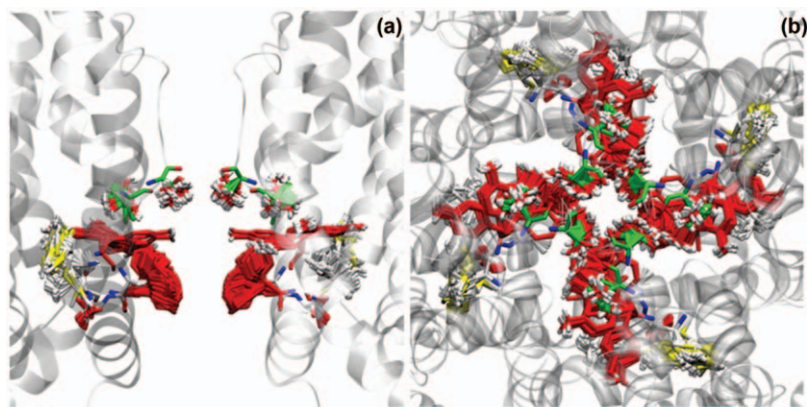


**Figure 12.4** Conformations generated for the blockers binding site in the hERG channel (a) Side and bottom (b) view of the channel conformations used to build the structure-based correlative models.

Reprinted with permission from G. P. Di Martino, M. Masetti, L. Ceccarini, A. Cavalli and M. Recanatini, *J. Chem. Inf. Model.*, 2013, 53, 159–175. Copyright (2013) American Chemical Society.



increases.<sup>85</sup> This has justified the use of hERG homology models for molecular docking studies. Of the many models published to date, one example is the model derived by Tseng *et al.*<sup>86</sup> This model included the voltage-gating domain (segments S1–S4), the pore-forming domain (S5–S6), and the long extracellular region (S5-P linker, more than 40 amino acids), usually called the ‘outer mouth’ or ‘turret’. The models have revealed that the cavity involved in interaction with blockers is, in many respects, an uncommon binding site, suggesting that docking results should be considered with caution.<sup>72</sup> Indeed, given the intrinsic symmetry of the channel, the conformational freedom of key aromatic residues (Tyr652 and Phe656), and the limited number of strong directional interactions, there are several problems for common docking software and the associated scoring functions. There is poor consensus in the binding modes obtained by different research groups, which mirrors the uncertainties associated with the intrinsic approximations of the homology models and the difficulties related to the nature of the binding site.<sup>72,87,88</sup> To better address these often overlooked aspects of hERG computational modelling, Di Martino *et al.* recently developed a strategy that uses massive sampling and minimal knowledge-based information to generate reliable structure-based correlative models. This is achieved by automatically taking into account binding site symmetries and the ligand’s configurational entropy and desolvation contributions upon binding.<sup>89</sup>

Taking a different approach, homology models of hERG have also been used for MD simulation studies. These studies have sought to investigate the channel’s functionality in realistic membrane environments and to gain further insights into the drug-induced blocking. For example, Masetti *et al.*<sup>85</sup> compared the dynamics of the closed and open state of the channel, and successfully docked one of the most potent blockers (astemizole) in channel configurations identified during the dynamics. On a more quantitative basis, Boukharta *et al.*<sup>90</sup> used an MD framework to assess the free energy of binding for a series of congeneric sertindole derivatives through the Linear Interaction Energy method. Finally, turning to more biophysically oriented studies, we note characterisation of the mechanism of ion permeation by Ceccarini *et al.*,<sup>91</sup> using metadynamics simulations and subsequently estimating the channels’ conductance. As another example, Köpfer *et al.*<sup>92</sup> probed a possible inactivation mechanism of the channel achieved through collapse of the selectivity filter.

In the last decade, researchers have begun to consider the channel as a more conventional therapeutic target, bringing a fresh perspective to an old subject.<sup>93</sup> Indeed, compounds that enhance the channel’s function by potentiating hERG activity (*i.e.* allosteric modulators) could become useful drugs for treating some forms of arrhythmias or simply for countering the channel blocking. The limitations in using computational tools to identify hERG blockers also affect the identification of hERG activators. Indeed, the outlook may be more dismal here, given that some activators have been hypothesised to bind the outer mouth of the channel, close to the selectivity

filter,<sup>94</sup> while other groups have reported two adjacent subunits in the proximity of the pore helix.<sup>95</sup> In this context, Durdagi *et al.*<sup>96</sup> exploited an elaborated topographic mapping to identify the binding site of a prototypical activator (NS1643). Additionally, Giacomini *et al.*<sup>97</sup> used a ligand-based virtual screening to discover an activator able to enhance hERG currents by 195%.

In summary, hERG modelling is a complex and multifaceted issue. While several computational tools have unquestionably helped drug discovery and development, knowledge of hERG's functionality and interaction with drugs is still far from comprehensive. Ultimately, it may be that only crystallisation of the channel's transmembrane portion will provide the information required for genuine progress.

### 12.3.3 GPCR: The Case of the Human Adenosine Receptor A<sub>2A</sub> Embedded in Neuronal-like Membrane

Human G-protein-coupled receptors (GPCRs) belong to the largest membrane-bound receptor family expressed by humans.<sup>98</sup> Remarkably, GPCRs participate in about 80% of signalling processes in the brain.<sup>99</sup> In particular, they are involved with perceiving chemicals from the environment. Indeed, one of their most important features is their ability to bind molecules of diverse shapes, sizes, and chemical properties.<sup>100,101</sup> Thus, understanding the mechanisms of GPCR function/dysfunction means characterising their interactions with their natural ligands. Increasingly, emerging experiments point to the role of membrane lipid composition in GPCR structure/function relationships.<sup>102–105</sup> This is a crucial issue for pharmacological applications because more than one quarter of FDA-approved drugs target GPCRs.<sup>106</sup>

Therefore, it is crucial to determine the effect of the native membrane environment on the conformational properties of pharmaceutically relevant GPCRs. Classical MD simulations can be instrumental in clarifying the molecular basis of these effects. For instance, MD studies of a class A GPCR (beta-2 adrenergic receptor) showed that the stability of the receptor's conformations varies with different lipid compositions.<sup>105</sup> Lipid composition was also found to be responsible for the stability of the receptor's secondary structural elements.<sup>104</sup> Additionally, MD simulation of rhodopsin embedded in a cholesterol-containing membrane has shown how specific cholesterol–rhodopsin interactions might modulate the receptor activation.<sup>107</sup> Furthermore, the progressive removal of cholesterol from bovine hippocampal membranes results in a cholesterol-concentration-dependent reduction of the binding of the agonist 8-OH-DPAT to the class A GPCR serotonin1A receptor.<sup>108</sup>

In a recent work,<sup>109</sup> MD simulations were used to elucidate the effect of membrane composition on an antagonist-bound neuronal GPCR. This is the class A GPCR human adenosine receptor type 2A (hA<sub>2A</sub>R)<sup>110</sup> in complex with the antagonist caffeine (CFF hereafter). CFF binding to hA<sub>2A</sub>R may lead to

neuroprotection.<sup>111–116</sup> hA<sub>2A</sub>R is highly localised in the so-called ‘striatum’ of the brain,<sup>117</sup> where the cholesterol presence varies from 33% to 50%.<sup>118</sup> In order to investigate the effect on moving from artificial lipid bilayers to conditions near to real neuronal membrane, 0.8  $\mu$ s-long MD simulations were performed of hA<sub>2A</sub>R embedded in three different membranes. In system **I**, the membrane comprised a pure 1-palmitoyl-2-oleoylphosphatidylcholin lipid bilayer (POPC). In system **II**, the bilayer comprised equally mixed POPC and POPE (1-palmitoyl-2-oleoylphosphatidylethanolamine) lipids. In system **III**, the membrane resembled the synaptic membrane (42% POPC, 34% POPE and 25% of cholesterol molecules), where hA<sub>2A</sub>R is expressed.<sup>119</sup> Notably, **III** markedly differs from the artificial membrane mimics (detergent *n*-nonyl- $\beta$ -D-glucopyranoside), where the CFF/hA<sub>2A</sub>R complex is embedded for X-ray structure determination.<sup>120</sup>

CFF exhibited multiple binding poses in the hA<sub>2A</sub>R ligand binding cavity across the three systems, similarly to the findings for adenosine in this receptor.<sup>121</sup> Most of the identified binding poses were similar to those found in the 0.069  $\mu$ s MD study of a H8-truncated CFF/hA<sub>2A</sub>R complex.<sup>122</sup> These poses still differ from those in the 0.005  $\mu$ s MD study,<sup>123</sup> most probably because of the relatively short timescale in their simulations. The artificial environment was found here to drastically affect the population of ligand poses: the pose found in the X-ray structure at 3.6 Å resolution<sup>120</sup> was not the most populated of the  $\mu$ s-long MD simulations. Moreover, in **I** and **II**, a binding pose preserved for more than 40% of the simulated time was never found. This is not true for system **III**, where CFF was in the same binding pose for  $\sim$ 92% of the simulated time.

This extra stabilisation was triggered by the diffusion of a cholesterol molecule to the cleft between H1 and H2 after 0.3  $\mu$ s. Indeed, this specific cholesterol binding induces conformational rearrangements of VAL57, LEU58, ILE66, and SER67. These, in turn, result in the enhanced stabilisation of the hydrophobic interaction between CFF and H2 residues. More specifically, this interaction involves the hydrophobic contacts between the CFF/C5 methyl group and the ILE66 and SER67 side chains on the extra-cellular side of H2 (Figure 12.5). Hence, cholesterol drives the specific pose for CFF and possibly for other CFF mimic/hA<sub>2A</sub>R complexes.

These results show that both hA<sub>2A</sub>R folding and CFF binding dynamics are sensitive to the lipid environment where hA<sub>2A</sub>R is embedded. Most importantly, the presence of cholesterol in the membrane was shown to drastically affect the population and mobility of the CFF binding pose. X-ray studies commonly crystallise ligand/receptor complexes in detergent mimics,<sup>124</sup> without the physiologically high concentration of cholesterol. This study thus suggests that computer-aided studies of hA<sub>2A</sub>R in near-physiological conditions may make key contributions to investigating the receptor's function. Specifically, during the computer-aided design of ligands to target hA<sub>2A</sub>R and possibly other GPCRs, it may be very important to include a correct description of the neuronal membranes.

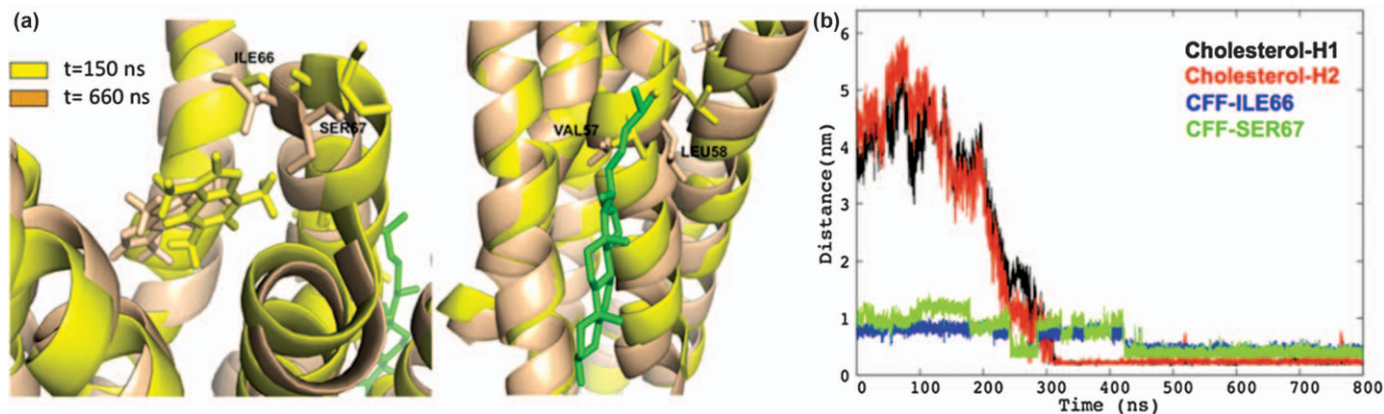


**Figure 12.5** Specific cholesterol binding to hA2AR (a) Conformational transition of H2 residues induced by specific cholesterol binding. Snapshots at  $t=150$  ns and 660 ns are shown in yellow and orange and cholesterol is shown in green. (b) The diffusion of cholesterol into the H1/H2 cleft enhances hydrophobic contacts between CFF and H2. The minimum distances between the specific cholesterol molecule and H1 (residues 5–34), between cholesterol and H2 (residues 41–67), between C5@CFF and the heavy atoms of the ILE66 and SER67 side chains, are shown in black, red, blue, and green, respectively. Reproduced with permission from R. Cao, G. Rossetti, A. Bauer and P. Carloni, *PLoS ONE*, 2015, **10**, e0126833. © 2015 Cao *et al.*

## 12.4 Tackling Target Flexibility Through Simulations

Flexibility is an intrinsic feature of proteins. It is required for most of their biological functions, including recognition of molecular partners. In structure-based drug discovery, using single receptor conformations would mean sticking with the outdated lock-and-key model of protein–ligand recognition proposed by Emil Fischer more than a century ago.<sup>125</sup> In a more realistic description, an ensemble of protein conformations will be accessible at finite temperature. By designing molecules to target specific conformations, researchers may increase the chances of success in identifying novel active compounds or improving target specificity. Although it is always advisable to include a certain amount of flexibility in structure-based drug discovery,<sup>126</sup> this is not always straightforward to implement from a computational standpoint. Nowadays, local structural changes upon binding can be (more or less) easily taken into account by docking search algorithms. However, it is much more challenging to deal with larger conformational changes, ranging from domain motions to loop rearrangements. To address this, one successful strategy is to combine the sampling capabilities of molecular dynamics (MD)-based methods with the efficient description of ligand binding provided by docking methods.<sup>125</sup>

### 12.4.1 Lactate Dehydrogenase

As an example, we can consider lactate dehydrogenase (LDH, EC 1.1.1.27). This enzyme is expressed in all tissues and is responsible for the NADPH-dependent (inter)conversion of pyruvate to lactate.<sup>127</sup> LDH plays a central role in the life of cells, but its best-characterised function is related to cell metabolism. It is well-known that, in normal cells at rest, the most efficient mechanism for energy production is the complete degradation of glucose through a stepwise process called oxidative phosphorylation. An important intermediate product, pyruvate, is further degraded to carbon dioxide *via* the TCA cycle. The resulting NADH is reconverted to NAD<sup>+</sup> by exploiting the mitochondrial respiratory chain. The whole process requires aerobic conditions and amounts to an energy equivalent of up to 36 ATPs per glucose molecule. LDH catalyses a reaction that is located at a branching point of this process, right before the TCA cycle is engaged by pyruvate. LDH thus provides an alternative metabolic pathway. Although much less efficient (only two ATPs/glucose), this alternative pathway can be exploited by cells in hypoxic conditions or when there are enhanced metabolic needs (*anaerobic glycolysis*).<sup>128</sup>

Drug discoverers are interested in LDH because of energy metabolism reprogramming, which is one of the emerging hallmarks of cancer.<sup>129</sup> Cancer cells tend to convert most glucose into lactate regardless of oxygen availability (*aerobic glycolysis*), even though this is less efficient in terms of providing ATP.<sup>130</sup> This process is achieved through an increased expression of the LDH-A isoform, and is advantageous for highly proliferating cells that

are in constant need of precursors to sustain growth and division. This need is met through a massive increase in glucose uptake, for which aerobic glycolysis is preferred in order to prevent saturation of the mitochondrial respiratory chain, which would otherwise lead to a potentially hazardous excess of NADH. Because of LDH's position in the metabolic pathway, inhibiting this enzyme would offer an effective therapeutic strategy against cancer without causing substantial damage to normal cells.<sup>131</sup>

From a structural standpoint, LDH is an oligomeric enzyme assembled as a tetramer. In each monomer, a Rossmann-fold domain delimits the NADH-co-factor binding site, and a mixed  $\alpha/\beta$  structure forms the substrate binding pocket close to the nicotinamide pocket. The substrate binding pocket is shielded from the solvent by a mobile loop of about fifteen residues including the highly conserved Arg105, which is directly involved in catalysis. Notably, the open to closed transition of the active site's loop has been hypothesised to act as the catalytic mechanism's rate-limiting step.<sup>132</sup> Despite their emerging role in cancer treatment, discovering LDH inhibitors has proven quite challenging. This can be partly imputed to the peculiar features of the binding site, which is highly polar and rather limited in size.<sup>133</sup>

To date, structure-based drug design approaches have been used successfully to identify compounds showing an inhibitory activity in the micromolar or nanomolar range. In particular, docking studies, molecular dynamics, and free-energy estimation (through the MM-GB/SA method) have been used to rationalise structure-activity relationships (SAR) for a series of N-hydroxyindole-based derivatives.<sup>134,135</sup> Additionally, Kohlmann *et al.*<sup>136</sup> used a molecular modelling approach to identify nanomolar inhibitors through guided fragment growing and linking. Taking a different approach, Manerba *et al.* used structure-based virtual screening to discover a novel LDH inhibitor (galloflavin), which can inhibit lactate production and ATP synthesis in human hepatocellular carcinoma cells.<sup>137</sup> Despite these successes, most computational studies seeking to discover LDH inhibitors have not considered the conformational flexibility of the LDH binding site. In other words, by only considering the closed conformation of the active site loop, one might argue that these investigations did not fully exploit the druggability potential created by this enzyme's conformational variability. To address this, researchers developed a novel computational strategy, which combined a hybrid-solvent variant of the replica-exchange method with network-based analysis to better classify the conformational space. Thus, representative structures of the sampled conformational ensemble were validated by retrospective virtual screening simulations.<sup>138</sup> Notably, these studies identified a protein conformation (referred to as 'seed 13', see Figure 12.6), which had striking potential for a successful prospective virtual screening campaign. In a subsequent ensemble-based virtual screening, this strategy allowed the researchers to discover a novel class of LDH-A inhibitors in a series of N-acylhydrazone derivatives.<sup>139</sup> With the same aim of including flexibility in docking and virtual screening, Granchi *et al.* recently developed an ensemble-based method that incorporates closed and open



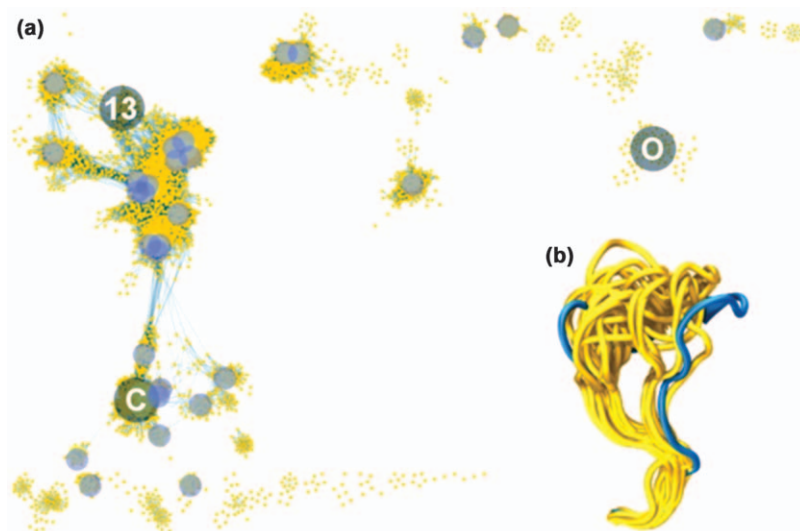


**Figure 12.6** Conformational space of the LDH-A binding site sampled through hybrid-solvent replica exchange molecular dynamics. (a) Network representation of the sampled conformations (yellow dots). The blue transparent dots indicate LDH-A conformations that were subject to retrospective virtual screening. Configuration 13 represents a binding site conformation, which was not disclosed through simulations, displaying a high potential for prospective virtual screening experiments. The location of the experimentally available closed (C) and open (O) structures on the network is also shown. (b) Relevant active site loop conformations sampled (yellow tubes) shown together with conformation C and O (blue tubes on the left and on the right, respectively). Reprinted with permission from R. Buonfiglio, M. Ferraro, F. Falchi, A. Cavalli, M. Masetti and M. Recanatini, *J. Chem. Inf. Model.*, 2013, **53**, 2792–2797. Copyright (2013) American Chemical Society.

conformations of the enzyme in the presence or absence of the cofactor (for a total of four structures).<sup>140</sup> Although most of the compounds identified with these methods show good potency in enzymatic assays, more experimental work is required to improve their physicochemical properties. Overall, the reported examples confirm the relevance of including flexibility in computationally driven drug discovery.

