In Silico Medicinal Chemistry

Computational Methods to Support Drug Design

RSC Theoretical and Computational Chemistry Series

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

ROYAL SOCIETY
OF CHEMISTRY

THE QUEEN'S AWARDS
FOR ENTERPRISE:
INTERNATIONAL TRADE
2013

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

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

The authors have sought to locate owners of all reproduced material not in their own possession and trust that no copyrights have been inadvertently infringed.

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

CHAPTER 2

Chemistry and Graph Theory

2.1 Overview

One of the most important aspects of using computers to answer questions in chemistry is graph theory. The chemical structure representation we all recognise today comes from the mathematical subfield of graph theory and is often termed a molecular graph. The molecular graph is the single-most important data structure in computational chemistry. Graph theory has been intrinsically linked with chemistry since the advent of atomistic theory in the early nineteenth century and has given us the representation we identify today.

2.2 Graph Theory and Chemistry

Graph theory and chemistry have had a long-standing partnership from the mid-eighteenth century until the present day. From their undergraduate training, many computer scientists and mathematicians know of graph theory, the data structures, algorithms and applications and how the concepts can be used to answer very complex problems in a very logical way. However, what might be less known is that the name, Graph Theory, originated directly from efforts in chemistry in the early 1800s at the advent of atomistic theory to determine a method by which the structures of molecules could be represented pictorially.[1]

The first example of a graph theoretic approach to solving a problem is from Leonhard Euler in 1735, and this work would even ultimately result in the field of topology. The problem that Euler faced was based in the then city of Königsberg in Prussia and considered just the landmasses in the city and the bridges that connected them. Königsberg is bordered on both sides by the river Pregel and consisted of two large landmasses. To enable transport

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

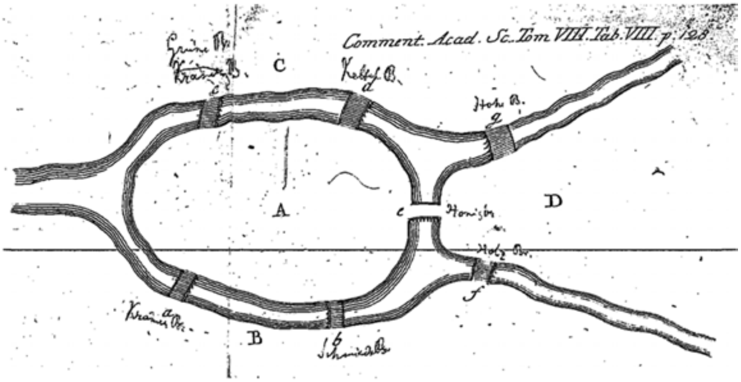


Figure 2.1 The original map of Königsberg used by Euler in his seminal work introducing the concepts of graph theory. It can be seen clearly that Euler labelled the landmasses as A, B, C, and D whereas the bridges connecting those four landmasses were labelled as a, b, c, d, e, f, and g. Accessed from https://math.dartmouth.edu/~euler/pages/E053.html.

and trade it was clearly necessary for a number of bridges to be built to cross the Pregel at various points. In fact, the landmasses were interconnected by a total of seven bridges (Figure 2.1). This set Euler wondering, would it be possible, starting from any landmass, to cross each one of the seven bridges once and only once? Euler stipulated that the river must only be crossed by one of the seven bridges and once someone has started walking across a bridge, they cannot turn back and count that as a crossing. It was not necessary that the start points and end points of the walk be the same. Euler's challenge was, could he solve this problem with sufficient abstract and mathematical thoroughness that the concepts could be adapted and solved for similar problems?

Considering the map, Euler realised that landmasses A and B were connected by bridges a and b; A and C were connected by bridges c and d; A and D were connected by c only; B and C had no directly connecting bridges; B and D were connected only by bridge f; and C and D were connected by only bridge g. Since the start and end points were irrelevant, Euler realised that the challenge was simply the sequence in which the bridges were crossed. This permitted Euler to understand that all that was important was how the bridges were connected to each other *via* the landmasses. This allowed Euler to abstract the map representation, eliminating all extraneous features other than the bridges and how they intersected landmasses. In graph theory today, we would call each bridge an *edge* or an *arc*, and each landmass a *node* or a *vertex*. The abstracted representation of the landmasses and bridges connecting them is what is today called a graph and was the foundation of graph theory (Figure 2.2).

Using his graph representation, Euler understood that when a node is entered *via* an edge, then another edge must be available that has not already been crossed, unless at the start or beginning of the walk. Therefore, the

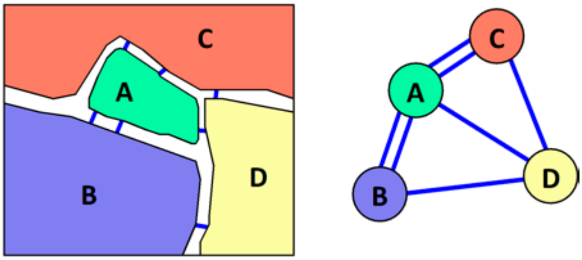


Figure 2.2 The labelled physical landmasses on the left can be reduced to the abstract graph representation of the nodes (landmasses) and edges (bridges) that represent all that Euler required in terms of information to solve whether it was possible to cross each of the bridges of Konigsberg once and only once.

number of times a vertex is entered, not at the start or end of the walk, must equal the number of times that it is left. Since every bridge must be crossed only once then each landmass, except for the start and finish, must have an even number of bridges connecting them. However, all four of the landmasses have an odd number of bridges, so it must be impossible to solve this problem for the Königsberg graph.

In the 1750s in Edinburgh, two scientists, William Cullen and Joseph Black, devised a form of graph theory on the theory of the relationships between chemical substances, called affinity diagrams. The nodes of these graphs were substances, and the edges the measured affinities between them. More than a century later, the new atomistic theory was being formalised and scientists were attempting to understand the structures of molecules. This is where Euler's earlier work on mathematical graph abstractions was noticed and used to define relationships between atoms in molecular structures.

A large number of scientists of the time explored ways of pictorially representing molecules, their atoms and how they are interconnected. Two scientists in particular probably did the most to give graph theory its name as we know it today: Alexander Crum Brown and James Joseph Sylvester. In 1864, Crum Brown devised his constitutional formulae where he represented atoms as nodes and bonds as edges.[2] However, Crum Brown was insistent that these were intended as an abstraction on not necessarily reality, merely representing the relationships between the atoms. Sylvester devised a very similar representation to Crum Brown, but whereas Crum Brown referred to his representations as molecular graphic notations, Sylvester called his molecular representation the chemicograph.[3] It is likely that it will never be known which of the two chemists actually gave rise to the name of the field of Graph Theory, but it is clear that both eminent scientists gave this old field its new name.

Arthur Cayley, a mathematician, used graph theory to study a particular subset, called trees, which are acyclic. Cayley used his research in this area to begin his interest in what today we would call theoretical chemistry.

Cayley used graph theory to mathematically enumerate all alkanes possible containing a given number of carbon atoms. At the time, it was thought that the enumeration of the possible structures of alkanes could also inform about their possible properties. Sylvester's work was the foundation of graph enumeration, which continues to be an area of great endeavor to this day.

The field of Graph Theory has continued progressing and is now a fully-fledged field within mathematics in its own right. Now graph theory is hugely influential in all areas of science, such as social network analysis, routing in communications networks, and understanding of biological and biochemical pathways. Indeed, the Six Degrees of Kevin Bacon, where every actor can be linked to any other actor within six degrees of connectivity, is an example of small world phenomena that provides an understanding of how communities can emerge. The theory was extended to the Erdős numbers, which enable scientists to see how closely related they are to Paul Erdős, one of the most prolific scientists in history.

However, graph theory is still applied, and new algorithms designed, in the field of chemistry, as will be seen in the remainder of this book. Graph theory and chemistry can be used together to understand chemical systems and make predictions from molecular structure. Before these data structures and algorithms are introduced, it might be worthwhile to review some of the concepts and terminology of graph theory.

2.3 Graph Theory in Chemistry

Graph theoretic techniques are widely applied in computer science; however, it is prudent here to provide a brief overview of graph theory and the terms and standards used in this article before moving on to the rest of the book.[4] A graph G is a collection of objects $V(G)$ and the relationships between those objects $E(G)$ called nodes (or vertices) and edges (or arcs), respectively. In the context of chemoinformatics, the nodes are the atoms of a molecule and the edges are the bonds. The nodes in G are connected if there exists an edge $(v_i, v_j) \in E(G)$ such that $v_i \in V(G)$ and $v_j \in V(G)$. The order of a graph G is given by the size of $|V(G)|$. A node v_i is incident with an edge if that edge is connected to the node, while two nodes, v_i and v_j, are said to be adjacent if they are connected by the edge $(v_i, v_j) \in E(G)$. Two edges are said to be incident if they have a node in common. A complete graph is where every node is connected to every other node in the graph. The edge density of a graph can then be calculated as the number of edges in a particular graph normalised between the number of edges in a connected graph $(|V(G)| - 1)$, and the number of edges in the complete graph $(|V(G)| \cdot (|V(G)| - 1)/2)$, with the given number of nodes, $|V(G)|$.

One of the most important applications of graph theory to chemoinformatics is that of graph-matching problems. It is often desirable in chemoinformatics to determine differing types of structural similarity between two molecules, or a larger set of molecules. This will be expanded upon later in Chapter 4 on Molecular Similarity. Two graphs are said to be isomorphic when they are structurally identical. Subgraph isomorphism of G_1, G_2 holds if G_1 is isomorphic to some subgraph in G_2. On the other hand, the identification of the maximum common subgraph between two graphs is the determination of the largest connected subgraph in common between the two. Last, the maximum overlap set is the set of the, possibly disconnected, largest subgraphs in common between two graphs. In chemoinformatics, the term structure is often used in place of graph. These graph-matching problems are thought to be NP-complete and therefore numerous methods have been applied to prune the search tree.

The molecular graph is a type of graph that is undirected and where the nodes are colored and edges are weighted. The individual nodes are colored according to the particular atom type they represent (carbon (C), oxygen (O), nitrogen (N), chlorine (Cl), *etc.*), while the edges are assigned weights according to the bond order (single, double, triple, and aromatic). Aromaticity is an especially important concept in chemistry. An aromatic system, such as the benzene ring, involves a delocalised electron system where the bonding system can be described as somewhere between single and double bonds, as in molecular orbital (MO) theory. In the case of the benzene ring—a six-member carbon ring—six π electrons are delocalised over the entire ring. A common approach to representing an aromatic system in a computer is to use resonant structures, where the molecule adopts one of two bonding configurations using alternating single and double bonds. However, this is an inadequate model for the representation of aromaticity and therefore the use of an aromatic bond type is also used. Molecular graphs also tend to be hydrogen depleted, that is, the hydrogens are implicitly represented in the graph since they are assumed to fill the unused valences of each of the atoms in the molecule. Each atom is ascribed a particular valence that is deemed at least to be indicative of the typical valence of the molecule: carbon has a valence of 4, oxygen has 2, and hydrogen has 1.

Graph theory offers an excellent foundation on which to build computer systems that store, manipulate and retrieve chemical data. Referring to the field of graph theory every so often is beneficial since many new algorithms have developed in graph theory that may have direct applicability to challenges in chemical structure analyses. Many of the algorithms and methods defined in this book will build on graph theoretic representations and their manipulation, and the terminology introduced in this chapter will be invaluable when reading through the remainder of the book.

2.4 Mathematical Chemistry and Chemical Graph Theory

The two fields of Mathematical Chemistry and Chemical Graph Theory will not be discussed explicitly in the remainder of this book, but these fields are integral to many of the molecular descriptors that will be discussed in Section 3. Pioneers in these fields, by the names of Balaban, Gutman, Hosoya, Randic, Wiener, and Trinajstic, developed many of the topological indices that are still used today.[5] Their assertion is that the simple chemical topology, the two-dimensional structure, can provide many insights into the many chemical phenomena that can be observed and measured. It will become clear that in many cases it is possible to correlate physical phenomena with aspects of the chemical structures under examination.

2.5 Summary

Graph theory is integral to chemistry as its data structure represents the *lingua franca* of chemistry in the two-dimensional chemical structures. The history of chemistry and graph theory are inextricably linked from the foundations of the latter. Once atomistic theory arose, chemists sought ways to represent the structures of their molecules in the belief that their structure would reveal explanations for their properties and that it would be possible to predict these properties without the need to make and test these compounds. Essentially, this is what we do in this field, make predictions for molecular structures that may or may not have ever been synthesised. This is the key to this field; we make predictions regarding the properties of molecules and use these to make better and rational decisions. In the remainder of this book, many of the most effective computational methods will be introduced and explained, with their context of application and effectiveness discussed. What is important to remember is that none of this would have been possible without the research and work of many scientists that came before us.

References

1. N. Biggs, E. Lloyd and R. Wilson, *Graph Theory*, Oxford University Press, 1986, pp. 1736–1936.
2. A. Crum Brown, On the theory of isomeric compounds, *Trans. R. Soc. Edinburgh*, 1864, **23**, 707–719.
3. J. J. Sylvester, Chemistry and Algebra, *Nature*, 1878, **17**, 284.
4. R. Diestel. *Graph Theory*, Springer-Verlag, New York, NY, 2nd edn, 2000.
5. D. Bonchev, *Chemical graph theory: introduction and fundamentals*, CRC Press, 1991, vol. 1.

CHAPTER 3

Structure Representation

3.1 Overview

The foundation of any computational manipulation of molecules requires a defined encoding scheme that permits the input into the computer in a machine-readable form, the molecular structure, which must also be represented in-memory, typically in a different format. The same format must also be adhered to when computer systems write out the resulting molecules to file, to ensure that the format is invariant and can be used interchangeably between software systems.

The simplest molecular structure representation is a list of the atoms in a molecule and how bonds connect those atoms; often the hydrogen atoms are not explicitly defined. The assumption here is that, with the molecular structure defined, everything else may be generated, or indeed regenerated, since many properties such as protonation states and three-dimensional structures are implicit in the topological structure. However, if this is not necessarily the case, and even if it may be, it can be more computationally time-consuming to recalculate properties of three-dimensional structures than to encode all of these additional properties in an appropriate file format.

This chapter provides a brief history and overview of some of the more common file formats used in modern chemistry software. Advantages and limitations of each will be considered.

A warning here is that small discrepancies between programs may lead to an inability to read files generated in one system in another current system. This is particularly an issue when considering files with large numbers of chemical structures since these discrepancies are more likely to occur. Oftentimes the associated errors can be overcome as a workaround by reading the file in an alternative software and writing out the file for reading into the

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

intended software. This is somewhat of a kludge, but is sadly necessary due to a lack of a defined common standard in many of our structure representation systems.

3.2 The Need for Machine-Readable Structure Representations

Substances have always been given names throughout history. Berzelius was the first to propose that chemicals should be named, not from where they came, but by what they are. We have seen already the close links between graph theory in mathematics and atomistic theory in chemistry, and how these two scientific branches merged with molecular graph notation being the progeny. However, there are many names for even a single substance.

Caffeine, a chemical compound close to the heart of anyone involved with computers, is an accepted name in general parlance for this chemical stimulant. However, it is known by many names. Runge first isolated caffeine in 1819, which he called *Kaffebase*, a base that exists in coffee. However, it was given its name by French scientists, hence caffeine. In 1827, Oudry isolated what he called theine from tea, but this was later found by Mulder to be caffeine. Caffeine was given its name because it is most commonly extracted from the seeds of the coffee plant, but this breaks the rule of Berzelius since it is named from where it came, hence theine, rather than what it is. Indeed, a number of alternative common names are also given to caffeine: theine, 1,3,7-trimethylxanthine, 1,3,7-triméthylxanthine, anhydrous caffeine, cafeina, caféine, caféine anhydre, caféine benzodate de sodium, caffeine sodium benzoate, caffeine anhydrous, caffeine citrate, caffeinum, citrate de caféine, citrated caffeine, methylxanthine, méthylxanthine, and trimethylxanthine, triméthylxanthine. All of these names describe precisely the same chemical. The structure of caffeine is given in Figure 3.1 and a number of its common representations in Figure 3.2.

Systematic names were developed by the International Union of Pure and Applied Chemistry (IUPAC) to provide standard naming conventions for chemical structures. While systematic names do not encode the connectivity of the chemical structure itself, the representation encodes the substituents, carbon chain lengths and chemical endings. For example, a systematic name ending in 'ane' defines a single bonded carbon chain, such as hexane. Systematic names can easily become unwieldy for large and more complex structures, defeating the point of having a standardised representation system that is easily written and understood. However, many chemical information

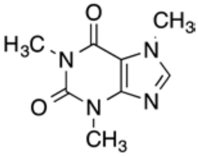
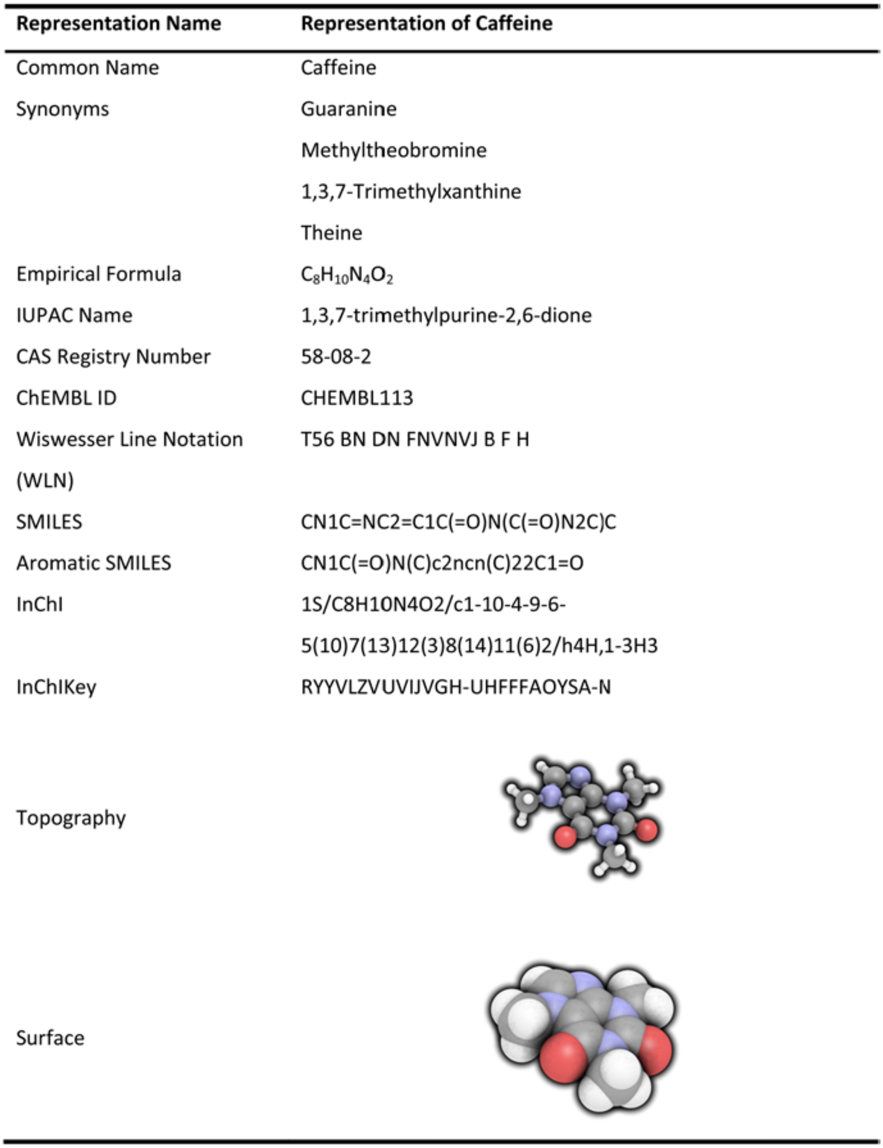


Figure 3.1 The chemical structure of caffeine.

systems can encode and decode systematic names such that they follow the naming conventions, but this representation is not commonly used in practice for computational work.

Chemical structures are the explicit representation and the chemist's *lingua franca*. Chemical structure drawings appear in many scientific publications,



Representation Name	Representation of Caffeine
Common Name	Caffeine
Synonyms	Guaranine Methyltheobromine 1,3,7-Trimethylxanthine Theine
Empirical Formula	$C_8H_{10}N_4O_2$
IUPAC Name	1,3,7-trimethylpurine-2,6-dione
CAS Registry Number	58-08-2
ChEMBL ID	CHEMBL113
Wiswesser Line Notation (WLN)	T56 BN DN FNVNVJ B F H
SMILES	CN1C=NC2=C1C(=O)N(C(=O)N2C)C
Aromatic SMILES	CN1C(=O)N(C)c2ncn(C)22C1=O
InChI	1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3
InChIKey	RYYVLZVUVIJVGH-UHFFFAOYSA-N
Topography	
Surface	

Figure 3.2 A list of commonly accepted different types of chemical structure representations, or simply names, for the chemical known commonly as caffeine.

whether in chemistry or in related fields of physics and biology. Although a chemist can look at a chemical structure picture, it is more difficult for a machine to process the graphical information into a chemical structure representation. However, software systems exist that can frequently and successfully extract the chemical structure information from structure images, such as CLiDE from Keymodule.

The free and open-access database of chemical patents uses a combination of name-to-structure and image-to-structure tools to extract chemical structure data. While this works well in practice, issues can occur in the translation of these data from simple issues in the way the representations have been written or drawn. This remains a challenge for systems such as SureChEMBL, a chemical patent database, since it applies an automated curation process. However, the ChEMBL and SureChEMBL team are working towards improving their processes.

