In Silico Medicinal Chemistry

Computational Methods to Support Drug Design

Nathan Brown
In Silico Medicinal Chemistry
Computational Methods to Support Drug Design
RSC Theoretical and Computational Chemistry Series

Editor-in-Chief:
Professor Jonathan Hirst, University of Nottingham, Nottingham, UK

Series Advisory Board:
Professor Joan-Emma Shea, University of California, Santa Barbara, USA
Professor Dongqing Wei, Shanghai Jiao Tong University, China

Titles in the Series:
2: Non-Covalent Interactions: Theory and Experiment
3: Single-Ion Solvation: Experimental and Theoretical Approaches to Elusive Thermodynamic Quantities
4: Computational Nanoscience
5: Computational Quantum Chemistry: Molecular Structure and Properties in Silico
6: Reaction Rate Constant Computations: Theories and Applications
7: Theory of Molecular Collisions
8: In Silico Medicinal Chemistry: Computational Methods to Support Drug Design

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
Visit our website at www.rsc.org/books
In Silico Medicinal Chemistry
Computational Methods to Support Drug Design

Nathan Brown
The Institute of Cancer Research, London, UK
Email: nathan.brown@icr.ac.uk
Preface

My aim with this book is to provide an introduction to all aspects of the field of in silico medicinal chemistry for the beginner, but this does not preclude its usefulness to the intermediate and expert in terms of offering quick guides on specific areas. To this end, the book does not give a deep-dive into the field, but instead emphasises the key concepts that are of importance to understand in context and the more abstract challenges. However, to offer some kind of completeness, each chapter has a list of key references to which the reader is referred for further information, including methodologies and case studies where appropriate.

Having edited two books recently, I did not want another commission, but I could not turn down this invitation to write the kind of book that I felt would be of benefit to scientists starting out in the field. I also felt that this might be the right time to write such a book.

I would like to extend my thanks primarily to Prof. Jonathan Hirst at The University of Nottingham, who commissioned me to write this book. Without the Royal Society of Chemistry’s publishing team, I probably would not have finally finished writing this book.

I would like to thank the members of my team, past and present, who, whether they are aware or not, have contributed positively to this book: Yi Mok, Mike Carter, Berry Matijssen, Caterina Barillari, Nick Firth, Sarah Langdon, Lewis Vidler, Josh Meyers and Fabio Broccatelli. I asked for some guidance from an early research scientist who probably best represents the audience of this book, William Kew, then at The University of St. Andrews, and now a PhD student in whisky analysis at The University of Edinburgh, Scotland. Will’s feedback was invaluable in understanding how I should pitch the book and what I should cover. A heartfelt thanks to all of the many scientists with whom I have worked and co-authored research papers since starting out in this field: Bob Clark, Ben McKay, François Gilardoni, Ansgar
Preface

Schuffenhauer, Peter Ertl, Gisbert Schneider, Val Gillet, John Holliday, George Papadatos, Mike Bodkin, Andreas Bender, Richard Lewis, Edgar Jacoby, Christos Nicolaou and Swen Hoelder. I apologise if I have missed anyone off the list. I would also like to thank my colleague and medicinal chemistry mentor, Prof. Julian Blagg, who allowed me the time to dedicate to writing this book and has been a constant inspiration from his medicinal chemistry background.

A special thanks to my two academic mentors, Prof. Peter Willett from The University of Sheffield and Prof. Johnny Gasteiger from The University of Erlangen-Nuremberg. They both took a chance on me early in my career and gave me thorough grounding in using computers to solve problems in chemistry, and also instilled in me a full appreciation of the pragmatism required in computational methods, the importance of adherence to the scientific method and the essential, yet highly appropriate, design of experiments with the absolutely necessary control experiments.

Lastly, I would like to thank my Mum and Dad who encouraged me from an early age to be inquisitive about the world and ask questions, which led me to a career in science. I would also like to thank them for letting me have a ZX81 at a very, very young age, and also for letting me play with Lego a lot, which helped me to understand combinatorial explosion.
# Contents

## Part 1: Introduction

<table>
<thead>
<tr>
<th>Chapter 1</th>
<th>Introduction</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Overview</td>
<td>3</td>
</tr>
</tbody>
</table>