#### 12.4.2 Intrinsically Disordered Proteins

Targeting intrinsically disordered proteins (IDPs) provides an extreme illustration of the absolute need to include protein flexibility in computationally driven drug discovery. IDPs are an important class of functional proteins, which are highly abundant in nature,<sup>141–144</sup> specifically in humans, where they represent almost one third of the genome.<sup>141–145</sup> Notably, it has been predicted that 30 or more consecutive disordered residues can be



found in 79% of cancer-associated proteins, 57% of cardiovascular-disease-associated proteins, and 55% of neurodegenerative-disease-associated proteins.<sup>146</sup>

Targeting this class of proteins is thus key to designing potential therapeutics for altered disease-associated pathways. Unfortunately, traditional computational and experimental approaches have been hampered by a variety of challenges.<sup>101</sup> IDPs do not adopt a well-defined native three-dimensional structure,<sup>147</sup> they lack stable tertiary and/or secondary structures when isolated in solution under near-physiological conditions,<sup>148,149</sup> and they exist in an ensemble of states both in solution and when unbound to a ligand *in vivo*.

Computational methods that use physics-based empirical molecular mechanics force fields are increasingly providing critical general insights into the behavior of IDPs.<sup>150–155</sup> However, the dynamic and heterogeneous nature of IDPs presents substantial challenges in terms of force field accuracy and conformational sampling capability. MD simulations are sensitive to the choice of protein force field.<sup>150,156,157</sup> Hence, an important caveat is that these force fields are typically parameterised to reproduce the behaviours of folded proteins rather than IDPs. Thus, they may fail to capture important aspects of IDP conformational ensembles.<sup>158–160</sup>

To investigate how the conformational space of the IDP  $\alpha$ -synuclein (AS) is affected by the binding of dopamine (DOP), an anti-aggregation drug, Dibenedetto *et al.*<sup>161</sup> proposed a computational protocol based on classical MD simulations with the amber parm99SB force field<sup>162</sup> plus the ildn modification.<sup>157,163–166</sup> To address the conformational sampling challenge, the researchers did not use randomly generated conformations for their MD simulations. Instead, they used realistic configurations of AS obtained by NMR spectroscopy combined with paramagnetic relaxation enhancement (PRE).<sup>167</sup> They also analysed the conformational ensemble of AS, alone and in the presence of the drug, with a newly developed tool based on the dihedral angle distributions visited during MD.<sup>161</sup> This latter aspect was truly innovative: the tool can not only detect and quantify backbone conformational transitions in proteins, it can also quantitatively detect the effects of the drug on the protein's dynamical spectra. Finally, the tool allows the interpretation of the 2D  $^1\text{H}$ - $^{15}\text{N}$  heteronuclear multiple-quantum correlation (HMQC) spectra of AS in the presence of the anti-aggregation drug. Heteronuclear 2D NMR is arguably one of the most versatile and powerful experimental tool for investigating drug binding to IDPs.<sup>168</sup> This is because it can detect fast conformational changes and binding events with high resolution and within a short time.<sup>168,169</sup> However, it is not easy to distinguish the differences in chemical shift that are due to direct contacts between the drug and the IDP from those that are due to the long-range effects of the binding.<sup>170</sup> In contrast, the tool proposed by Dibenedetto *et al.*<sup>161</sup> could distinguish variations in chemical shifts due to direct contacts with the drug from those variations that were due to AS conformational changes induced by the long-range effects of the binding (Figure 12.7).

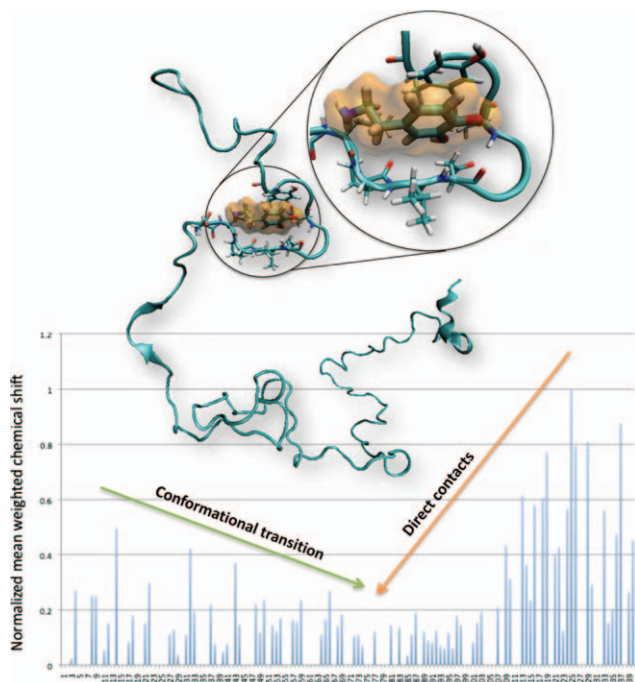


**Figure 12.7** 2D  $^1\text{H}$ - $^{15}\text{N}$  Heteronuclear multiple-quantum correlation (HMQC) spectra of AS in the presence of DOP and relative mixed cartoon/liquorice representation of the AS-DOP complex. 2D  $^1\text{H}$ - $^{15}\text{N}$  HMQC spectra of free AS and AS-DOP were measured at 303 K on a 200  $\mu\text{M}$  AS sample in a buffer that consisted of 20 mM sodium phosphate and 150 mM NaCl (pH 6.4), with 10 mg of DOP. The normalised mean-weighted chemical shift variations per residue are calculated as:  $[(\Delta H)^2 + (\Delta N/10)^2]^{0.5}$ , where  $\Delta$  indicates the difference between the chemical shift between the bound and free state, given for each backbone amide. The detail of the interaction between AS and DOP is magnified. Different contributions to the spectra coming from direct contacts between DOP and AS and conformational transitions are highlighted with orange and green arrows, respectively.

Reprinted from *Archives of Biochemistry and Biophysics*, Volume 582, G. Rossetti, D. Dibenedetto, V. Calandrini, A. Giorgetti and P. Carloni, Structural predictions of neurobiologically relevant G-protein coupled receptors and intrinsically disordered proteins, 91–100, Copyright (2015), with permission from Elsevier.

### 12.4.3 Targeting RNA in Trinucleotide Repeats Diseases

Tandem repeats are a short motif of repeated nucleotides that occur frequently in the human genome, on average every 2000 base pairs. These repeats are instable throughout meiosis and have the tendency to vary in length. Beyond a certain threshold, the expansion of these short tandem repeats (for example CAG repeats) is linked to disease development. Huntington's disease (HD) is a devastating autosomal dominant neurodegenerative human

disease,<sup>171</sup> for which there is currently no cure.<sup>172</sup> HD affects 4–10 individuals per 100 000 in the Western world.<sup>172</sup> It is characterised by progressive motor, cognitive, and psychiatric disorders.<sup>171,173</sup> The gene responsible for the disease (HTT) encodes the ubiquitously expressed Huntingdon protein (Htt),<sup>174</sup> which is essential for brain development.<sup>175</sup> The disease is caused by an expanded CAG repeat in the 5'-end of the gene HTT, which encodes an abnormally long polyglutamine (polyQ) tract in the Htt protein<sup>171,174</sup> beyond the first 17 amino acids, followed by a polyproline region (polyP).<sup>174</sup> HD penetrance is correlated to the number  $n$  of CAG repeats.<sup>176–178</sup> Similar features are also present in a variety of other rarer CAG-expansion-based neurodegenerative diseases (Figure 12.8).<sup>179†</sup>

Much research has focused on the polyglutamine protein and its involvement in disease development.<sup>181</sup> However, other factors can contribute to the pathogenesis of CAG disorders.<sup>182</sup> We will consider some research that has focused on the mutant HTT messenger RNA (mRNA) transcripts.<sup>183,184</sup> When  $n \geq 36$ , these mRNA bind to proteins and protein complexes in a manner that depends on the number of the CAG repeats,<sup>179,185–189</sup> by aberrantly regulating several cellular mechanisms.<sup>184</sup> One potential strategy against HD and other CAG-related disorders may therefore be to discover ligands with high affinity for mRNA transcripts of pathogenic length.<sup>184,185,190</sup> Disney and co-workers recently published a small library of ligands that target CAG repeats in RNA oligonucleotides.<sup>186</sup> The most potent ligand was ((diaminomethylene)amino)phenyl 4-((diaminomethylene)amino)benzoate (ligand **1**, hereafter), with a  $K_d$  of 60(30) nM. It is also the most potent ligand so far discovered for CAG repeats.<sup>191–193</sup>

Unfortunately, the binding poses of all these ligands are unknown, limiting the scope for improving their efficacy against CAG repeats and/or their selectivity for pathogenic CAG repeats. Researchers have reported computational studies that seek to predict the binding poses of small ligands targeting r(CUG) and r(CCUG) repeats by combining molecular docking and MD simulations in implicit solvent.<sup>191–193‡</sup> These calculations could, in principle, produce structural insights into ligand/CAG repeats complexes. Unfortunately, the prediction accuracy is known to be severely affected by the absence of explicit solvent when calculating RNA-containing systems.<sup>197</sup>

Molecular simulations in explicit solvent, based on the latest version of the AMBER force field<sup>196,198</sup> are emerging as increasingly powerful tools for investigating RNA in the free form and in complex with ligands.<sup>199–202</sup> They have reproduced, with sufficient accuracy, structure, (sub-) microsecond conformational fluctuations, and energetics of RNA oligonucleotides in the free state<sup>203–207</sup> (including one containing CAG repeats<sup>208</sup>) and in complex with ligands.<sup>196,209,210</sup>

†These affect a further 2–3 cases per 100 000 of the population worldwide.<sup>180</sup>

‡The protocol used a molecular docking based on a series of back and forth pulling of the ligands, followed by molecular dynamics based on the amber99 force field<sup>194</sup> with Yildirim-revised  $\chi^{195}$  and  $\alpha/\gamma^{196}$  torsional parameter sets in implicit solvent.

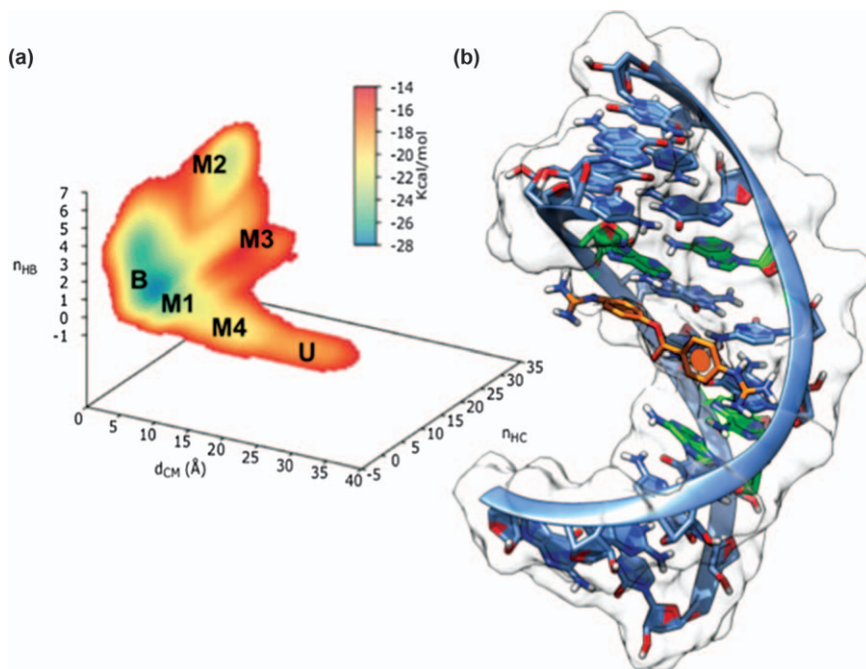


**Figure 12.8** (a) Unbinding free-energy profiles ( $\text{kcal mol}^{-1}$ ) of CAG-repeat-containing RNA in complex with compound **1** as a function of  $d_{\text{CM}}$ ,  $n_{\text{HC}}$ , and  $n_{\text{HB}}$ .  $d_{\text{CM}}$  is the distance between the centres of mass of RNA and **1**;  $n_{\text{HC}}$  is the number of hydrophobic contacts at interface between the two moieties;  $n_{\text{HB}}$  is the number of H-bonds between the two moieties. B denotes the bound states, while U represents unbound state minimum. M1–M4 are local minima. For the bound state B, a detail of the structure is offered in panel (b): ligand **1** is in orange, A bases are in green, and the remaining RNA bases are in blue. Reproduced from A. Bochicchio, G. Rossetti, O. Tabarrini, S. Kraubeta and P. Carloni, *J. Chem. Theory. Comput.*, 2015, **11**, 4911–4922. Copyright (2015) American Chemical Society.

In this study, MD and well-tempered metadynamics (WT-meta)<sup>211</sup> were run in explicit solvent, with the above-mentioned AMBER force field, to predict binding pose and affinities of both the highest affinity compound **1** and a lower affinity ligand: 6-(4,5-dihydro-1H-imidazol-2-yl)-2-(4-(4,5-dihydro-1H-imidazol-2-yl)phenyl)-1H-indol-3-amine (ligand **2**,  $K_d$  of 700 (80)).<sup>212</sup> The WT-meta approach provides the free energy of the process as a function of selected reaction coordinates (or collective variables, CVs). These calculations do not depend on the choice of the initial structure<sup>213</sup> and are particularly suitable for targeting biomolecular flexibility.<sup>186</sup>

Here, the model RNA system was a double-stranded decamer containing two CAG repeats ( $r(5'-G_1GCAGCAGGCC_{10})_2$ ).<sup>185</sup> It involved A–A mismatches between canonical C–G base pairs, as in CAG-repeat-containing RNAs.<sup>179,214</sup> Through calculations, the molecular basis of the different ligands' specificity

toward CAG repeats was uncovered. The calculated free energies are not too dissimilar from those derived by the experimental  $K_d$ . The simulations provide a molecular basis for the increase in affinity on passing from 2 to 1. Compound **1** features a dominant binding mode in which it intercepts both RNA<sub>10</sub> strands, binding in the major groove of the r(5'-A<sub>4</sub>G<sub>5</sub>C<sub>6</sub>A<sub>7</sub>)<sub>2</sub> region into a wedge formed by the two A-A pairs. It features five main 'anchoring' interactions to the RNA<sub>10</sub>, the salt bridges between the charged guanidine tails and the G<sub>15</sub> and G<sub>5</sub> phosphates, the H-bond of the N23 nitrogen with the O5'(G<sub>5</sub>) atom, as well as the two  $\pi$ -stacking interaction, one parallel ( $\beta$  ring) and one T-shaped ( $\alpha$  ring), with the A<sub>4</sub> and A<sub>14</sub> bases, respectively. In contrast, ligand **2** binds in the upper major groove to the A<sub>4</sub>-A<sub>17</sub> pair only. It forms a salt bridge with the G<sub>5</sub> phosphate group, H-bonds with the N1(A<sub>4</sub>) atom,  $\pi$ -stacking of its  $\delta$  ring with A<sub>4</sub>, and a hydrophobic interaction with the C<sub>6</sub>. Hence, **1** establishes a higher number of molecular interactions with RNA<sub>10</sub> than **2**. This is likely a key factor affecting the binding affinity, which is suggested here to be enthalpy-driven.

These predictions, validated against comparison with experimental data, could help design ligands that have high affinity for CAG-repeat-containing RNA and that can interfere with pathological mRNA-protein complexes in cells.

## 12.5 Conclusions

By presenting and discussing these selected examples of computational studies of challenging drug discovery targets, we have sought to describe how molecular modelling and simulations can nowadays be used to effectively study and dissect complex events involved in regulating proteins of pharmaceutical interest, such as membrane proteins, RNA, and metalloenzymes. These examples ultimately show how traditional *in silico* drug discovery tools, like docking and virtual screening, are today routinely flanked by molecular simulations at both the classical and quantum levels. Importantly, this computational strategy can fully account for structural flexibility, generating valuable information to aid the design of probes and potent inhibitors.

## References

1. M. De Vivo, M. Masetti, G. Bottegoni and A. Cavalli, *J. Med. Chem.*, 2016, **59**, 4035–4061.
2. W. L. Jorgensen, *Science*, 2004, **303**, 1813–1818.
3. J. Durrant and J. A. McCammon, *BMC Biol.*, 2011, **9**, 71.
4. G. Palermo and M. De Vivo, in *Encyclopedia of Nanotechnology*, ed. B. Bhushan, Springer Netherlands, 2015, pp. 1–15.
5. R. Abagyan and M. Totrov, *Curr. Opin. Chem. Biol.*, 2001, **5**, 375–382.
6. M. De Vivo, *Front. Biosci.*, 2011, **16**, 1619–1633.