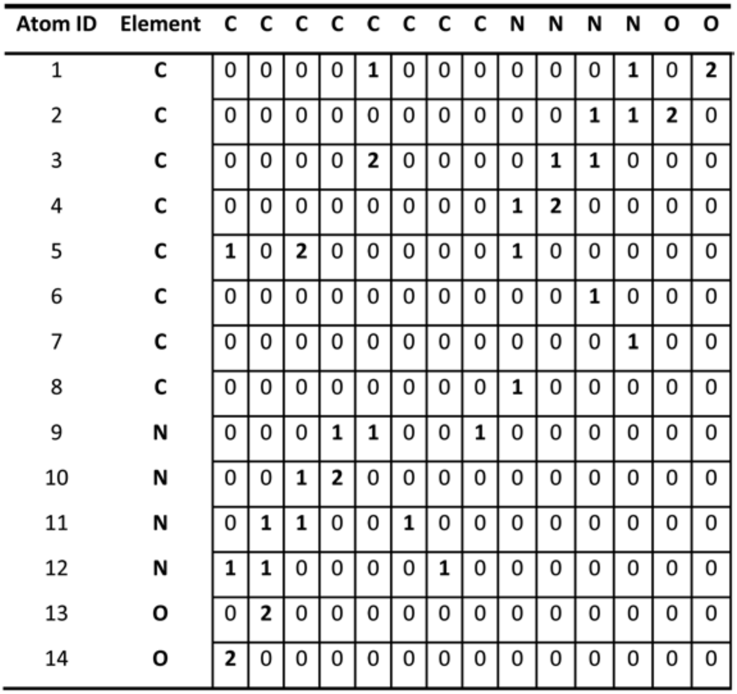
Sadly, much of the structure data information available in the scientific literature has largely been lost due to systematic names and chemical structure images, but recently the open-access and free database from ChEMBL has abstracted much of the chemical structure data available in the literature, together with associated metadata, such as biological assay readouts. This was no mean feat and has been on going for many years. Moves are now afoot in the chemical community to not lose these data but instead associate the chemical structure data explicitly with scientific publications and chemistry patents to avoid the challenges in reverse engineering of the chemical structure information from text and images.

The machine-readable structure representation systems are manifold, but it is a fundament of chemical information systems. What follows is an overview of the some of the more commonly used chemical structure representation systems today, with Molfiles (MOL or SDF) and SMILES arguably being the most common. This is by no means an exhaustive list, but the interested reader can find many different chemical structure representations online.

3.3 Adjacency Matrix

The adjacency matrix (AM) is perhaps the simplest computer-readable representation for molecules, although it is not often used as a file format, but is frequently used as an in-memory representation allowing rapid access to atom connectivity in a given molecule.

Given a molecule that contains *n* atoms, the adjacency matrix is a square ($n \times n$) matrix of variables (typically integers) that define the bonding between each atom. Typically, an *n*-length array would also be associated with an adjacency matrix to encode atom properties—often simply the elemental atom type, but additional properties may also be included, such as *xyz* co-ordinate information. An adjacency matrix representation of caffeine is given in Figure 3.3.



Atom ID	Element	C	C	C	C	C	C	C	C	N	N	N	N	O	O
1	C	0	0	0	0	1	0	0	0	0	0	0	1	0	2
2	C	0	0	0	0	0	0	0	0	0	0	1	1	2	0
3	C	0	0	0	0	2	0	0	0	0	1	1	0	0	0
4	C	0	0	0	0	0	0	0	0	1	2	0	0	0	0
5	C	1	0	2	0	0	0	0	0	1	0	0	0	0	0
6	C	0	0	0	0	0	0	0	0	0	0	1	0	0	0
7	C	0	0	0	0	0	0	0	0	0	0	0	1	0	0
8	C	0	0	0	0	0	0	0	0	1	0	0	0	0	0
9	N	0	0	0	1	1	0	0	1	0	0	0	0	0	0
10	N	0	0	1	2	0	0	0	0	0	0	0	0	0	0
11	N	0	1	1	0	0	1	0	0	0	0	0	0	0	0
12	N	1	1	0	0	0	0	1	0	0	0	0	0	0	0
13	O	0	2	0	0	0	0	0	0	0	0	0	0	0	0
14	O	2	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 3.3 An adjacency matrix representation of caffeine. The adjacency matrix is redundant, in that $a[i][j] = a[j][i]$, and $a[i][i]$ is always zero, since an atom may not be connected to itself. The total number of possible connections for any molecule, is given by the general formula $n(n - 1)/2$, in this case 91.

The adjacency matrix is typically represented in-memory as a redundant symmetric matrix that is equal to its transpose. This may be considered wasteful in memory requirements, since the memory required is twice that of a non-redundant implementation and both $a[i][j]$ and $a[j][i]$ must be updated consistently to retain validity of representation.

3.4 Connection Table

The connection table (CT) is a much more common file format representation used in everyday chemoinformatics applications. A connection table consists of lines of information regarding the atoms in a particular molecule, such as elemental type and *xyz* co-ordinate information. In addition to the atom lines, or the atom block, is the bond block that encodes the connections between each of the atoms according to the index.

By far the most used connection table format in modern day use is the MDL Molfile (extension *.mol) or structure-data file (SDF, extension *.sdf), the latter for multi-record files containing more than one structure.

The Molfile contains the atoms and the bonding patterns between those atoms, but also includes *xyz* co-ordinate information so the 3D structure can be explicitly encoded and stored for subsequent use. The file format was originally developed by MDL Information Systems, which through a number of acquisitions and mergers, Symyx Technologies and Accelrys, respectively, is now subsumed with Biovia, a subsidiary of Dassault Systems.

The Molfile is split into distinct lines of information, referred to as blocks. The first three lines of any Molfile contain header information: molecule name or identifier; information regarding its generation, such as software, user, *etc.*; and the comments line for additional information, but in practice this is often blank. The next line always encodes the metadata regarding the connection table and must be parsed to identify the numbers of atoms and bonds, respectively. The first two digits of this line encode the numbers of atoms and bonds, respectively.

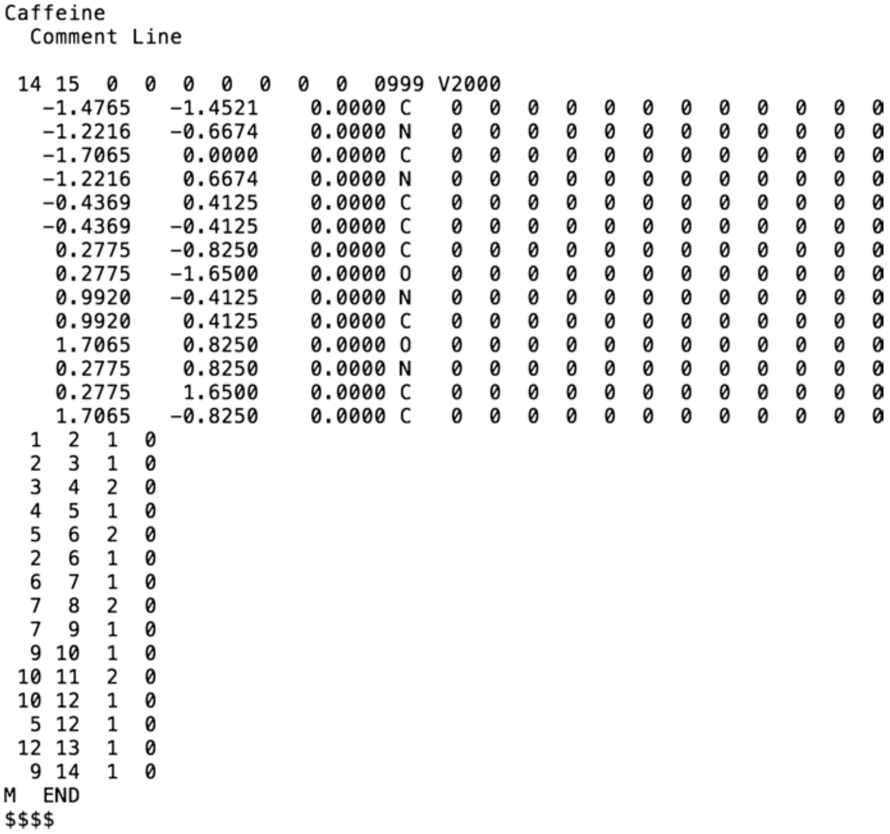
The atom block contains each of the atoms encoded in the Molfile, one atom per line. The standard format first encodes the *xyz* co-ordinates as real-valued data, followed by the elemental atom type of this particular atom. The atom type is followed by 12 atom property fields that can encode for a number of properties depending on the software used.

Directly after the last atom line in the atom block, the bond block begins. The first two values in each bond line inform the source and target atoms of each bond, the index given implicitly by the atom position in the atom block. The following digit encodes the bonding order or type: 1 = single bond, 2 = double bond, *etc.* The subsequent four bond property fields can encode for a number of properties depending on the software used.

For a Molfile, which always encodes a single molecular structure, although each record may contain many disconnected molecular structures, the structure record would then end. The SDF format wraps and extends the Molfile format. There are two distinct advantages in using SDF: incorporation of additional metadata and encoding multiple chemical structures in a single file.

In the SDF format, additional data is included by first defining the field name on a single line according to this format "> <NAME>", followed by a single line that contains the actual field data. The field name and data pairs can repeated as required for the number of data fields you may wish to encode. Since the SDF format is designed to contain multiple molecules, it is necessary to have a record delimited line so that the file parser can detect that one structure record has finished and a new one has begun. The record delimited in the SDF format is simply an additional line containing four dollar signs, "$$$$". An example of the SDF file format for the chemical structure of caffeine is given in Figure 3.4.

A key advantage of the Molfile and SDF formats is the inclusion of geometric information regarding the spatial arrangement of atoms in three-dimensional space. Furthermore, the hydrogen atoms may be defined explicitly to obviate the need for recalculation. These additional data, coupled with the additional metadata recorded in the SDF format, makes this file format ideal



```
Caffeine
  Comment Line

 14 15  0  0  0  0  0  0  0  0999 V2000
   -1.4765   -1.4521    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.2216   -0.6674    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
   -1.7065    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.2216    0.6674    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
   -0.4369    0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.4369   -0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.2775   -0.8250    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.2775   -1.6500    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.9920   -0.4125    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    0.9920    0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.7065    0.8250    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.2775    0.8250    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    0.2775    1.6500    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.7065   -0.8250    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0
  2  3  1  0
  3  4  2  0
  4  5  1  0
  5  6  2  0
  2  6  1  0
  6  7  1  0
  7  8  2  0
  7  9  1  0
  9 10  1  0
 10 11  2  0
 10 12  1  0
  5 12  1  0
 12 13  1  0
  9 14  1  0
M  END
$$$$
```

Figure 3.4 MOL/SDF file format for the caffeine molecule.

for any computational method that relies on geometry, such as pharmacophore or shape search (Chapter 7) and virtual ligand docking (Chapter 13).

However, this increased flexibility can come at the cost of storage space since the files will tend to be significantly larger than other file formats such as line notations (*vide infra*). Furthermore, manipulation of the data within the SDF format requires specialist chemoinformatics software, whereas it may often be easier to manipulate these data in generic data editors, such as a spreadsheet editor, but these lack the chemistry parsing ability.

3.5 Line Notations

Line notations are highly desirable structure representation methods as they fit within the alphanumeric string data often used in spreadsheets and rudimentary database systems. They tend to offer a compact representation of the constitution and connectivity of a topological representation of chemical structures, but tend to lack additional information, such as protonation and geometry, that is necessary for many modelling techniques. Here, three line notations will be introduced: Wiswesser Line Notation (WLN), Simplified Molecular-Input Line-Entry Specification (SMILES), and IUPAC International Chemical Identifier (InChI). Other line notations of note are Representation of Organic Structures Description Arranged Linearly (ROSDAL) and SYBYL Line Notation (SLN) from Tripos, Inc.

The linearisation of molecular structures means that one structure may have many different line notations in the same encoding scheme. This depends on the choice of starting atom, and decisions of direction and branching while translating the molecular graph into the line notation. Indeed, this is the case for any representation, including the previous ones described in this chapter, but is more pressing an issue in line notations. However, unambiguous representations are desirable, particularly for rapid molecular identity matching in, for instance, identifying duplicate molecules. Therefore, a number of canonicalisation schemes were investigated, with the Morgan algorithm becoming the *de facto* standard for structure canonicalisation.[1]

3.5.1 WLN: Wiswesser Line Notation

One of, if not the first, chemical structure line notations was that defined by William J. Wiswesser in 1949.[2] The WLN, rather than encode explicit bonding patterns, encoded fragments or groups. Wiswesser's reasoning for this was that he believed that the valence bond model would eventually be superseded by the molecular orbital representation.

WLN uses 41 symbols in its encoding system: the 10 numerals, 26 uppercase alphabetic characters, four punctuation symbols (&, -, /, *), and the blank space. All international elemental symbols are used, except for K, U, V, W, Y, Cl, and Br. Elemental symbols that contain two characters are enclosed within hyphens. Cl and Br are represented as G and E, respectively.

WLN was used mainly in registration, search and retrieval in chemical information systems. Perhaps the largest system that used WLNs was the CROSSBOW database system at ICI. With the advent of structure and substructure search systems, WLN fell out of favour and is now used in very few systems.

3.5.2 SMILES: Simplified Molecular-Input Line-Entry Specification

Arguably the most commonly used line notation is the SMILES string. The SMILES representation uses alphanumeric characters that closely mimic atoms and bonds as drawn in two-dimensional chemical structures. By mimicking these structural elements, it is easy to explain the SMILES encoding scheme and typically simple for an experienced human to roughly understand the chemical structure represented by a particular SMILES string at a simple glance.[3,4]

Atoms in a SMILES string are represented by their elemental symbol in the periodic table of the elements, within square brackets. However, the square brackets can be implicit for the organic subset of elements: ‘B’, ‘C’, ‘N’, ‘O’, ‘S’, ‘P’, ‘Br’, ‘Cl’, ‘F’, and ‘I’. The hydrogens are typically implicit, but can be defined in certain cases. An atom that contains one or more charges must be enclosed in square brackets followed by the ‘H’ symbol and number of hydrogens bonded to it—if only one then it may be omitted. Following this, a plus symbol represents a positive charge and a subtraction symbol represents a negative charge. The number of charges can be included after the charge symbol, with one charge again being implicit. The number of charges can also be included explicitly by additional charge symbols. Therefore, methane is simply ‘C’ and water ‘O’.

Bonds in a SMILES string are represented by symbols that mimic the chemical structure diagram representations: a single bond is ‘-’; a double bond is ‘=’; a triple bond is ‘#’; a quadruple bond is ‘$’; and an aromatic bond is ‘:’. However, bonds in a SMILES string are implied in a large number of cases. Bonds between aliphatic atoms are implicitly assumed to be single bonds and therefore the single bond symbol is not required. Therefore, ethanol, starting the SMILES string from the monovalent carbon, is written as ‘CCO’, but is equally valid as ‘C–C–O’. Bonds between aromatic atoms are implicitly assumed to be aromatic.

Branching in a SMILES string is defined by round brackets. Therefore, ethanol, starting from the divalent carbon in the middle of the structure, would be ‘C(C)O’, to indicate that the first carbon is bonded to both the second carbon atom and the oxygen atom.

Ring systems in a SMILES string are encoded by ring closure tags, which indicate that two atoms in the string are connected and therefore form a ring system. So, hexane would be ‘CCCCCC’, whereas cyclohexane would be ‘C1CCCCC1’. For a second ring, the ring closure tag would be ‘2’, and so on. If the number of ring closure tags needed exceeds ‘9’ then a percentage symbol must be used in front of the symbol. This is important since a single atom may encode two different ring closures, *e.g.* ‘−C12−’.

Aromaticity in a SMILES string is encoded by using the lowercase characters for carbon, nitrogen, oxygen, and sulphur: ‘c’, ‘n’, ‘o’, ‘s’, respectively. Therefore, cyclohexane, as we have already seen, is ‘C1CCCCC1’, whereas benzene is ‘c1ccccc1’. Aromatic bonds are implied between aromatic atoms, but may be explicitly defined using the ‘:’ symbol. An aromatic nitrogen bonded to a hydrogen must be explicitly defined as ‘[nH]’: pyrrole is ‘c1cc[nH]c1’ and imidazole is ‘c1cnc[nH]1’.

Stereochemistry in a SMILES string is encoded by the special characters ‘\’, ‘/’, ‘@’, and ‘@@’. Around two double bonds, the configuration specifies the *cis* and *trans* configurations. Therefore, valid SMILES strings of *cis*- and *trans*-butene are ‘C\C=C\C’ and ‘C\C=C/C’, respectively. The configuration around a tetrahedral carbon is specified by ‘@’ or ‘@@’. Therefore, the more common enantiomer of alanine, L-alanine, is ‘N[C@@H](C)C(=O)O’ and D-alanine is ‘N[C@H](C)C(=O)O’. The order

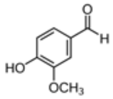
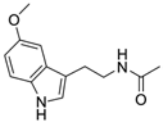
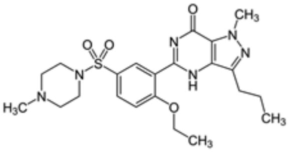
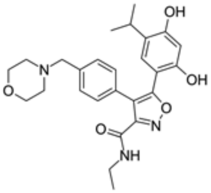
Name	Structure	SMILES
Vanillin		O=Cc1ccc(O)c(OC)c1
Melatonin		CC(=O)NCCC1=CNc2c1cc(OC)cc2
Sildenafil		CN1CCN(S(=O)(C2=CC=C(OCC)C(C3=NC4=C(N(C)N=C4CCC)C(N3)=O)=C2)=O)CC1
NVP-AUY922		CC(C)c1cc(c(O)cc1O)c2onc(C(=O)NCC)c2c3ccc(cc3)CN4CCOCC4

Figure 3.5 Examples of SMILES string representations for a number of common molecular structures.

of the substituents is also important, so D-alanine may also be written as 'N[C@@H](C(=O)O)C'.

Examples of some typical chemical structures encoded as SMILES with common names are provided in Figure 3.5.

Another language, based on conventions in SMILES, has also been developed for rapid substructure searching, called SMiles ARbitrary Target Specification (SMARTS). Similarly, SMIRKS has also been defined as a subset of SMILES that encodes reaction transforms. SMIRKS does not have a definition, but plays on the SMILES acronym. SMARTS and SMIRKS will be considered in more detail in later chapters.

3.5.3 InChI: IUPAC International Chemical Identifier

The IUPAC International Chemical Identifier (InChI™) is an international standard in structure representation based on an open standard, as opposed to the Chemical Abstracts Service (CAS) number. The first release date for the InChI standard was 15th April 2005 and it is now supported by the InChI Trust, a not-for-profit organisation.[5]

The InChI identifier provides a layered representation of a molecule to allow for the representation of differing levels of resolution depending on the application in mind. The layers defined by InChI are as follows:

- Main layer
    - Chemical formula, no prefix
    - Atom connections, prefix ‘c’
    - Hydrogen atoms, ‘h’
- Charge layer
    - Proton sublayer, ‘p’
    - Charge sublayer, ‘q’
- Stereochemical layer
    - Double bonds and cumulenes, ‘b’
    - Tetrahedral stereochemistry of atoms and allenes, ‘t’ or ‘m’
    - Stereochemistry information type, ‘s’
- Isotope layer, ‘I’, ‘h’, and ‘b’, ‘t’ and ‘m’ for stereochemistry of isotopes
- Fixed-H layer, ‘f’

In addition to the InChI representation, the standard providers have also published an InChIKey. An InChIKey is a hashed representation of an InChI and, as such, is not machine readable as a structure representation but is rather a structure identifier for rapid structure identity searching. Furthermore, there is the additional chance, albeit minimal, that two molecules will be represented by a single InChIKey, known as a hash collision. The InChIKey is a 27-character version of the InChI representation using the SHA-256 hashing algorithm. The InChIKey consists of 14 characters representing the hash code of the connectivity data in a given InChI, followed by a minus sign and a further 9 characters representing the hash code of the remaining layers in that InChIKey. While the InChI representation is normally too complex for a human to decode, it is impossible for even a computer to extract the chemical structure from the InChIKey. Therefore, it is important that the InChI representation is also included in any database. InChIKey resolutions to InChI representations are also available from NCI, PubChem and ChemSpider.[6]

3.6 Summary

Structure representations are one of the fundamental concepts in computational chemistry and chemoinformatics. Without an appropriate structure representation, ambiguities can arise as to the actual chemical structure being represented. This chapter has defined some of the more common chemical structure representations used in modern chemical information systems. Subsequent chapters will consider what one can do with the chemical structures once they are represented in a machine-readable form, but the structure representations themselves are fundamental to all of these applications. Many of the popular chemical structure representations are supported by both open-source and commercial software, including the open-source RDKit API.

References

1. H. L. Morgan, The Generation of a Unique Machine Description for Chemical Structures – A Technique Developed at Chemical Abstracts Service, *J. Chem. Doc.*, 1965, **5**(2), 107–113.
2. W. J. Wiswesser, How the WLN began in 1949 and how it might be in 1999, *J. Chem. Inf. Comput. Sci.*, 1982, **22**(2), 88–93.
3. D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules,*J. Chem. Inf. Comput. Sci.*, 1988, **28**(1), 31–36.
4. D. Weininger, A. Weininger and J. L. Weininger, SMILES. 2. Algorithm for the generation of unique SMILES notation,*J. Chem. Inf. Comput. Sci.*, 1989, **29**(2), 97–101.
5. A. McNaught, The IUPAC International Chemical Identifier: InChI – A New Standard for Molecular Informatics, *Chem. Int.*, 2006, **28**(6), 12–15.
6. S. Heller, A. McNaught, S. Stein, D. Tchekhovski and I. Pletnev, InChI – the worldwide chemical structure identifier standard,*J. Cheminf.*, 2013, **5**(7), DOI: 10.1186/1758-2946-5-7.