## Part 2: Molecular Representations

<table>
<thead>
<tr>
<th>Chapter 2</th>
<th>Chemistry and Graph Theory</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Overview</td>
<td>13</td>
</tr>
<tr>
<td>2.2</td>
<td>Graph Theory and Chemistry</td>
<td>13</td>
</tr>
<tr>
<td>2.3</td>
<td>Graph Theory in Chemistry</td>
<td>16</td>
</tr>
<tr>
<td>2.4</td>
<td>Mathematical Chemistry and Chemical Graph Theory</td>
<td>17</td>
</tr>
<tr>
<td>2.5</td>
<td>Summary</td>
<td>18</td>
</tr>
<tr>
<td>References</td>
<td></td>
<td>18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 3</th>
<th>Structure Representation</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Overview</td>
<td>19</td>
</tr>
<tr>
<td>3.2</td>
<td>The Need for Machine-Readable Structure Representations</td>
<td>20</td>
</tr>
<tr>
<td>3.3</td>
<td>Adjacency Matrix</td>
<td>22</td>
</tr>
<tr>
<td>3.4</td>
<td>Connection Table</td>
<td>23</td>
</tr>
<tr>
<td>3.5</td>
<td>Line Notations</td>
<td>25</td>
</tr>
<tr>
<td>3.5.1</td>
<td>WLN: Wiswesser Line Notation</td>
<td>26</td>
</tr>
<tr>
<td>3.5.2</td>
<td>SMILES: Simplified Molecular-Input Line-Entry Specification</td>
<td>26</td>
</tr>
<tr>
<td>3.5.3</td>
<td>InChI: IUPAC International Chemical Identifier</td>
<td>28</td>
</tr>
<tr>
<td>3.6</td>
<td>Summary</td>
<td>29</td>
</tr>
<tr>
<td>References</td>
<td></td>
<td>30</td>
</tr>
</tbody>
</table>
## Chapter 4 Molecular Similarity

4.1 Overview 31  
4.2 Molecular Similarity 33  
4.3 Similar Property Principle 35  
4.4 Molecular Descriptors 36  
4.5 Calculation of Molecular Similarity 37  
4.5.1 Similarity Coefficients 37  
4.6 Molecular Diversity 40  
4.7 Summary 40  
References 41

## Part 3: Molecular Descriptors

### Chapter 5 Molecular Property Descriptors

5.1 Overview 45  
5.2 Molecular Weight (MW or MWt) 45  
5.3 Octanol/Water Partition Coefficient (ClogP) 46  
5.4 Topological Polar Surface Area (TPSA) 47  
5.5 Hydrogen Bond Acceptors and Donors (HBA and HBD) 48  
5.6 Lipinski’s Rule-of-Five 48  
5.7 Summary 50  
References 51

### Chapter 6 Topological Descriptors

6.1 Overview 52  
6.2 Topological Indices 52  
6.2.1 Wiener Index 53  
6.2.2 Randić Index 53  
6.2.3 Petitjean Index 55  
6.2.4 Zagreb Indices 55  
6.3 Molecular Fingerprints 55  
6.3.1 Structure-Key Fingerprints 56  
6.3.2 Hash-Key Fingerprints 57  
6.3.3 Ligand-Based Topological Pharmacophores 60  
6.4 Summary 64  
References 64

### Chapter 7 Topographical Descriptors

7.1 Overview 66  
7.2 Topographic Descriptors 67  
7.3 Pharmacophores 71  
7.4 ROCS: Rapid Overlay of Chemical Structures 75  
7.5 USR: Ultrafast Shape Recognition 76  
7.6 XED: Cresset Group 77  
7.7 Conformer Generation and the Conformer Problem 77  
7.8 Summary 80  
References 81
# Part 4: Statistical Learning

## Chapter 8  Statistical Learning  85

8.1 Overview  85  
8.2 Statistical Learning  85  
8.3 Unsupervised Learning  86  
  8.3.1 Overview  86  
  8.3.2 Cluster Analysis  87  
  8.3.3 \(k\)-Means Clustering  88  
  8.3.4 Stirling Numbers of the Second Kind  89  
  8.3.5 Self-Organising Maps  89  
  8.3.6 Principal Component Analysis  91  
8.4 Supervised Learning  92  
  8.4.1 Naive Bayesian Classification  92  
  8.4.2 Support Vector Machine  93  
  8.4.3 Partial Least Squares  93  
8.5 Best Modelling Practice  94  
8.6 Summary  95  
References  96

# Part 5: Modelling Methodologies

## Chapter 9  Similarity Searching  101

9.1 Overview  101  
9.2 Similar Property Principle  102  
9.3 Molecular Similarity and Virtual Screening  104  
9.4 Data Fusion  104  
9.5 Enrichment  106  
  9.5.1 Lift Plots  106  
  9.5.2 Confusion Matrix  107  
  9.5.3 Receiver Operating Characteristic Curves  108  
  9.5.4 Enrichment Factors  108  
9.6 Summary  110  
References  110