7. A. Lodola and M. De Vivo, *Adv. Protein Chem. Struct. Biol.*, 2012, **87**, 337–362.
8. I. Buch, T. Giorgino and G. De Fabritiis, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**(25), 10184–10189.
9. Y. Shan, E. T. Kim, M. P. Eastwood, R. O. Dror, M. A. Seeliger and D. E. Shaw, *J. Am. Chem. Soc.*, 2011, **133**, 9181–9183.
10. S. Decherchi, A. Berteotti, G. Bottegoni, W. Rocchia and A. Cavalli, *Nat. Commun.*, 2015, **6**, 6155.
11. B. Ensing, M. De Vivo, Z. W. Liu, P. Moore and M. L. Klein, *Acc. Chem. Res.*, 2006, **39**, 73–81.
12. H. Lu and P. J. Tonge, *Curr. Opin. Chem. Biol.*, 2010, **14**, 467–474.
13. A. Warshel, *Annu. Rev. Biophys. Biomol. Struct.*, 2003, **32**, 425–443.
14. A. J. Mulholland, *Drug Discovery Today*, 2005, **10**, 1393–1402.
15. M. De Vivo, *Front. Biosci.*, 2011, **16**, 1619–1633.
16. A. Bottoni, G. P. Miscione and M. De Vivo, *Proteins*, 2005, **59**, 118–130.
17. G. Palermo, D. Branduardi, M. Masetti, A. Lodola, M. Mor, D. Piomelli, A. Cavalli and M. De Vivo, *J. Med. Chem.*, 2011, **54**, 6612–6623.
18. S. V. Vummaleti, D. Branduardi, M. Masetti, M. De Vivo, R. Motterlini and A. Cavalli, *Chemistry*, 2012, **18**, 9267–9275.
19. M. Cascella, M. Dal Peraro and M. De Vivo, in *Frontiers in Computational Chemistry*, ed. J. D. Madura, Bentham Science Publishers, 2015, pp. 290–343.
20. A. Lodola, D. Branduardi, M. De Vivo, L. Capoferri, M. Mor, D. Piomelli and A. Cavalli, *PloS One*, 2012, **7**, e32397.
21. M. De Vivo, G. Bottegoni, A. Berteotti, M. Recanatini, F. L. Gervasio and A. Cavalli, *Future Med. Chem.*, 2011, **3**, 1551–1559.
22. C. M. Dupureur, *Curr. Opin. Chem. Biol.*, 2008, **12**, 250–255.
23. G. Palermo, A. Cavalli, M. L. Klein, M. Alfonso-Prieto, M. Dal Peraro and M. De Vivo, *Acc. Chem. Res.*, 2015, **48**, 220–228.
24. M. Pavlin, G. Rossetti, M. De Vivo and P. Carloni, *Biochemistry*, 2016, **55**, 2772–2784.
25. M. H. Ho, M. De Vivo, M. Dal Peraro and M. L. Klein, *J. Chem. Theory Comput.*, 2009, **5**, 1657–1666.
26. M. De Vivo, A. Cavalli, G. Bottegoni, P. Carloni and M. Recanatini, *Proteins: Struct., Funct., Bioinf.*, 2006, **62**, 89–98.
27. M. De Vivo, A. Cavalli, P. Carloni and M. Recanatini, *Chemistry*, 2007, **13**, 8437–8444.
28. M. De Vivo, M. Dal Peraro and M. L. Klein, *J. Am. Chem. Soc.*, 2008, **130**, 10955–10962.
29. W. Yang, J. Y. Lee and M. Nowotny, *Mol. Cell*, 2006, **22**, 5–13.
30. S. Broccoli, F. Rallu, P. Sanscartier, S. M. Cerritelli, R. J. Crouch and M. Drolet, *Mol. Microbiol.*, 2004, **52**, 1769–1779.
31. K. Klumpp, J. Q. Hang, S. Rajendran, Y. Yang, A. Derosier, P. Wong Kai In, H. Overton, K. E. Parkes, N. Cammack and J. A. Martin, *Nucleic Acids Res.*, 2003, **31**, 6852–6859.



32. M. Nowotny, S. A. Gaidamakov, R. J. Crouch and W. Yang, *Cell*, 2005, **121**, 1005–1016.
33. M. Nowotny and W. Yang, *EMBO J.*, 2006, **25**, 1924–1933.
34. C. Sissi and M. Palumbo, *Nucleic Acids Res.*, 2009, **37**, 702–711.
35. D. Branduardi, M. De Vivo, N. Rega, V. Barone and A. Cavalli, *J. Chem. Theory Comput.*, 2011, **7**, 539–543.
36. S. D. Lahiri, G. F. Zhang, D. Dunaway-Mariano and K. N. Allen, *Science*, 2003, **299**, 2067–2071.
37. B. Elsasser, M. Valiev and J. H. Weare, *J. Am. Chem. Soc.*, 2009, **131**, 3869–3871.
38. E. Rosta, M. Nowotny, W. Yang and G. Hummer, *J. Am. Chem. Soc.*, 2011, **133**, 8934–8941.
39. E. Rosta, H. L. Woodcock, B. R. Brooks and G. Hummer, *J. Comput. Chem.*, 2009, **30**, 1634–1641.
40. C. A. Shaw-Reid, B. Feuston, V. Munshi, K. Getty, J. Krueger, D. J. Hazuda, M. A. Parniak, M. D. Miller and D. Lewis, *Biochemistry*, 2005, **44**, 1595–1606.
41. M. De Vivo, B. Ensing and M. L. Klein, *J. Am. Chem. Soc.*, 2005, **127**, 11226–11227.
42. M. De Vivo, B. Ensing, M. Dal Peraro, G. A. Gomez, D. W. Christianson and M. L. Klein, *J. Am. Chem. Soc.*, 2007, **129**, 387–394.
43. M. F. Moghaddam, D. F. Grant, J. M. Cheek, J. F. Greene, K. C. Williamson and B. D. Hammock, *Nat. Med.*, 1997, **3**, 562–566.
44. K. Node, Y. Q. Huo, X. L. Ruan, B. C. Yang, M. Spiecker, K. Ley, D. C. Zeldin and J. K. Liao, *Science*, 1999, **285**, 1276–1279.
45. K. R. Schmelzer, L. Kubala, J. W. Newman, I. H. Kim, J. P. Eiserich and B. D. Hammock, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 9772–9777.
46. A. Cronin, S. Mowbray, H. Durk, S. Homburg, I. Fleming, B. Fisslthaler, F. Oesch and M. Arand, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 1552–1557.
47. G. A. Gomez, C. Morisseau, B. D. Hammock and D. W. Christianson, *Biochemistry*, 2004, **43**, 4716–4723.
48. J. W. Newman, C. Morisseau and B. D. Hammock, *Prog. Lipid Res.*, 2005, **44**, 1–51.
49. F. Bernardi, A. Bottoni, M. De Vivo, M. Garavelli, G. Keseru and G. Naray-Szabo, *Chem. Phys. Lett.*, 2002, **362**, 1–7.
50. G. Palermo, M. Stenta, A. Cavalli, M. Dal Peraro and M. De Vivo, *J. Chem. Theory Comput.*, 2013, **9**, 857–862.
51. G. Palermo, E. Minniti, M. L. Greco, L. Riccardi, E. Simoni, M. Convertino, C. Marchetti, M. Rosini, C. Sissi, A. Minarini and M. De Vivo, *Chem. Commun.*, 2015, **51**, 14310–14313.
52. M. H. Ho, M. De Vivo, M. Dal Peraro and M. L. Klein, *J. Am. Chem. Soc.*, 2010, **132**, 13702–13712.
53. V. Genna, R. Gaspari, M. Dal Peraro and M. De Vivo, *Nucleic Acids Res.*, 2016, **44**, 2827–2836.



54. V. Genna, P. Vidossich, E. Ippoliti, P. Carloni and M. Vivo, *J. Am. Chem. Soc.*, 2016, **138**, 14592–14598.
55. G. Palermo, U. Rothlisberger, A. Cavalli and M. De Vivo, *Eur. J. Med. Chem.*, 2015, **91**, 15–26.
56. G. Palermo, A. D. Favia, M. Convertino and M. De Vivo, *ChemMedChem*, 2016, **11**, 1252–1258.
57. S. Petrosino and V. Di Marzo, *Curr. Opin. Invest. Drugs*, 2010, **11**, 51–62.
58. M. H. Bracey, M. A. Hanson, K. R. Masuda, R. C. Stevens and B. F. Cravatt, *Science*, 2002, **298**, 1793–1796.
59. L. Bertolacci, E. Romeo, M. Veronesi, P. Magotti, C. Albani, M. Dionisi, C. Lambruschini, R. Scarpelli, A. Cavalli, M. De Vivo, D. Piomelli and G. Garau, *J. Am. Chem. Soc.*, 2013, **135**, 22–25.
60. M. Mileni, S. Kamtekar, D. C. Wood, T. E. Benson, B. F. Cravatt and R. C. Stevens, *J. Mol. Biol.*, 2010, **400**, 743–754.
61. M. Seierstad and J. G. Breitenbucher, *J. Med. Chem.*, 2008, **51**, 7327–7343.
62. A. D. Favia, D. Habrant, R. Scarpelli, M. Migliore, C. Albani, S. M. Bertozzi, M. Dionisi, G. Tarozzo, D. Piomelli, A. Cavalli and M. De Vivo, *J. Med. Chem.*, 2012, **55**, 8807–8826.
63. G. Palermo, D. Branduardi, M. Masetti, A. Lodola, M. Mor, D. Piomelli, A. Cavalli and M. De Vivo, *J. Med. Chem.*, 2011, **54**, 6612–6623.
64. G. Palermo, I. Bauer, P. Campomanes, A. Cavalli, A. Armirotti, S. Giroto, U. Rothlisberger and M. De Vivo, *PLoS Comput. Biol.*, 2015, **11**, e1004231.
65. G. Palermo, P. Campomanes, M. Neri, D. Piomelli, A. Cavalli, U. Rothlisberger and M. De Vivo, *J. Chem. Theory Comput.*, 2013, **9**, 1202–1213.
66. G. Palermo, P. Campomanes, A. Cavalli, U. Rothlisberger and M. De Vivo, *J. Phys. Chem. B*, 2015, **119**, 789–801.
67. S. K. Bagal, A. D. Brown, P. J. Cox, K. Omoto, R. M. Owen, D. C. Pryde, B. Sidders, S. E. Skerratt, E. B. Stevens, R. I. Storer and N. A. Swain, *J. Med. Chem.*, 2013, **56**, 593–624.
68. C. A. Hubner and T. J. Jentsch, *Hum. Mol. Genet.*, 2002, **11**, 2435–2445.
69. J. W. Warmke and B. Ganetzky, *Proc. Natl. Acad. Sci. U. S. A.*, 1994, **91**, 3438–3442.
70. M. C. Trudeau, J. W. Warmke, B. Ganetzky and G. A. Robertson, *Science*, 1995, **269**, 92–95.
71. M. C. Sanguinetti and M. Tristani-Firouzi, *Nature*, 2006, **440**, 463–469.
72. M. Recanatini, A. Cavalli and M. Masetti, *ChemMedChem*, 2008, **3**, 523–535.
73. A. Coi, I. Massarelli, L. Murgia, M. Saraceno, V. Calderone and A. M. Bianucci, *Bioorg. Med. Chem.*, 2006, **14**, 3153–3159.
74. L. Du-Cuny, L. Chen and S. Zhang, *J. Chem. Inf. Model.*, 2011, **51**, 2948–2960.
75. G. M. Keseru, *Bioorg. Med. Chem. Lett.*, 2003, **13**, 2773–2775.

76. B. H. Su, M. Y. Shen, E. X. Esposito, A. J. Hopfinger and Y. J. Tseng, *J. Chem. Inf. Model.*, 2010, **50**, 1304–1318.
77. K. Yoshida and T. Niwa, *J. Chem. Inf. Model.*, 2006, **46**, 1371–1378.
78. A. M. Aronov, *J. Med. Chem.*, 2006, **49**, 6917–6921.
79. A. M. Aronov and B. B. Goldman, *Bioorg. Med. Chem.*, 2004, **12**, 2307–2315.
80. A. Cavalli, R. Buonfiglio, C. Ianni, M. Masetti, L. Ceccarini, R. Caves, M. W. Chang, J. S. Mitcheson, M. Roberti and M. Recanatini, *J. Med. Chem.*, 2012, **55**, 4010–4014.
81. A. Cavalli, E. Poluzzi, F. De Ponti and M. Recanatini, *J. Med. Chem.*, 2002, **45**, 3844–3853.
82. S. Durdagi, H. J. Duff and S. Y. Noskov, *J. Chem. Inf. Model.*, 2011, **51**, 463–474.
83. S. Ekins, W. J. Crumb, R. D. Sarazan, J. H. Wikel and S. A. Wrighton, *J. Pharmacol. Exp. Ther.*, 2002, **301**, 427–434.
84. R. A. Pearlstein, R. J. Vaz, J. Kang, X. L. Chen, M. Preobrazhenskaya, A. E. Shchekotikhin, A. M. Korolev, L. N. Lysenkova, O. V. Miroshnikova, J. Hendrix and D. Rampe, *Bioorg. Med. Chem. Lett.*, 2003, **13**, 1829–1835.
85. M. Masetti, A. Cavalli and M. Recanatini, *J. Comput. Chem.*, 2008, **29**, 795–808.
86. G. N. Tseng, K. D. Sonawane, Y. V. Korolkova, M. Zhang, J. Liu, E. V. Grishin and H. R. Guy, *Biophys. J.*, 2007, **92**, 3524–3540.
87. F. Osterberg and J. Aqvist, *FEBS Lett.*, 2005, **579**, 2939–2944.
88. P. J. Stansfeld, P. Gedeck, M. Gosling, B. Cox, J. S. Mitcheson and M. J. Sutcliffe, *Proteins*, 2007, **68**, 568–580.
89. G. P. Di Martino, M. Masetti, L. Ceccarini, A. Cavalli and M. Recanatini, *J. Chem. Inf. Model.*, 2013, **53**, 159–175.
90. L. Boukharta, H. Keranen, A. Stary-Weinzinger, G. Wallin, B. L. de Groot and J. Aqvist, *Biochemistry*, 2011, **50**, 6146–6156.
91. L. Ceccarini, M. Masetti, A. Cavalli and M. Recanatini, *PLoS One*, 2012, **7**, e49017.
92. D. A. Kopfer, U. Hahn, I. Ohmert, G. Vriend, O. Pongs, B. L. de Groot and U. Zachariae, *PLoS One*, 2012, **7**, e41023.
93. E. Raschi, V. Vasina, E. Poluzzi and F. De Ponti, *Pharmacol. Res.*, 2008, **57**, 181–195.
94. X. Xu, M. Recanatini, M. Roberti and G. N. Tseng, *Mol. Pharm.*, 2008, **73**, 1709–1721.
95. R. S. Hansen, T. G. Diness, T. Christ, E. Wettwer, U. Ravens, S. P. Olesen and M. Grunnet, *Mol. Pharm.*, 2006, **70**, 1319–1329.
96. S. Durdagi, J. Guo, J. P. Lees-Miller, S. Y. Noskov and H. J. Duff, *J. Pharmacol. Exp. Ther.*, 2012, **342**, 441–452.
97. E. Giacomini, R. Buonfiglio, M. Masetti, Y. Wang, G. N. Tseng, M. Roberti and M. Recanatini, *Comb. Chem. High Throughput Screening*, 2015, **18**, 269–280.
98. T. Schoneberg, A. Schulz, H. Biebertmann, T. Hermsdorf, H. Rompler and K. Sangkuhl, *Pharmacol. Ther.*, 2004, **104**, 173–206.
99. A. Sali and J. P. Overington, *Protein Sci.*, 1994, **3**, 1582–1596.

100. F. Musiani, G. Rossetti, A. Giorgetti and P. Carloni, *Adv. Exp. Med. Biol.*, 2014, **805**, 441–457.
101. G. Rossetti, D. Dibenedetto, V. Calandrini, A. Giorgetti and P. Carloni, *Arch. Biochem. Biophys.*, 2015, **582**, 91–100.
102. S.-L. Niu, D. C. Mitchell and B. J. Litman, *J. Biol. Chem.*, 2002, **277**, 20139–20145.
103. W. Stone, C. Farnsworth and E. Dratz, *Exp. Eye Res.*, 1979, **28**, 387–397.
104. A. Bruno, G. Costantino, G. de Fabritiis, M. Pastor and J. Selent, *PLoS One*, 2012, **7**, e42023.
105. M. I. Mahmood, X. Liu, S. Neya and T. Hoshino, *Chem. Pharm. Bull.*, 2013, **61**, 426–437.
106. J. P. Overington, B. Al-Lazikani and A. L. Hopkins, *Nat. Rev. Drug Discovery*, 2006, **5**, 993–996.
107. G. Khelashvili, A. Grossfield, S. E. Feller, M. C. Pitman and H. Weinstein, *Proteins: Struct., Funct., Bioinf.*, 2009, **76**, 403–417.
108. T. J. Pucadyil and A. Chattopadhyay, *Biochim. Biophys. Acta Biomembr.*, 2004, **1663**, 188–200.
109. R. Cao, G. Rossetti, A. Bauer and P. Carloni, *PLoS One*, 2015, **10**, e0126833.
110. K. A. Jacobson, D. Ukena, W. Padgett, J. W. Daly and K. L. Kirk, *J. Med. Chem.*, 1987, **30**, 211–214.
111. G. Ross, R. D. Abbott, H. Petrovitch, D. M. Morens, A. Grandinetti, K. H. Tung, C. M. Tanner, K. H. Masaki, P. L. Blanchette, J. D. Curb, J. S. Popper and L. R. White, *JAMA: J. Am. Med. Assoc.*, 2000, **283**, 2674–2679.
112. C. V. Gomes, M. P. Kaster, A. R. Tom, P. M. Agostinho and R. A. Cunha, *Biochim. Biophys. Acta Biomembr.*, 2011, **1808**, 1380–1399.
113. W. Liu, E. Chun, A. A. Thompson, P. Chubukov, F. Xu, V. Katritch, G. W. Han, C. B. Roth, L. H. Heitman, A. P. Ijzerman, V. Cherezov and R. C. Stevens, *Science*, 2012, **337**, 232–236.
114. R. B. Postuma, A. E. Lang, R. P. Munhoz, K. Charland, A. Pelletier, M. Moscovich, L. Filla, D. Zanatta, S. Rios Romenets, R. Altman, R. Chuang and B. Shah, *Neurology*, 2012, **79**, 651–658.
115. H. Douna, B. M. Bavelaar and H. Pellikaan, *Open Pharmacol. J.*, 2012, **6**, 12–26.
116. K. Nakaso, S. Ito and K. Nakashima, *Neurosci. Lett.*, 2008, **432**, 146–150.
117. J. S. Fink, D. R. Weaver, S. A. Rivkees, R. A. Peterfreund, A. E. Pollack, E. M. Adler and S. M. Reppert, *Mol. Brain Res.*, 1992, **14**, 186–195.
118. F. W. Pfrieger, *Biochim. Biophys. Acta Biomembr.*, 2003, **1610**, 271–280.
119. A. Mori, T. Shindou, M. Ichimura, H. Nonaka and H. Kase, *The Basal Ganglia V*, Springer, 1996, pp. 119–122.
120. A. S. Doré, N. Robertson, J. C. Errey, I. Ng, K. Hollenstein, B. Tehan, E. Hurrell, K. Bennett, M. Congreve, F. Magnani, C. G. Tate, M. Weir and F. H. Marshall, *Structure*, 2011, **19**, 1283–1293.
121. J. Y. Lee and E. Lyman, *Biophys. J.*, 2012, **102**, 2114–2120.
122. D. Sabbadin, A. Ciancetta and S. Moro, *J. Chem. Inf. Model.*, 2014, **54**, 169–183.