CHAPTER 4

Molecular Similarity

4.1 Overview

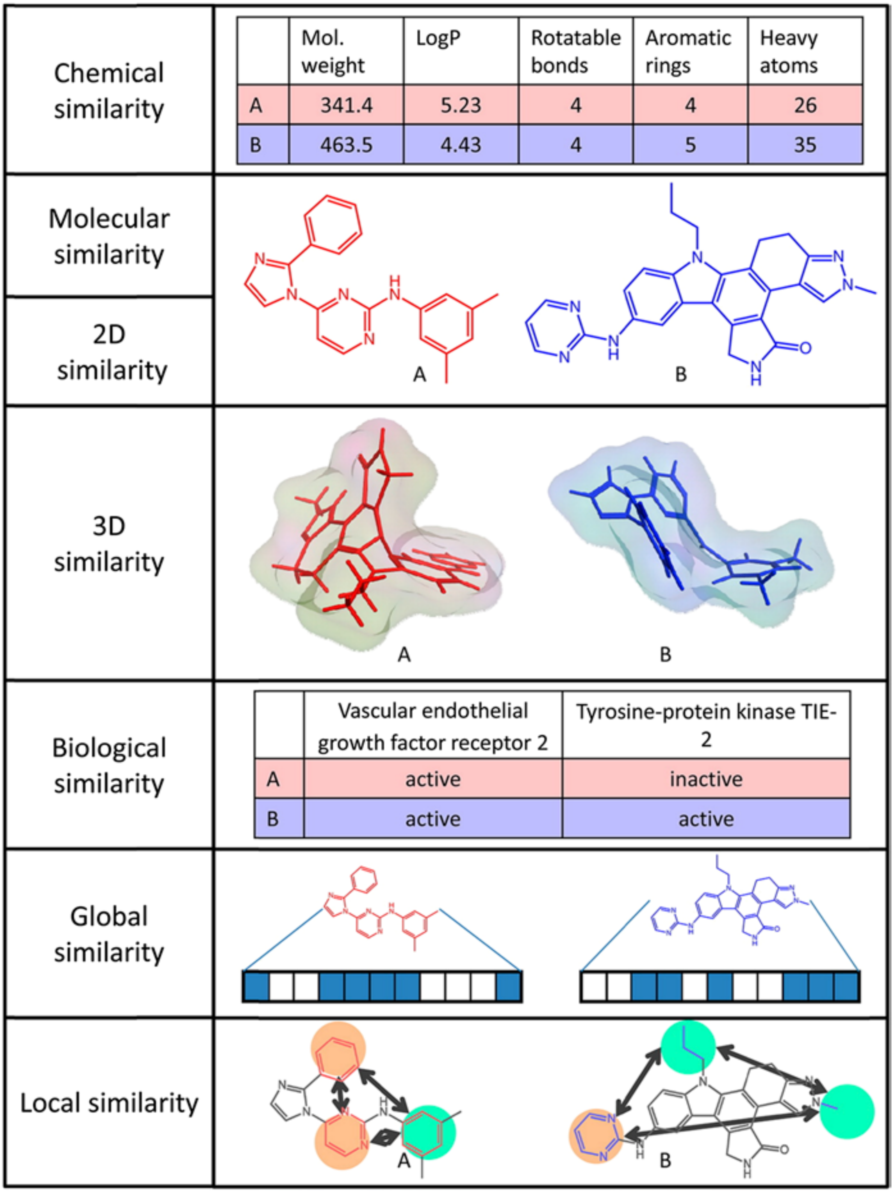
The concept of molecular similarity is important and arguably core to the field of computational medicinal chemistry. However, as is often the norm for such key concepts, molecular similarity is highly subjective and context dependent. Ultimately, the only type of molecular similarity that counts in drug discovery is the biological response. Therefore, two molecules may be significantly different in structure, but so long as they interact biologically in a similar way, they will said to be similar. However, this type of molecular similarity assumes more knowledge than is likely to be derived from the ligands alone.

When designing new drugs, one must often rely on similarity to a reference ligand alone when a protein–ligand crystal structure complex is not available. The structural or functional similarity of two ligands can assist in understanding whether they are likely to exhibit similar biological similarity. The *similar property principle* is a concept that defines that, if two ligands are similar, they will also tend to have similar properties. Of course, there are many situations where this *rule-of-thumb* (or *heuristic*) breaks down, but in general the rule holds and can be applied effectively to many challenges in drug discovery. When the similar-property principle does break down, the effect is often referred to as an *activity cliff*. When an activity cliff occurs where the only difference between two chemical structures is a single methyl group, the effect is unscientifically defined as a *magic methyl*. The different classes of molecular similarity are illustrated in Figure 4.1.[1]

The inherent subjectivity of molecular similarity necessitates more objective measures of similarity that are invariant in application. For example,

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



	Mol. weight	LogP	Rotatable bonds	Aromatic rings	Heavy atoms
A	341.4	5.23	4	4	26
B	463.5	4.43	4	5	35

	Vascular endothelial growth factor receptor 2	Tyrosine-protein kinase TIE-2
A	active	inactive
B	active	active

Figure 4.1 Similarity perception and concepts. Two exemplary vascular endothelial growth factor receptor 2 ligands are shown, and different ways to assess their similarity are illustrated. Reprinted with permission from G. Maggiora, M. Vogt, D. Stumpfe and J. Bajorath, Molecular Similarity in Medicinal Chemistry: Miniperspective, *J. Med. Chem.*, 2012, **57**(8), 3186–3204. Copyright 2012 American Chemical Society.

two chemists can easily disagree on the molecular similarity of two chemical structures. One chemist may be more concerned with synthetic accessibility and see the similarity, or lack thereof, in this light. However, the second chemist may be more concerned with how the structure may perform in the physiological environment and observe similarities or dissimilarities that relate to its solubility, potential for it to be rapidly metabolised, or likely issues in toxicity. Indeed, one can often ask the same chemist, at different times, to provide a measure of similarity, and they will often suggest differing measures of similarity.

4.2 Molecular Similarity

The concept of molecular similarity is highly subjective, even philosophical in nature. Two expert medicinal chemists may disagree on the degree of molecular similarity of a set of chemical structures. Surprisingly still, the same expert medicinal chemist may even disagree with themselves on different days and given specific challenges. However, when dealing with computational approaches, we seek objective and invariant methods. Objectivity in the form of seeking unbiased approaches and invariance such that the results will not change over time so that comparisons may be made.[2]

Molecular similarity may be considered in terms of the connectivity of the atoms, which is called topological similarity. Many similarity methods work on the principle of topological similarity. Topological similarity is very useful for identifying chemical structures that are similar, or analogues, in terms of the structural space they occupy. Therefore, similar topological structures will naturally sit together in chemistry space allowing such methods as analogue-by-catalogue in which one can purchase or synthesise compounds to explore the structure–activity relationship (SAR) of chemical structures around a given chemotype or molecular scaffold.

Property similarity is another approach to molecular similarity in which the chemical properties define the similarity of two given molecules. For example, two chemical structures may have the same molecular weight, but could be entirely different in their chemical constitution. The extreme case here would be classing all Lipinski-rule (Chapter 5) compliant drugs as similar since they fulfil the property criteria, which in some ways is a truth given that they match these criteria. However, this is not a form of similarity that is terribly useful.

Topological similarity is an effective measure of structural similarity between two molecules, but molecules also have shape, which is a very important consideration in drug design. For example, two structures may be structurally similar, differing in only one position, a methyl group for example. Although ostensibly highly similar, the methyl group could force a conformation change or stabilisation due to intramolecular clashes or interactions, respectively. It is often desirable to pre-organise the conformations of molecules so as to minimise the entropic penalty upon binding. However, the consideration of geometry introduces an additional challenge, often referred to as the conformer problem. Here, one must explicitly consider the different shapes that the molecules under investigation may adopt, vastly increasing the number of comparisons needed to identify similarity and therefore concomitantly increasing the runtimes of the process. In this way, molecules are not only three-dimensional, but also four-dimensional, moving in time as well as space. The conformer problem is a significant ongoing challenge in drug discovery and careful application of conformer search and analysis must be undertaken to ensure that one is not simply introducing noise into your model system.

A further approach to measuring intermolecular similarity is through the use of pharmacophores. A pharmacophore is the hypothesised or known requirements for a molecule to interact with, for example, its protein target binding site. A pharmacophore is therefore an abstract model defining the necessary functional binding elements. Pharmacophores are often defined geometrically, that is in the spatial arrangement that one would require for favourable interactions. However, pharmacophores may also be defined topologically with the geometric arrangement being implied in the through-graph distances between structural features. These *Ligand-Based Topological Pharmacophores* will be discussed in more detail in Chapter 6 on Topological Descriptors.

The last form of similarity that we will consider in this chapter is that of biological similarity, which is essentially our goal in the context of drug discovery. Here, we look at the relevant biological endpoints, such as enzyme potency in terms of its inhibitory concentration, or IC_{50}. In the case of biological similarity, the structures may be entirely unrelated, potentially not being recognised as similar by any of the previously discussed measures of similarity. However, the assumption must be that they act on their target in a similar way so as to still remain comparable. The identification of biological similarity with substantially diverse chemical matter is very important in drug discovery. Firstly, it is often prudent to have at least one back-up chemical series in a drug discovery programme for situations where late-stage attrition may lead to the primary chemical series failing, for example, for toxicity reasons. Secondly, having diverse molecules that act on the same target is beneficial as chemical tools to assist in chemical biology and deconvoluting the effect on the target of interest.

Topological similarity considers only the molecular structure of the molecules being considered and highlights structures that are significantly similar in appearance. This is useful in identifying close analogues to a known compound of interest. Property similar can lead to identifying very diverse structures that share the same properties. This can be useful in designing focussed or diverse molecular libraries. 3D similarity considers the actual shape, or shapes when considering multiple conformers, of the molecules under consideration. This is useful when identifying similar or dissimilar shapes for library designs, and can help identify structures that would otherwise be reported as topologically similar, but do not adopt similar conformations due to a steric hindrance. Pharmacophoric similarity can be calculated from both topological and geometric structures and encode the pharmacophores, or feature points that are likely to contribute positively, typically, to binding. This, often including geometric similarity as in ROCS (Rapid Overlay of Chemical Structures), can often be a highly effective virtual screening strategy for identifying both analogues and structurally diverse molecules. The conformer problem is also a challenge here since multiple and appropriate conformers must be generated to appropriately consider similarity. The last method reported here is the biological similarity, which is simply the difference between the biological endpoints of interest. This can be useful for identifying scaffold hops (see Chapter 10), for example, where structures are identified that are structurally different but are similar in terms of biological activity.

4.3 Similar Property Principle

The similar property principle states that chemical structures that are highly structurally similar will tend to exhibit similar properties.[3] These properties may be biological affinity or physicochemical properties, such as aqueous solubility. The similar property principle was also discussed as neighbourhood behaviour, but in the context of evaluating diversity measures and selection methods.[4] A schematic diagram of neighbourhood behaviour is provided in Figure 4.2.

In Figure 4.3, the chemical structures of (a) morphine, (b) codeine, and (c) heroin exhibit highly similar chemical structures and also have similar therapeutic effects. These offer an illustrative example of how molecular structure similarity can be related to functional or therapeutic similarity.

While the similar property is straightforward and intuitive in concept, it is merely a concept and therefore qualitative at best. Therefore, much research has been undertaken into quantifying the degree of similarity that is sufficient between chemical structures to give an acceptable probability that they will exhibit the characteristics of this principle.

Brown and Martin reported an empirically derived threshold cut-off for molecular similarity in structural clustering studies.[5] The authors report a Tanimoto coefficient cut-off of 0.85 to represent the similarity (or dissimilarity) value above which the structure pairs are deemed to be highly structurally similar with an 80% probability of similar activity. However, one must express caution in using such a cut-off, as they will tend to not take into account structural decoration, which could introduce a steric or electronic clash that is unfavourable to binding.[6] Furthermore, the cut-off is derived empirically for the Tanimoto coefficient and for a single molecular descriptor, the UNITY fingerprint as implemented by Tripos (now Certara). Recently, Maggiora *et al.* have dismissed this cut-off as the ‘0.85 myth’ mainly due to the lack of understanding of molecular similarity coefficients and descriptors in medicinal chemistry.[1] Therefore, as always, due caution and appropriate experimental controls should be employed.

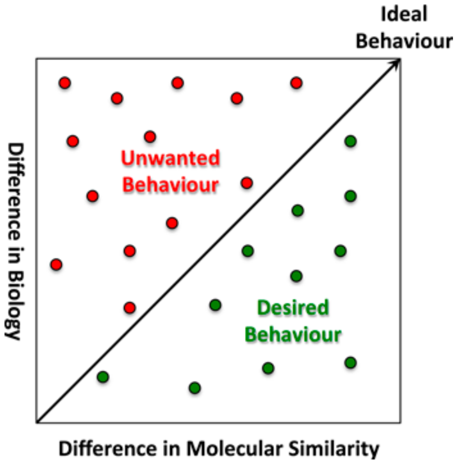


Figure 4.2 A schematic diagram of neighbourhood behaviour demonstrating the ideal case, the line of identity; desirable characteristics, the lower triangle of the plot; and undesirable characteristics, the triangle above the line of identity. Ideally, small changes in molecular similarity should represent small changes in biological endpoint too. Large changes in descriptor *versus* small or large changes in the biological endpoint are desirable in a medicinal chemistry programme since the molecular descriptor can be tuned to the biological end point more easily. This would lead to the lower half of the plot being populated. If large changes in the biological endpoint only relate to small changes in molecular similarity, it would be difficult to optimise the molecular structures using the biological endpoint since the optimisation would be conducted within the noise of the descriptor.

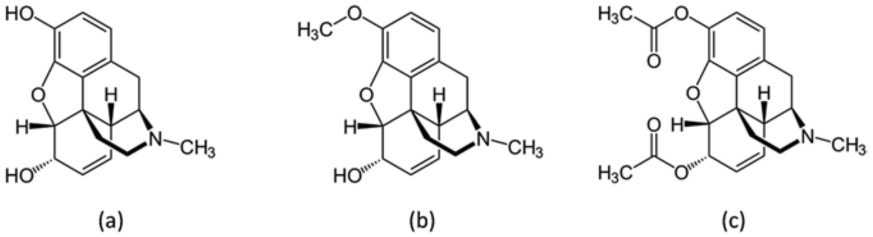


Figure 4.3 The chemical structures of (a) morphine, (b) codeine and (c) heroin exhibit highly similar chemical structures and also have similar therapeutic effects.

4.4 Molecular Descriptors

The focus of this chapter is the concept of molecular similarity, but molecular similarity is often calculated from molecular descriptors, rather than the actual molecular structures. Exceptions to this exist, such as the degree of molecular graph overlap between two chemical structures, or the maximum common substructure (MCS). Molecular descriptors will be covered extensively in the following section, but it is important to refer to some aspects of molecular descriptors in the context of molecular similarity.

Many different molecular descriptors exist and have been developed for different motivations. The Daylight fingerprint system was not originally designed as a molecular descriptor, but as a vector representation that could be used for efficient screen-out in substructure searching. Latterly, however, scientists have recognised the power of fingerprint descriptors, such as the Daylight fingerprint, as effective and efficient molecular descriptors that encode entirely and invariantly the molecule they represent. A great deal of the molecular fingerprint analysis work was conducted in the laboratory of Peter Willett at The University of Sheffield.

4.5 Calculation of Molecular Similarity

4.5.1 Similarity Coefficients

A graph-based calculation of two chemical structures can be achieved using graph-theoretic algorithms, but this is only one approach to calculating molecular similarity, and can often be too computationally intensive when considering large chemical structure databases. For this reason, molecular similarity is often calculated on molecular descriptors that encode aspects of the chemical structure. Oftentimes, molecular similarity is calculated on binary molecular fingerprints, of which there are many (Chapter 6). The calculation of the similarity between two molecular fingerprints is achieved by means of a similarity coefficient and is important in many chemoinformatics applications.[7]

The most commonly used molecular similarity coefficient in chemical information systems is the Tanimoto (or Jaccard) coefficient, although many have been reported in the literature to be useful in different circumstances. The Tanimoto coefficient is, strictly speaking, an association coefficient, which are most commonly considered with binary data and tend to be normalised in the range zero (no similarity) to one (complete identity). A summary of the most commonly applied similarity and dissimilarity coefficients is provided in Table 4.1 for both the continuous and dichotomous descriptors for integer or real-valued data vectors and binary descriptors, respectively.

One common misunderstanding in the medicinal chemistry community is the use of the term Tanimoto similarity. This is largely due to the '0.85 myth' discussed earlier, where the Tanimoto cut-off was considered as a method to identify molecules that are likely to maintain biological activity. The issue with referring to the Tanimoto similarity as an approach is that it does not specify the descriptor under comparison. Different fingerprints can give vastly different similarity values, mainly due to the numbers of bits set (in fingerprints) or non-zero variables (in continuous data) leading to very sparse fingerprints, in the case of Morgan fingerprints, as compared

Table 4.1 List of similarity coefficients used widely in calculating molecular similarity from molecular descriptors.

Name	Continuous	Dichotomous
Tanimoto coefficient	$T(x_a,x_b) = \dfrac{\sum_{i=0}^{N} x_{ai}\cdot x_{bi}}{\sum_{i=0}^{N} x_{ai}^{2} + \sum_{i=0}^{N} x_{bi}^{2} - \sum_{i=0}^{N} x_{ai}\cdot x_{bi}}$	$T(x_a,x_b) = \dfrac{c}{a+b-c}$
Euclidean distance	$D(x_a,x_b) = \sqrt{\sum_{i=0}^{N} (x_{ai}\cdot x_{bi})^2}$	$D(x_a,x_b) = \sqrt{a+b-2c}$
Hamming distance	$D(x_a,x_b) = \sum_{i=0}^{N} \lvert x_{ai} - x_{bi}\rvert$	$D(x_a,x_b) = a+b-2c$
Cosine coefficient	$C(x_a,x_b) = \dfrac{\sum_{i=0}^{N} x_{ai}\cdot x_{bi}}{\sqrt{\sum_{i=0}^{N} x_{ai}^{2}}\cdot\sqrt{\sum_{i=0}^{N} x_{bi}^{2}}}$	$C(x_a,x_b) = \dfrac{c}{\sqrt{ab}}$
Dice coefficient	$D(x_a,x_b) = \dfrac{2\sum_{i=0}^{N} x_{ai}\cdot x_{bi}}{\sqrt{\sum_{i=0}^{N} x_{ai}^{2}}\cdot\sqrt{\sum_{i=0}^{N} x_{bi}^{2}}}$	$D(x_a,x_b) = \dfrac{2c}{a+b}$
Soergel	$S(x_a,x_b) = \dfrac{\sum_{i=0}^{N} \lvert x_{ai} - x_{bi}\rvert}{\sum_{i=0}^{N} \max(x_{ai},x_{bi})}$	$S(x_a,x_b) = \dfrac{a+b-2c}{a+b-c}$

with Daylight-style path fingerprints, which tend to be much denser in terms of bits set.

The differences between similarity measures used together with different molecular descriptors are illustrated in Figure 4.4.[1] Here, the first graph illustrates the differences in similarity distributions between using the Tanimoto (blue) and Dice (green) coefficients using only the MACCS structure key fingerprint. It is clear in this figure that with Tanimoto it would be expected that, on average molecular similarities would be approximately 0.2 less similar than would be assumed when compared with the Dice coefficient on the same fingerprint. Similarly, in the second graph, the difference in distributions of the similarities, using Tanimoto (blue) and Dice (green) again, illustrates a substantial difference between the distributions when using the Extended Connectivity Fingerprint (ECFP4) fingerprint, but here the average difference in similarity is only around 0.1. Furthermore, the distribution of the Tanimoto similarities is much tighter, which explains to some extent the smaller difference in similarity distributions, but also highlights how the descriptor can offer very different values for similarity and also the expected ranges. Clearly, from these two examples it is important to understand what is being compared and in what way. Both the similarity coefficient and the molecular descriptor have a marked effect on the level of similarity, which highlights that caution should be used when working with molecular similarity methods. Additionally, it is important to ensure that when such values are presented to other scientists the values and what they

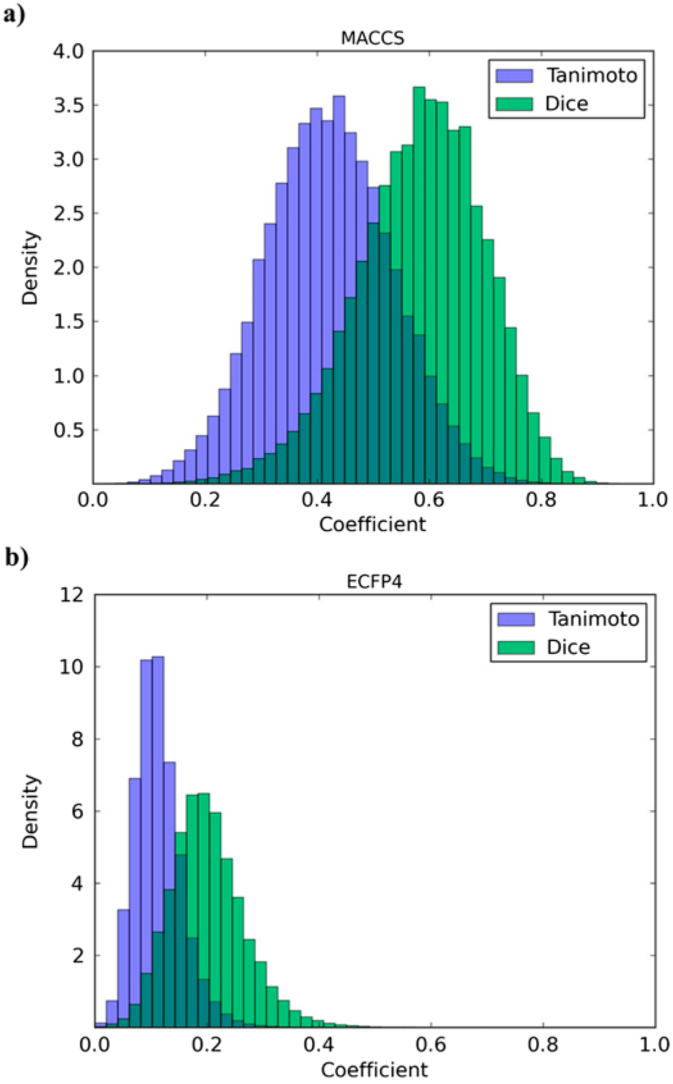


Figure 4.4 Similarity coefficient distributions. Distributions of similarity values resulting from 10 million comparisons of randomly chosen ZINC compounds are reported for the Tanimoto and Dice coefficient and the (a) MACCS and (b) ECFP4 fingerprint. Reprinted with permission from G. Maggiora, M. Vogt, D. Stumpfe and J. Bajorath, Molecular Similarity in Medicinal Chemistry: Miniperspective, *J. Med. Chem.*, 2012, **57**(8), 3186–3204. Copyright 2012 American Chemical Society.

mean are clearly articulated to ensure that the 'Tanimoto similarity' myth is not propagated.

One must express caution when considering what Tanimoto similarity may mean in its context of application and ensure that the descriptor is also included in the understanding and dissemination of any analysis.

