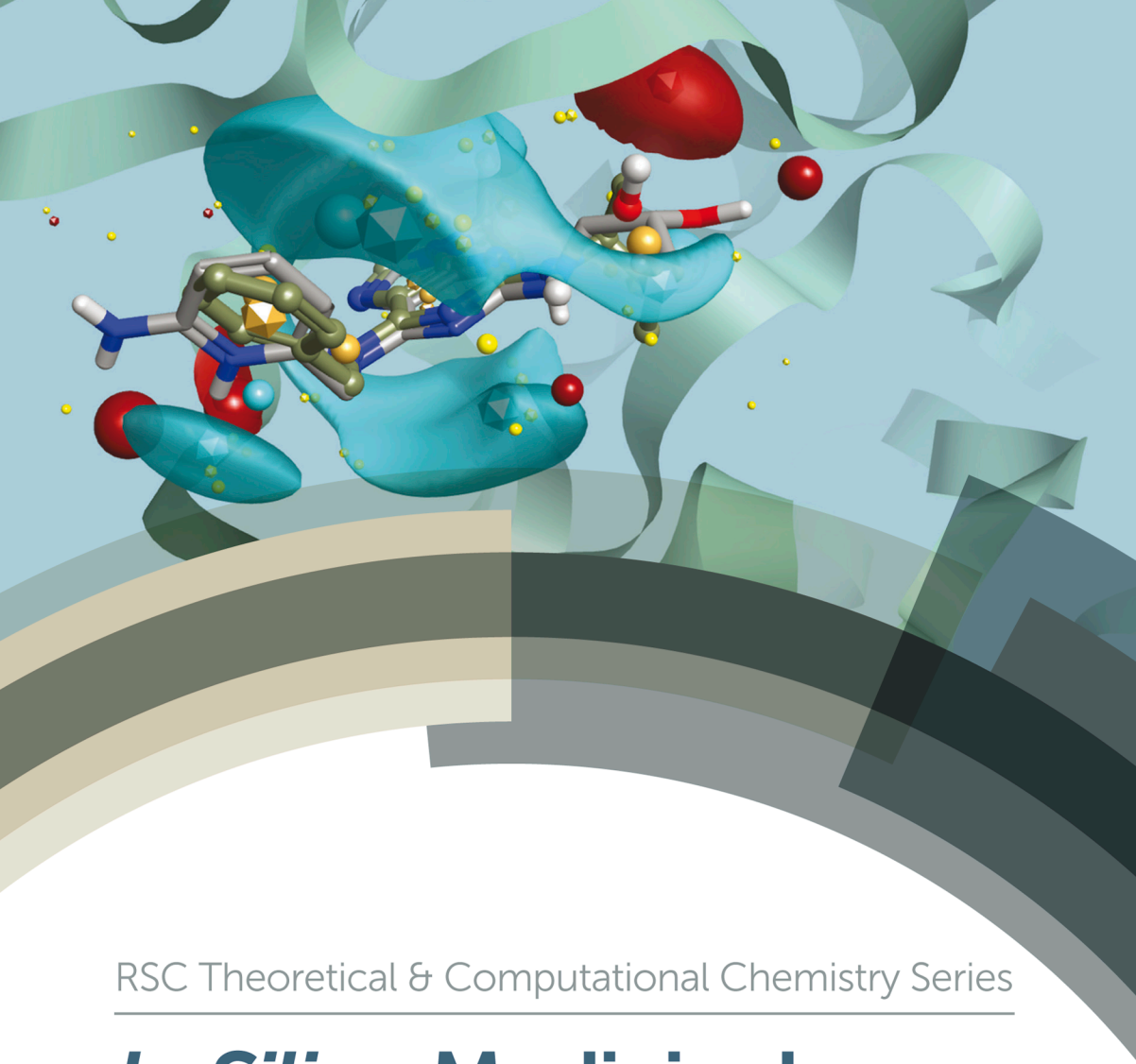

RSC Theoretical & Computational Chemistry Series

In Silico Medicinal Chemistry

Computational Methods to Support Drug Design

Nathan Brown

ROYAL SOCIETY OF CHEMISTRY

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:
1: Knowledge-based Expert Systems in Chemistry: Not Counting on Computers
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

RSC Theoretical and Computational Chemistry Series No. 8

Print ISBN: 978-1-78262-163-8
PDF eISBN: 978-1-78262-260-4
ISSN: 2041-3181

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

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

Registered Charity Number 207890

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

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

RSC Theoretical and Computational Chemistry Series No. 8
In Silico Medicinal Chemistry: Computational Methods to Support Drug Design
By Nathan Brown