## Chapter 10  Bioisosteres and Scaffolds  112

10.1 Overview  112  
10.2 A Brief History of Bioisosterism  113  
10.3 Bioisosteric Replacement Methods  115  
  10.3.1 Knowledge-Based Bioisosteric Replacements  115  
  10.3.2 Information-Based Bioisosteric Replacements  116  
10.4 Scaffold Representations  117  
10.5 Scaffold Diversity Analysis  118  
10.6 Summary  121  
References  122
Chapter 11  Clustering and Diversity 124
11.1 Overview 124
11.2 Dissimilarity-Based Compound Selection 125
11.3 Sphere-Exclusion 127
11.4 Cell-Based Diversity Selection 127
11.5 Hierarchical Clustering 128
11.6 Non-Hierarchical Clustering 129
11.7 Summary 130
References 131

Chapter 12  Quantitative Structure–Activity Relationships 132
12.1 Overview 132
12.2 Free–Wilson Analysis 132
12.3 Hansch–Fujita Analysis and Topliss Trees 133
12.4 QSAR Model Generation 133
12.5 Feature Selection 136
12.6 Methods for Estimating Model Validity, Predictive Power, and Applicability Domains 139
12.7 Automated Model Generation, Validation and Application 140
12.8 Summary 141
References 142

Chapter 13  Protein–Ligand Docking 143
13.1 Overview 143
13.2 Search Algorithms 143
13.3 Scoring Functions 144
13.4 GOLD: Genetic Optimisation for Ligand Docking 145
13.5 Model Validation 147
13.6 Docking in Prospective Studies 147
13.7 Summary 149
References 151

Chapter 14  De Novo Molecular Design 153
14.1 Overview 153
14.2 Receptor-Based Methods 154
14.3 Fragment-Linking Methods 155
14.4 Fragment-Growing Methods 156
14.5 Sampling Chemistry Space 156
14.6 Atom-Based De Novo Design 157
14.7 Fragment-Based De Novo Design 158
14.8 Reaction-Based De Novo Design 158
14.9 Multiobjective Optimisation 160
14.10 Summary 161
References 161
## Contents

### Part 6: Applications in Medicinal Chemistry

Chapter 15  Applications in Medicinal Chemistry  165  
15.1  Overview  165  
15.2  Early Stage Targets  166  
15.3  Hit Generation  168  
15.4  Hits-to-leads  172  
15.5  Lead Optimisation  175  
15.6  Summary  176  
References  177

### Part 7: Summary and Outlook

Chapter 16  Summary and Outlook  181  
16.1  The Past  181  
16.2  The Present  182  
16.3  The Future  182  
16.4  Summary  184

### Appendices

Appendix A  Glossary of Terms  187
Appendix B  Professional Societies  192
Appendix C  Journals  195
Appendix D  Resources for Computational Drug Discovery  199
Subject Index  205
Part 1
Introduction
CHAPTER 1

Introduction

1.1 Overview

The discovery and design of new drugs is an endeavour that humanity has undertaken only in more recent history thanks to the scientific advances made by scientists from many different fields. Chemists have been able to isolate, synthesise and characterise potential therapeutic agents. Biologists can then test the safety and efficacy of those agents in multiple biological models, and clinicians can test the agents in humans. However, there are more potential new chemical structures that could be synthesised than time allows. Some estimates have put the potential space of druglike molecules at $10^{20}$ and others up to $10^{200}$. Regardless of how precisely vast that space is and how much of it is actually worthy of exploration, I think we can agree that it is truly, astronomically vast.

Computers have transformed our lives in recent times, with a standard smartphone carried in our pockets having more computing power than all of the computing power that NASA (National Aeronautics and Space Administration) had in 1969 when we put a man on the moon. The chip in a modern iPhone has more than two billion transistors and is capable of running tens of billions of instructions per second. However, the ability to process more data does not necessarily mean that we automatically start making better decisions. Indeed, there is a misguided assumption that increased computer power means that we can get the right answers faster, but without careful thought and experimental design with appropriate controls, we will only find the wrong answers faster and still waste a great deal of time in physical experiments based on inappropriate predictions made using computational methods.

The computer is a tool, like any other. One would not go into a chemistry or biology laboratory and simply start moving things around and think