For a coefficient to be formally referred to as a metric, it must obey a number of defined conditions. The four conditions necessary to be confirmed as a metric are:

1. Non-negativity: $d_{xy} \geq 0$
2. Identity of discernibles: $d_{xy} = 0$, if and only if $x = y$
3. Symmetry: $d_{xy} = d_{yx}$
4. Triangle inequality: $d_{xz} \leq d_{xy} + d_{yz}$

Should a coefficient fulfil each of these conditions, it can be said that the comparisons between objects are embedded in metric space and induce a topology on this space.

The Euclidean, Hamming and Soergel distances all fulfil these conditions. Additionally, the complements of the Tanimoto, Dice and Cosine coefficients also fulfil all four conditions except for the triangle inequality, although the binary complement of the Tanimoto coefficient does.

The Tanimoto coefficient is by far the most widely used coefficient in molecular similarity search, although there is no clear reason for this being the case. There is some benefit in terms of a lack of size dependence where larger molecules may be scored more highly than smaller molecules, due to the numbers of bits set in each.

4.6 Molecular Diversity

The opposite of molecular similarity, molecular diversity (or dissimilarity), is also a key concept in drug design. Often there are many more possible chemical structures—virtual or physical—than could possibly be synthesised and tested. Here, molecular diversity methods can be applied to select chemical structures that represent the diversity of the chemistry space under consideration, but with far fewer actual structures. We will return to clustering and diversity selection in Chapter 11, but we will discuss it in brief here in the context of molecular similarity.

Molecular diversity is important in many different endeavours in drug design: screening library design, triaging hitlists from High-Throughput Screening (HTS), and also selecting structures from virtual libraries to prioritise for purchase or synthesis. Here, the approach is not to identify a set of similar molecules to prioritise for further analysis, but to select a subset that represents the entirety of the space under consideration. The anticipation is that the set will represent the distribution of chemical matter over the chemistry space to a greater extent, compared to a random sample, and therefore allow a greater exploration of the space with few molecular structures considered.

4.7 Summary

The concept of molecular similarity is key in the field of drug discovery and fundamental in computational methods that deal with chemical structures. There is no single measure of molecular similarity that is appropriate for all applications and the users of these methods select and appropriately apply methods that meet their necessary criteria.

Molecular similarity is an important concept that is used in many of the methods discussed in greater detail throughout this book. Therefore, it is important that one takes time to thoroughly understand what the different methods offer. The next section covers a wide range of molecular descriptors, from which one may calculate molecular similarities for application to new challenges in drug discovery.

References

1. G. Maggiora, M. Vogt, D. Stumpfe and J. Bajorath, Molecular Similarity in Medicinal Chemistry: Miniperspective, *J. Med. Chem.*, 2012, **57**(8), 3186–3204.
2. P. Willett, J. M. Barnard and G. M. Downs, Chemical similarity searching, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 983–996.
3. M. A. Johnson and G. M. Maggiora, *Concepts and Applications of Molecular Similarity*, John Wiley & Sons, New York, 1990.
4. D. E. Patterson, R. D. Cramer, A. M. Ferguson, R. D. Clark and L. E. Weinberger, Neighborhood behavior: a useful concept for validation of “molecular diversity” descriptors, *J. Med. Chem.*, 1996, **39**, 3049–3059.
5. R. D. Brown and Y. C. Martin, Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 572–584.
6. Y. C. Martin, J. L. Kofron and L. M. Traphagen, Do structurally similar molecules have similar biological activity? *J. Med. Chem.*, 2002, **45**, 4350–4358.
7. A. Bender and R. C. Glen, Molecular similarity: a key technique in molecular informatics, *Org. Biomol. Chem.*, 2004, **2**, 3204–3218.

Part 3
Molecular Descriptors

CHAPTER 5

Molecular Property Descriptors

5.1 Overview

Molecular and physicochemical properties are some the most important molecular descriptors that are used in drug discovery projects. Physicochemical properties are calculated from measured phenomena of extant compounds, from which statistical models can be empirically derived. Physicochemical properties can also provide the basis for further statistical models through combination of these descriptors with others and using statistical learning methods (Chapter 8). Furthermore, physicochemical descriptors are often used as rules-of-thumb in drug discovery when linked to empirically observed undesirable traits in druglike molecules.

In this chapter, a number of physicochemical descriptors will be described, along with their utility in drug discovery and methods by which they are calculated. The importance of understanding the limitations of physicochemical descriptors, and indeed any empirical model, will be covered and approaches to ensuring these limitations are taken into account in their application.

5.2 Molecular Weight (MW or MWt)

The calculation of the molecular weight of a given molecular structure is arguably the most accurate of all physicochemical properties in chemistry, since it is the sum of the mass of each atom in a molecule. Molecular weight has been shown to be correlated with liabilities in the development of a new drug, with larger molecules typically more likely to have undesirable properties. This is simply because larger molecules will have more chance of containing undesirable moieties.

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

Table 5.1 Atomic properties of common organic chemistry elements found in drugs, used as a lookup table to calculate molecular weights from molecular structures.

Atomic no.	Element	Symbol	Atomic mass
6	Carbon	C	12.011
7	Nitrogen	N	14.007
8	Oxygen	O	15.999
9	Fluorine	F	18.998
15	Phosphorus	P	30.974
16	Sulphur	S	32.065
17	Chlorine	Cl	35.453
35	Bromine	Br	79.904

MW is a key physicochemical descriptor used in the heuristics known as Lipinski's Rule-of-Five (Chapter 6) and also used in calculating a type of Ligand Efficiency (LE), known as the Binding Efficiency Index (BEI), which normalises the potency of a particular ligand by its mass. The argument in drug design is that every atom added should contribute efficiently to the potency of the ligand and not simply add weight without any perceived benefit.

Simple rules-of-thumb based on relatively simple physicochemical descriptors are very useful, but it is important to ensure they are applied appropriately and are not used without thought and consideration.

The calculation of molecular weight is achieved by simply summing the molecular masses of each of the different heavy atoms, therefore omitting hydrogen, contained in a molecular structure. The masses of the common atoms present in synthetic organic chemistry are given in Table 5.1. The resultant molecular weight value can then be applied in a number of applications from simple cut-offs, such as the Lipinski heuristic of 500 Da, to incorporation in statistical learning models for the prediction of other properties.

5.3 Octanol/Water Partition Coefficient (ClogP)

Perhaps one of the most over-used physicochemical descriptors is logP. It has been identified as a property that can indicate whether a particular chemical structure may have liabilities in late-stage development, based on historical data. Regardless of the application of logP as a surrogate for other properties, whether they be other physicochemical properties or historical trends, the descriptor itself is very useful. logP is the partition coefficient as a ratio of the two concentrations of an unionised compound in two liquid phases, typically water and octanol. The logP value for a particular compound is given by the following equation:

$$\log P_{\text{octanol/water}} = \log\left(\frac{[\text{solute}]_{\text{octanol}}}{[\text{solute}]_{\text{water}}}\right) \tag{5.1}$$

Simply put, the logP measurement indicates the extent to which a given molecule is hydrophilic (water-loving) and hydrophobic [water-fearing, or lipophilic (grease-loving)]. The logP of a molecule can have an effect on drug administration, absorption, transport and excretion. logP is often used as a crude surrogate descriptor for aqueous solubility.

Many different approaches to calculate the logP of a given molecular structure have been proposed. Any calculated logP value is generically called the ClogP, but it is important to state the specific method used in any communication to ensure that it is clear which method has been employed and any limitations identified.

Here, only one method of calculating logP will be discussed, the Wildman–Crippen model,[1] as implemented in the RDKit API.[2] The Wildman–Crippen ClogP model is an atom contribution method of 68 empirically derived atom contributions from a training set of 9920 molecules. The correlation of the experimentally derived logP values and the ClogP using the Wildman–Crippen model was reported to be $R^2 = 0.918$ with $\sigma = 0.677$, representing a highly predictive model. The 68 atomic contributions were derived using only the atom patterns of atoms commonly seen in drug molecules, C, H, N, O, S, P and halogens, and also includes noble gases and metals. Each atom present in a molecule will only match a single atom type by design to ensure there is no ambiguity in the typing system. The atom types are given in the original paper by Wildman and Crippen [1999], including their SMARTS representations for easy re-implementation in another system.

The ClogP model and the Molecular Refractivity model, derived by Wildman and Crippen, have been implemented in the RDKit API, which is freely available.

5.4 Topological Polar Surface Area (TPSA)

Polar Surface Area (PSA) is an important physicochemical descriptor used in drug discovery as a surrogate descriptor for cell permeability. PSA is defined as the sum of the surface area of all polar atoms in a molecule, typically oxygen and nitrogen, including their connected hydrogens. PSA is most frequently used in drug design as a surrogate property for cell permeability with a rule-of-thumb that molecules with a PSA of less than 140 Å^2 would be able to permeate cells. PSA is also used as a surrogate for penetrating the blood–brain barrier (BBB), where a PSA of less than 90 Å^2 is often needed. BBB penetration is a key property in central nervous system (CNS) drug development.[4]

The PSA of a molecule can, of course, be calculated from the three-dimensional structure of a molecule, but its calculation can be quite computationally intensive, particularly due to the need for a 3D conformation. Ertl *et al.* therefore developed a predictive model for PSA that only requires the topological structure of a molecule for its calculation.[3] The result is a rapid descriptor calculator and, more importantly, a highly predictive and reliable statistical model that can be performed over many millions of structures very quickly. The PSA model from Ertl *et al.* is called the Topological Polar Surface Area (TPSA) and has been implemented in the RDKit API.

The TPSA of a given molecular structure is calculated simply by first identifying those atoms that contribute to the polar surface area using a simple look-up table of these atoms. For each occurrence of a particular atom and its environment (or bonding pattern), the values in this look-up table (Table 5.2) are simply summed.

Topological polar surface area is a rapidly calculable descriptor that can be applied in the drug discovery setting as an empirically appropriate surrogate for cell permeability. As with all descriptors, caution must be demonstrated in its application, but this TPSA has been demonstrated to be of great utility in drug design and it is now used commonly in designing new compounds as well as new screening libraries for high-throughput screening.

5.5 Hydrogen Bond Acceptors and Donors (HBA and HBD)

There are many way of calculating the number of hydrogen bond acceptors and donors in a given molecular structure. The actual numbers are dependent on the context of the potential acceptors and donors in the molecule, and the pH (pK_a), and this can complicate the calculations. Strictly speaking, the H-bond donor is the electron lone pair acceptor, and the H-bond acceptor is the electron lone-pair donor.

For simplicity, Lipinski,[5] for his druglike heuristics, defined the numbers of hydrogen bond donors as the sum of nitrogen–hydrogen and oxygen–hydrogen bonds, whereas the number of hydrogen bond acceptors was defined as the sum of all nitrogen and oxygen atoms.

5.6 Lipinski's Rule-of-Five

In the late 1990s, Chris Lipinski, then at Pfizer, began investigating historical data regarding the oral bioavailability of drugs. In his studies, Lipinski identified that the vast majority of orally bioavailable drugs were small and moderately lipophilic. Therefore, Lipinski was able, using this historical data, to define the following heuristics for druglikeness (or, more properly, oral bioavailability):

1. Molecular mass (molecular weight) of less than 500 daltons.
2. Octanol–water partition coefficient (logP) no greater than five.
3. No more than five hydrogen bond donors, counted as the sum of all nitrogen–hydrogen and oxygen–hydrogen bonding pairs.
4. No more than ten hydrogen bond acceptors, counted as the sum of all nitrogen and oxygen atoms.

These rules (or heuristics) are called the Lipinski Rule-of-Five, since each of the parameters is a multiple of five. Oftentimes, the number of rotatable bonds is also included as a parameter, with ten or fewer rotatable bonds

Table 5.2 Atom and bonding pattern contributions (Å^2) to polar surface area.[a]

Atom type	Contribution (Å^2)
[N](-*)(-*)-*	3.24
[N](-*)=*	12.36
[N]#*	23.79
[N](-*)(=*)=*[b]	11.68
[N](=*)#*[c]	13.60
[N]1(-*)-*-*-1[d]	3.01
[NH](-*)-*	12.03
[NH]1-*-*-1[d]	21.94
[NH]=*	23.85
[NH2]-*	26.02
[N+](-*)(-*)(-*)-*	0.00
[N+](-*)(-*)=*	3.01
[N+](-*)#*[e]	4.36
[NH+](-*)(-*)-*	4.44
[NH+](-*)=*	13.97
[NH2+](-*)-*	16.61
[NH2+]=*	25.59
[NH3+]-*	27.64
[n](:*):*	12.89
[n](:*)(:*):*	4.41
[n](-*)(:*):*	4.93
[n](=*)(:*):*[f]	8.39
[nH](:*):*	15.79
[n+](:*)(:*):*	4.10
[n+](-*)(:*):*	3.88
[nH+](:*):*	14.14
[O](-*)-*	9.23
[O]1-*-*-1[d]	12.53
[O]=*	17.07
[OH]-*	20.23
[O-]-*	23.06
[o](:*):*	13.14
[S](-*)-*	25.30
[S]=*	32.09
[S](-*)(-*)=*	19.21
[S](-*)(-*)(=*)=*	8.38
[SH]-*	38.80
[s](:*):*	28.24
[s](=*)(:*):*	21.70
[P](-*)(-*)-*	13.59
[P](-*)=*	34.14
[P](-*)(-*)(-*)=*	9.81
[PH](-*)(-*)=*	23.47

[a] An asterisk (*) stands for any non-hydrogen atom, a minus sign for a single bond, an equals sign for a double bond, a hash or pound sign for a triple bond, and a colon for an aromatic bond; atomic symbol in lowercase means that the atom is part of an aromatic system.

[b] As in nitro group.

[c] Middle nitrogen in azide group.

[d] Atom in a three-membered ring.

[e] Nitrogen in isocyano group.

[f] As in pyridine *N*-oxide.

being preferred. The calculation of rotatable bonds is simply the sum of the acyclic single bonds in the molecule.

It is most important to consider the limitations of the Lipinski Rule-of-Five, as has been highlighted by the original author. The rules are more properly called heuristics or rules-of-thumb. They are trends that often appear to discriminate good from bad, in this case druglike molecules from those unlikely to be drugs.

Many enhancements have been proposed to these rules, but they are still commonly applied, often as a crude filter, in many chemoinformatics systems. One extension of the druglike parameters of the Lipinski Rule-of-Five are the leadlike parameters of the Astex Rule-of-Three, where, as one would expect, the multiplier is now three rather than five.[6] The reduction of this parameter necessarily leads to smaller and less lipophilic compounds. These types of compounds are the ones commonly found through fragment-based screening strategies and worked on at the early stages of a drug discovery programme, with the original rules written as:

> *"The study indicated that such hits seem to obey, on average, a 'Rule of Three', in which molecular weight is <300, the number of hydrogen bond donors is ≤3, the number of hydrogen bond acceptors is ≤3 and ClogP is ≤3. In addition, the results suggested number of rotatable bonds (NROT) (≤3) and PSA (≤60) might also be useful criteria for fragment selection."*

As for the Lipinski rules, the Astex rules are also heuristics and must be employed sensibly with appropriate consideration. However, these heuristics are effective at designing fragment screening libraries, but it is important to say that these heuristics are not set in stone and they may be adjusted for a particular application.

5.7 Summary

The set of physicochemical descriptors are some of the most important and used, sometimes over-used and abused, in drug discovery. Their judicious application can often be informative of general trends in the physicochemical property space and its relationship with specific parameters relevant to drug discovery, such as enzyme potency, cell permeability and other properties that are difficult to predict, such as aqueous solubility. However, it is important to remember that these are very simple parameters, and their over-interpretation and over-reliance on them are not advised.

Many different physicochemical descriptors exist that are beneficial in drug discovery, and only a few have been introduced here. Further physicochemical properties that are important to consider are pK_a and logD, which describe the distribution coefficients, as opposed to the partition coefficient (logP). Here, the focus was on the Lipinski Rule-of-Five and its associated physicochemical descriptors as an introduction to the area. Further information on other physicochemical parameters can be found in the literature and this is an ever-changing field with many advances being made in the computational predictions.

References

1. S. A. Wildman and G. M. Crippen, Prediction of Physicochemical Parameters by Atomic Contributions,*J. Chem. Inf. Comput. Sci.*, 1999, **39**, 868–873.
2. G. A. Landrum, The RDKit, Open-Source Chemoinformatics, http://www.rdkit.org.
3. P. Ertl, B. Rohde and P. Selzer, Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties, *J. Med. Chem.*, 2000, **43**, 3714–3717.
4. S. A. Hitchcock and L. D. Pennington, Structure - Brain Exposure Relationships,*J. Med. Chem.*, 2006, **49**, 7559–7583.
5. C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, *Adv. Drug Delivery Rev.*, 2001, **46**, 3–26.
6. M. Congreve, R. Carr, C. Murray and H. Jhoti, A ‘rule of three’ for fragment-based lead discovery? *Drug Discovery Today*, 2003, **8**, 876–877.

CHAPTER 6

Topological Descriptors

6.1 Overview

Topological descriptors are widely used and highly regarded in the field of chemoinformatics. The exemplified use of topological descriptors in a wide range of applications is well known, from similarity search and clustering algorithms through to statistical modelling and *de novo* design. One of the most desirous aspects of many topological descriptors is their unbiased information content. These descriptors are not biased to empirical models, such as those we have seen in physicochemical descriptors, and also do not succumb to the conformer problem, although this itself can have limitations in some problem domains.

Perhaps the most beneficial aspect of topological descriptors is their sole reliance on the molecular graph notation, and derivations thereof, and their typical rapidity in calculation permitting a huge number of molecules to be considered in a particular analysis. This especially makes them a popular go-to descriptor for many applications, including those that have a computational complexity in the algorithmic method that cannot be rationalised.

The topological descriptors that will be covered in this chapter use graph-theoretic representations of molecular structures, as covered in earlier chapters. The graph theoretic notation is of great value to the field of chemistry and has a very closely connected history to the foundations of atomistic theory in the early nineteenth century.

6.2 Topological Indices

The family of topological indices (or TIs), also known as graph theoretic invariants, are graph theoretic in nature and result in a single value that characterises certain aspects of a molecular structure. As to be expected with a

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

single numerical value that encodes a great deal of topological information, the topological indices can often convolute other properties.

A selection of topological index values for a set of simple molecular structures to more complex drug molecules is given in Figure 6.1. The topological indices are calculated from the molecular graph representation of the molecules using graph theoretic approaches. Perhaps the simplest topological index is the edge density of a given molecular structure. Given a molecular structure with $|V(G)|$ atoms (nodes or vertices), the maximum number of bonds (edges or arcs) that would be possible, assuming one bond possible between each atom (remember the implicit multigraph representation in molecular structures) is given by $|V(G)| \times (|V(G)| - 1)/2$. A graph, like this molecular structure, where all atoms are connected to every other atom, is called a complete graph. The edge density of a given molecular structure would then be calculated by taking the actual number of bonds in the molecular structure ($|E(G)|$) and dividing this by the maximum number of bonds possible theoretically, but this theoretical maximum may not actually be feasible in terms of adhering to the valence bond model. Therefore, the edge density would be calculated by the following equation: $|E(G)|/(|V(G)| \times (|V(G)| - 1)/2)$; more casually written as $e/(n \times (n - 1)/2)$.

6.2.1 Wiener Index

The first topological index to gain much traction in the field was the Wiener index.[1–4] Indeed, this index is the oldest topological index that encodes molecular branching. Developed by Harry Wiener in 1947, the path number (now called the Wiener index) W, is "the sum of distances between any two carbon atoms in a molecule, in terms of carbon–carbon bonds":

$$W = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij} \tag{6.1}$$

6.2.2 Randić Index

The Randić index, which is also known as the connectivity index, is a graph invariant that was introduced in 1975 by Milan Randić.[5] The Randić index uses the bond contributions to describe the connectivity by summing the products of the atom degrees of the atoms connected by each bond. The square root is taken of the summed values and the reciprocal calculated:

$$C(G) = \sum \frac{1}{\sqrt{\deg_i - \deg_j}} \tag{6.2}$$

The Randić index has been shown to correlate with a number of chemical properties, including boiling point, Kovats constants and a calculated surface.[6]

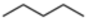
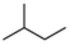
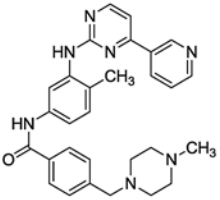
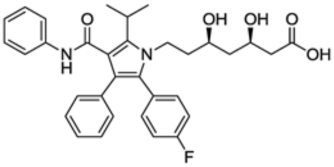
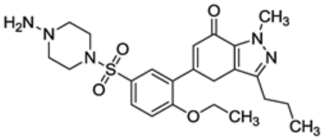
Structure	$V(G)$	$E(G)$	Edge Density	Diameter	Radius	Petitjean Shape Index	Zagreb index
 Pentane	5	4	0.4	4	2	0.5	14
 Isopentane	5	4	0.4	3	2	0.333	16
 Neopentane	5	4	0.4	2	1	0.5	20
 Cyclopentane	5	5	0.5	2	2	0.0	20
 Imatinib	37	41	0.0616	22	11	0.5	194
 Atorastatin	41	44	0.0537	16	8	0.5	210
 Sildenafil	33	36	0.0682	15	8	0.4667	178

Figure 6.1 Some examples of molecular graphs and their graph-theoretic properties.

6.2.3 Petitjean Index

Another popular topological descriptor is the Petitjean index, which is a type of shape descriptor that is calculated from the distance matrix of a molecular structure.[7] This index uses the longest through-graph distance between each atom in the structure using a variant of the all-pairs shortest path algorithm, such as the Floyd–Warshall or Dijsktra algorithm. Petitjean defined the eccentricity of an atom in a molecular structure as the longest path between that atom and any other atom in the structure. Petitjean thus defined the radius (R) of a molecular structure as the smallest atom eccentricity and the diameter (D) as the largest eccentricity. Using the radius and diameter of the molecular structure, Petitjean defined a topological index of shape as $I = (D - R)/R$, which represents a balance between its cyclic and acyclic parts. In a molecular structure where $I = 0$ the graph must be strictly cyclic. However, if $I = 1$ then the graph must be acyclic and have even diameter.