123. Y. Liu, S. K. Burger, P. W. Ayers and E. Vöhringer-Martinez, *J. Phys. Chem. B*, 2011, **115**, 13880–13890.
124. E. Serebryany, G. A. Zhu and E. C. Yan, *Biochim. Biophys. Acta Biomembr.*, 2012, **1818**, 225–233.
125. R. Buonfiglio, M. Recanatini and M. Masetti, *ChemMedChem*, 2015, **10**, 1141–1148.
126. G. Bottegoni, W. Rocchia, M. Rueda, R. Abagyan and A. Cavalli, *PLoS One*, 2011, **6**.
127. J. A. Read, V. J. Winter, C. M. Eszes, R. B. Sessions and R. L. Brady, *Proteins*, 2001, **43**, 175–185.
128. G. Di Stefano, M. Manerba, L. Di Ianni and L. Fiume, *Future Med. Chem.*, 2016, **8**, 713–725.
129. D. Hanahan and R. A. Weinberg, *Cell*, 2011, **144**, 646–674.
130. M. G. Vander Heiden, L. C. Cantley and C. B. Thompson, *Science*, 2009, **324**, 1029–1033.
131. R. Scatena, P. Bottoni, A. Pontoglio, L. Mastrototaro and B. Giardina, *Expert Opin. Invest. Drugs*, 2008, **17**, 1533–1545.
132. L. Qiu, M. Gulotta and R. Callender, *Biophys. J.*, 2007, **93**, 1677–1686.
133. L. Fiume, M. Manerba, M. Vettraino and G. Di Stefano, *Future Med. Chem.*, 2014, **6**, 429–445.
134. C. Granchi, E. C. Calvaresi, T. Tuccinardi, I. Paterni, M. Macchia, A. Martinelli, P. J. Hergenrother and F. Minutolo, *Org. Biomol. Chem.*, 2013, **11**, 6588–6596.
135. C. Granchi, S. Roy, C. Giacomelli, M. Macchia, T. Tuccinardi, A. Martinelli, M. Lanza, L. Betti, G. Giannaccini, A. Lucacchini, N. Funel, L. G. Leon, E. Giovannetti, G. J. Peters, R. Palchaudhuri, E. C. Calvaresi, P. J. Hergenrother and F. Minutolo, *J. Med. Chem.*, 2011, **54**, 1599–1612.
136. A. Kohlmann, S. G. Zech, F. Li, T. Zhou, R. M. Squillace, L. Comodore, M. T. , Greenfield, X. Lu, D. P. Miller, W. S. Huang, J. Qi, R. M. Thomas, Y. Wang, S. Zhang, R. Dodd, S. Liu, R. Xu, Y. Xu, J. J. Miret, V. Rivera, T. Clackson, W. C. Shakespeare, X. Zhu and D. C. Dalgarno, *J. Med. Chem.*, 2013, **56**, 1023–1040.
137. M. Manerba, M. Vettraino, L. Fiume, G. Di Stefano, A. Sartini, E. Giacomini, R. Buonfiglio, M. Roberti and M. Recanatini, *ChemMedChem*, 2012, **7**, 311–317.
138. R. Buonfiglio, M. Ferraro, F. Falchi, A. Cavalli, M. Masetti and M. Recanatini, *J. Chem. Inf. Model.*, 2013, **53**, 2792–2797.
139. S. Rupiani, R. Buonfiglio, M. Manerba, L. Di Ianni, M. Vettraino, E. Giacomini, M. Masetti, F. Falchi, G. Di Stefano, M. Roberti and M. Recanatini, *Eur. J. Med. Chem.*, 2015, **101**, 63–70.
140. C. Granchi, A. Capecchi, G. Del Frate, A. Martinelli, M. Macchia, F. Minutolo and T. Tuccinardi, *Molecules*, 2015, **20**, 8772–8790.
141. H. Dyson and P. Wright, *Nat. Rev. Mol. Cell Biol.*, 2005, **6**, 197–208.
142. P. Tompa, *Trends Biochem. Sci.*, 2002, **27**, 527–533.

143. P. E. Wright and H. J. Dyson, *J. Mol. Biol.*, 1999, **293**, 321–331.
144. V. N. Uversky, *Protein Sci.*, 2002, **11**, 739–756.
145. A. K. Dunker, Z. Obradovic, P. Romero, E. C. Garner and C. J. Brown, *Genome Inf. Int. Conf. Genome Inf.*, 2000, **11**, 161–171.
146. V. N. Uversky, C. J. Oldfield and A. K. Dunker, *J. Mol. Recognit.*, 2005, **18**, 343–384.
147. V. N. Uversky and D. Eliezer, *Curr. Protein Pept. Sci.*, 2009, **10**, 483–499.
148. P. H. Weinreb, W. Zhen, A. W. Poon, K. A. Conway and P. T. Lansbury, *Biochemistry*, 1996, **35**, 13709–13715.
149. A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. Kang, C. R. Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner and Z. Obradovic, *J. Mol. Graph. Modell.*, 2001, **19**, 26–59.
150. J. W. Ponder and D. A. Case, *Adv. Protein Chem.*, 2003, **66**, 27–85.
151. A. D. Mackerell Jr., *J. Comput. Chem.*, 2004, **25**, 1584–1604.
152. A. K. Dunker, I. Silman, V. N. Uversky and J. L. Sussman, *Curr. Opin. Struct. Biol.*, 2008, **18**, 756–764.
153. J. M. Bourhis, B. Canard and S. Longhi, *Curr. Protein Pept. Sci.*, 2007, **8**, 135–149.
154. A. I. Bartlett and S. E. Radford, *Nat. Struct. Mol. Biol.*, 2009, **16**, 582–588.
155. D. Eliezer, *Curr. Opin. Struct. Biol.*, 2009, **19**, 23–30.
156. R. B. Best, N.-V. Buchete and G. Hummer, *Biophys. J.*, 2008, **95**, L07–L09.
157. K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror and D. E. Shaw, *Proteins: Struct., Funct., Bioinf.*, 2010, **78**, 1950–1958.
158. A. E. Aliev and D. Courtier-Murias, *J. Phys. Chem. B*, 2010, **114**, 12358–12375.
159. O. F. Lange, D. van der Spoel and B. L. de Groot, *Biophys. J.*, 2010, **99**, 647–655.
160. S. Piana, K. Lindorff-Larsen and D. E. Shaw, *Biophys. J.*, 2011, **100**, L47–L49.
161. D. Dibenedetto, G. Rossetti, R. Caliendo and P. Carloni, *Biochemistry*, 2013, 130821094755001.
162. V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg and C. Simmerling, *Proteins*, 2006, **65**, 712–725.
163. K. Lindorff-Larsen, P. Maragakis, S. Piana, M. P. Eastwood, R. O. Dror and D. E. Shaw, *PLoS One*, 2012, **7**, e32131.
164. R. G. Perez, J. C. Waymire, E. Lin, J. J. Liu, F. Guo and M. J. Zigmond, *J. Neurosci.*, 2002, **22**, 3090–3099.
165. L. Yavich, H. Tanila, S. Vepsäläinen and P. Jäkälä, *J. Neurosci.*, 2004, **24**, 11165–11170.
166. V. Lehmsiek, E.-M. Tan, J. Schwarz and A. Storch, *NeuroReport*, 2002, **13**, 1279–1283.

167. M. M. Dedmon, K. Lindorff-Larsen, J. Christodoulou, M. Vendruscolo and C. M. Dobson, *J. Am. Chem. Soc.*, 2005, **127**, 476–477.
168. P. Schanda and B. Brutscher, *J. Am. Chem. Soc.*, 2005, **127**, 8014–8015.
169. P. Schanda, E. Kupce and B. Brutscher, *J. Biomol. NMR*, 2005, **33**, 199–211.
170. L. Salmon, G. Nodet, V. Ozenne, G. Yin, M. R. Jensen, M. Zweckstetter and M. Blackledge, *J. Am. Chem. Soc.*, 2010, **132**, 8407–8418.
171. G. Huntington, *Med. Surg. Rep.*, 1872, **26**, 317–321.
172. C. A. Ross and S. J. Tabrizi, *Lancet Neurol.*, 2011, **10**, 83–98.
173. F. O. Walker, *Lancet*, 2007, **369**, 218–228.
174. M. MacDonald, C. Ambrose, M. Duyao, R. Myers, C. Lin, L. Srinidhi, G. Barnes, S. Taylor, M. James, N. Groot, H. MacFarlane, B. Jenkins, M. Anderson, N. Wexler, J. Gusella, G. Bates, S. Baxendale, H. Hummerich, S. Kirby, M. North, S. Youngman, R. Mott, G. Zehetner, Z. Sedlacek, A. Poustka, A.-M. Frischauf, H. Lehrach, A. Buckler, D. Church, L. Doucette-Stamm, M. O'Donovan, L. Riba-Ramirez, M. Shah, V. Stanton, S. Strobel, K. Draths, J. Wales, P. Dervan, D. Housman, M. Altherr, R. Shiang, L. Thompson, T. Fielder, J. Wasmuth, D. Tagle, J. Valdes, L. Elmer, M. Allard, L. Castilla, M. Swaroop, K. Blanchard, F. Collins, R. Snell, T. Holloway, K. Gillespie, N. Datson, D. Shaw and P. Harper, *Cell*, 1993, **72**, 971–983.
175. A. Reiner, I. Dragatsis, S. Zeitlin and D. Goldowitz, *Mol. Neurobiol.*, 2003, **28**, 259–276.
176. C. A. Ross, *Neuron*, 1995, **15**, 493–496.
177. J. F. Gusella and M. E. MacDonald, *Nat. Rev. Neurosci.*, 2000, **1**, 109–115.
178. A. Rosenblatt, K. Y. Liang, H. Zhou, M. H. Abbott, L. M. Gourley, R. L. Margolis, J. Brandt and C. A. Ross, *Neurology*, 2006, **66**, 1016–1020.
179. M. de Mezer, M. Wojciechowska, M. Napierala, K. Sobczak and W. J. Krzyzosiak, *Nucleic Acids Res.*, 2011, **39**, 3852–3863.
180. L. Schols, P. Bauer, T. Schmidt, T. Schulte and O. Riess, *Lancet Neurol.*, 2004, **3**, 291–304.
181. D. Housman, *Nat. Genet.*, 1995, **10**, 3.
182. D. D. Rudnicki, R. L. Margolis, C. E. Pearson and W. J. Krzyzosiak, *PLoS Genet*, 2012, **8**, e1002545.
183. M. Wojciechowska and W. J. Krzyzosiak, *RNA Biol.*, 2011, **8**, 565–571.
184. R. Nalavade, N. Griesche, D. P. Ryan, S. Hildebrand and S. Krauss, *Cell Death Dis.*, 2013, **4**, e752.
185. A. Kiliszek, R. Kierzek, W. J. Krzyzosiak and W. Rypniewski, *Nucleic Acids Res.*, 2010, **38**, 8370–8376.
186. A. Kumar, R. Parkesh, L. J. Sznajder, J. L. Childs-Disney, K. Sobczak and M. D. Disney, *ACS Chem. Biol.*, 2012, **7**, 496–505.
187. L. P. W. Ranum and T. A. Cooper, *Annu. Rev. Neurosci.*, 2006, **29**, 259–277.
188. A. Fiszer and W. J. Krzyzosiak, *J. Mol. Med.*, 2013, **91**, 683–691.
189. S. Krauss, N. Griesche, E. Jastrzebska, C. Chen, D. Rutschow, C. Achmuller, S. , Dorn, S. M. Boesch, M. Lalowski, E. Wanker, R. Schneider and S. Schweiger, *Nat. Commun.*, 2013, **4**, 1511.



190. A. Fiszer and W. J. Krzyzosiak, *Nucleic Acids Res.*, 2014, **42**, 6787–6810.
191. S. G. Rzuczek, H. Park and M. D. Disney, *Angew. Chem., Int. Ed.*, 2014, **53**, 3.
192. J. L. Childs-Disney, I. Yildirim, H. Park, J. R. Lohman, L. Guan, T. Tran, P. Sarkar, G. C. Schatz and M. D. Disney, *ACS Chem. Biol.*, 2014, **9**, 538–550.
193. J. L. Childs-Disney, E. Stepniak-Konieczna, T. Tran, I. Yildirim, H. Park, C. Z. Chen, J. Hoskins, N. Southall, J. J. Marugan, S. Patnaik, W. Zheng, C. P. Austin, G. C. Schatz, K. Sobczak, C. A. Thornton and M. D. Disney, *Nat. Commun.*, 2013, **4**, 2044.
194. W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell and P. A. Kollman, *J. Am. Chem. Soc.*, 1995, **117**, 18.
195. I. Yildirim, H. A. Stern, S. D. Kennedy, J. D. Tubbs and D. H. Turner, *J. Chem. Theory Comput.*, 2010, **6**, 11.
196. A. Perez, I. Marchan, D. Svozil, J. Sponer, T. E. Cheatham 3rd, C. A. Loughton and M. Orozco, *Biophys. J.*, 2007, **92**, 3817.
197. J. Lipfert, S. Doniach, R. Das and D. Herschlag, *Annu. Rev. Biochem.*, 2014, **83**, 813–841.
198. M. Zgarbova, M. Otyepka, J. Sponer, A. Mladek, P. Banas, T. E. Cheatham, 3rd and P. Jurecka, *J. Chem. Theory Comput.*, 2011, **7**, 2886.
199. P. T. Lang, S. R. Brozell, S. Mukherjee, E. F. Pettersen, E. C. Meng, V. Thomas, R. C. Rizzo, D. A. Case, T. L. James and I. D. Kuntz, *Rna*, 2009, **15**, 1219–1230.
200. S. Fulle and H. Gohlke, *J. Mol. Recognit.*, 2009, **23**, 220–231.
201. P. Daldrop, F. E. Reyes, D. A. Robinson, C. M. Hammond, D. M. Lilley, R. T. Batey and R. Brenk, *Chem. Biol.*, 2011, **18**, 324–335.
202. L. Chen, G. A. Calin and S. Zhang, *J. Chem. Inf. Model.*, 2012, **52**, 2741–2753.
203. F. Musiani, G. Rossetti, L. Capece, T. M. Gerger, C. Micheletti, G. Varani and P. Carloni, *J. Am. Chem. Soc.*, 2014, **136**, 15631.
204. P. Banas, D. Hollas, M. Zgarbova, P. Jurecka, M. Orozco, T. Chetam, J. Sponer and M. Otyepka, *J. Chem. Theory Comput.*, 2010, **6**, 3836.
205. I. Faustino, A. Perez and M. Orozco, *Biophys. J.*, 2010, **99**, 1876–1885.
206. N. J. Deng and P. Cieplak, *Biophys. J.*, 2010, **98**, 627–636.
207. P. Banáš, P. Sklenovský, J. E. Wedekind, J. Šponer and M. Otyepka, *J. Phys. Chem. B*, 2013, **116**, 13.
208. I. Yildirim, H. Park, M. D. Disney and G. C. Schatz, *J. Am. Chem. Soc.*, 2013, **135**, 3528.
209. S. Fulle, N. A. Christ, E. Kestner and H. Gohlke, *J. Chem. Inf. Model.*, 2010, **50**, 1489–1501.
210. T. N. Do, E. Ippoliti, P. Carloni, G. Varani and M. Parrinello, *J. Chem. Theory Comput.*, 2012, **8**, 688–694.



- 211. A. Barducci, G. Bussi and M. Parrinello, *Phys. Rev. Lett.*, 2008, **100**, 020603.
- 212. A. Bochicchio, G. Rossetti, O. Tabarrini, S. Kraubeta and P. Carloni, *J. Chem. Theory Comput.*, 2015, **11**, 4911–4922.
- 213. M. Masetti, A. Cavalli, M. Recanatini and F. L. Gervasio, *J. Phys. Chem. B*, 2009, **113**, 4807.
- 214. K. Sobczak and W. J. Krzyzosiak, *J. Biol. Chem.*, 2005, **280**, 3898–3910.

## CHAPTER 13

# *The Polypharmacology Gap Between Chemical Biology and Drug Discovery*

ALBERT A. ANTOLIN AND JORDI MESTRES\*

Research Group on Systems Pharmacology, Research Program on Biomedical Informatics (GRIB), IMIM Hospital del Mar Medical Research Institute and Universitat Pompeu Fabra, Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain

\*Email: jmestres@imim.es

## **13.1 Introduction: Chemical Biology and the Limits of Reductionism**

### **13.1.1 Polypharmacology in Drug Discovery**

Our understanding of drug action has always been limited by our understanding of biology and by technology and data availability. In the early days of drug discovery, drugs were discovered directly using phenotypic screening assays and the mechanism of action of drugs was largely unknown.<sup>1</sup> Subsequent advances in pharmacology proved that drugs exert their therapeutic action by ‘uniting with certain specific receptors of cells’ leading to the drug receptor theory becoming fully accepted in the 1950s.<sup>2</sup> At that time, a reductionist approach to science was necessary due to the limitations in available knowledge and technologies. Therefore, it was often assumed that

drugs exerted their therapeutic action *via* the modulation of a single protein, although this was often not proved.