Published by the Royal Society of Chemistry, www.rsc.org

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

Part 2: Molecular Representations

RSC Theoretical and Computational Chemistry Series No. 8
In Silico Medicinal Chemistry: Computational Methods to Support Drug Design
By Nathan Brown

Published by the Royal Society of Chemistry, www.rsc.org

Part 3: Molecular Descriptors

Part 4: Statistical Learning

Part 5: Modelling Methodologies

Part 6: Applications in Medicinal Chemistry

Part 7: Summary and Outlook

Appendices

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

RSC Theoretical and Computational Chemistry Series No. 8
In Silico Medicinal Chemistry: Computational Methods to Support Drug Design
By Nathan Brown

Published by the Royal Society of Chemistry, www.rsc.org

we are conducting good science, and hope to leave the lab without causing considerable harm to oneself. Conducting good science requires a significant amount of expert training. The same can be said for the computer, it is essentially a molecular modeller's laboratory. It is a facile assumption that because we can install molecular modelling software, then this will make us a modeller. To become an effective and successful modeller requires as much time as becoming an effective and successful laboratory scientist. It is not sufficient to believe that installing software and clicking buttons will make you a good molecular modelling scientist; it may give rise to that being the case, but this is merely an illusion.

This book is an attempt to provide some of the history and popular methods applied in modern day medicinal chemistry and drug discovery using computers and informatics platforms, a discipline for which an appropriate title may be: *in silico* medicinal chemistry. In this title, the aim is to define a field of endeavour and scientific rigour that contributes positively in every appropriate aspect of medicinal chemistry and drug discovery, from the design of high-throughput screening libraries to providing predictions of molecular properties required for drug compounds and understanding how those molecules interact with biological macromolecules. It is always my primary concern to contribute positively to the many projects I work on. By 'contribute positively' I mean that it is important for everyone involved to understand what the predictions or analyses tell us, as well as having a thorough understanding of the limitations of these methods. With understanding comes control, and this can only assist in designing experiments and prioritising possible decisions. It is important as a practicing molecular modeller to be fully aware that, despite taking relatively little time, *in silico* experiments can lead to a huge amount of wasted resource, both in chemistry and biology laboratories, if best practice and appropriate checks and balances are not put in place.

Molecular modellers should be hypothesis-driven scientists. The hypothesis is the core of science: just because we can do something does not mean that we should. We must have a specific question in mind. Once the hypothesis has been formalised then we can consider how we might tackle the challenge. It is important to understand the commitment required from the laboratory scientists and project budgets to ensure that expectations are managed.

Drug discovery and design takes place in large pharmaceutical companies, biotechnology start-ups, and increasingly academicians are being ever more effective and demonstrably capable of drug discovery. Anyone in a career in drug discovery, or with the intention of developing a career in this area, will be exposed to computational methods in chemistry regardless of where they sit in the organisation. Molecular modellers assist high-throughput screening (HTS) teams in designing their compound libraries and analysing their hit matter through HTS triaging. Medicinal chemists work most closely with molecular modellers and chemoinformaticians on aspects ranging from compound registration of new molecular entities into databases to designing vast virtual compound libraries from which targets for synthesis can be

prioritised. Working with structural biologists and crystallographers we can enable structure-based drug design, where we have experimental evidence for binding modes of potential drugs in protein binding sites allowing the project teams to design compounds that should, or sometimes should not, work to test specific hypotheses. Working with computational biologists we can assist in identifying and validating therapeutic targets *in silico*. And this is without considering the impact we can have in basic biology, genetics, metabolism and pharmacokinetics.

It is clear that the field of *in silico* medicinal chemistry is truly interdisciplinary, working across many different teams. Furthermore, the *in silico* medicinal chemists of today increasingly come from different backgrounds and not just chemistry. Many computer scientists, mathematicians, statisticians, physicists and scientists from other disciplines work very effectively and contribute positively to the discovery of new drugs.

In addition to working with multidisciplinary teams in the context of drug discovery, we are still making fundamental advances and discoveries in the field of *in silico* medicinal chemistry. That is to say that the field is not a solved problem and we still have many challenges to work on. A good molecular modeller is agile and adaptable to these new challenges and can see opportunities for contributing fundamentally to the community.

Computers, although all-pervasive nowadays, are actually a very modern advance. However, the advent of modern computation and all that it offers has been included in drug design for many more years than one might expect.