6.2.4 Zagreb Indices

The extent to which a given molecular structure is branched has been approached by two different topological indices by Balaban, called Zagreb indices.[8] M1 is the first Zagreb index and is calculated using the atom degrees of each atom in a molecular structure. M1 is therefore calculated as the sum of the squares of the atom degrees in a given molecular structure. M2 is the second Zagreb index and is calculated again using the atom degrees, but this time using the atom degrees of adjacent atoms to the atom under consideration. Therefore, M2 is given as the sum of the products of the atom degrees of adjacent atoms. The Zagreb indices encode the extent to which a molecular structure is branched and provide a descriptor of structural complexity.

6.3 Molecular Fingerprints

The indices that we have discussed so far are very useful, but gross, descriptors of molecular structure. It is clearly impossible to encode the complexity of a molecular structure as a single number, but other descriptors are available that can encode more information about a molecular structure at a greater resolution.[9,10]

The molecular fingerprint descriptor has many different embodiments, but simply put can be described as the transformation of a molecular graph into a string of variables, most often binary variables. Originally, molecular fingerprints were developed to rapidly speed-up accurate screen-out of chemical structures that, for certain, do not contain a particular chemical substructure. This screen-out procedure meant that far fewer of the much more computationally intensive subgraph isomorphism algorithm calls needed to be made in a substructure search.

Here, we will consider the two classes of molecular fingerprint that are most often used: structure-key and hash-key fingerprints. These can be

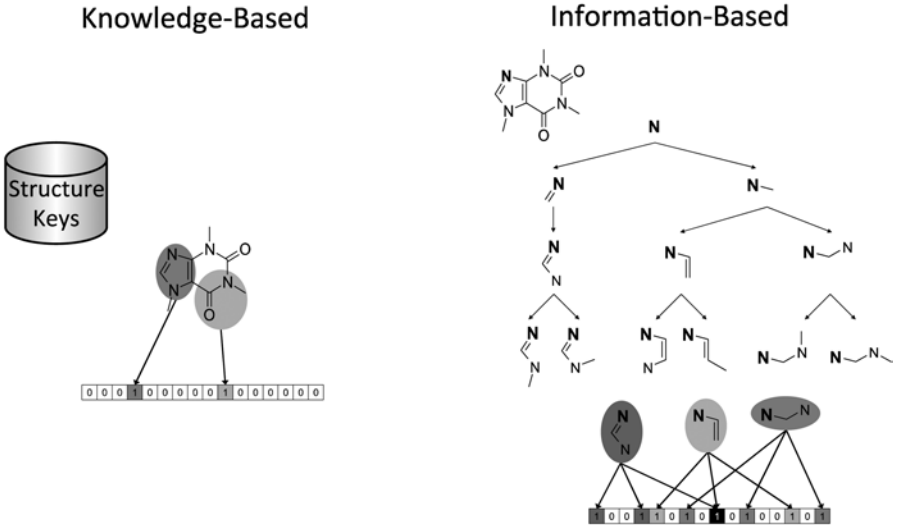


Figure 6.2 An illustration of the differences between knowledge-based and information-based fingerprints. Knowledge-based or structure-key fingerprints use predefined substructural keys that have been defined for the domain of interest, in this case small-molecule druglike molecules. Information-based or hash-key fingerprints encode the information present in the molecular structure being encoded.

more generalised as knowledge-based and information-based descriptors (Figure 6.2). A knowledge-based fingerprint is one that encodes known molecular structure moieties—or sometimes with some extent of fuzziness—and their presence or absence recorded. One issue with knowledge-key fingerprints is that they can be brittle to new and unknown chemistries, which can limit their applicability. On the other hand, information-based fingerprints encode, using substructure enumeration techniques to encode typically linear or circular fragments from the molecular structure. The enumeration technique in information-based fingerprints reduces the brittleness seen in knowledge-based fingerprints, but the use of the often-necessary hashing algorithm can lead to mapping collisions into a fingerprint representation that can reduce their effectiveness. In general, however, both general classes of fingerprint representation are very useful.

6.3.1 Structure-Key Fingerprints

Although we have emphasised that topological descriptors tend to be based solely on the information content of the molecular structure being encoded, structure-key fingerprints are an exception to this. Structure-key fingerprints (such as MACCS keys) are based on an explicitly defined dictionary of molecular keys or substructures. Since a human decision, albeit sometimes by committee, has been made in selecting these particular structural features, it is inevitable that biases may be incorporated into the resultant fingerprint. However, this does not necessarily make their application less useful. One must remember, however, that not only are biases likely included, but each feature will typically be weighted equally in the calculation of similarity measures, although it is likely that each of the features does not contribute equally to the degree of similarity.

6.3.2 Hash-Key Fingerprints

An alternative type of fingerprint that has been found to be very useful in many applications is the hash-key fingerprint. Rather than taking a structural dictionary as for structure-key fingerprints, the hash-key fingerprint is an attempt to wholly encode the molecular graph structure into a fingerprint more akin to a mathematical graph-vector transform. Interestingly, due the nature of the transformation, it is often relatively simple to reconstruct the original molecular graph representation using simple graph optimisation tools (such as iterative optimisation in *de novo* design software). Therefore, great care must be taken when disclosing these fingerprints, and indeed any molecular descriptors, where confidentiality is an issue.

The hash-key fingerprint was originally designed solely for the purpose of rapid screen-out of molecules when conducting a substructure search. The aim was to reduce as much as possible, but not too much, the number of calls to the substructure search algorithm, which was much more computationally expensive at the time. The substructure of interest was encoded as a fingerprint and then applied as a substructure screen against a large molecular database. A screen, in this instance, means that any fingerprint in the database that has the same bits set in its fingerprint as the substructure screen fingerprint will be passed to the substructure search algorithm. Since the substructure screen fingerprint is therefore entirely represented in the discovered fingerprint, it must also entirely contain the substructure of interest. However, it must be noted that bit collisions that occur with over-ambitious hashing and folding of the fingerprint, may bring about spurious matches, the beauty of this approach is that it will never lead to a false negative, it will only lead to more molecules that need to be considered with the substructure search algorithm. Therefore, each new implementation can be tuned for screen-out performance without any risk of generating false negatives.

6.3.2.1 Fingal Fingerprints

The Fingerprinting algorithm (Fingal) fingerprint was developed as a hash-key fingerprint based on the algorithm published online by Daylight Chemical Information Systems.[11] In this way, Fingal was not novel, but some algorithmic extensions and applications demonstrated the use of this fingerprint in new domains. The fingerprint was originally developed for the Compound Generator (CoG) *de novo* design system as a rapid molecular descriptor generator that could be used in proof-of-principle studies, but was latterly used to design median molecules as well as highly predictive QSAR and QSPR models. Fingal was extended from binary fingerprints to integer fingerprints, such that the frequency of occurrence of molecular features is also included. In addition, the Fingal fingerprint was implemented to encode geometric information of molecules, such that conformational information could be included in the fingerprint.

Fingal fingerprints are generated by iterating over each atom in a molecule and enumerating paths from these atoms. For each atom, every possible path, from zero up to a user-specified length in edges, is extracted and represented as a character string with bonding information. This provides a large number of possible paths for each atom. An example of how fingerprint keys are generated from a given molecule using the path-based approach is given in Figure 6.3 and the pseudocode is given below.

```
function makeFingerprint(Graph molecule, Size d, Int length)
    fingerprint = initializeFingerprint(d)

    paths = getPaths(molecule, length)

    for each atom in molecule

        for each path from atom
            seed = hash(path)
            indices = random(seed)

            for each value in indices
                index = value mod d
                fingerprint[index] = TRUE

    return fingerprint
```

6.3.2.2 Morgan Fingerprints

A more recent extension to the hash-key fingerprint is the Morgan fingerprint, which looks at atom environments as circular substructures, as opposed to the path-based structural keys of Fingal, *etc.* The Morgan fingerprint is similar to the Extended Connectivity Fingerprint (ECFP) that was originally popularised in the PipelinePilot software from SciTegic (latterly Accelrys, and now BIOVIA from Dassault Systèmes).[12] However, the Morgan fingerprint has a long and illustrious history in chemoinformatics from the very early days of computers.

Named after Henry Morgan, who was working for Chemical Abstracts at the time, the Morgan algorithm forms the basis of the Morgan fingerprint and was published in 1965.[13] The Morgan algorithm is a method to canonicalise a molecular structure such that the linear atom ordering is always the same for that particular structure. This algorithm was introduced for use in chemical database systems to allow for rapid structure search. The result is a systematic and unique numbering order for each atom in the molecule.

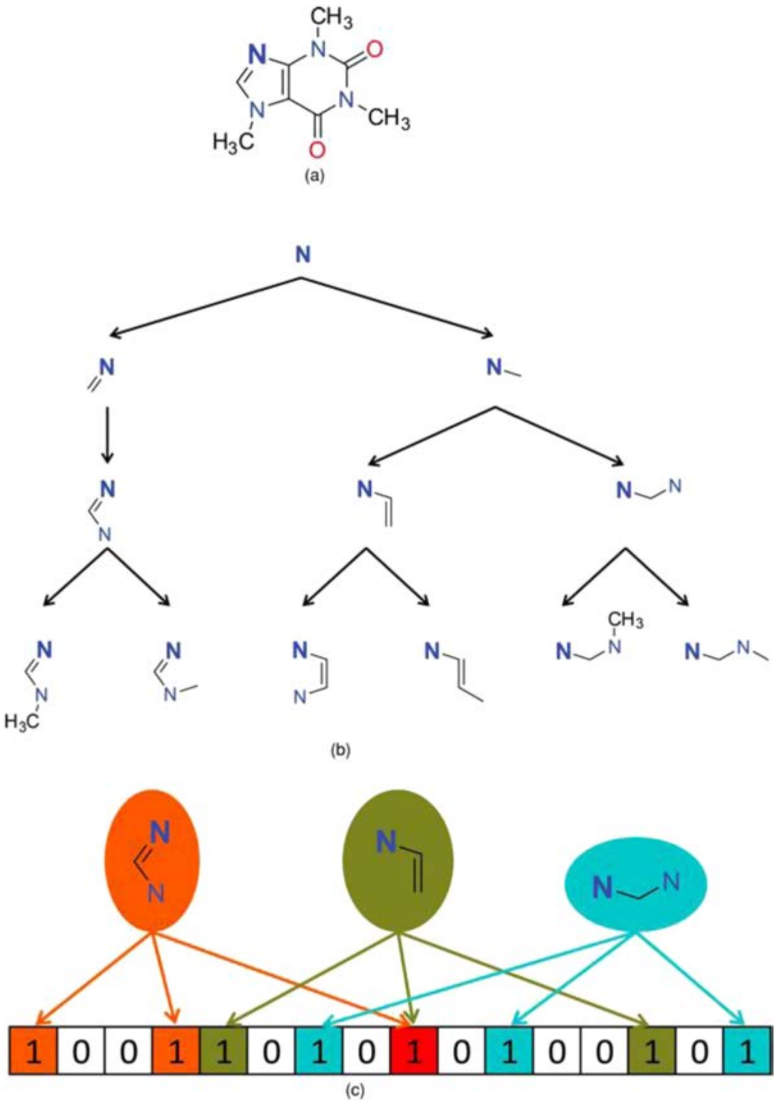


Figure 6.3 A graphical representation of the generation of a molecular fingerprint using the Fingal algorithm. (a) The original molecule, caffeine. (b) Illustration of the path tree through the molecule for one nitrogen atom up to a path length of three edges. (c) Three paths being encoded into the fingerprint using the hashed indices. Note that the element highlighted in red in (c) is an example of a hashing collision where different paths hash into the same value. Multiple indices are often used to alleviate this problem as the chance of different paths hashing into identical sets of values is greatly reduced.

The algorithm initialises the chemical structure by first assigning a numerical identifier to each non-hydrogen atom in the structure based on the node degree of each atom, the number of non-hydrogen atoms to which it is connected. The next iteration of the algorithm assigns a new numerical identifier to each atom as the sum of the previous identifiers of its direct neighbours. This procedure continues iterating until there is no increase the number of different identifiers.

The next step of the algorithm is to assign the ordering of the atoms using the generated numerical identifiers. The atom with the highest identifier is assigned as the first atom. The atoms connected to the first atom are then considered and the one with the highest identifier is assigned as atom number two, and the remaining neighbour atoms of atom one are labelled sequentially. Once all neighbours of atom one are numbered, the algorithm moves to number two and numbers its atoms accordingly. The algorithm then proceeds through each numbered atom in order, so atom number three is next and its neighbours numbered according to the same process. The final numbering is the ordered listing of each atom.

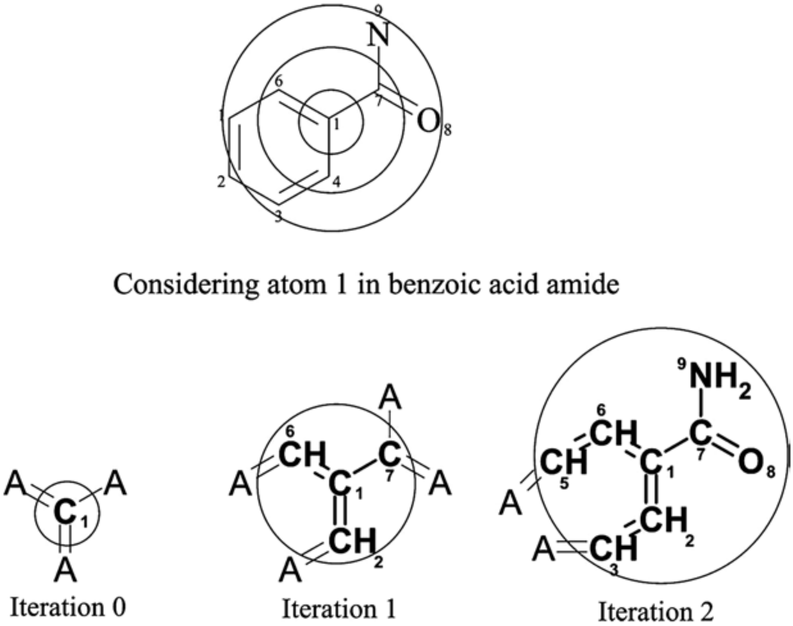
An illustration of the general extended connectivity principle is given in Figure 6.4 using the example given on the published algorithm for the Extended Connectivity Fingerprint.[12]

The simple algorithm works well in practice but it does have some limitations for which extensions of the algorithm have been published. However, the process of the algorithm is the same in principle as the other algorithms, an example of a relaxation algorithm from computer science.

The Morgan algorithm forms the basis of the Extended Connectivity and Morgan fingerprints. The Morgan fingerprint is a member of a class known as circular fingerprints. Each atom in a given molecular structure is assigned a value based on a number of atomic invariants, which can be user-defined, and are based on neighbourhood characteristics of the substructure, such as bonding orders or atomic numbers. For each atom in a molecule, substructures are induced up to a specified radius, usually two or three, with the substructure at each radius from zero being retained. Iterations are then performed to update the initial atom identifiers and its neighbouring atoms until the specified radius is reached. The sets of numbers are then hashed into a single key that provides an identifier for each substructure. The resulting hash keys can then be folded into a fixed-length fingerprint using a hash function, which is where bit collisions may occur.

6.3.3 Ligand-Based Topological Pharmacophores

Another often-used class of topological descriptor is the ligand-based topological pharmacophore. We will cover geometric pharmacophores in the following chapter, where the explicit spatial information of molecular conformations is used, but the topological equivalent uses through-graph distances as a surrogate for explicit geometry; the through-graph distances were discussed earlier in this chapter with regard to the Petitjean index.


Considering atom 1 in benzoic acid amide

Iteration 0

Iteration 1

Iteration 2

Figure 6.4 Illustration of the effect of iterative updating on the information represented by an atom identifier. Here, we consider atom 1 in benzoic acid amide. Each iteration has the effect of creating an identifier that represents larger and larger circular substructures around the central atom, as shown at the top of the figure. At iteration 0 (that is, the initial atom identifier), the atom only represents information about atom 1 and its attached bonds, and can be represented by the substructure on the bottom left ('A' represents an atom of any type other than hydrogen). After one iteration, the identifier now contains information about the immediate neighbours of atom one, as shown in the bottom centre substructure. After two iterations, the represented substructure has grown further, now fully incorporating the amide group as well as much of the aromatic ring, as shown in the bottom right. Reprinted with permission from D. Rogers and M. Hahn, Extended-connectivity fingerprints, *J. Chem. Inf. Model.*, 2010, **50**(5), 742–754. Copyright 2010 American Chemical Society.

Ligand-based topological pharmacophores use the through-graph distances to represent the potential three-dimensional arrangements of the atoms in a given molecular structure. Furthermore, each atom is further abstracted to represent the potential for molecular interactions, rather than as the elemental label.

6.3.3.1 CATS Vectors

One of the most popular ligand-based topological pharmacophores is the CATS vector (Chemically Advanced Template Search) published by Schneider *et al.* in 1999.[14] Originally, the CATS vector representation was published for the application of scaffold hopping, a specific subset of bioisosteric replacement, which we will cover in greater detail in Chapter 11. In scaffold hopping, the objective is to identify molecular structures that are likely to be functionally equivalent to a known molecular structure, but sufficiently different in their underlying chemical structure. The CATS fingerprint introduces two levels of abstraction from the typical topological fingerprints: disconnecting the representation from the underlying connectivity of the molecule, and the abstraction of atom types into functional (or pharmacophoric) features. This is a useful approach to maintain potency while modulating other properties, such as solubility and likely sites of metabolism. Another, more prosaic, application is to move away from patented core scaffolds in an otherwise encumbered region of chemical space. In this way, the original objective of CATS vectors was to identify different molecular structures while retaining the key functional requirements for macromolecular recognition.

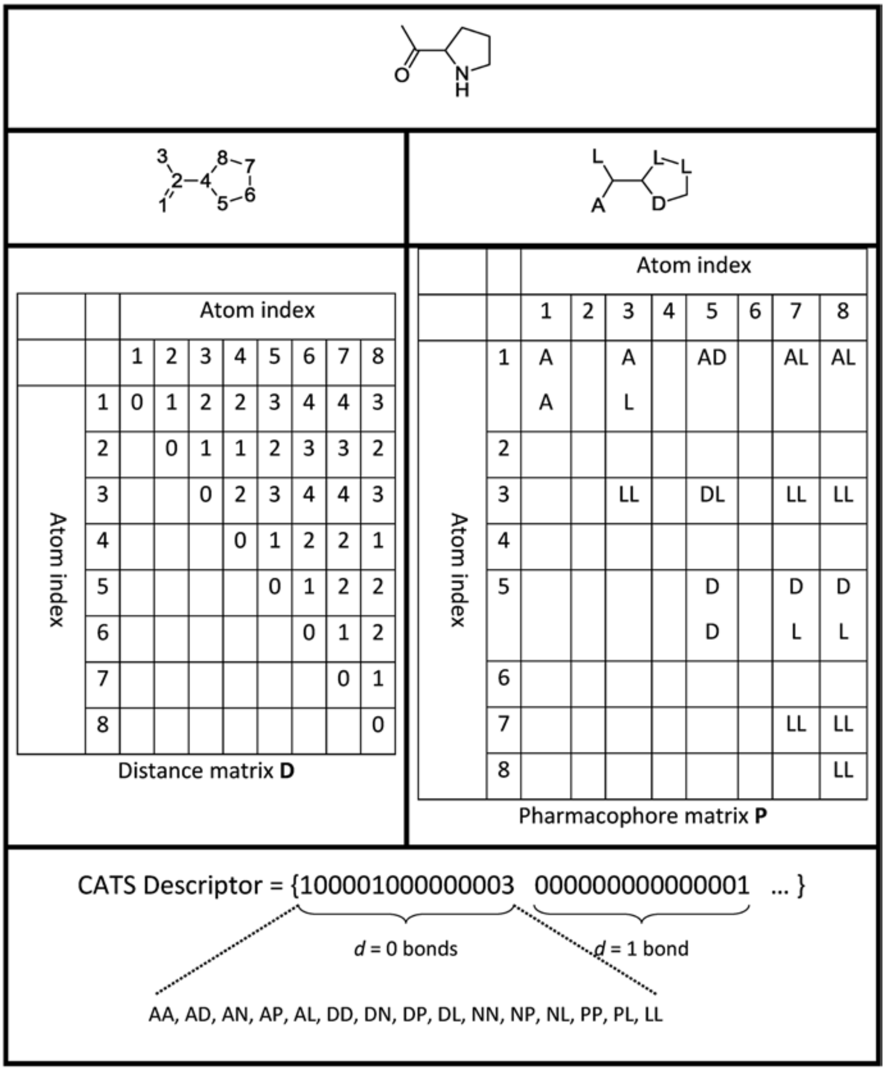
The algorithm to generate a CATS vector requires a number of independent steps to be conducted. The first step is to use an atom abstraction scheme that takes each atom in a given molecular structure in turn and encodes them as one of the following six pharmacophoric types: lipophilic (L), aromatic (R), hydrogen bond donor (D), hydrogen bond acceptor (A), positively charged or ionisable (P), and negatively charged or ionisable. The second abstraction of CATS is to reduce the explicit reliance on the topology of the features, but attempt to retain the general spatial (at least in a through-graph nature) constraints. Through-graph distances are then calculated using an all-pairs, shortest path algorithm, such as the Floyd–Warshall or Dijkstra algorithms, with atoms being encoded between distances of typically one to ten atoms. An exemplar calculation of the calculation of a CATS vector is given in Figure 6.5.

Scaffold hopping, as mentioned, is the intended application of CATS vectors and the descriptor has been applied successfully in this aim, including variants of the algorithm.[15,16] Although the main intended application of CATS vectors is to the scaffold hopping problem, they are a useful general topological descriptor and can be used in clustering and diversity selection.

The Schneider laboratory has made a version of CATS available online here: http://modlab-cadd.ethz.ch/. An implementation of CATS is also available, with greater options in parameterisation, in The RDKit API.