The molecular biology revolution of the 1980s enabled a paradigm shift in drug discovery, from discovering drugs using phenotypic screening assays to developing drugs as specific inhibitors of a single protein believed to be solely responsible of a given disease phenotype using miniaturised biochemical assays on isolated proteins obtained through biotechnology techniques.<sup>3</sup> This new drug discovery approach was termed target-based drug discovery. However, due to insufficient time and resources, drugs were not screened systematically through a large panel of protein targets to acquire knowledge about their complete pharmacological profile, and thus the selectivity of both target-based and phenotypic drugs was largely an assumption.<sup>4</sup>

In the 2000s, the sequence of the human genome prompted the embracement of a more comprehensive approach to studying biology, termed systems biology.<sup>5</sup> Moreover, it was also confirmed that some old phenotypic antipsychotic drugs exerted their therapeutic action through binding to several proteins of the GPCR family.<sup>6,7</sup> In addition, target-based kinase inhibitors such as imatinib, designed to selectively inhibit a unique kinase, were also found to inhibit several proteins of the kinase family and this promiscuity was beneficial to extend the uses of these drugs without unacceptable toxicity.<sup>8,9</sup>

Overall, a more complex picture of the mechanism of action of drugs was emerging. Several public and private initiatives to collect and store drug-target interaction data also contributed to change our biased perception of drug selectivity by enabling the performance of the first global analysis of drug action.<sup>10,11</sup> Moreover, the first drug-target network showed that drugs were already known to bind two targets on average.<sup>12</sup> However, that initial drug-target network was highly incomplete and the information highly biased towards certain areas of interest. When more data or computational predictions were added, the average number of protein targets per drug increased to six.<sup>13</sup> These discoveries prompted the appearance of the term polypharmacology, coined to refer to the binding of a chemical molecule to several proteins.<sup>10</sup>

Drug polypharmacology was shown to be a common phenomenon among drugs, challenging previous reductionist assumptions of their mechanism of action (Figure 13.1).<sup>14</sup> It is now widely recognised that selective drugs are more the exception rather than the rule and a recent analysis showed that only 15% of current drugs are known to bind solely to one protein target, uncovering an increasingly complex drug-target network.<sup>15</sup> Today, the use of experimental and computational methods to identify drug polypharmacology is becoming increasingly essential in drug discovery and development, in order to identify new targets of drugs to both maximally exploit all repurposing opportunities and alert earlier on possible side-effects or toxicities.<sup>15-17</sup> This notwithstanding, the current understanding of drug action remains incomplete and the implications of drug polypharmacology for drug

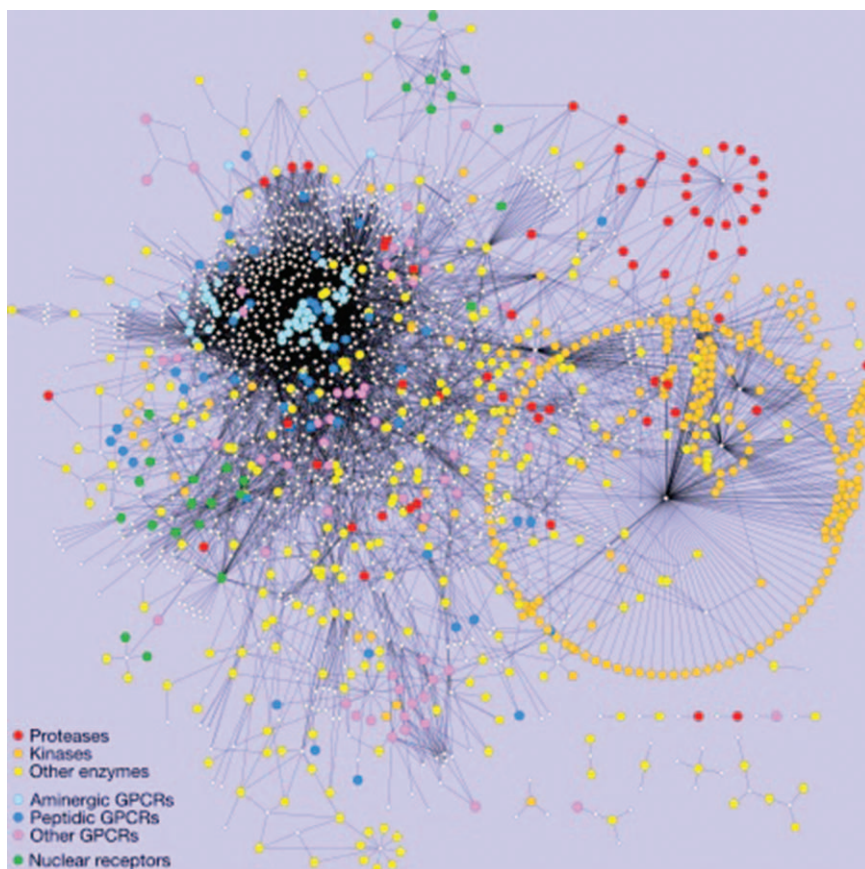


**Figure 13.1** Drug-target network. Drugs (small white circles) are linked to protein targets if their affinity is more potent than 1  $\mu\text{M}$ .<sup>14</sup>

efficacy and safety in the clinic are still, to a worrying extent, mostly unknown.<sup>18–20</sup> In this respect, however, more comprehensive system approaches to understanding pharmacology are being adopted in drug discovery, whereas chemical biology is still narrowly focussed on the reductionist view of selectivity.

### 13.1.2 Selectivity in Chemical Biology

Small molecules have long been used as a non-invasive means to study biological systems given their capacity to bind and modulate biomolecules.<sup>21,22</sup> Today, new small molecule tools (also referred to as standard inhibitors, chemical probes or chemical tools) continue to be strongly needed in order to expand our understanding of the functions of proteins in health and disease, contribute to the annotation of the human genome, and enable the validation of new targets for drug discovery.<sup>23–25</sup> Accordingly, the

development of chemical probes continues to be one of the cornerstones of chemical biology.<sup>26</sup>

Chemical tools are highly complementary to more invasive biological tools such as RNA interference or the recently developed CRISPR-Cas technology, since they do not eliminate the target from the system but inhibit its function.<sup>22</sup> Moreover, chemical probes provide proof that the target can be modulated pharmacologically and are key pathfinder molecules in follow-up drug discovery projects.<sup>22</sup> Accordingly, the development of high-quality chemical probes is paramount to the exploration of new areas of biology and quickly translating them into new therapeutic opportunities.<sup>27</sup> Given their importance, during the last decade several initiatives have discovered and developed new chemical probes and have made them available to the wide chemical biology community. The National Institute of Health (NIH) Molecular Libraries Program (MLP) enabled the access of high-throughput screening—usually restricted to pharmaceutical industry—to academic institutions and this initiative facilitated the identification of almost 200 chemical probes.<sup>28</sup> More recently, the Structural Genomics Consortium was launched as a public–private partnership that has enabled the discovery and widespread access of several chemical probes for epigenetic targets, illuminating new areas of biology that have been quickly translated into new drug discovery programs.<sup>29</sup>

However, these initiatives have sparked discussions surrounding the definition of what constitutes a chemical probe and which quality criteria should be used.<sup>22,30–33</sup> An early revision of the first phase of the MLP program led to more stringent criteria being used in the second phase of the MLP.<sup>31</sup> But it quickly became apparent that over-restrictive chemical probe requirements stifle innovation, as probe quality tends to increase over time.<sup>22</sup> Despite the on-going discussions surrounding chemical probe criteria, it is widely accepted that chemical probes are only useful if they are potent, have known selectivity and a proven mechanism of action.<sup>27</sup> Interestingly, although drug polypharmacology is widely acknowledged to be beneficial in many cases, chemical probes are considered to be only useful if they are selective against their intended specific target, illustrating the disconnect that currently exists between chemical biology and drug discovery.<sup>24</sup>

To ensure target selectivity, chemical probes are usually tested *in vitro* on a panel of target proteins phylogenetically related to the primary target or on a larger panel of pre-selected diverse targets. However, due to limitations in our current screening technologies, the percentage of the proteome that is currently screened is relatively low. Accordingly, the risk of non-obvious polypharmacology to distantly related targets that can confound the effects observed using chemical probes is always there. Given the devastating consequences for biomedical research that can originate from the use of poor quality chemical probes,<sup>27</sup> it is of paramount importance that recently developed computational methods to predict polypharmacology are used to identify potentially unknown targets of chemical probes and thus de-risk the practise of chemical biology while we advance towards a more

comprehensive and systems understanding of chemical biology.<sup>34</sup> In this chapter, we review systems pharmacology computational methods to predict polypharmacology and present two case studies where the use of *in silico* target profiling enabled the identification of non-obvious off-targets for chemical probes.

## 13.2 Systems Pharmacology: Databases and Methods

### 13.2.1 Databases of Chemical, Biological and Pharmacological Data

The development of computational methods to predict polypharmacology relies on the availability of chemical, biological and/or pharmacological data in the public domain. The first chemical database retrieval systems, such as the Chemical Abstract Service (CAS), were established in the 1950s. However, they were not publicly available and this information would not be easily accessible to the wide scientific community for many years.<sup>35</sup> In 1971, the Protein Data Bank (PDB) began its activity and its public accessibility established it as the reference database for experimental crystallographic data (3D structures) of macromolecules.<sup>35</sup> During the 2000s, chemistry started to embrace the openness that had characterised biology databases and several public databases hosting chemical, biological and/or pharmacological information were launched.<sup>36</sup> First, the International Union of Basic and Clinical Pharmacology (IUPHAR) launched its database, closely followed by BindingDB that hosted diverse types of compound–target affinity data.<sup>37,38</sup> Shortly afterwards, PubChem was launched to host the data generated by the NIH MLP program, followed by the launch of DrugBank and ChemBank.<sup>39–41</sup> The effort culminated with the launch of ChEMBL, a high quality literature-curated database hosting medicinal chemistry information.<sup>11</sup> More recently, the canSAR knowledgebase, that integrates pharmacology data with a wide diversity of biological data, was also launched to facilitate drug discovery.<sup>42</sup> The release of all this high-quality information in the public domain has recently boosted the development of computational methods to predict polypharmacology (Table 13.1).

**Table 13.1** Representative public sources of ligand–target interaction data.

| Database  | Web address                                                                                     | Data description                            |
|-----------|-------------------------------------------------------------------------------------------------|---------------------------------------------|
| ChEMBL    | <a href="https://www.ebi.ac.uk/chembl/">https://www.ebi.ac.uk/chembl/</a>                       | > 2 036 000 compounds<br>11 224 targets     |
| BindingDB | <a href="https://www.bindingdb.org/bind/index.jsp">https://www.bindingdb.org/bind/index.jsp</a> | > 581 000 compounds<br>6755 protein targets |
| PubChem   | <a href="https://pubchem.ncbi.nlm.nih.gov/#">https://pubchem.ncbi.nlm.nih.gov/#</a>             | > 2 395 000 compounds<br>10 341 targets     |
| DrugBank  | <a href="https://www.drugbank.ca/">https://www.drugbank.ca/</a>                                 | 8261 drugs<br>4338 targets                  |

### 13.2.2 Computational Methods to Predict Polypharmacology

The aforementioned revolution in chemical, biological and pharmacological data availability in the public domain is making it increasingly possible to predict *in silico* the affinity of any given small molecule across a large number of targets.<sup>43</sup> Historically, these methods have been divided between structure-based and ligand-based methods depending on the source of information being used. Structure-based methods use information on the 3D protein structure and computational methods such as docking or binding site similarity to predict new targets of given molecules.<sup>43–45</sup> They have been successfully used to identify new targets of drugs, such as in the identification of carbonic anhydrase as a nanomolar off-target of celecoxib.<sup>46</sup> However, they are limited by 3D protein structure availability in the PDB that has been reported to be highly biased.<sup>47</sup> In contrast, ligand-based methods use pharmacological information to generate target models based on the chemical molecules that are known to bind to these protein targets. Several strategies have been used to generate these target models, including Neural Networks, Bayesian statistics and other machine learning algorithms.<sup>45</sup> However, the strategy that has been most successfully used exploits the basic principle that similar molecules should bind to the same target(s), and they rely on chemical similarity calculated using molecular descriptors in order to predict new targets of chemical molecules. The first descriptors used for *in silico* target profiling were topological fingerprints encoding the presence of sub-structural fragments in molecules using a binary fingerprint such as MDL MACCS or Daylight. These fingerprints were then compared, generally using Tanimoto distances, in a fast and efficient manner.<sup>48</sup>

In our research, we use the ligand-based *in silico* target profiling approach implemented in the CT-link software.<sup>18</sup> Given the two-dimensional structure of a molecule, CT-link returns the predicted affinities for those targets for which ligand information is available in public sources of pharmacological data. Six independent ligand-based approaches are implemented in the current version of CT-link that rely on descriptor-based similarities,<sup>18</sup> an adapted implementation of the similarity ensemble approach,<sup>16</sup> machine learning methods (including support vector machine, random forest, and neural networks), fuzzy fragment-based mapping based on the identification of a simple active subgraph, quantitative structure–activity relationships, and probabilities based on target cross-pharmacology.<sup>18</sup> The software has been successfully validated both retrospectively, on its ability to predict the entire experimental interaction matrix between 13 antipsychotic drugs and 34 protein targets,<sup>49</sup> to identify cancer-relevant targets from selective cytotoxic compounds in tumour cells,<sup>50</sup> and to predict the target space relevant to malaria,<sup>51</sup> but also prospectively, on its capacity to identify the correct targets for all molecules contained in a biologically-orphan chemical library,<sup>52</sup> to correctly anticipate the affinity profile of the drug cyclo-benzaprine,<sup>53</sup> to identify confounding off-targets of chemical probes,<sup>54</sup>



to predict the target of novel inhibitors of amyloid  $\beta$ -induced neuronal apoptosis,<sup>55</sup> and to anticipate the off-target pharmacology of phenotypic antimalarials.<sup>56</sup>

## **13.3 Case Study 1: The Impact of Chemical Probe Polypharmacology on PARP Drug Discovery**

### **13.3.1 The History of PARP Biology: From Probes to Drugs**

Poly(ADP-ribose) (PAR) is a natural branched polymer of adenosine diphosphoribose that serves as a post-translational modification of proteins.<sup>57</sup> Although the polymer was first discovered in 1963 and subsequent work led to the identification of the enzyme responsible for its synthesis—later renamed poly(ADP-ribose)polymerase-1 or PARP1—the function of PAR remained elusive. Despite their relatively weak affinity, the identification of the first PARP1 small-molecule inhibitors during the 1970s, particularly 3-aminobenzamide (3-AB, PARP1  $IC_{50}$  = 30  $\mu$ M) (Figure 13.2), demonstrated that PARP1 participated in cellular recovery from DNA damage and to validate PARP1 as a promising cancer target.<sup>58</sup>

During the 1990s, bicyclic and tricyclic lactams were identified as more potent PARP1 inhibitors and the crystal structure of PARP1 was solved, enabling the refinement of the benzamide PARP1 pharmacophore that would characterise all future PARP inhibitors.<sup>59</sup> Some of these lactams, including PJ34 (PARP1  $IC_{50}$  = 0.02  $\mu$ M) (Figure 13.2), became widely used PARP1 chemical tools due to their higher potency and solubility as compared to 3-AB. Moreover, this knowledge base enabled the launch of the first PARP1 drug discovery programs. Initially, PARP1 drug discovery aimed at identifying drugs that could potentiate chemotherapy or radiation.<sup>59</sup> Given the reductionist view that dominated science at that time, the synthesis of the PAR polymer was attributed exclusively to a single protein, PARP1, and all chemical tools were believed to be completely selective for PARP1, although available evidence was very limited. However, the identification of PARP residual activity in PARP-1 knockout cells prompted the identification of a new poly(ADP-ribose) polymerase enzyme, named PARP2.<sup>60</sup> Moreover, evidence started suggesting that PARP1 had a more complex role than previously anticipated and participated in a wide range of biological processes, including cellular differentiation and chromatin remodelling.<sup>60</sup> The chemical biology of PARPs was more complex than previously anticipated.

The human genome project enabled the advancement in the study of PARPs even further. When the human genome was sequenced, it became quickly apparent that PARPs were a family formed of 18 members sharing a very similar catalytic domain (17 PARPs and one PAR glycohydrolase or PARG).<sup>61</sup> However, the efforts to characterise the biological functions of this enzyme family continued to be highly biased towards PARP1, PARP2 and PARP5—further subdivided into two tankyrases, namely TNKS1 and TNKS2. Biological tools would start to be used to investigate the functions of this

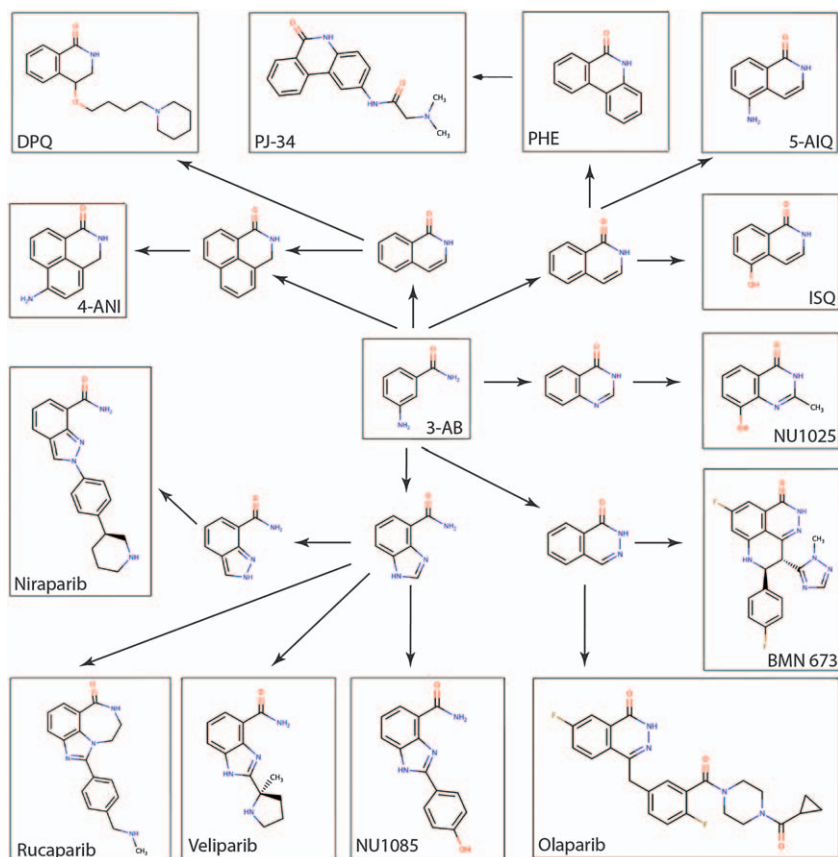


**Figure 13.2** Evolution of PARP chemical tools and clinical candidates (in boxes) from the classical benzamide pharmacophore of the early PARP chemical tool 3-AB (in the centre) highlighted in bold.

enzyme family further but chemical tools would continue to be used as well, such as the use of 3-AB to identify that PARP1 played a key role in Tumour Necrosis Factor (TNF)-mediated necroptosis.<sup>62</sup> It was not until 2004, when the crystal structure of the murine catalytic domain of PARP2 was published, that the possible lack of selectivity of PARP inhibitors started being considered.<sup>63</sup> It would soon become apparent that all existing PARP inhibitors were unselective between PARP1 and PARP2, but their selectivity against other members of the PARP family would remain unknown.<sup>59</sup> Despite the unknown selectivity of PARP inhibitors across the PARP family, a major discovery would boost their clinical application.