A more recent advance in *in silico* medicinal chemistry is the availability of toolkits implemented to allow for the quick development of software programs to tackle challenges quickly and easily, such as the RDKit API. Workflow tools have become available that enable many non-expert scientists to quickly generate simple processes using visual programming techniques. One such workflow tool is KNIME. Data analysis is also becoming more achievable on large data sets thanks to interactive data exploration and analysis tools such as DataWarrior. Lastly, all these methods and software would be worthless without data. Again, recently datasets have become available that represent marketed drugs, clinical candidates, medicinal chemistry compounds from journals, commercially available compounds, and those structures contained in patents: ChEMBL, DrugBank, SureChEMBL. The most amazing aspect of all of these advances is that everything mentioned in this paragraph is free. Free to download, free to install, free to use, with no limits.

This truly is a golden age of *in silico* medicinal chemistry as a data science, which is essentially what it is, working with lots of heterogeneous data (so-called big data) and various modelling techniques from structure-based modelling through to statistical learning methods. All of these and more are covered in this book.

The title of this book, *In Silico* Medicinal Chemistry, is intended as an umbrella term for all approaches to using computers in chemistry to benefit

medicinal chemistry and drug discovery. In this way, one can see *in Silico* Medicinal Chemistry as covering aspects of: chemoinformatics (also called cheminformatics), molecular modelling and computational chemistry. This book is not intended to be all-inclusive and exhaustive, but rather to make a solid foundation from which the reader can pursue aspects that most interest them or are relevant to a particular scientific challenge. Each chapter concludes with an inexhaustive list of key references to which the interested reader is directed for more in-depth information around specific subject areas from leading monographs in those areas.

The book covers the fundamentals of the field first: how we represent and visualise those molecules in the computer, and how we compare them. The section begins, though, with a brief history and introduction to mathematical graph theory and its close links with chemistry and molecular representations going back to the advent of atomistic theory and even earlier. Representing molecules in the computer is essential for whatever subsequently needs to be achieved in the computer. For some applications it may be possible to have more complex representations, but more complex representations will typically require more complex calculations to analyse and make best use of the data. The methods by which we compare molecules also lie at the heart of computational chemistry. Similarity is a philosophical concept, but it is essential to consider the different types of similarities that may be measured and how they may be applied. All of these topics are covered in the first section of the book.

The second section of the book considers the many different ways we can describe molecules in the computer. The old parable of the 'Six Blind Men and the Elephant' written by John Godfrey Saxe, from ancient tales, highlights challenges in measuring similarity and understanding differences. In the parable, six blind men were each asked to describe an elephant. The first blind man suggested that the elephant was like a wall because he felt its body. The second thought it like a snake, having touched its trunk. The third identified it as like a spear when feeling its tusk, and so on. This parable highlights the importance of recognising and understanding the concept of similarity and why it is important. The section begins with physicochemical descriptors, from which it possible to calculate properties that are measurable, with a high degree of accuracy. The second chapter moves onto topological descriptors that encode aspects of the molecular graph representation, whether through the calculation of a single value that encapsulates an aspect of the molecular graph but is interpretable, or large quantities of complex descriptors that do not tend to be so interpretable, but are highly efficient and effective. The third class of molecular descriptor is the topographical or geometric descriptor that encodes information about the shapes and geometries of molecules, since clearly they are typically not flat, or static, entities.

The third section of the book considers statistical learning methods, an integral aspect of computational drug discovery, and some of the best methods we have to investigate different properties. An introduction to statistical

learning will be given, prior to breaking off into two different aspects of statistical learning: unsupervised and supervised learning. Unsupervised learning uses statistical methods to understand the structure of data and how different objects, described by variables, relate to each other. This is important in understanding the proximity or otherwise of our data points, in our case molecules, and is integral to the concepts of molecular similarity and diversity in chemical databases and techniques used in many methods. Supervised learning still uses the descriptions of our objects, molecules, but attempts to relate these to another variable or variables. In chemistry, supervised learning can be used to make predictions about molecules before they are synthesised. This predictive learning can be very powerful in computational chemistry since we can explore that vast space of possible small molecules discussed earlier in a much more effective and rapid way. Lastly, a discussion and some advice on best practices in statistical learning are given to assist the modern scientist using computers to make statistical analyses or summaries.