6.3.3.2 Hopfen Fingerprints

The Hopfen fingerprints were designed as a hybrid between the Morgan fingerprints and the CATS vectors by combining the atom environments and the through-graph distances in atom pairs of triplets.[17] Hopfen fingerprints first enumerate the atom environments of each atom in a given molecule and encode each substructure as structural keys, as per Morgan Fingerprints. The next stage is to calculate the all-pairs, shortest-paths between each atom as a surrogate for interatomic distances in three-dimensional space. The last stage is to enumerate all shortest-path triplets between each atom and


O
N
H
3
8
7
2
4
1
5
6
L
L
L
A
D
CATS Descriptor = {100001000000003 000000000000001 ... }
d = 0 bonds
d = 1 bond
AA, AD, AN, AP, AL, DD, DN, DP, DL, NN, NP, NL, PP, PL, LL

		Atom index							
		1	2	3	4	5	6	7	8
Atom index	1	0	1	2	2	3	4	4	3
	2		0	1	1	2	3	3	2
	3			0	2	3	4	4	3
	4				0	1	2	2	1
	5					0	1	2	2
	6						0	1	2
	7							0	1
	8								0

Distance matrix **D**

		Atom index							
		1	2	3	4	5	6	7	8
Atom index	1	A A		A L		AD		AL	AL
	2								
	3			LL		DL		LL	LL
	4								
	5					D D		D L	D L
	6								
	7							LL	LL
	8								LL

Pharmacophore matrix **P**

Figure 6.5 Calculation of the CATS descriptor. The two-dimensional graph is given first, followed by the graphs representing the atom indices and pharmacophoric types.

each of the atom environment structural keys that have been generated. The triplets are then canonicalised and encoded as a key for incorporation into a fingerprint. The generation of a Hopfen key for three atoms is given in Figure 6.6.

Hopfen fingerprints were designed and implemented by the author and first reported in a prospective case study for scaffold hopping.[17] In this study, the Hopfen fingerprints were used to generate a Quantitative Structure–Activity Relationship (QSAR) model using Partial Least Squares (PLS) on a

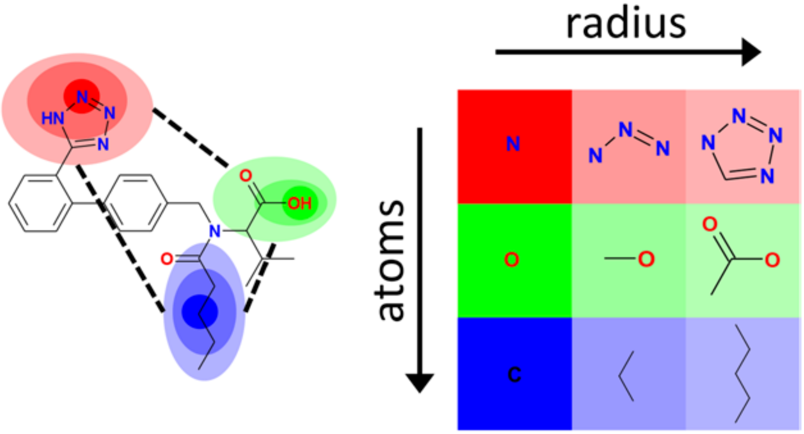


Figure 6.6 Illustration of the Hopfen neighbourhood enumeration and triplet encoding of those enumerated substructures.

training and test set of known actives against the target of interest, MDM2/p53 a protein–protein interaction (PPI). The approach was found to be favourable in finding good quality hits through experiment and the molecular structures were more diverse than those found through other methods considered.

6.4 Summary

Topological descriptors are some of the most used, and possibly least understood, of the family of available molecular descriptors. Whenever you use a similarity searching system, such as that found in ChEMBL or in the online vendor catalogues of eMolecules and Sigma-Aldrich, invariably it is a topological descriptor that is being calculated or used behind the scenes and the similarity calculated according to the Tanimoto coefficient. Often, scientists will refer to the Tanimoto similarity of molecules, but it is important to know and understand the actual molecular descriptor that is being used behind the scenes. The Morgan fingerprint tends to have a much narrower distribution of similarities compared to the much wider similarity distributions observed with the Fingal fingerprint. The changes in the shape of the similarity distributions can lead to misinterpretation of the similarity measure.

References

1. H. J. Wiener, Structural Determination of Paraffin Boiling Points, *J. Am. Chem. Soc.*, 1947, **69**, 17–20.
2. H. J. Wiener, Influence of Interatomic Forces on Paraffin Properties, *J. Chem. Phys.*, 1947, **15**, 766.
3. H. J. Wiener, Vapor Pressure-Temperature Relationships Among the Branched Paraffin Hydrocarbons, *J. Phys. Chem.*, 1948, **52**, 425–430.

4. H. J. Wiener, Relation of the Physical Properties of the Isomeric Alkanes to Molecular Structure. Surface Tension, Specific Dispersion, and Critical Solution Temperature in Anilene,*J. Phys. Chem.*, 1948, **52**, 1082–1089.
5. M. Randić, Characterization of molecular branching,*J. Am. Chem. Soc.*, 1975, **97**(23), 6609–6615.
6. L. B. Kier and L. H. Hall, *Molecular Connectivity in Chemistry and Drug Research*, Academic Press, San Francisco, 1976.
7. M. Petitjean, Applications of the Radius-Diameter Diagram to the Classification of Topological and Geometrical Shapes of Chemical Compounds,*J. Chem. Inf. Comput. Sci.*, 1992, **32**, 331–337.
8. A. T. Balaban, I. Motoc, D. Bonchev and O. Mekenyan, Topological indices for structure-activity correlations, *Top. Curr. Chem.*, 1983, **114**, 21–55.
9. N. Brown, Chemoinformatics – an introduction for computer scientists, *ACM Comput. Surv.*, 2009, **41**, 1–38.
10. N. Brown, Algorithms for Chemoinformatics, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2011, **1**, 716–726.
11. N. Brown, B. McKay and J. Gasteiger, Fingal: A Novel Approach to Geometric Fingerprinting and a Comparative Study of Its Application to 3D-QSAR Modelling, *QSAR Comb. Sci.*, 2005, **24**(4), 480–484.
12. D. Rogers and M. Hahn, Extended-connectivity fingerprints,*J. Chem. Inf. Model.*, 2010, **50**(5), 742–754.
13. H. L. Morgan, The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service,*J. Chem. Doc.*, 1965, **5**(2), 107–113.
14. G. Schneider, W. Neidhart, T. Giller and G. Schmid, Scaffold-Hopping” byTopologicalPharmacophoreSearch:AContributiontoVirtualScreening, *Angew. Chem., Int. Ed. Engl.*, 1999, **38**, 2894–2986.
15. S. Renner and G. Schneider, Scaffold-Hopping Potential of Ligand-Based Similarity Concepts, *ChemMedChem*, 2006, **1**, 181–185.
16. M. Wagener and J. P. M. Lommerse, The Quest for Bioisosteric Replacements,*J. Chem. Inf. Model.*, 2006, **46**, 677–685.
17. E. Jacoby, A. Boettcher, L. M. Mayr, N. Brown, J. L. Jenkins, J. Kallen, C. Engeloch, U. Schopfer, P. Furet, K. Masuya and J. Lisztwan, Knowledge-Based Virtual Screening: Application to the MDM4/p53 Protein-Protein Interaction, in *Chemogenomics, Methods in Molecular Biology*, ed. E. Jacoby, Humana Press, 2009, vol. 575.

CHAPTER 7

Topographical Descriptors

7.1 Overview

Physicochemical and topological descriptors, describing physical phenomena and molecular connectivity, respectively, are highly effective at tackling many modern challenges in modelling being fast to calculate yet still offering significant predictive power. However, as with all methods, they do have their limitations since they do not explicitly consider the three-dimensional, not to mention the four-dimensional, character of molecules, in that they have three-dimensional shape and move in geometric space and time.

Molecular shape is clearly important in invoking binding events since binding sites exist in three dimensions. Therefore, the three-dimensionality of a molecular structure will be important in filling the binding site sufficiently to have the potential to make the appropriate interactions with the protein: shape complementarity. Without this shape complementarity, intermolecular interactions could not be formed with the protein even if the appropriate electronic interactions are present. Therefore, topographical, or molecular shape, descriptors must appropriately describe the delicate balance of shape and electronics necessary for a binding event.

The similar-property principle also applies to three-dimensional (3D) shape and this has been known for decades. However, it is only relatively recently that the 3D methods have become mainstream. First, with the advent of the ability to computationally generate appropriate single conformations of given molecular structures. Tools such as CORINA and CONCORD were the first computer programs that could rapidly generate the 3D co-ordinates of a molecule using empirical data. With the advent of these tools, molecular shape searching could be expanded into many more applications than had previously been considered.

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

Soon, though, practitioners became aware that a single conformation, perhaps simulating the 3D co-ordinates observed in small molecular crystallography, was not necessarily sufficient for shape matching since the multiple conformers that a molecule may adopt are also important: known as the conformer problem. The conformer problem is a significant challenge and will be addressed in this chapter with solutions to overcome potential limitations.

In this chapter, we will explore the history of molecular shape descriptors and comparisons, through to the present day in terms of how they are used in the context of drug discovery. A few specific examples of molecular shape descriptors will be introduced with simple explanations and discussion.

7.2 Topographic Descriptors

Many different geometric descriptors have been defined over time, and a summary of many of them is available.[1] Geometric descriptors can be calculated from conformations of molecular structures using the co-ordinate data of the atoms in the structures. Two different classes of three-dimensional (3D) descriptor are possible using these co-ordinate data: Cartesian descriptors (external 3D [x3D]) and internal 3D (i3D). Descriptors that use Cartesian co-ordinates are placed within a reference co-ordinate frame and typically require alignment or some form of normalisation, such as placing the centre-of-mass of the structure on the origin in this space. i3D descriptors, however, are calculated based on the relative distances between the atoms within its own internal reference co-ordinate frame. Generally, x3D and i3D descriptors can be summarised as alignment-dependent and alignment-free, respectively.

Recently, a good deal of interest has been generated in the area of the three-dimensionality and quantifying the three-dimensionality of ligands.[2] The motivation to identify more three-dimensional molecular structures is to adapt screening libraries to new challenges in drug discovery, such as protein–protein interactions. There is also good evidence to demonstrate that more three-dimensional molecules will also tend to have better physicochemical profiles, such as solubility, due to disrupting planarity and therefore less likely to pack tightly in crystal lattices. Two descriptors will be described briefly here: Principal Moments of Inertia (PMI) and Plane of Best Fit (PBF).

The PMI were introduced into computational drug discovery to measure the diversity of combinatorial libraries in terms of the shapes covered.[3] The PMI is a calculation of the first three principal moments of inertia, which essentially specify how rod-like, disc-like, and sphere-like the molecular shape is. Therefore, it is necessary to calculate the geometric arrangement of atoms, or a conformer. A single conformer was used in this, and the following descriptor. Using these three descriptors from PMI, it is possible to plot the coverage of shape space on a so-called PMI plot (Figure 7.1). A PMI plot is a ternary plot where the three vertices of the triangle represent the three extremes of the shapes as described above. The space is continuous, such that as a point on the PMI plot is moved along the edge from disc-like to sphere-like, the molecular structure represented by those points becomes more sphere-like and less disc-like. One limitation identified with

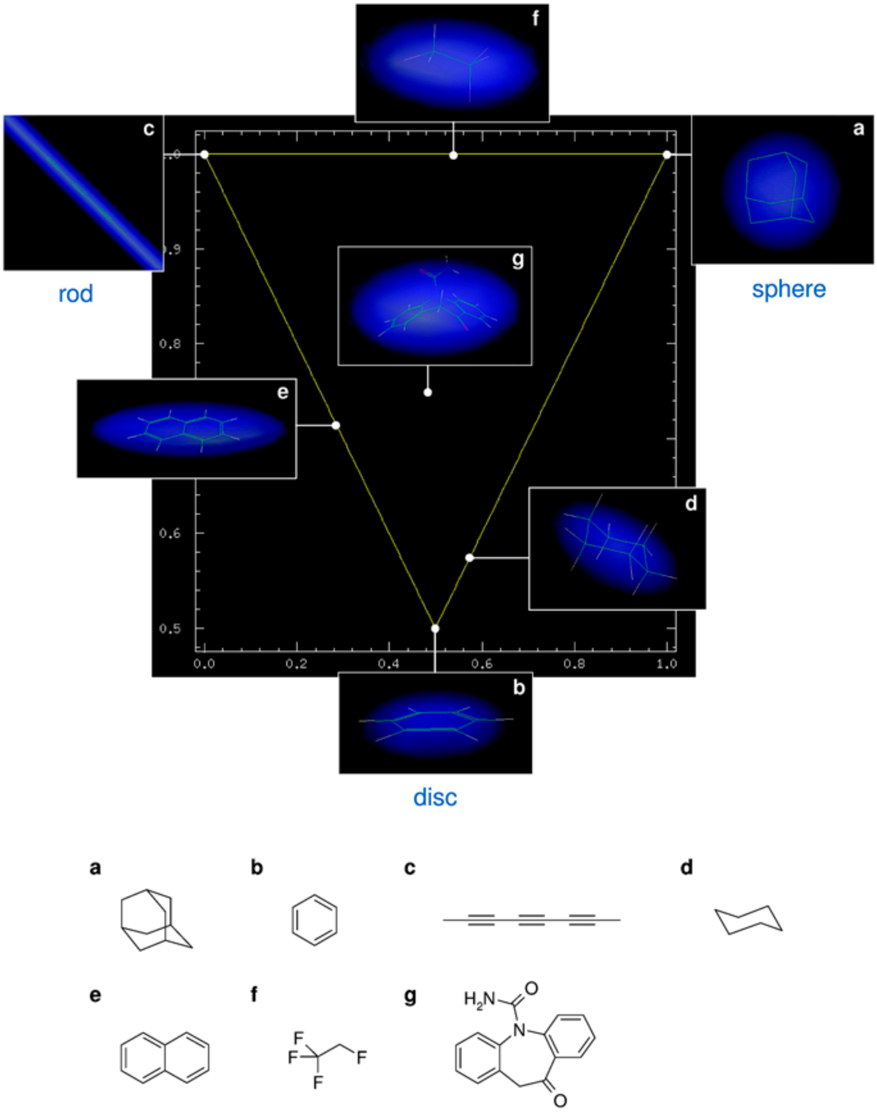


Figure 7.1 Normalised PMI ratios as shape descriptors: the position within the triangle reveals the "envelope shape". Structures (a), (b), and (c), represent spherical, planar, and rod-like shapes, respectively. Structures (d), (e), and (f), exhibit shapes that are between the three basic shapes above. Lastly, (g), represents a structure that is balance between the three general shapes of rods, discs, and spheres. Reprinted with permission from W. H. B. Sauer and M. K. Schwarz, Molecular shape diversity of combinatorial libraries: a prerequisite for broad bioactivity, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 987–1003. Copyright 2012 American Chemical Society.

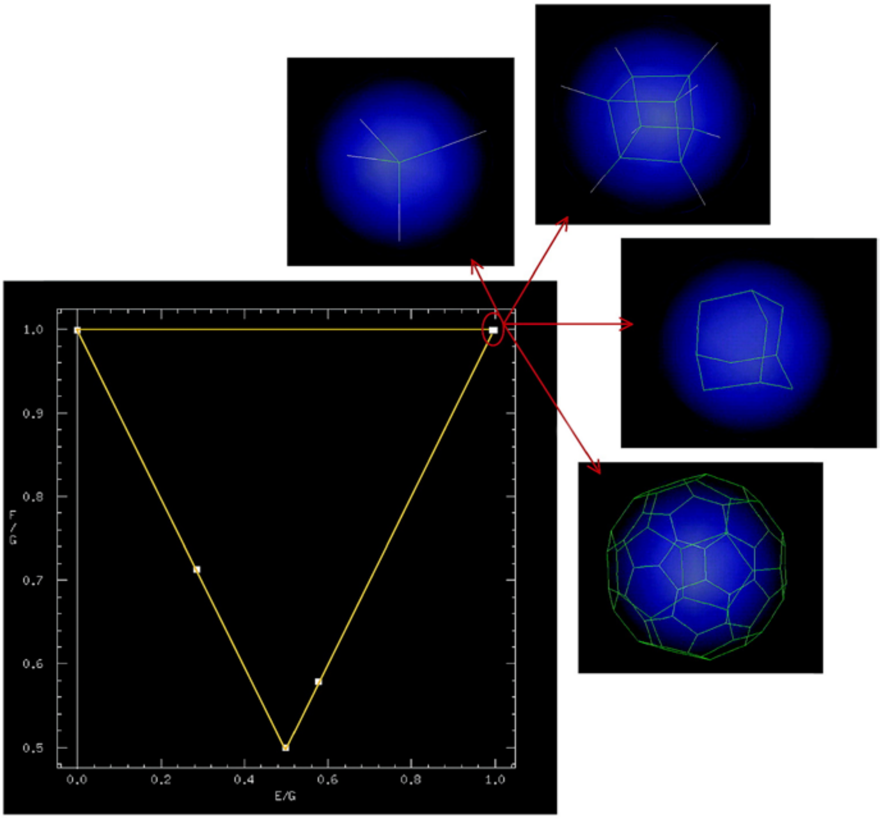


Figure 7.2 Limitations of the envelope shape analysis: degenerate situations. Reprinted with permission from W. H. B. Sauer and M. K. Schwarz, Molecular shape diversity of combinatorial libraries: a prerequisite for broad bioactivity, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 987–1003. Copyright 2012 American Chemical Society.

PMI, illustrated in Figure 7.2, is that there is no size dependence in the shape analysis, but this can be corrected by normalising for size by the number of heavy atoms in the molecule, for example.

The second descriptor that quantifies the three-dimensionality of molecular structure is the Plane of Best Fit (PBF).[4] PBF also requires the calculation of a conformer prior to calculation. PBF is most easily explained as an extension the line of best fit in two dimensions, but expanded into three dimensions. Each heavy atom in a given molecule is treated as a data point in three-dimensional space. A plane is then defined that is optimised using a least squares fit algorithm that minimises the distances of each atom from the plane. Once the optimisation has completed, the average of the errors from the plane is calculated giving the PBF value for that conformer of that structure.

Exemplar PBF values for a range of structures from the fragment screening library are given in Figure 7.3. It can be observed that the structures (given in 2D and 3D representations) at the bottom end of the PBF range are much

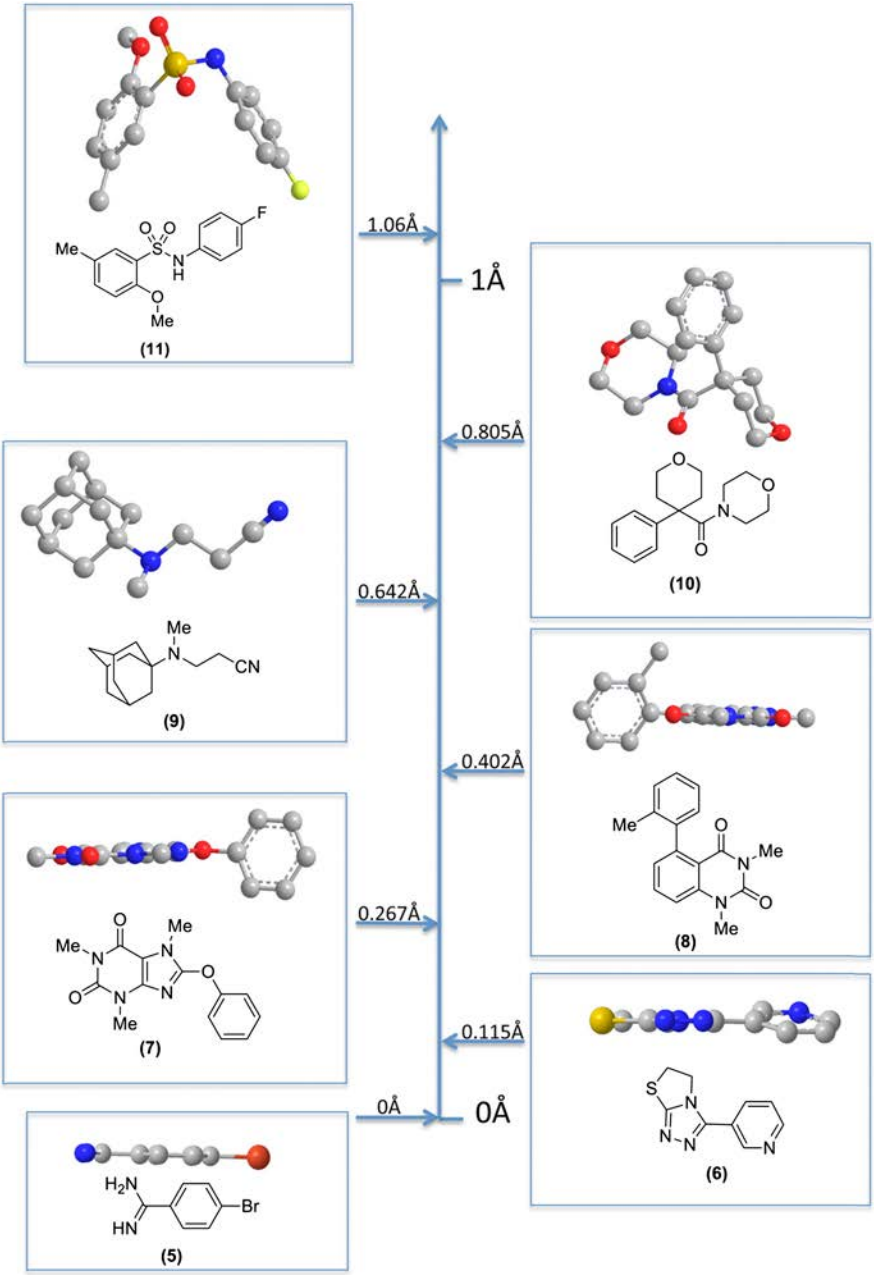


Figure 7.3 Example molecules selected from the fragment library data set and their respective Plane of Best Fit scores depicted on a linear scale in ångströms. It is observable that as the PBF score increases the molecules that are representative of the increasing scores become visually more three-dimensional. Reprinted with permission from N. C. Firth, N. Brown and J. Blagg, Plane of Best Fit: A novel method to characterize the three-dimensionality of molecules, *J. Chem. Inf. Model.*, 2012, **52**(10), 2516–2525. Copyright 2012 American Chemical Society.

flatter than those towards the higher end, and that the continuum of calculated PBF values are intuitive in terms of increasing three-dimensionality.