In 2005, two independent studies published in *Nature* found that PARP inhibitors were selectively killing cancer cells diploid of BRCA activity, thus identifying the first cancer application of synthetic lethality.<sup>64,65</sup> The underlying principle was that the redundancy of DNA repair mechanisms in

human cells was compromised in BRCA-defective cancer cells that relied only on one DNA repair mechanism. When this mechanism was abrogated with PARP inhibitors, cancer cells died due to their inability to repair their DNA while normal cells could repair their defects through BRCA-related pathways.<sup>66</sup> This breakthrough was quickly translated to the clinic in a landmark Phase I clinical trial that would boost the clinical development of PARP inhibitors as single agents in BRCA-mutated cancers.<sup>67</sup> However, the selectivity of PARP inhibitors remained unknown and some reports started to challenge the exact molecular mechanism driving the effect of PARP inhibitors.<sup>65</sup>

It was not until 2012 that the selectivity of a wide selection of PARP inhibitors was evaluated against 17 members of the PARP family, uncovering widespread polypharmacology.<sup>68</sup> Most PARP inhibitors were potently inhibiting PARP1-4 and some showed weaker affinity for other PARP family members. The wide promiscuity of PARP inhibitors challenged many of the functions of PARP1 attributed using exclusively chemical tools. Moreover, since 2010 several studies started to report differential effects among PARP inhibitors. First, several studies suggested that the widely used PARP chemical tool PJ34 had 'PARP1-independent' effects on cell cycle arrest, centrosome de-clustering and in *Helicobacter pylori* preneoplasia.<sup>69–71</sup> While some of these effects could be attributed to other members of the PARP family, its similar target profile against members of the PARP family to other PARP inhibitors also used in these studies suggested that PJ34 could be inhibiting other proteins beyond the PARP family. Later on, further differential effects were reported between PARP drug candidates including different cancer cell line profiling activities, pharmacogenomics biomarkers, cell cycle arrest, siRNA sensitivity and their capacity to trap PARP1 at the DNA damage site.<sup>72–76</sup> Despite these differences were poorly understood, olaparib would be the first PARP inhibitor to be approved by the FDA in December 2014 for BRCA-mutated ovarian cancer. Rucaparib recently received FDA approval also for BRCA-mutated ovarian cancer in December 2016 while several other PARP inhibitors are currently in late-stage clinical trials for this and other cancer indications.<sup>77</sup> Accordingly, the clarification of the complete mechanism-of-action of these drug candidates is increasingly important to take the most out of these drugs in personalised and precision oncology.<sup>20</sup>

### 13.3.2 PJ34: A PARP Chemical Tool Binding to PIM Kinases

Given the reported PARP1-independent effects of the PJ34 chemical tool, its differential effects as compared to other PARP inhibitors, and its wide use in more than 150 publications since its discovery in 2001, we decided to use the computational target profiling method implemented in CT-link to investigate the target profile of PJ34 further.<sup>78</sup> In addition to recovering many of the already known PARP targets, putative novel affinities to PIM1 and PIM2 serine/threonine kinases were predicted based on the similarity between

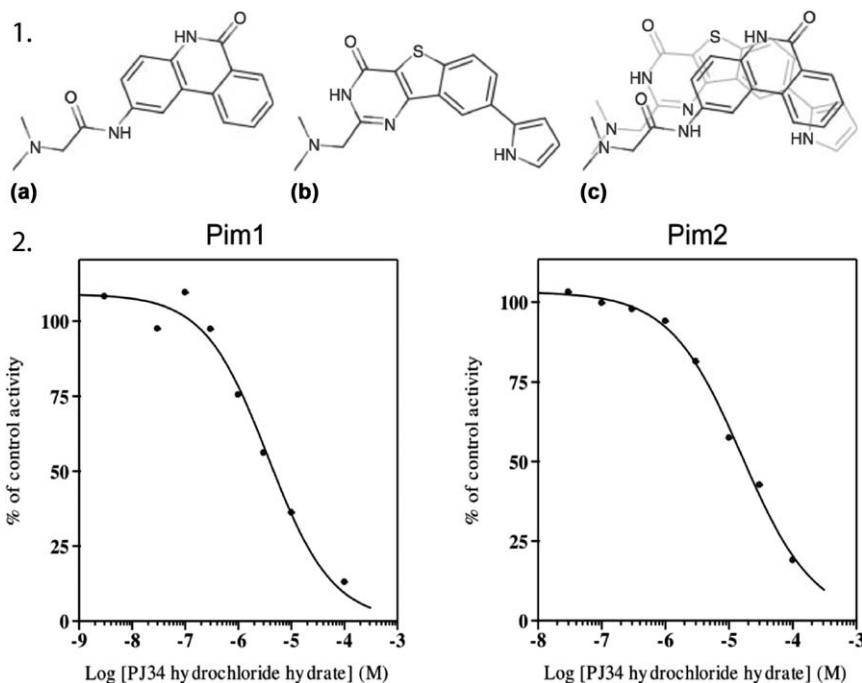


**Figure 13.3** Identification of PIM kinase off-targets among the PARP chemical tool PJ34. 1. Chemical structures of PJ34 (a), the PIM kinase inhibitor CHEMBL572783 (b) and their superposition (c). 2. Dose-response curves of PJ34 against PIM1 (squares) and PIM2 (triangles) kinases using an *in vitro* biochemical assay.

Reprinted from A. A. Antolin, X. Jalencas, J. Yelamos and J. Mestres, *ACS Chem. Biol.*, 2012, 7, 1962. Copyright (2012) American Chemical Society.

PJ34 and the PIM kinase inhibitor CHEMBL572783 (Figure 13.3). Follow-up *in vitro* testing confirmed that PJ34 is a *bona fide* PIM kinase inhibitor with an IC<sub>50</sub> of 3.7 and 16 μM for PIM1 and PIM2, respectively (Figure 13.3). Therefore, using computational methods we were able to extend the target profile of PJ34 beyond the PARP protein family.

It is important to stress that PIM kinases had never been suspected off-targets of PARP inhibitors, as they do not even bind to NADH. Originally identified as oncogenes, PIM kinases are overexpressed in a wide-range of cancers and they are involved in several biological processes including regulation of cell cycle progression, survival signalling and protein transcription, all of them being overlapping functions with PARPs.<sup>78</sup> Therefore, the possibility that the use of PJ34 as a PARP chemical tool could have confounded PARP biology ought to be considered, as some of the functions attributed to PARP1 through PJ34 could have been achieved *via* PIM kinase modulation. However, the difference in the affinity of PJ34 for PARP1 and PIM1 was high enough to ensure selective modulation of PARPs if

PJ34 had been used at low concentrations (PJ34 PARP1  $IC_{50}$  = 0.02  $\mu$ M; PIM1  $IC_{50}$  = 3.7  $\mu$ M). Therefore, we revised the concentrations of PJ34 being used in chemical biology publications indexed in PubMed over the years 2010 and 2011. Strikingly, over 60% of the publications used PJ34 at concentrations higher than 5  $\mu$ M, unaware that at this concentration six different targets could be inhibited, PIM1 among them.<sup>78</sup> Although *in vitro* activities cannot be extrapolated to a cellular context, where several factors including membrane penetration and subcellular localisation play important roles, without careful control of intracellular concentration it was difficult to rule out confounding effects due to PIM kinase inhibition at concentrations of PJ34 higher than 1  $\mu$ M. Interestingly, several of the 'PARP1-independent' effects of PJ34 that had been reported could be ascribed to the newly identified off-targets. The p21-dependent G2/M cell cycle arrest observed using PJ34 at 10  $\mu$ M could be ascribed to PIM1 inhibition since this kinase is known to regulate the cell cycle *via* cdc25c modulation at the G2/M checkpoint.<sup>78</sup> Moreover, PIM1 kinase phosphorylates the mitotic apparatus NuMA and could potentially be responsible of the unique centrosome de-clustering effects observed using 20–30  $\mu$ M PJ34.<sup>78</sup> Therefore, the confusion of functions attributed to PARP1 *via* the use of biologically promiscuous chemical tools was a real possibility.

In 2013, the first function wrongly attributed to PARP1 using a promiscuous tool compound was reported.<sup>79</sup> Since 2001, PARP1 was thought to participate in the TNF-mediated necroptosis pathway, a function attributed to PARP1 using the chemical tool 3-AB. However, more than ten years later Adam and co-workers clarified that PARP and TNF represent two distinct necroptosis pathways and that only PJ34, but not the rest of PARP inhibitors used, was able to reduce TNF-mediated necroptosis.<sup>79</sup> Importantly, our report of the off-target effects of PJ34 contributed to the clarification of a function wrongly attributed to PARP1 for more than a decade.

To our knowledge, this was the first application of a computational method to predict polypharmacology in chemical biology to identify new confounding targets of a chemical probe. Our results, clearly illustrated the cost-effectiveness of computational methods at de-risking the practise of chemical biology and that profiling chemical probes only across members of their protein family was not sufficient to guarantee their selectivity, as proteins from totally different families can also be potential off-targets. Today, new PARP-independent functions of PARP chemical tools continue being reported, questioning further those biological functions attributed to PARP1, such as its role in ultraviolet apoptosis, that make it increasingly important to comprehensively characterise the target profile of all chemical tools as we advance towards a more systems approach to chemical biology.<sup>80,81</sup> In summary, caution should be taken when using chemical tools to study the biological functions of proteins as their selectivity across the entire human proteome is always a big unknown. In this respect, computational methods to predict polypharmacology emerge as a cost-effective de-risking strategy in

chemical biology and should be widely applied to identify potentially confounding targets of chemical probes.<sup>78</sup>

### 13.3.3 Differential Off-target Kinase Pharmacology Between Clinical PARP Inhibitors

Following the identification of PIM kinase off-targets for PJ34, it was important to clarify the impact that this polypharmacology had on follow-up PARP inhibitors in late-stage clinical trials. Moreover, several differential effects between clinical PARP inhibitors had recently been reported, including differential DNA-break repair, differential anti-proliferative activities, differential PARP trapping effects at the DNA damage site and differential cancer cell line profiles.<sup>82</sup> Since it was impossible to explain all these differences among PARP drug candidates solely on the basis of their different affinity profile across members of the PARP protein family, we decided to explore the possibility that polypharmacology to distantly-related off-targets could be partly responsible for those differences.

Taking the off-target activity of PJ34 against PIM1 kinase as a starting point, we used an *in-silico* approach to identify kinases that could be potential off-targets of clinical PARP inhibitors. Based on a recently developed computational approach to organise proteins by ligand similarity, we identified 15 kinases that shared >60% of their active ligands with PIM1 kinase ( $IC_{50} < 10 \mu M$ ).<sup>83</sup> From these kinases, we selected the 11 kinases that could be tested at the contract research organisation Cerep ([www.cerep.fr](http://www.cerep.fr)). We added ABL and ALK as they had been reported to be differential pharmacogenomics biomarkers of some PARP inhibitors in a large-scale cancer cell line profiling.<sup>82</sup> Finally, we also included PIM1, CDK1 and PRKD2, as they had been already reported as weak off-targets of rucaparib.<sup>82</sup> Overall, the clinical PARP inhibitors olaparib, veliparib and rucaparib were tested *in vitro* across 16 kinases, namely PIM1, PIM2, STK17A, DYRK1A, CDK9, HIPK2, CK2, CDC7, AKT3, PRKCG, PRKCI, CDK1, PRKD2, ABL, ALK and PKA.

The experimental kinase profiling uncovered that clinical PARP inhibitors, albeit showing a similar profile for members of the PARP family, had a totally distinct target profile across kinases (Figure 13.4). Rucaparib, the most promiscuous of the PARP inhibitors tested, showed micromolar affinities for nine kinases, namely PIM1 (1.2  $\mu M$ ), DYRK1A (1.4  $\mu M$ ), CDK1 (1.4  $\mu M$ ), CDK9 (2.7  $\mu M$ ), HIPK2 (4.4  $\mu M$ ), PIM2 (7.7  $\mu M$ ), CK2 (7.8  $\mu M$ ), PRKD2 (9.7  $\mu M$ ), and ALK (18  $\mu M$ ). In contrast, olaparib did not inhibit any of the kinases tested. Finally, veliparib showed a polypharmacology in between olaparib and rucaparib, with off-target affinities only for PIM1 (17  $\mu M$ ) and CDK9 (8.2  $\mu M$ ). Interestingly, the higher the affinity for PIM1 the higher the number of kinases inhibited, as expected given the computational method used (Figure 13.4). Overall, these results clearly illustrate how the unknown polypharmacology form a chemical probe can differentially translate to drug candidates inspired on this chemical probe, compromising the link between



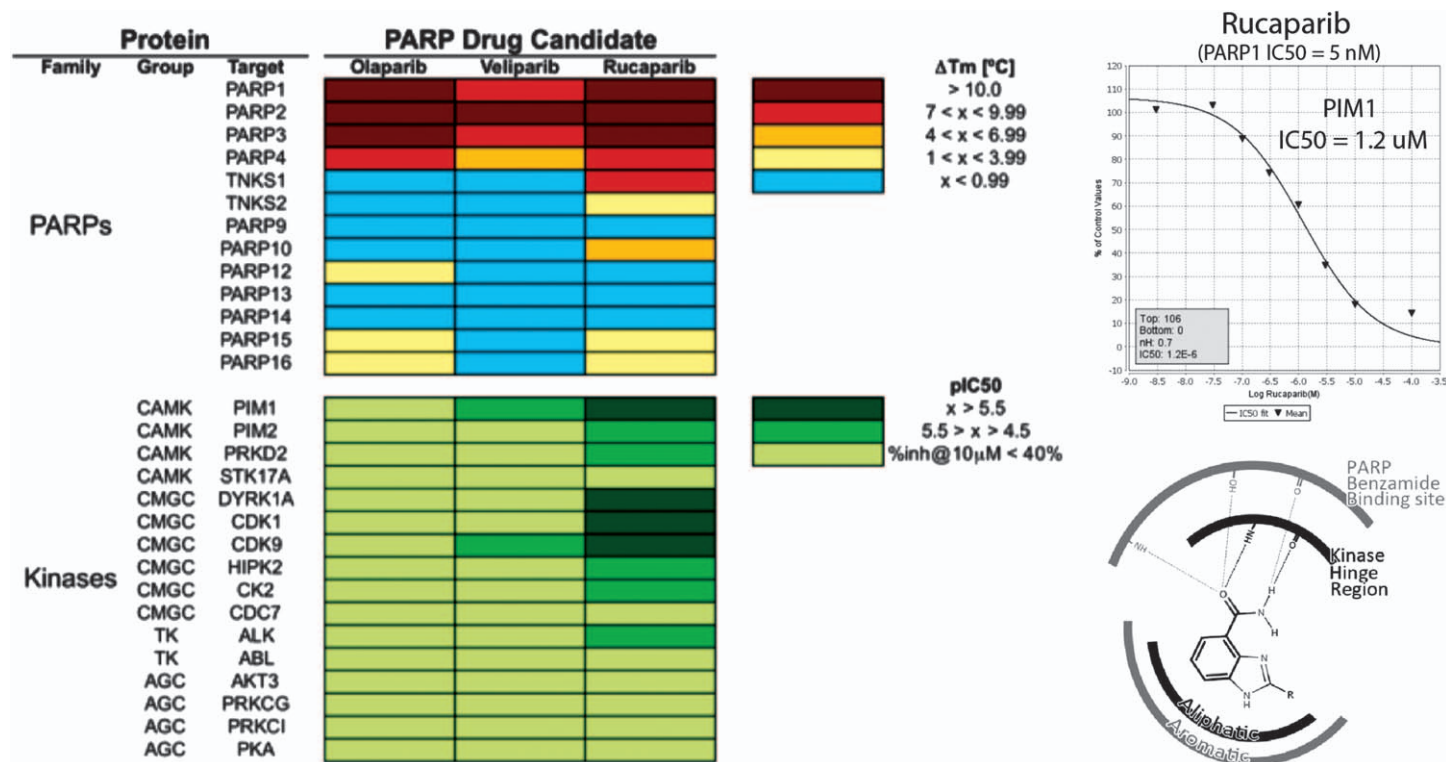


**Figure 13.4** Differential kinase polypharmacology among PARP inhibitors. 1. Pharmacological profile of olaparib, veliparib and rucaparib against members of the PARP and kinase families. 2. Dose-response of rucaparib against PIM1 kinase. 3. Schematic representation of the benzamide pharmacophore binding to the PARP and the kinase binding sites. Reprinted from A. A. Antolín and J. Mestres, *Oncotarget*, 2014, 5, 3023. Originally published in *Oncotarget* under a Creative Commons CC BY Licence (<https://creativecommons.org/licenses/by/3.0/>).



chemical biology and drug discovery and confounding the effects and the clinical development of these inhibitors.

Having confirmed that three clinical PARP inhibitors had a different kinase profile, we hypothesised that this could be a common phenomenon among PARP inhibitors. A recent report had identified that the PARP chemical tool Nu1085 also inhibited S6K1 kinase off-target ( $IC_{50} = 0.56 \mu M$ ).<sup>84</sup> The crystal structure of Nu1085 on S6K1 clearly showed that the benzamide that characterises all PARP inhibitors interacted with the kinase hinge region, a highly conserved structure in the ATP binding site of kinases. This interaction enabled us to speculate that all PARP inhibitors shared an inherent capacity to inhibit kinases and that they all had a unique kinase profile depending on their size and decoration as they grew from the hinge region towards the selectivity pockets of kinases (Figure 13.4). To further strengthen this hypothesis, we searched the database ChEMBL<sup>11</sup> for dual inhibitors of PARPs and kinases and we identified four dual inhibitors that had already been reported including CHEMBL539474 that inhibited PARP-1 ( $K_i = 0.698 \mu M$ ), PLK1 ( $K_i = 0.079 \mu M$ ) and GSK3A ( $K_i = 0.2 \mu M$ ) with sub-micromolar affinities. In light of these results, the off-target kinase pharmacology of PARP inhibitors was likely to expand as more PARP inhibitors were tested on kinases.

Importantly, the differential profile of clinical PARP inhibitors was able to explain some of the different effects reported between these inhibitors. For instance, the unique capacity of rucaparib to inhibit STAT3 phosphorylation at  $5 \mu M$  could be ascribed to its micromolar affinity for DYRK1A or CDK1, both known to phosphorylate STAT3 directly.<sup>82</sup> Similarly, CDK1 and PIM1 both regulate G2/M cell cycle arrest and thus explained the higher G2/M cell cycle arrest observed upon rucaparib treatment. Finally, the weak off-target inhibition of ALK was, to our knowledge, the first genomic biomarker of a drug being identified as an off-target of this drug. Overall, our results suggested that confounding effects due to kinase inhibition ought to be considered when the clinical PARP inhibitors rucaparib and veliparib were used at micromolar concentrations in preclinical models.

The newly identified kinase off-targets of clinical PARP inhibitors could also offer a new clinical perspective of these agents. First, these new off-targets could be used to repurpose these drugs to new cancer indications, as previously done for a few other multi-target cancer drugs like imatinib.<sup>20</sup> Second, these new targets could be potential biomarkers, to better identify the patients more likely to respond in the cancer subtypes that were already being investigated. Finally, due to the reported synergy between PARP and kinase inhibitors, including PI3K, CDK1 and MET, clinical PARP inhibitors should not be considered as equivalent anymore, particularly regarding drug combinations.<sup>82</sup> The newly identified affinity of rucaparib and veliparib for kinases may affect their capacity to synergise with specific kinase inhibitors and thus caution should be taken when translating clinical or preclinical drug combinations from one PARP inhibitor to another. However, the differential kinase polypharmacology would only be relevant if the unbound drug concentration at the tumour site was above the  $IC_{50}$ s.<sup>82</sup> Unfortunately, the intra-tumour

concentration of these drugs remains unknown, but their low micromolar peak plasma concentrations at maximum therapeutic dose makes it plausible that some of these off-targets are actually being modulated in the clinic, at least in circulating tumour cells exposed to higher drug concentrations.