The next section moves on to explicit applications of computational methods in drug discovery. These methods are well known in the field and use aspects of all of the previously discussed concepts and methods. Similarity searching is up first, which is focussed on the identification of molecules that are similar to those that are already known, but also comparing large numbers of molecules for similarity and diversity. One of the most important aspects of similarity searching is the introduction of the concept of virtual screening, where new and interesting molecules can be identified by using ones that are already known, but with a similarity measure that is relevant to the challenge being addressed.

The second chapter in this section covers the twin concepts of bioisosteric replacements and scaffold hopping. These two concepts are related to similarity searching, which was mentioned previously, but instead of trying to identify molecules that tend to have structural similarities, this approach looks for functional similarity, with little regard for the underlying structure. This is becoming increasingly important in drug discovery as it allows projects to move away from troublesome regions of chemistry space that, although important for potency, may exhibit other issues that are undesirable in drugs.

The third chapter covers clustering and diversity analysis, which are essentially two sides of the same coin. Cluster analysis permits the identification of natural groupings of objects, molecules, based on molecular descriptors and example of the application of unsupervised learning. Using cluster analysis it is possible to select clusters of interesting molecules for follow-up or, using molecular diversity to select a subset of molecules that are different to each other.

Whereas cluster analysis is an example of unsupervised learning, Quantitative Structure–Activity Relationships (QSARs) are an example of supervised statistical learning methods. Here, the objective is to correlate molecular structure with known biological endpoints, such as enzyme potency, and

build a statistical model. The benefit of such a model, a QSAR, is that it may, with care and caution, be applied to predict for molecules that have not been tested, and have not even been synthesised. This allows vast virtual libraries to be analysed and prioritised to allow the focus to rest on those molecules that are most likely to succeed.

Since proteins began being crystallised and their structures identified through X-ray crystallography, the structures have held the promise of allowing the optimisation of new molecular entities *in silico* that are predicted to be enhanced in potency against the enzyme-binding site of interest. Protein–ligand docking methods have been developed for more than 30 years to model virtual molecules that are more optimal in interactions and potency than their predecessors. Many new methods and developments have been made and the predictive abilities of docking have improved greatly over the years. Still, however, challenges remain. This chapter considers the methods that have been developed, an understanding of how to validate docking models and finally how best to use the methods.

The last chapter in this section covers *de novo* design, arguably the pinnacle of computational drug discovery. The grand objective in *de novo* design is to design molecules in the computer that are entirely optimised for each of the objectives of interest. Clearly, the discipline is not that close to being able to achieve such a grand challenge, but much headway has been made, particularly in recent years, utilising all of the methods that go before in this book. A brief history of *de novo* design is given in structure- and ligand-based methods, with a final view towards the future and the incorporation of multiple objectives in *de novo* design workflows.

The penultimate section of the book looks at a few successful case studies and methods that have been applied in every stage of drug discovery, from aspects of target validation in terms of druggability analyses and hit discovery, through to moving from hit compounds to leads and the optimisation of those leads. Some examples of methods that have or can be used in these in these are covered to set the context of the field and its level of importance through the drug discovery pipeline.

Lastly, the book concludes with the 'Ghosts of Christmases Past, Present and Yet to Come'. This chapter represents the importance of remembering where we came from and respecting the contributions of the giants that came before us; it reflects on where we are, how we got here and what has been achieved in recent years; and lastly, the chapter discusses what needs to be addressed in future, how can we achieve this and what we all need to do to prepare for the future.

This book is intended as an overview of a vast field, with thousands of scientists working in it worldwide. Each chapter has a set of key, yet not extensive, references as guides to where the interested reader may go next in the development of their skills and expertise in this complex field, no matter what it may be called.

Finally, it is important as you read through this book to remember the two mantras of anyone involved in modelling real-world systems:

"*In general we look for a new law by the following process. First we guess it. Then we compute the consequences of the guess to see what would be implied if this law that we guessed is right. Then we compare the result of the computation to nature, with experiment or experience, compare it directly with observation, to see if it works. If it disagrees with experiment it is wrong. In that simple statement is the key to science. It does not make any difference how beautiful your guess is. It does not make any difference how smart you are, who made the guess, or what his name is—if it disagrees with experiment it is wrong. That is all there is to it.*"

Richard P. Feynman
Chapter 7, Seeking New Laws. *The Character of Physical Law*, 1965.

"*Since all models are wrong the scientist cannot obtain a 'correct' one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity.*"

George E. P. Box
Science and Statistics. *J. Am. Statist. Assoc.* 1976, **71**, 791–799.

Part 2
Molecular Representations