PMI and PBF both compare the three-dimensionality of 3D molecular structures, but do they offer different information, or put another way, are they highly correlated? In Figure 7.4, a density plot correlation of PMI (sum of normalised PMIs) and PBF is given where it is clear that there is some weak correlation between the two descriptors but they do offer differences, especially in resolution. The PBF descriptor is much more discriminatory at the lower end of three-dimensionality than PMI, indicated by the heavy black lines. The increased discriminatory quality at the lower end of three-dimensionality suggests that PBF may outperform PMI in discriminating the flatter molecular structures, which is where most of the molecular structures in synthetic libraries occur.

7.3 Pharmacophores

The pharmacophore is a relatively recent term with many people having been credited for its introduction in the past, including the father of modern drug discovery, Paul Ehrlich in the early 1900s. However, extensive research by John van Drie[5] has revealed that the first definition of pharmacophore was only described relatively recently, by Monty Kier in 1971.[6] However, Kier had been using the approach on muscarinic agonists since 1967, but called them a ‘proposed receptor pattern’.[7]

The IUPAC now defines a pharmacophore to be “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.”[8]

There are many different implementations of pharmacophore elucidation and search algorithms, and they function approximately similarly. However, the generation of a pharmacophore model often adopts a similar workflow. It is possible, and simple, to generate a pharmacophore model from a single ligand, but this may result in poor results of the final pharmacophore model in terms of consensus of the features that are desirable. Structurally diverse ligands are preferable, but oftentimes these will not be available early in drug discovery programmes. Additionally, it is highly beneficial to include inactive ligands in the model generation stage since they will often add additional information regarding what is required to design more optimal ligands. However, if a bound ligand from a protein–ligand complex is available, it would make sense to derive a pharmacophore from this, obviating the next two steps.

The next step in generating a pharmacophore model is to enumerate sets of conformers for each of the ligands in the training set. The preference is for low energy conformers since the assumption is that bound conformations will typically be closer in shape to the bound conformations.

Once the conformers have been generated, the conformers of all of the ligands need to be aligned to each other using a superimposition algorithm.

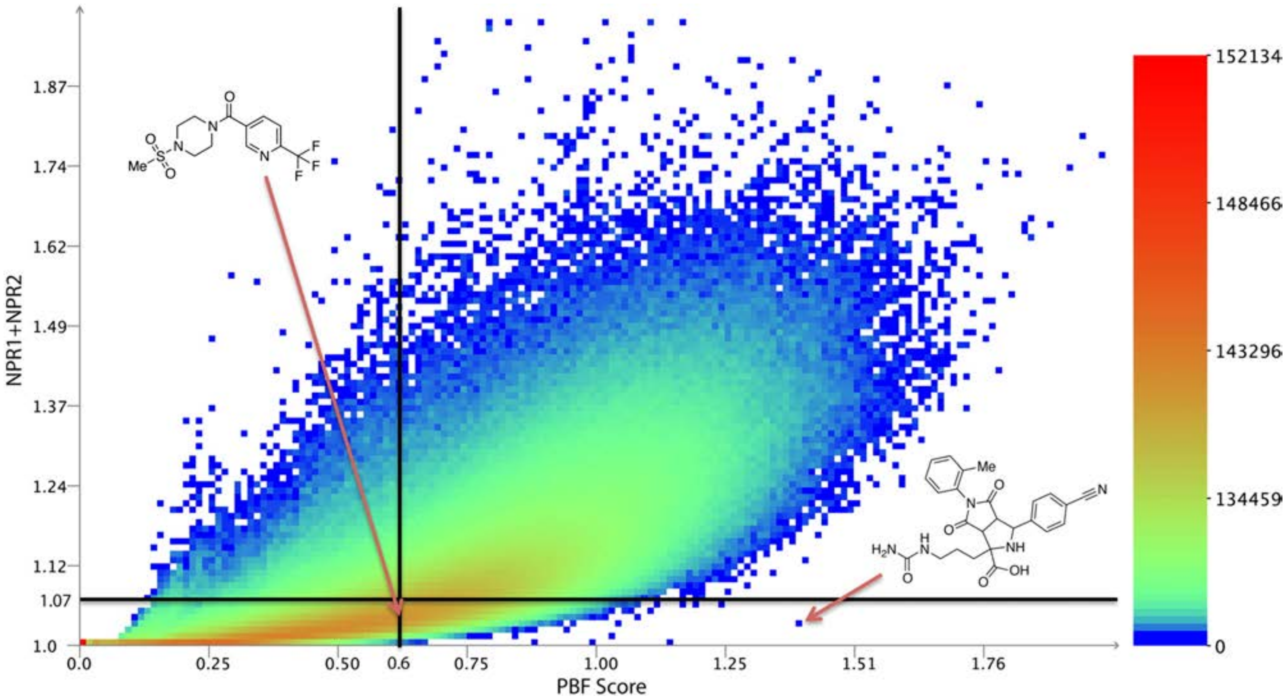


Figure 7.4 Density plot of PBF score *versus* the sum of normalised principal moments of inertia (NPR) for the eMolecules data set with acyclic and Ro5 noncompliant compounds removed. The horizontal black line represents a cutoff for 3D molecules for NPR1 + NPR2, and the vertical line, a corresponding cutoff for PBF. Reprinted with permission from N. C. Firth, N. Brown and J. Blagg, Plane of Best Fit: A novel method to characterize the three-dimensionality of molecules, *J. Chem. Inf. Model.*, 2012, **52**(10), 2516–2525. Copyright 2012 American Chemical Society.

Superimposition will attempt to align the low energy conformations of the ligands so that similar functional or bioisosteric features overlay (Chapter 10). The best overlay with a single conformer of each of the active ligands is taken as the optimal overlay.

The abstraction step is the final one in generating a pharmacophore model. The task of abstraction is to reduce the aligned atomic and substructural features of the ligands into feature points or spheres. Abstraction approaches vary between software, but benzene rings should be collapsed into a single aromatic ring feature and hydroxyl groups as acceptor/donor motifs.

Most modern molecular modelling software packages have protocols to derive pharmacophores automatically, including the alignment and abstraction steps. The flexible alignment, or multi-conformer alignment, task can be quite computationally intensive and therefore it is best to be pragmatic and build up from a few ligands to investigate the effect that the amount of data has on the quality of the pharmacophore models. The abstraction is also often implemented as an automated process in which pharmacophoric features, such as aromatic ring, acceptor/donor and hydrophobic, are assigned based on the overlays of structural features of the ligands. However, there is a tendency for the autofit methods to over-specify the pharmacophore, and therefore many software packages offer a manual curation step for the model. Additionally, it is often possible to include any protein information that may be available to define exclusion spheres to indicate where the protein is present and therefore any ligand discovered that extended into those exclusion spheres would likely cause a steric clash. As additional ligand and structural information becomes available, it may be necessary to refine and update the pharmacophore model to adapt it to new hypotheses to test. An illustration of the general workflow to generate pharmacophore models and subsequently compare them is given Figure 7.5.

Once a pharmacophore of appropriate quality has been generated, it is necessary to validate it retrospectively prior to prospective application in a virtual screen. Retrospective studies are intended to ensure that a model fulfils its intention to separate actives from inactives, so that those predicted more likely to be active are prioritised for synthesis and/or testing. Therefore, a sufficient dataset of experimentally identified active and inactive structures should be applied as a test case and the model statistics assessed for suitability. Virtual Screening will be covered in more detail in Chapter 9.

The preceding has outlined the typical protocol for pharmacophore model elucidation from multiple ligands and their conformers, and the pharmacophore model abstracted from the overlays. However, as mentioned, many more recent software programs for shape-based searching have automated or streamlined many of these processes, but it is important to still be aware of the computation being performed. A number of more recent advances in shape-based molecular descriptors have been generated to permit both pharmacophore and shape search in virtual screening and other applications. A selection of the more popular methods is presented in brief below: Rapid Overlay of Chemical Structures (ROCS), Ultrafast Shape Recognition (USR), and software from the Cresset Group.

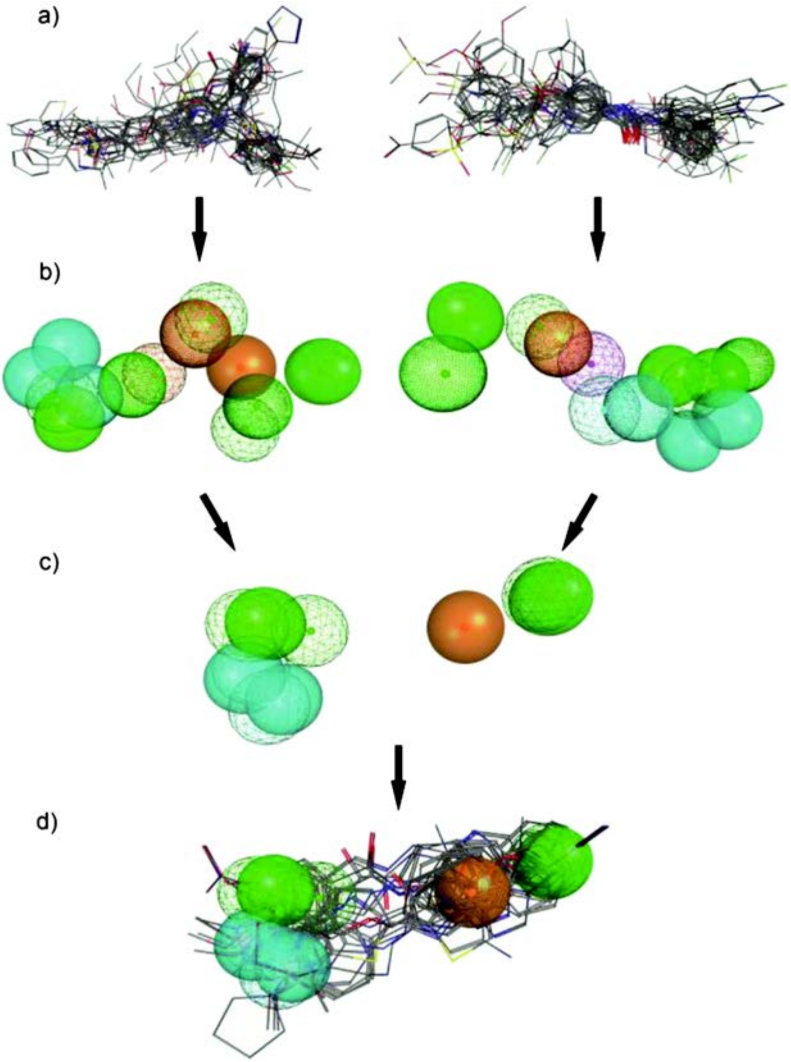


Figure 7.5 Virtual screening process. On the basis of multiple conformations of known ligands for both targets (a), a number of different pharmacophore models are generated (b). To find models sharing the same features at a similar spatial distance, pairwise alignments are computed (c). Using the aligned models, a pharmacophore search for molecules matching both models is performed (d). The potential "dual" compounds are scored by a shape-based comparison with the known active ligands (not shown). Different models are drawn in solid, as mesh, and as wireframe. The colours represent different pharmacophore features: green, hydrophobic; orange, aromatic; blue, H-bond acceptor; and purple, H-bond donor. Reprinted with permission from D. Moser, J. M. Wisniewska, S. Hahn, J. Achenbach, E. L. Buscató, F. M. Klingler, B. Hofmann, D. Steinhilber and E. Proschak, Dual-target virtual screening by pharmacophore elucidation and molecular shape filtering, *ACS Med. Chem. Lett.*, 2012, 3(2), 155–158. Copyright 2012 American Chemical Society.

7.4 ROCS: Rapid Overlay of Chemical Structures

The Rapid Overlay of Chemical Structures (ROCS) software is a well-regarded and effective program that searches for optimal shape overlays and matching chemical types, or colour, mapped to each of the ROCS features.[9] The ROCS algorithm uses Gaussians overlaid onto each atom in a given molecule to define its 'shape'. The Gaussians are used to represent the volumes of the atoms and also significantly soften the sensitivity of using hard sphere cut-offs as in many other shape overlay and pharmacophore matching methods. The method also represents functional features, akin to pharmacophoric features, for each of the atoms or groups of atoms that are assigned these features. Again, each of these features is a Gaussian, to define its 'colour'. The types of features that the colour spheres encode are rings, hydrogen bond donors and acceptors, *etc.* Ring centroids are given an 'extra credit' if the centroid aligns regardless of ring type. An example of an ROCS representation and its overlay using shape and colour is given in Figure 7.6.

ROCS has been demonstrated to be very effective at a variety of virtual screening campaigns, including in the challenge of scaffold hopping (see Chapter 10) where the aim is to replace the core scaffolds of molecules that maintain functional and/or geometric scaffolding properties.[10]

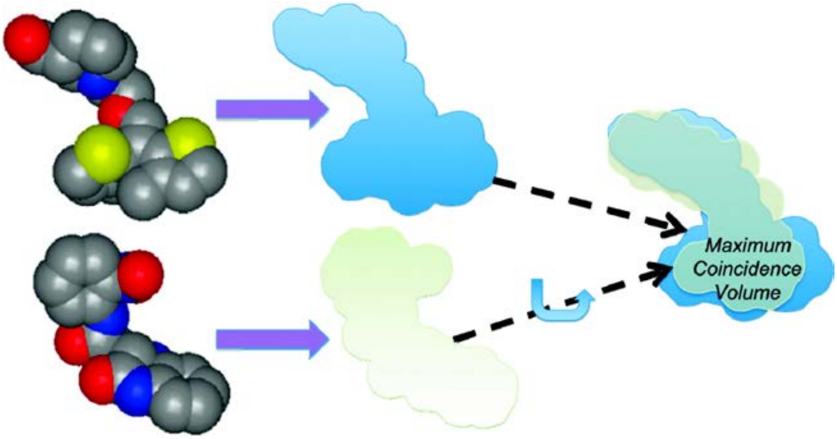


Figure 7.6 Illustration of a fundamental definition of shape similar, derived from the alignment that achieves an optimal overlap of objects. The mismatch volume between two objects is a true mathematical metric distance, *i.e.*, obeys the triangle inequality that says the distance from object A to object C cannot be greater than the distance from A to B plus B to C nor less than the difference between these distances. However, the optimal overlap leads to the more intuitive shape Tanimoto, *i.e.*, the ratio of the overlap to the absolute difference of the sum of the self-overlaps and optimal overlap. It has the useful character of ranging from 1.0 (perfect overlap) to 0.0 (no overlap). Reprinted from A. Nicholls, G. B. McGaughey, R. P. Sheridan, A. C. Good, G. Warren, M. Mathieu, S. W. Muchmore, S. P. Brown, J. A. Grant, J. A. Haigh, N. Nevins, A. N. Jain and B. Kelley, Molecular shape and medicinal chemistry: a perspective, *J. Med. Chem.*, 2010, **53**(10), 3862–3886.

7.5 USR: Ultrafast Shape Recognition

The Ultrafast Shape Recognition (USR) algorithm was introduced as a very fast method for calculating the similarities of 3D molecules.[11] The method generates a series of discrete distance distributions based on the centroid atom and a variety of other atoms. These distributions are then used to calculate statistical moments that encode the shape of the molecule. Once the vector of these moments is calculated it is possible to calculate intermolecular similarity using a normalised Manhattan distance (Figure 7.7).

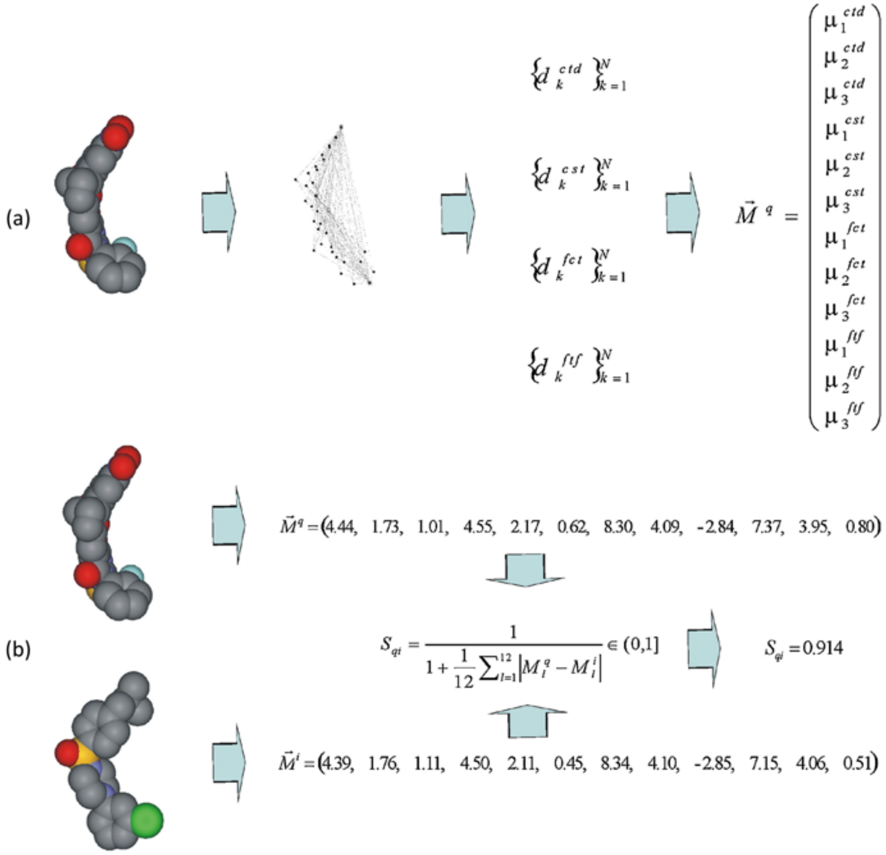

(a)
$\{d_k^{ctd}\}_{k=1}^{N}$
$\{d_k^{cst}\}_{k=1}^{N}$
$\{d_k^{fct}\}_{k=1}^{N}$
$\{d_k^{ftf}\}_{k=1}^{N}$
$\vec{M}^q = (\mu_1^{ctd}, \mu_2^{ctd}, \mu_3^{ctd}, \mu_1^{cst}, \mu_2^{cst}, \mu_3^{cst}, \mu_1^{fct}, \mu_2^{fct}, \mu_3^{fct}, \mu_1^{ftf}, \mu_2^{ftf}, \mu_3^{ftf})$
(b)
$\vec{M}^q = (4.44,\ 1.73,\ 1.01,\ 4.55,\ 2.17,\ 0.62,\ 8.30,\ 4.09,\ -2.84,\ 7.37,\ 3.95,\ 0.80)$
$S_{qi} = \frac{1}{1+\frac{1}{12}\sum_{l=1}^{12}|M_l^q - M_l^i|} \in (0,1]$
$S_{qi} = 0.914$
$\vec{M}^i = (4.39,\ 1.76,\ 1.11,\ 4.50,\ 2.11,\ 0.45,\ 8.34,\ 4.10,\ -2.85,\ 7.15,\ 4.06,\ 0.51)$

Figure 7.7 (a) USR encoding. The shape of the molecule is characterised by the distributions of atomic distances to four strategic reference locations. In turn, each of these distributions is described through its first three moments. In this way, each molecule has associated a vector of 12 shape descriptors. (b) Comparing the shape of two conformers with USR. Each database conformer has a vector of 12 USR descriptors associated, which are used to compare them through a normalised similarity score. Reprinted with permission from P. J. Ballester, P. W. Finn and W. G. Richards, Ultrafast shape recognition: evaluating a new ligand-based virtual screening technology, *J. Mol. Graphics Modell.*, 2009, **27**(7), 836–845. Copyright 2009 Elsevier Inc.

USR was shown to outperform ROCS in terms of speed—1546 times speed-up was cited—mainly due to the removal of the more computationally intensive 3D alignment step required in ROCS.

7.6 XED: Cresset Group

Cresset Group tools use the Extended Electron Distribution (XED) model that was developed by Andy Vinter as a more complex and accurate description of the charge around atoms, making it possible to discern the lone pairs in the charge model making the model much richer and informative than many others available.[12]

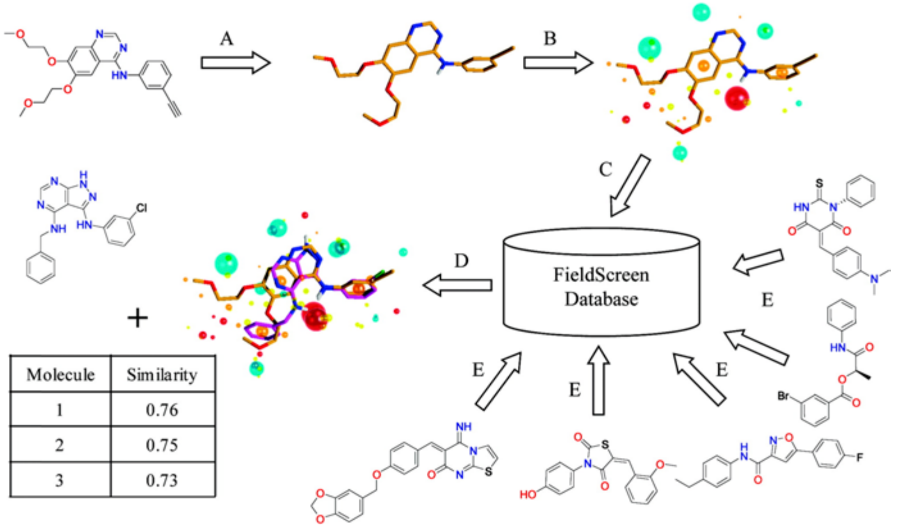
Using the XED model to generate the 3D field patterns from a 2D structure, it is necessary to reduce the field pattern to field points to make the computational search much faster for matching with other molecules. The field pattern represents electrostatic, hydrophobic and shape properties of a molecule. A schematic of the generation of the Cresset Group virtual screening workflow is given in Figure 7.8.