In summary, the clinical PARP inhibitors olaparib, veliparib and rucaparib had a distinct off-target profile against kinases that made them essentially unique and it may be translated to other PARP inhibitors. Therefore, a comprehensive understanding of the mechanism of action of PARP inhibitors and a broader target profile against members of the kinase family was urgently needed to ensure maximum clinical exploitation of this exciting new drug class for the ultimate benefit of cancer patients.<sup>82</sup>

### 13.4 Case Study 2: Distant Off-target Pharmacology among MLP Chemical Probes

The demonstration of the profound impact that the undetected polypharmacology of a chemical tool can have for chemical biology and follow-up drug discovery observed in the PARP case study prompted us to clarify whether polypharmacology was a common phenomenon among chemical tools and thus warranted higher attention during chemical probe development. To this end, we used CT-link to predict the target profile of the NIH MLP collection of 133 target-based chemical probes.<sup>85</sup> We focused on the 35 probes that were commercially available from a selected list of priority chemical suppliers. Biologically relevant affinities to new targets phylogenetically unrelated to their intended primary target were detected for 30 out of the 35 probes (86%). Next, we selected the supplier from which we could obtain the largest number of probes and selected eight probes for further experimental validation of the computational predictions. *In vitro* results confirmed that four of these probes, ML006, ML123, ML141 and ML204, inhibited a non-obvious off-target with biologically relevant affinities that could compromise its use as selective probe for its intended primary target (Figure 13.5).<sup>85</sup> Overall, these results illustrated the widespread polypharmacology of chemical probes and the cost-effectiveness of *in silico* target profiling to de-risk chemical probe utilisation.

ML141 was developed as an inhibitor of Cdc42 GTPase with an *in vitro* IC<sub>50</sub> of 200 nM and selective over other members of the Rho GTPase family, although an unexplained inhibition of EGF-stimulated Rac1 was also identified and ascribed to putative off-target effects.<sup>85</sup> We predicted (*in vitro* confirmed) that ML141 inhibited carbonic anhydrase II (CA2) with an IC<sub>50</sub> of 530 nM (Figure 13.5). Given the low selectivity between both targets, the protective anticancer effect ascribed to CA2 and its expression being regulated by Ras GTPases, we warned on the use of ML141 as a Cdc42 chemical tool and we recommended that the involvement of CA2 on the biological process being studied should always be carefully considered when using ML141 as a tool compound.<sup>85</sup>

| chemical probe | canonical target | IC <sub>50</sub> /EC <sub>50</sub> (μM) | distantly related off-target | predicted IC <sub>50</sub> (μM) | IC <sub>50</sub> /EC <sub>50</sub> (μM) | fold-selectivity (off-target/target) |
|----------------|------------------|-----------------------------------------|------------------------------|---------------------------------|-----------------------------------------|--------------------------------------|
| ML006          | S1P3             | 6.63                                    | mTOR kinase                  | 1.6                             | 7.4                                     | 1.1                                  |
| ML123          | TRPML3/2         | 0.873                                   | Sigma-1 receptor             | 0.008                           | 31                                      | 35.5                                 |
| ML141          | Cdc42 GTPase     | 0.2                                     | Carbonic Anhydrase 2         | 0.08                            | 0.53                                    | 2.7                                  |
| ML204          | TRPC4/5          | 0.96                                    | AChE                         | 0.5                             | 0.84                                    | 0.9                                  |

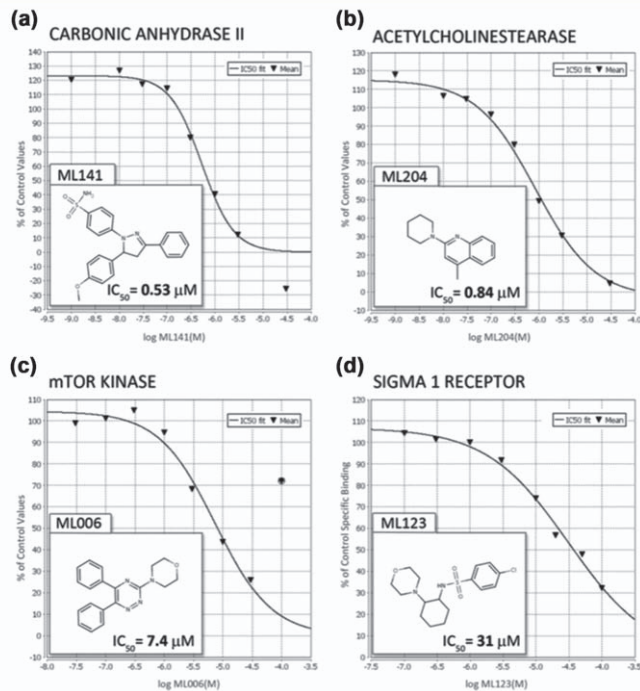


**Figure 13.5** Affinity values for the primary targets and off-targets identified (top). Dose-response curves showing the interaction between the probes and the newly identified off-targets (bottom).<sup>8</sup>  
 Reprinted from A. A. Antolín and J. Mestres, *ACS Chem. Biol.*, 2015, **10**(2), 395–400, Copyright (2015) American Chemical Society.

ML123 was developed as a TRPML2 and TRPML3 chemical tool with  $EC_{50}$  values of 0.9  $\mu$ M and 1.4  $\mu$ M, respectively, and demonstrated to be selective over other ion channels.<sup>85</sup> We predicted (and experimentally confirmed) that ML123 also inhibited *in vitro* the human sigma-1 receptor with an  $IC_{50}$  of 31  $\mu$ M (Figure 13.5). Given the difference in affinity between the primary target and the newly identified off-target, we acknowledged that in this case ML123 could continue being used as a TRPML2/3 chemical tool, but attention should be paid when using it at concentrations higher than 10  $\mu$ M, as sigma-1 inhibition could start to confound the observed biological effects.

ML006 was identified as a Shingosine 1-phosphate receptor 3 (S1P3) chemical probe with an  $EC_{50}$  of 6.6  $\mu$ M. We validated that ML006 is also an inhibitor of mTOR kinase with an  $IC_{50}$  of 7.4  $\mu$ M (Figure 13.5).<sup>85</sup> Given the low selectivity between both targets and their shared signalling pathways, we warned on the use of this tool compound to probe for S1P3 activity and questioned the development of probes with weak micromolar affinity. However, in this case ML249 was developed after ML006 as an improved S1P3 probe with higher *in vitro* affinity, illustrating the importance of developing several chemical probes for the same target.<sup>85</sup>

Finally, ML204 was discovered as the first TRPC4/C5 chemical probe with an  $EC_{50}$  of 0.96  $\mu$ M and shown to be selective over other ion channels. *In silico* profiling and *in vitro* validation enabled us to identify that ML204 also inhibited acetylcholinesterase (AChE) with an  $IC_{50}$  of 0.84  $\mu$ M (Figure 13.5). Although it was not possible to formally compare the cellular  $EC_{50}$  for the primary target with the *in vitro*  $IC_{50}$  for the newly identified off-target, they were sufficiently similar to anticipate selectivity issues.<sup>85</sup> Moreover, the fact that TRPC4 transduced smooth muscle contraction evoked by muscarinic acetylcholine receptor activation while AChE was transported outside cells to hydrolyse acetylcholine prompted us to warn on the use of MLP204 to probe for TRPC4/5 as its effects could be confounded due to AChE inhibition. In the light of these results, we questioned the use of ML204 alone as a TRPC4/5 chemical tool without careful control of AChE implications on the biological system under study.<sup>85</sup>

In summary, the results demonstrated that chemical probes show widespread polypharmacology including targets distantly related to the primary target of the chemical probe. Accordingly, distant polypharmacology should carefully be considered during chemical probe development and use. *In silico* target profiling emerges as an efficient de-risking strategy in chemical biology and the utility of multi-target probes must be recognised in the context of multi-target drug discovery and a much needed embracement of a more comprehensive and systems approach to chemical biology.<sup>85</sup>

## 13.5 Conclusions and Outlook

In this chapter, we have shown that chemical probe polypharmacology has a profound impact in chemical biology and follow-up drug discovery. The use of *in silico* target profiling enabled us to uncover that the PARP chemical probe PJ34 had micromolar affinity for PIM kinase off-targets and we showed

how this unknown polypharmacology had already potentially confounded a biological function wrongly attributed to PARP1. Moreover, this unknown polypharmacology was translated into differential polypharmacology among clinical PARP inhibitors. The new differential pre-clinical effects of PARP inhibitors that had been reported were ascribed to some of the newly identified off-targets and may have implications for the clinical exploitation of this new drug class. Finally, we have also shown that distant polypharmacology is common among chemical probes using the NIH MLP chemical probe collection as a case study.

Overall, the results presented here have strong implications for the future practise of chemical biology. First, the use of computational methods to predict the target profile of chemical probes should be widely used in chemical probe development as they represent a cost-effective method for identifying new potentially confounding targets of chemical tools that de-risk the practise of chemical biology. Second, chemical probes should always be used considering that their selectivity across the entire human proteome is always a big unknown, and thus the involvement of the target of interest in the biological process under study should always be carefully controlled using other probes and biological controls, particularly when probes are used as pathfinder inhibitors in new drug discovery projects. The failure to identify the off-targets of tool compounds in which new drug discovery projects are inspired may have profound consequences for the development of the new drug candidates. Finally, we have also warned about the disconnection that currently exists between drug discovery, where polypharmacology is believed to be common and potentially beneficial, and chemical biology, where polypharmacology is mostly ignored or believed to be rare and detrimental. It is paramount that we embrace a more systematic approach to studying both chemical biology and drug discovery to fully untangle the complex interactions between chemical molecules and proteins to both boost our understanding of biology using chemical tools and to develop safer and more effective small-molecule therapeutics.

## Acknowledgements

Albert A. Antolin is funded by the People Programme (Marie Curie Actions) of the 7th Framework Programme of the European Union (FP7/2007-2013) under REA grant agreement no. 600388 (TECNIOspring programme), and from the Agency of Business Competitiveness of the Government of Catalonia, ACCIO. Jordi Mestres is funded by the Spanish Ministerio de Economía y Competitividad (project BIO2014-54404-R).

## References

1. J. Drews, *Science*, 2000, **287**, 1960.
2. A. Maehle, C. Prüll and R. F. Halliwell, *Nat. Rev. Drug Discovery*, 2002, **1**, 637.

3. D. Brown, *Drug Discovery Today*, 2007, **12**, 1007.
4. J. Mestres, E. Gregori-Puigjane, S. Valverde and R. V. Sole, *Mol. Biosyst.*, 2009, **5**, 1051.
5. H. Kitano, *Science*, 2002, **259**, 1662.
6. B. L. Roth, D. J. Sheffler and W. K. Kroeze, *Nat. Rev. Drug Discovery*, 2004, **3**, 353.
7. H. Kubinyi, *Nat. Rev. Drug Discovery*, 2003, **2**, 5.
8. T. Hampton, *JAMA*, 2004, **292**, 419.
9. S. Frantz, *Nature*, 2007, **44**, 32.
10. G. V. Paolini, R. H. B. Shapland, W. P. van Hoorn, J. S. Mason and A. L. Hopkins, *Nat. Biotechnol.*, 2006, **24**, 805.
11. A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani and J. P. Overington, *Nucleic Acids Res.*, 2012, **40**, 1100.
12. M. A. Yildirim, K. I. Goh, M. E. Cusick, A. L. Barabási and M. Vidal, *Nat. Biotechnol.*, 2007, **25**, 1119.
13. J. Mestres, E. Gregori-Puigjané, S. Valverde and R. V. Solé, *Nat. Biotechnol.*, 2008, **26**, 983.
14. I. Vogt and J. Mestres, *Mol. Inf.*, 2010, **29**, 10.
15. X. Jalencas and J. Mestres, *Medchemcomm*, 2013, **4**, 80.
16. M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijer, R. C. Matos, T. B. Tran, R. Whaley, R. A. Glennon, J. Hert, K. L. H. Thomas, D. D. Edwards, B. K. Shoichet and B. L. Roth, *Nature*, 2009, **462**, 175.
17. T. I. Oprea and J. Mestres, *AAPS J.*, 2012, **14**, 759.
18. R. Garcia-Serna, D. Vidal, N. Remez and J. Mestres, *Chem. Res. Toxicol.*, 2015, **28**, 1875.
19. N. Remez, R. Garcia-serna, D. Vidal and J. Mestres, *Chem. Res. Toxicol.*, 2016, **28**, 637.
20. A. A. Antolin, P. Workman, J. Mestres and B. Al-Lazikani, *Curr. Pharm. Des.*, 2016, **22**, 6235.
21. S. L. Schreiber, *Nat. Chem. Biol.*, 2005, **1**, 64.
22. P. Workman and I. Collins, *Chem. Biol.*, 2010, **17**, 561.
23. S. Knapp, P. Arruda, J. Blagg, S. Burley, D. H. Drewry, A. Edwards, D. Fabbro, P. Gillespie, N. S. Gray, B. Kuster, K. E. Lackey, P. Mazzafera, N. C. O. Tomkinson, T. M. Willson, P. Workman and W. J. Zuercher, *Nat. Chem. Biol.*, 2013, **9**, 3.
24. Editorial, *Nat. Chem. Biol.*, 2013, **9**, 193.
25. M. E. Bunnage, E. L. P. Chekler and L. H. Jones, *Nat. Chem. Biol.*, 2013, **9**, 195.
26. A. D. Thompson, L. N. Makley, K. Mcmenimen and J. E. Gestwicki, *ACS Chem. Biol.*, 2012, **7**, 791.
27. C. H. Arrowsmith, J. E. Audia, C. Austin, J. Baell, J. Bennett, J. Blagg, C. Bountra, P. E. Brennan, P. J. Brown, M. E. Bunnage, C. Buser-Doepner, R. M. Campbell, A. J. Carter, P. Cohen, R. A. Copeland, B. Cravatt, J. L. Dahlin, D. Dhanak, A. M. Edwards, M. Frederiksen, S. V. Frye,



- N. Gray, C. E. Grimshaw, D. Hepworth, T. Howe, K. V. M. Huber, J. Jin, S. Knapp, J. D. Kotz, R. G. Kruger, D. Lowe, M. M. Mader, B. Marsden, A. Mueller-Fahrnow, S. Müller, R. C. O'Hagan, J. P. Overington, D. R. Owen, S. H. Rosenberg, R. Ross, B. Roth, M. Schapira, S. L. Schreiber, B. Shoichet, M. Sundström, G. Superti-Furga, J. Taunton, L. Toledo-Sherman, C. Walpole, M. A. Walters, T. M. Willson, P. Workman, R. N. Young and W. J. Zuercher, *Nat. Chem. Biol.*, 2015, **11**, 536.
28. C. P. Austin, L. S. Brady, T. R. Insel and F. S. Collins, *Science*, 2004, **306**, 1138.
29. P. Filippakopoulos and S. Knapp, *Nat. Rev. Drug Discovery*, 2014, **13**, 337.
30. M. E. Bunnage, E. L. P. Chekler and L. H. Jones, *Nat. Chem. Biol.*, 2013, **9**, 195–199.
31. T. I. Oprea, C. G. Bologa, S. Boyer, R. F. Curpan, R. C. Glen, A. L. Hopkins, C. A. Lipinski, G. R. Marshall, Y. C. Martin, L. Ostropovici-Halip, G. Rishton, O. Ursu, R. J. Vaz, C. Waller, H. Waldmann and L. A. Sklar, *Nat. Chem. Biol.*, 2009, **5**, 441–447.
32. S. V. Frye, *Nat. Chem. Biol.*, 2010, **6**, 159.
33. R. Garcia-serna and J. Mestres, *Drug Discovery Today*, 2011, **16**, 99.
34. R. T. Peterson, J. Starr and C. Tracey, *Nat. Chem. Biol.*, 2008, **4**, 635.
35. W. L. Chen, E. Mdl, C. Ramon and S. Ramon, *J. Chem. Inf. Model.*, 2006, **46**, 2230.
36. A. Bender, *Nat. Chem. Biol.*, 2010, **6**, 309.
37. W. A. Catterall, J. Striessnig, T. P. Snutch and E. Perez-reyes, *Pharmacol. Rev.*, 2003, **55**, 579.
38. X. Chen, M. Liu and M. Gilson, *Comb. Chem. High Throughput Screening*, 2001, **4**, 719.
39. D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, L. Y. Geer, W. Helmberg, Y. Kapustin, D. L. Kenton, O. Khovayko, D. J. Lipman, T. L. Madden, D. R. Maglott, J. Ostell, K. D. Pruitt, G. D. Schuler, L. M. Schriml, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, T. O. Suzek, R. Tatusov, T. A. Tatusova, L. Wagner and E. Yaschenko, *Nucleic Acids Res.*, 2006, **34**, 3.
40. D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang and J. Woolsey, *Nucleic Acids Res.*, 2006, **34**, 668.
41. K. P. Seiler, G. A. George, M. P. Happ, N. E. Bodycombe, H. A. Carrinski, S. Norton, S. Brudz, J. P. Sullivan, J. Muhlich, M. Serrano, P. Ferraiolo, N. J. Tolliday, S. L. Schreiber and P. A. Clemons, *Nucleic Acids Res.*, 2008, **36**, 351.
42. J. E. Tym, C. Mitsopoulos, E. A. Coker, P. Razaz, A. C. Schierz, A. A. Antolin and B. Al-Lazikani, *Nucleic Acids Res.*, 2016, **44**, 938.
43. S. Ekins, J. Mestres and B. Testa, *Br. J. Pharmacol.*, 2007, **152**, 9.
44. X. Jalencas and J. Mestres, *J. Chem. Inf. Model.*, 2013, **53**, 279.
45. A. Koutsoukas, B. Simms, J. Kirchmair, P. J. Bond, A. V. Whitmore, S. Zimmer, M. P. Young, J. L. Jenkins, M. Glick, R. C. Glen and A. Bender, *J. Proteomics*, 2011, **74**, 2554.



46. A. Weber, A. Casini, A. Heine, D. Kuhn, C. T. Supuran, A. Scozzafava and G. J. Klebe, *Med. Chem.*, 2004, **47**, 550.
47. J. Mestres, *Drug Discovery Today*, 2005, **10**, 1629.
48. J. Hert, P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby and A. Schuffenhauer, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1177.
49. D. Vidal, R. Garcia-Serna and J. Mestres, *Methods Mol. Biol.*, 2011, **672**, 489.
50. B. Flachner, Z. Lörincz, A. Carotti, O. Nicolotti, P. Kuchipudi, N. Remez, F. Sanz, J. Tóvári, M. J. Szabó, B. Bertók, S. Cseh, J. Mestres and G. Dormán, *PLoS One*, 2012, **7**, e35582.
51. A. Spitzmüller and J. Mestres, *J. Cheminf.*, 2012, **6**, 8003.
52. F. M. Areias, J. Brea, E. Gregori-Puigjané, M. E. Zaki, M. A. Carvalho, E. Domínguez, H. Gutiérrez-de-Terán, M. F. Proença, M. I. Loza and J. Mestres, *Bioorg. Med. Chem.*, 2010, **18**, 3043.
53. J. Mestres, S. A. Seifert and T. I. Oprea, *Clin. Pharmacol. Ther.*, 2013, **10**, 54.
54. A. A. Antolin, X. Jalencas, J. Yelamos and J. Mestres, *ACS Chem. Biol.*, 2012, **7**, 1962.
55. M. Montolio, E. Gregori-puigjane, D. Pineda, J. Mestres and P. Navarro, *J. Med. Chem.*, 2012, **55**, 9838.
56. W. C. Voorhis, *et al.*, *PLoS Pathog.*, 2016, **76**, 1.
57. B. A. Gibson and W. L. Kraus, *Nat. Rev. Mol. Cell Biol.*, 2012, **13**, 411.
58. B. Durkacz, O. Omidiji, D. A. Gray and S. Shall, *Nature*, 1980, **283**, 593.
59. D. V. Ferraris, *J. Med. Chem.*, 2010, **1**, 4561.
60. S. Shall and G. De Murcia, *Mutat. Res.*, 2000, **460**, 1.
61. C. Spenlehauer, G. De Murcia and J. Ame, *Bioessays*, 2004, **26**, 882.
62. M. Los, *et al.*, *Mol. Biol. Cell*, 2002, **13**, 978.
63. S. M. Roe, *et al.*, *Nucleic Acids Res.*, 2004, **32**, 456.
64. A. The, *et al.*, *Nature*, 2005, **7**, 913.
65. H. Farmer, N. McCabe and C. J. Lord, *et al.*, *Nature*, 2005, **239**, 236.
66. J. D. Iglehart and D. P. Silver, *N. Engl. J. Med.*, 2009, **361**, 189.
67. P. C. Fong, M. D. David, S. Boss, T. A. Yap, A. Tutt, P. Wu and M. Mergui-roelvink, *et al.*, *N. Engl. J. Med.*, 2009, **361**, 123.
68. P. Jones, *Nat. Biotechnol.*, 2012, **30**, 249.
69. D. L. Madison, D. Stauffer and J. R. Lundblad, *DNA Repair*, 2011, **10**, 1003.
70. A. Castiel, L. Visochek, L. Mittelman, F. Dantzer and S. A. Izraeli, *BMC Cancer*, 2011, **11**, 412.
71. I. M. Toller, M. Altmeyer, E. Kohler, M. O. Hottiger and A. Müller, *Cancer Res.*, 2010, **70**, 5912.
72. M. J. Garnett, *et al.*, *Nature*, 2012, **483**, 570.
73. J. Murai, *et al.*, *Mol. Cancer Ther.*, 2014, **13**, 433.
74. P. Jelinic and D. A. Levine, *Mol. Cancer Ther.*, 2014, **13**, 1645.
75. Y. Shen, *et al.*, *Cancer Res.*, 2013, **19**, 5003.
76. J. Murai, *et al.*, *Cancer Res.*, 2012, **72**, 5588.
77. A. Mullard, *Nat. Rev. Drug Discovery*, 2016, **15**, 523.

78. A. A. Antolin, X. Jalencas, J. Yelamos and J. Mestres, *ACS Chem. Biol.*, 2012, **7**, 1962.
79. J. Sosna, *et al.*, *Cell Mol. Life Sci.*, 2014, **71**, 331.
80. P. Lakatos, *et al.*, *Biochim. Biophys. Acta*, 2013, **1833**, 743.
81. P. Lakatos, C. Hegedűs, N. Salazar Ayestarán, A. Juarranz, K. E. Kövér, E. Szabó and L. Virág, *Mutat. Res.*, 2016, **790**, 31.
82. A. A. Antolín and J. Mestres, *Oncotarget*, 2014, **5**, 3023.
83. H. Lin, M. F. Sassano, B. L. Roth and B. K. Shoichet, *Nat. Methods*, 2013, **10**, 140.
84. S. Couty, I. Westwood, A. Kalusa and C. Cano, *Oncotarget*, 2013, **4**, 1647.
85. A. A. Antolin and J. Mestres, *ACS Chem. Biol.*, 2015, **10**, 395.

# *Subject Index*

- absorption, distribution,
  - metabolism, excretion and toxicity (ADMET) modelling, 282, 304
- accelerated molecular dynamics (aMD), 222
- activation free energy, 139–141
- adaptive biasing force method (ABF), 252
- adaptive steered molecular dynamics (ASMD), 223
- allosteric communication, modelling
  - in GPCRs, 205–212
  - in G proteins, 213–216
- allostery, 75, 77
- ambiguous interaction restraints (AIRs), 227
- AMOEBA force field, 250
- anomeric effect, 148
- Anton Supercomputer, 266
- Armadillo (ARM)-repeat proteins, 224
- atomistic computer simulations, 60
- atomistic force-field simulations
  - basic methodological description, 170–172
  - force-field refinements, 172–175
  - nucleic acids force-field studies, examples, 175–177
- basic data analysis, 8–11
  - membranes, 10
  - nucleic acids, 10
  - proteins, 8–9
  - small molecules, 10–11
- basic simulation techniques, 4–8
  - boundary conditions, 7
  - enhanced sampling techniques, 7–8
  - molecular dynamics (MD), 5–6
  - Monte Carlo (MC) simulations, 6–7
- bending modulus, 53
- big data, 81–83
- bimolecular Diels–Alder reaction, 90
- biological catalysts, modelling, 108–141
- biomolecular simulations, elastic properties, 50–54
- biomolecular simulations, techniques, 2–13
  - basic data analysis, 8–11
  - basic simulation techniques, 4–8
  - examples, 12–13
  - molecular mechanics and force fields, 2–4
  - software, 11–12
- BLAST (Basic Logical Alignment Search Tool), 14
- bovine pancreatic trypsin inhibitor, 58
- butyrylcholinesterase, 96–99
- calcium-mediated binding, 158
- carbohydrate–protein interactions, 145–159
  - modelling, 155–159
- carbohydrates, 145–147
  - computational methodologies for, 151–153
  - force fields for, 153–155

- carbon nanotubes, 235–236
- Car–Parrinello based methods, 109
- CASP/CAPRI experiment, 228
- catalytic proficiency, 88
- catechol *O*-methyltransferase (COMT), 127
- cell surface receptors, signalling, 264–268
- chaperone, 72
- chaperone–ligand interactions, 78
- charge effects, 46
- CHARMM force field, 45, 46
- chemical biology
  - reductionism limits and, 349–353
  - selectivity in, 351–353
- chemical tools
  - bioinformatics, 79–81
  - chemical biology, structure based approaches, 72–74
  - chemoinformatics, 79–81
  - designing, 69–84
  - discovery of, 81–83
  - protein functions, 79–81
  - structural dynamics, source, 74–78
  - structural information, 79–81
- coarse-grain (CG) approach, 43–44, 61, 177–184, 250
  - basic methodological description, 178–182
  - for predicting RNA structures, 182–184
- cocaine detoxification, 96–99
- communication propensity (CP), 76
- complex protein systems, computational studies of, 229–232
- compound optimisation, 27–29
- computational docking
  - protocols, 153
- computational enzymology, 108–141
- computational methods, 248–252
- computer-aided drug design (CADD), 284. *See also* computer-based drug design (CBDD)
- computer-based drug design (CBDD), 20–30
  - compound optimisation, 27–29
  - ligand based drug design (LBDD), 26
  - pharmacophores, 26–27
  - software and web based servers, 29–30
  - structure-based drug design (SBDD), 20–26
- configurational space sampling, 139–141
- Coulomb potentials, 4
- CRISPR/Cas9 system, 230
- dasatinib, 59
- deep networks, 81–83
- de novo* drug design, 291–293
- diffusion, by transporters, 259–264
- dimethyl sulfoxide, 255
- disaccharides, 148–149
- docking, 226–229
- drug discovery process, 282–284
  - molecular modelling contribution, 284–285
  - polypharmacology in, 349–351
  - quantum and molecular mechanics in, 285
- drug discovery targets
  - membrane proteins, simulating, 323–330
  - metalloenzyme catalysis, deciphering, 319–322
  - molecular modelling and simulations, 317–338
- drug polypharmacology, 349–351
- dynamic histogram analysis method (DHAM), 125–126
- elastic network model paired with normal mode analysis (ENM-NMA), 199
- elastic network models (ENMs), 230
- electron transfer reactions, 99–101
- electrostatic embedding, 111
- empirical valence bond (EVB), 109

- ensemble-averaged variational transition state theory with multi-dimensional tunnelling (EA-VTST/MT), 136–139
- enzyme-catalysed chemical reaction
  - QM/MM methods, 110–114
  - transition state and energy barrier, 109–110
- enzyme design, QM/MM
  - approaches, 94–101
  - electron transfer reactions, metalloproteins, 99–101
  - engineered
    - butyrylcholinesterase, cocaine detoxification, 96–99
  - methods description, 94–96
- enzyme–substrate complexes
  - molecular dynamics simulations, 114–115
  - structure and system setup, 114
- epoxide hydrolase, 321–322
- equations of motion (EOM), 248, 249
- exo-anomeric effect, 148–149
- fatty acid amide hydrolase (FAAH), 323–324
- fibroblast growth factor (FGF), 158
- force fields, 2–4
- fragment-based drug design (FBDD)
  - approach, 291, 292
- free energy perturbation (FEP)
  - theory, 127–132, 251
  - QM/MM simulations, 129–132
- free energy simulations, 122–136
  - free energy perturbation theory, 127–132
  - minimum free energy paths, 132–136
  - string method, 132–136
  - umbrella sampling method, 123–127
- genetic algorithms (GA), 22, 289
- Gillespie's algorithm, 239
- Glu177, 258
- Glu290, 262, 263
- glycobiology, computational chemistry tools in, 145–159
- glycosaminoglycans (GAGs), 150, 151
- glycosidic linkage, 148–149
  - conformations around, 149
- gold nanoparticles (AuNPs), 233–234
- G protein-coupled receptors (GPCRs), 198–216, 264, 265, 267, 268, 328–330
  - allosteric communication, modelling, 205–212
  - PSN-ENM analysis of, 210–212
  - PSN-MD analysis of, 206–210
- G proteins, 198–216
  - allosteric communication, modelling in, 213–216
  - PSN-MD analysis of, 214–216
- GROMACS tool, 54
- guanine nucleotide exchange factors (GEFs), 199
- Helfrich Hamiltonian (HH), 51
- heme peroxidase, 100
- HG3, 93
- high ambiguity driven biomolecular docking (HADDOCK), 227
- Hsp90 functions, 72, 74, 77
- hub proteins, 70
- human adenosine receptor, 328–330
- hybrid QM/MM
  - basic methodological description, 167–168
  - examples of use, 168–170
- hydrogen bond interactions, 155
- hydrolysis reactions, of cocaine, 97
- inside-out design protocol, 89–94
  - enzymes designed, 92–94
  - Kemp eliminases, 92–94
  - method description, 89–91
- intrinsically disordered proteins (IDPs), 223, 333–335
- ion channels, passive transport, 255–259

- ionic interactions, 158
- Irving–Kirkwood–Noll theory, 54
- KE70, 93
- Kemp elimination reaction, 92–94
- laccase oxidoreductase, 100
- lactate dehydrogenase, 331–333
- lead discovery process, 281–305
- Lennard–Jones (LJ) potential, 3, 4, 23, 41, 42
- LeuT, 260, 261, 263
- Library of Integrated Network-Based Cellular Signatures (LINCS), 83
- ligand-based drug design (LBDD), 26, 285, 286
- ligand-based molecular modelling, 296–305
  - molecular descriptors, 297–298
  - molecular fingerprints, 299
  - pharmacophore modelling, 299–300
  - QSAR model, 300–304
  - in silico ligand-based approaches, antitubercular agents, 304–305
  - similarity, calculating, 298–299
- ligand–protein binding, 214–216
- lipid–nanomaterial interaction, 60
- lipopolysaccharides (LPS), 232
- Markov State Model (MSM), 49
- MARTINI model, 4, 44
- Maxwell–Demod discrete MD (MDdMD), 213
- mechanical embedding, 111
- membrane enzymes, 323–324
- membrane proteins, 323–330
  - GPCR, 328–330
  - human adenosine receptor, neuronal-like membrane, 328–330
  - ion channels, Kv11.1 channel, 324–328
  - membrane enzymes, FAAH, 323–324
- membrane transport, novel insights, 247–269
- mesoscopic models, 184–188
  - basic methodological description, 185–186
  - chromosome simulations, 187–188
  - nucleosome fibre simulations, 186–187
- metalloenzyme catalysis, deciphering, 319–322
  - epoxide hydrolase, 321–322
  - ribonuclease H (RNase H), 319–321
- metalloproteins, 99–101
- Metropolis–Hastings test, 6
- minimum free energy paths, 132–136
- MLP chemical probes, off-target pharmacology, 363–365
- molecular descriptors, 81, 297–298
- molecular dynamics, 5–6
  - advanced force fields, 45–46
  - basics of, 40–46
  - biomolecular simulations, elastic properties, 50–54
  - biomolecular simulations, force fields, 41–44
  - enhanced configurational sampling, 47–49
  - extracting information from, 46–54
  - free energy difference between states, 47
  - multiscale modelling, 44–45
  - NMR and, 55–57
  - phase problem, 57–58
  - simulating rare events, 49–50
  - simulation vs. experiment, 54–61
- molecular fingerprints, 299
- molecular mechanics, 2–4
- molecular modelling application, 281–305
  - basics, 1–30
  - structure- and ligand-based, 285–286

- molecular recognition processes
  - computational modelling of, 221–240
  - homology modelling, 224–225
  - proteins dynamics modelling, 221–223
  - three-dimensional structure prediction, 224–225
- monosaccharides, 147–148
- mutation effects, 214–216
- M-ZDOCK, 228
- nanomachines, 236–237
- nanostructures, computational modelling, 232–237
  - carbon nanotubes, 235–236
  - gold nanoparticles (AuNPs), 233–234
  - nanomachines, 236–237
  - nanowires, 234–235
- nanowires, 234–235
- new drugs discovery, 81–83
- new molecular entities (NMEs), 69, 282
- N*-glycans, 150
- 5-nitrobenzo[*d*]isoxazole, 92
- 6-nitrobenzotriazole, 94
- nodulation (Nod) factors, 151
- nucleic acids molecular modelling, 165–189
  - atomistic force-field simulations, 170–177
  - coarse-grain approach, 177–184
  - hybrid QM/MM, 167–170
  - mesoscopic models, 184–188
  - QM methods, 166–167
- oligosaccharides, 149–150
- parmbc0*, 172, 173, 175
- PARP drug discovery
  - chemical probe
    - polypharmacology, 355–363
  - chemical tool binding, PIM kinases, 357–360
  - clinical inhibitors, off-target kinase pharmacology, 360–363
  - history of, 355–357
  - PJ34, 357–360
- particle-based CG (pbCG) models, 177
- particle mesh Ewald (PME) method, 4
- passive transport, ion channels, 255–259
- PatchDock algorithm, 228
- Pauling's paradigm, 93
- pharmaceutical crisis, 281–282
- pharmacophores, 26–27
  - defined, 299
  - modelling, 299–300
- polarised embedding, 112
- polypharmacology gap, 349–366
- poly(ADP-ribose)polymerase-1 (PARP1), 355, 366
- polysaccharides, 150–151
- potential energy methods, 115–122
  - analysis, 118–122
  - electrostatic contributions analysis, 118–119
  - reaction path calculation, 115–117
  - steric effects analysis, 119–122
  - transition state localisation, 117–118
- principal component analysis (PCA), 9
- protein connectivity maps, 71
- protein data bank (PDB), 91, 224
- protein flexibility, 289
- protein function, computational design of, 87–102
- protein landscape exploration (PELE), 101
- protein myeloid differentiation factor 2 (MD-2), 232
- protein–protein interactions (PPIs), 291
  - docking, 226–229
  - modelling, 225–226
  - prediction of, 226–229



- protein-protein recognition
  - process, 78
- proteins dynamics modelling, 221–223
- protein structure graph (PSG), 201
  - building, 203–204
- protein structure network (PSN)
  - analysis, 198–216
- protein structure prediction, 13–19
  - comparative modelling
    - approaches, 15–17
  - function prediction, 18
  - modelled structure quality,
    - analysing, 18–19
  - secondary structure prediction, 13–15
  - sequence alignment, 13–15
  - software and web based servers, 19
- PSN-ENM method, 201
- PSN-MD workflow, 201–202
- quantitative structure-activity relationships (QSAR) model, 300–304
  - for clinical PK parameters, 302–304
- quantitative structure-pharmacokinetic relationships (QSPkR) modelling, 302
- quantum mechanical/ molecular mechanical (QM/MM) methods, 94–101, 110–114, 166–167
  - basic methodological description, 166–167
  - boundary and partition, 113
  - choice of, 113–114
  - examples of use, 167
  - general aspects, 110–113
  - practical issues, 113–114
- reaction rate constant, 136–139
  - ensemble-averaged variational transition state theory with multi-dimensional tunnelling (EA-VTST/MT), 136–139
- reductionism limits, chemical biology and, 349–353
- replica-exchange molecular dynamics (REMD), 49, 223, 236
- ribonuclease H (RNase H), 319–321
- root mean squared deviation (RMSD), 8–10, 17, 22, 78
- RosettaMatch, 90, 91
- SABER (Selection of Active/binding Sites for Enzyme Redesign), 91
- sampling algorithms, 21–22
  - simulation methods, 22
  - stochastic methods, 21–22
  - systematic methods, 21
- SciQSAR, 304
- scoring functions, 22–24, 290
  - consensus, 24
  - empirical, 23
  - force field-based, 23
  - knowledge-based, 23–24
- shortest communication pathways, 204–205
- signalling networks models, 237–240
- single-wall carbon nanotube field effect transistors (SWNT-FETs), 235
- Src kinase, 59
- state-of-the-art computational algorithms, 248
- state of the art nucleic acid simulation, 166
- steered molecular dynamics (SMD), 251
- stereoelectronic (gauche effect) effects, 148
- string method, 132–136
- structure-based drug design (SBDD), 20–26, 285
  - sampling algorithms, 21–22
  - scoring functions, 22–24
- structure-based molecular modelling, 286–296
  - de novo* drug design, 291–293
  - docking, 289–291

- dynamics, introducing, 293–296
- 3D structures source, 286–289
- systems pharmacology
  - chemical, biological and pharmacological data, databases, 353
  - prediction, computational methods, 354–355
- Tabu search (TS), 22
- target flexibility, tackling, 331–338
  - intrinsically disordered proteins (IDPs), 333–335
  - lactate dehydrogenase, 331–333
  - trinucleotide repeats diseases, RNA, 335–338
- theozyme, 89, 90
- three-dimensional structure
  - prediction, 224–225
- Toll-like receptor 4 (TLR4), 232
- transporters, diffusion, 259–264
- trinucleotide repeats diseases, 335–338
- two-dimensional free energy surface (2D FES), 126–127
- umbrella sampling (US) method, 123–127, 251
  - dynamic histogram analysis method (DHAM), 125–126
  - two-dimensional free energy surface (2D FES), 126–127
  - weighted histogram analysis method (WHAM), 124–125
- unassisted diffusion, lipid bilayers, 252–255
- van der Waals interactions, 158
- weighted histogram analysis method (WHAM), 124–125, 251
- Wordom software, 201, 203, 216
- ZDOCK, 228
- ZRANK, 228