Cresset Group tools using XEDs have been applied successfully to a wide range of challenges in drug discovery and their use in virtual screening was demonstrated to out-perform the well-known DOCK algorithm for ligand docking in terms of retrieval of novel active scaffolds using the Directory of Useful Decoys (DUD).[13]

7.7 Conformer Generation and the Conformer Problem

Conformer generation remains a challenging and time-consuming task and can require substantial storage requirements, given the number of conformers (in addition to tautomers and stereoisomers) that may need to be stored. There exist a number of methods to automatically generate conformers from connection tables (2D structures). Software packages such as CORINA generate a single conformation based on 3D data mined from small molecular crystal structure databases and exhibit good agreement with those structures in terms of root-mean-square deviation (RMSD) of the heavy atoms when aligning two conformations of the same molecule.[14] However, CORINA is limited to the generation of a single conformer by itself and typically many conformers are required to perform an effect shape similarity search. Conformer generation can be achieved by a number of methods, but two of the most popular are systematic search and stochastic (or random) search. However, given an illustration of the conformer problem in Figure 7.9, it is clear that many conformers can be generated for typical drug like molecules and it is not clear which, if any, of those that are generated is the one that is most appropriate. This is the definition of the conformer problem.

Systematic search takes each rotatable bond in term and discretely explores the torsion angles of each bond. The simplest approach to systematic conformer generation is to all conformers possible with each permitted torsion



Molecule	Similarity
1	0.76
2	0.75
3	0.73

Figure 7.8 Schematic representation of the steps involved in searching the FieldScreen database: (A) select an active molecule and convert it to a relevant conformation; (B) add field points to the search ligand in the specified conformation to produce the FieldScreen search query, which consists of a ligand and its field points in a specified conformation; (C) search the FieldScreen database by alignment of every structure using field points; (D) retrieve the top scoring compounds (score expressed as a molecular similarity) as 3D alignments to the search query or as 2D structures. The FieldScreen database (E) is populated by exploration of conformations of all molecules with field point patterns added to and stored with each conformation. Reprinted with permission from T. J. Cheeseright, M. D. Mackey, J. L. Melville and J. G. Vinter, FieldScreen: virtual screening using molecular fields. Application to the DUD data set, *J. Chem. Inf. Model.*, 2008, **48**(11), 2108–2117. Copyright 2008 American Chemical Society.

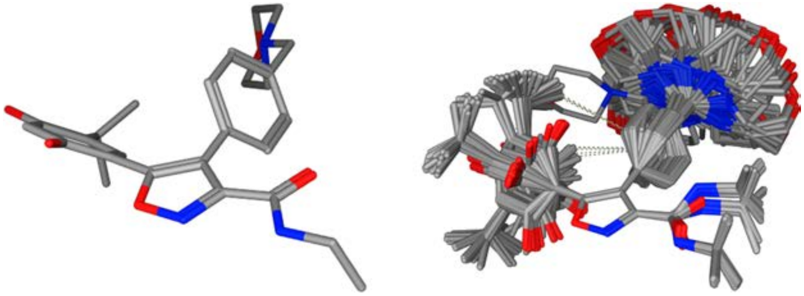


Figure 7.9 An illustration of the conformer problem using AUY922. On the left is a single energy-minimised conformation, but on the right hand side, a few hundred conformations have been generated. It is clear based on the distribution, and the shear infinity of torsion bond angles, that the conformer space is vast and it is important to use appropriate cut-offs to generate a sufficient quantity of appropriate conformations for the intended analysis.

angle. Typically, the torsion angle discretisation can be user-defined in most software packages. In systematic search, the main challenge to overcome is combinatorial explosion since all torsions are explored at all discretised angles. However, this exhaustive enumeration of the conformers will lead to many conformers that have steric clashes or are in very high predicted energy states. A number of heuristics have been introduced to avoid enumerating all conformers when situations such as those stated previously arise. One such method is to use a depth-first search with tree pruning that employs backtracking. Once an undesirable conformer has been identified, the tree is pruned and the algorithm backtracks to the next node on the tree not yet explored. Another approach to limit the size of the conformer space being considered is to treat the molecular structures as larger fragments to explicitly reduce the search space.

Stochastic search randomly explores new conformers using the conformers that have been generated already. Once a random conformer has been generated and the energy minimised so that it is suitable to retain, in terms of user-definable parameters such as energy, a copy of that conformer is made and then randomly perturbed at one torsion angle, and then minimised and retained if it fulfils the defined parameters. The process is repeated using a randomly selected conformer by perturbing it until some stop condition is met. Since this process is heuristic in nature, it is difficult to clearly understand when the process should be terminated. Typically, a number of user-defined parameters are specified to invoke a termination condition, such as: maximum number of conformations reached or no new conformations have been generated after a number of cycles, where a new conformation would be classed as one that falls inside an RMSD window and therefore deemed too similar.

Once generated, it is difficult if at all possible to identify whether any of those conformations is appropriate in terms being similar to that expected in a binding event: this is referred to as the Conformer Problem. The conformational space of a single ligand is very complex and the energy landscape potentially even more complex, with many low energy wells that could be missed with extant conformer generation methods. The diversity of the conformations is also important and it is important to cover the diversity of the conformers available but this can be quite wasteful in compute-time. However, how can the appropriate conformer be identified without the relevant information? Low energy conformers are preferred, but this may be a dogma in modelling. And what does low energy mean? Certainly, energy calculators have errors and therefore relevant conformers may be missed in the search.

Another question is to what are the generated conformers being compared to suggest that the appropriate conformations are not being generated? The two comparators that are used most are small molecule crystal structures, and the ligand conformation in complex with a biological macromolecule. The small molecular crystal structure databases are an excellent resource of experimentally determined structural information, but the conformations in small molecule crystal structures are not necessarily in what could be called a biologically relevant conformation since they are packed together and therefore equilibrate over the entire system. Protein–ligand co-crystals are another seemingly appropriate reference set for conformer generators, but this assumes that these ligand geometries are appropriate. Furthermore, crystallography offers only a snapshot of protein and ligand conformations, and the molecular dynamics that are undergone in solution are only beginning to be investigated and understood.

There is no simple answer to the conformer problem and therefore it remains an open problem in molecular modelling. The best response that is possible at present is to put in place appropriate controls as best as is possible and understand the errors, both in models and experiments.

7.8 Summary

In drug discovery, the ligands being designed make three-dimensional interactions with biological macromolecules and therefore it follows that characterising the three-dimensionality and shape of those molecules is important. Since the advent of automated structure generators in the 1980s, it has been possible to rapidly generate reasonable ligand conformations, allowing for the comparison of 3D molecules rather than the 2D representations that preceded them. Many descriptors have been explored and their ability at tackling pressing challenges in drug discovery assessed.

Descriptors regarding how 3D a molecule is have recently become a scientific challenge since it has been reported the molecules that exhibit more 3D character will also exhibit better properties than those that would be expected to be observed in drugs. A number of descriptors of three-dimensionality have been reported and studies conducted that demonstrate that much of the ligand space in which research is focussed tends to be quite flat in nature, which is perhaps unsurprising. However, with the advent of descriptors such as PMI and PBF, the anticipation is that library design can be tweaked over time to the desired level of three-dimensionality.

Pharmacophore representations have already been introduced in Chapter 6 in the context of topology only, but these ligand-based topological pharmacophores arose from 3D pharmacophores. The geometric arrangement of key interaction features that are relevant to binding is clearly important and these methods are able to abstract these key features successfully for use in virtual screening.

The tools of ROCS, USR and those from Cresset Group offer highly effective methods to conduct rapid virtual screening of large 3D databases and have been shown to be highly effective in a number of projects, both retrospective and prospective. Their use is now routine in most drug discovery programmes and they have had a significant impact.

The main challenge going forward is to improve conformer generation methods, or at least better understand what the challenge is and how to know when an improvement is made. Currently, the lack of understanding of what makes an appropriate control experiment that indicates success is a high priority.

References

1. R. Todeschini and V. Consonni. *Handbook of molecular descriptors*, John Wiley & Sons, 2008.
2. A. D. Morley, A. Pugliese, K. Birchall, J. Bower, P. Brennan, N. Brown, T. Chapman, M. Drysdale, I. H. Gilbert, S. Hoelder, A. Jordan, S. V. Ley, A. Merritt, D. Miller, M. E. Swarbrick and P. G. Wyatt, Fragment-based hit identification: thinking in 3D, *Drug Discovery Today*, 2013, **18**(23–24), 1221–1227.
3. W. H. B. Sauer and M. K. Schwarz, Molecular shape diversity of combinatorial libraries: a prerequisite for broad bioactivity,*J. Chem. Inf. Comput. Sci.*, 2003, **43**, 987–1003.
4. N. C. Firth, N. Brown and J. Blagg, Plane of Best Fit: A novel method to characterize the three-dimensionality of molecules,*J. Chem. Inf. Model.*, 2012, **52**(10), 2516–2525.
5. J. H. van Drie, Monty Kier and the origin of the pharmacophore concept, *Internet Electron. J. Mol. Des.*, 2007, **6**(9), 271–279.
6. L. B. Kier. *Molecular orbital theory in drug research*, Academic Press. Boston, 1971, pp. 164–169.
7. L. B. Kier, Molecular orbital calculation of preferred conformations of acetylcholine, muscarine, and muscarone, *Mol. Pharmacol.*, 1967, **3**(5), 487–494.
8. C. G. Wermuth, C. R. Ganellin, P. Lindberg and L. A. Mitscher, Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998), *Pure Appl. Chem.*, 1998, **70**(5), 1129–1143.
9. J. A. Grant, M. A. Gallardo and B. T. Pickup, A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape,*J. Comput. Chem.*, 1996, **17**, 1653.
10. T. S. Rush, J. A. Grant, L. Mosyak and A. Nicholls, A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction,*J. Med. Chem.*, 2005, **48**(5), 1489–1495.
11. P. J. Ballester and W. G. Richards, Ultrafast shape recognition for similarity search in molecular databases, *Proc. R. Soc. A*, 2007, **463**(2081), 1307–1321.
12. J. G. Vinter, Extended electron distributions applied to the molecular mechanics of some intermolecular interactions. II. Organic complexes, *J. Comput.-Aided Mol. Des.*, 1996, **10**(5), 417–426.
13. T. J. Cheeseright, M. D. Mackey, J. L. Melville and J. G. Vinter, FieldScreen: virtual screening using molecular fields. Application to the DUD data set,*J. Chem. Inf. Model.*, 2008, **48**(11), 2108–2117.
14. J. Sadowski and J. Gasteiger, From atoms and bonds to three-dimensional atomic coordinates: automatic model builders, *Chem. Rev.*, 1993, **93**(7), 2567–2581.

Part 4
Statistical Learning

CHAPTER 8

Statistical Learning

8.1 Overview

The field of statistical learning is essential in many disciplines and can be characterised as the science of learning from data. Using this learning, one can make informed and objective decisions about the future or, more simply, what we should do next. The emerging and highly desirable skillsets, such as data mining, big data and deep learning, are underpinned by a thorough and appropriate application of statistics and statistical learning methods. Increasingly, data is becoming openly available that enables statistical learning in many forms, such as the ChEMBL database of chemical structures and biological data. Computer power has also increased to such an extent that most statistical methods can now be applied routinely. The statistical learning methods have been implemented and made available at no cost in packages such as scikit-learn, an API in Python and the R Project for Statistical Computing. However, the availability of the data, the compute power and the methods only reinforce the importance of thorough analysis, design and implementation of statistical experiments.

8.2 Statistical Learning

The computational scientist can consider many millions of potential data points, in our case virtual chemical structures, any of which may be of interest to a drug discovery project. However, the speed at which these virtual experiments can be reduced to practice, synthesis and biological testing, is rate limiting. Therefore, it of the utmost importance that the modeller does the best job they can to ensure the predictions are reliable. A morning of routine modelling can easily translate into weeks, months or even years of

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

substantial real experimental work and it is the duty of the modeller to do their best to ensure that as little time as possible is wasted. However, it must also be borne in mind that appropriate negative experiments are as valuable, if not more valuable in some cases, than positive data, although there is of course a reluctance to generate negative results.

In this section the general concepts of statistical learning will be introduced, specifically: unsupervised learning and supervised learning. These methods are the most relevant to the *in silico* medicinal chemist and are arguably some of the most applied approaches in the field. The more general approaches of clustering, classification and regression will be discussed with specific reference to some of the more frequently used algorithms. Best practices are advice only and not prescriptive, since it is difficult to write a successful recipe that can be applied in every eventuality. It is still the responsibility of the modeller to critically determine the most appropriate methods to apply in each case, weighing the advantages and limitations of each of the respective methods. As such, this guidance is worthy of consideration, but one must still think how best to apply statistical learning methods and ensure that any potential limitations are communicated appropriately.

8.3 Unsupervised Learning

Unsupervised learning methods are one of the key statistical learning techniques applied in computational drug design. An introduction to unsupervised learning and where it may be applied in the drug design setting is provided. A number of examples of unsupervised learning methods are introduced and discussed with examples from the literature. Cluster analysis is introduced using two different clustering algorithms: Sequential, Agglomerative, Hierarchical, Non-overlapping (SAHN) clustering and *k*-means clustering. Subsequently, two different projection methods are introduced and discussed: Self-Organising Maps (SOMs) and Principal Component Analysis (PCA). Importantly in this chapter, consideration is given to when and where one would apply these methods. The advantages and relative limitations of each method are discussed to provide the reader with an understanding of how they might apply similar methods in their own work.

8.3.1 Overview

Unsupervised learning, or ‘learning without a teacher’, is frequently used in drug discovery in a wide range of applications. Unsupervised learning methods assist in understanding the ‘natural’ structure of our data, in our case typically chemical structure data. Unsupervised learning methods seek to identify the relationships between given data points, which can then enable objective decision making in compound selection for purchase or testing. Unsupervised learning also allows chemical space analysis, or where our chemical structures lie in the enormous virtual space of feasible structures.

These methods link in with the applications that will be introduced later in this book of subset selection, whether islands of highly similar chemical structures or diversity selection where the objective is to identify the fewest number of representative chemical structures that sufficiently explores chemical space for the particular scientific question under consideration.

In this chapter, we will focus on some of the most often used unsupervised learning methodologies applied in drug design. The methods introduced will not be exhaustive, but an introductory guide to ensure a thorough understanding required for application, therefore some other methods will be necessarily omitted.

8.3.2 Cluster Analysis

Cluster analysis is not a single algorithm, but a multitude of different methods that fall under the same umbrella. The objective in cluster analysis is to group objects, in our case chemical structures, based on defined properties, or molecular descriptors, according to their inter-object similarities. The resultant groups are referred to as clusters.

Cluster analysis has its foundations in anthropology where it was used for the analysis of cultural relationships between tribal groups by Driver and Kroeber in 1932.[1] Zubin later introduced cluster analysis methods to psychology in 1938, and Tryon in 1939, for the analysis and identification of personality types. These methods were later applied in 1943 by Cattell for trait theory classification in psychology of personalities.

One of the most used algorithms in cluster analysis in chemoinformatics is called Sequential, Agglomerative, Hierarchical, Non-overlapping (SAHN). Here, the core concept is that objects that are proximate in the given similarity in descriptor space will be more similar in general to each other. Agglomerative clustering techniques proceed with each object being contained in its own cluster. Pairs of these clusters are then combined based on similarity, so the two most similar objects are combined into one cluster, and this is continued until all clusters are combined into a single cluster. The result of this agglomerative process is an entire hierarchy of clusters and can be visualised as a dendrogram (Figure 8.1).

One of the main challenges in cluster analysis is the selection of the representative clusters. One of the most common methods is the application of a stopping rule. In the dendrogram representation (Figure 8.1), a simple method is to move a line down from the top of the dendrogram to the bottom. Given a particular line position, the points below that line are defined as the clusters. The decision where to stop can be given by a distance criterion where clusters are too distal to each other to be merged. An alternative approach to identify the clusters is a stopping rule based on the number criterion, where there are an appropriate number of clusters. A common heuristic for identifying the number of clusters is the square root of the number of objects divided by two (sqrt(n)/2), which is also commonly used in k-means clustering.

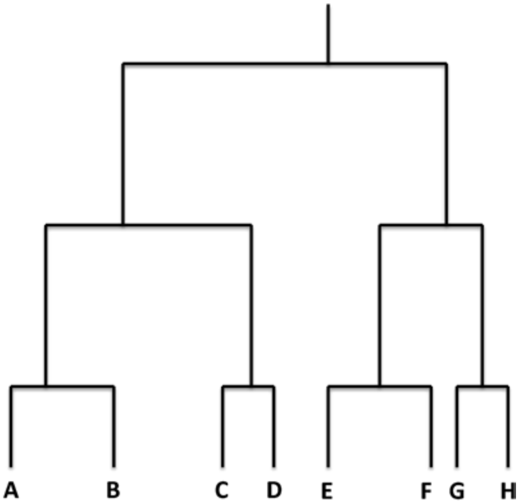


Figure 8.1 An exemplar cluster dendrogram with the individual compounds {A...H} at the leaves of the dendrogram, and each cluster being redefined as the clustering moves up the tree.

While hierarchical clustering methods are an excellent method for clustering data, it is very computationally intensive in the general case, since it has computational complexity of $O(n^3)$, although an $O(n^2)$ method has been reported called SLINK.[2] Therefore, this approach is appropriate for small datasets, such as hit list analysis from high-throughput screening (typically a few thousand objects), but unsuitable for larger datasets such as high-throughput screening libraries themselves (typically hundreds of thousands of objects).

8.3.3 *k*-Means Clustering

k-Means clustering was first developed and applied in signal processing, but was not published outside Bell Labs until 1982.[3] Hugo Steinhaus described the method in 1957 and the standard algorithm was designed by Stuart Lloyd in the same year.[4] However, the method was not called *k*-means clustering until some ten years later in 1967 by James MacQueen.[5]

An alternative clustering algorithm to the hierarchical methods is *k*-means clustering. In this clustering algorithm, the objective is to cluster *n* objects into *k* clusters, where *k* is predefined. The objective function in *k*-means clustering is to partition a given dataset into a specific number of clusters where the within-cluster sum of squares (WCSS) is minimised.

Initially, *k* points, or centroids, are placed into the same space occupied by the input objects; these are the initial group centroids. Each object is then assigned to the closest centroid in the space and becomes one of its cluster members. Once all of the objects have been assigned to the centroids, the positions of the centroids are recalculated. Each of the *k* centroids are updated to positions represented by the arithmetic mean of the values in the objects previously assigned to that centroid. The process is iterated until no, or little, variation is observed in the positions of the centroids.

k-Means clustering is somewhat faster than SAHN methods described above and can often be applied to much larger datasets. Although it is Non-deterministic Polynomial-time hard (NP-hard), numerous heuristics exist that lead to much faster runtimes in general. However, *k*-means clustering is very sensitive to the starting conditions, the initial centroid position assignments. Numerous algorithms exist that allow for the *greedy* initialisation of these positions based on the data.

8.3.4 Stirling Numbers of the Second Kind

It is often the case that an apparently simple procedure is intractable when considered in terms of enumeration. The relevant chemistry space of interest in drug design is one such example: a few atoms, from even fewer possible atoms, combined in multiple ways leads to vast space of potential chemical structures. This is referred to as combinatorial explosion. A similarly dramatic explosion occurs when considering how many ways are possible of partitioning *n* objects into *k* different clusters, the output from cluster analysis. These are called Stirling Numbers of the Second Kind. The number is calculated given the recurrence relation

$$S(N,k) = k{\cdot}S(N-1,k) + S(N-1,k-1) \tag{8.1}$$

As an example, if one wanted to cluster 1006 objects into 196 clusters, such that no cluster was empty, there are approximately 6.294×10^{1939} possible ways of achieving this, and this is a relatively trivial clustering problem. Therefore, it is perhaps not a surprise that there exist so many different clustering algorithms and heuristics therein that have been developed due to an identified limitation in the existing methods. Furthermore, given the potential space of possible solutions, it is difficult to determine what is optimal in terms of this partitioning and often a solution that appears appropriate is likely the best one can achieve. However, it should be noted that it is important, as with all modelling methodologies, to critically appraise the output the analysis to identify whether the solution is appropriate and, if not, make suitable changes to the experimental design to achieve what is required.

8.3.5 Self-Organising Maps

A different type of unsupervised learning algorithm to clustering algorithms is the Self-Organising Map (SOM) or Kohonen Map (KM). The SOM is a multidimensional scaling method that takes high-dimensional data and maps it onto a lower-dimensional space, typically two dimensions. A major advantage of SOMs is for the visualisation of groupings of data. SOMs are also described as a type of Artificial Neural Network (ANN). A key benefit of SOMs is the preservation of the topology of the data being processed. By analogy, SOMs are often compared to the visual recognition system in humans since the data is processed from a multidimensional stimulus into one-dimensional or two-dimensional neuronal structures in the brain. The Kohonen map was introduced by Teuvo Kohonen in the 1980s,[6] but is based on neurological models from the 1970s and even the work from Turing on morphogenesis models in the 1950s.[7]

The structure of a SOM is a discretised grid of cells called neurons. Each neuron consists of a weighting vector equal in length to the input vectors of the objects under consideration. In our case, the input vector length is the length of the molecular descriptor vector. Each neuron also has a defined position in the space of the map. Typically, the map itself is a grid of squares or hexagons tessellated with neighbouring neurons. The map grid itself can have a periodic boundary, in which the map edges are defined, and also a toroidal topology in which the edges wrap around, so that bottom of the map connects to the top of the map, and the right connects to the left. The toroidal topology provides a continuous space that avoids potential edge conditions that would be present with the periodic boundary.

The weight vector of each neuron (or node) is initially assigned a small random number. Another approach is to seed the weight vectors with values sampled evenly from the two principal component eigenvectors. The latter approach provides for much more rapid learning since the weight vectors already approximate the weights. However, as with many statistical learning methods, caution must be observed since this may lead to overtraining. Training examples, the molecular descriptors, are then iteratively fed into the SOM and the similarity calculated, typically Euclidean distance, between it and each neuron in the map. The training example is then mapped to the most similar neuron, or Best Matching Unit (BMU), and the weight vector of the neuron updated. The BMU weight vector is updated to reflect the training example vector, but the neighbouring neurons are also updated, but by a diminishing magnitude as the distance from the BMU increases. This process continues for each input vector for a large number of iterations. It may be necessary to fine tune the iteration limit to ensure that a reliable map of weight vectors has been generated.

Once trained with the training example vectors, the SOM is ready to use in the production mode. As for the training of the map, your dataset vectors are fed into the map and the Euclidean distance calculated between it and each and every neuron in the map. The neuron weight vector that is closest to the dataset vector is said to be the *winning vector* and the data point can be assigned to that neuron. This process is repeated for each data point in the dataset. Visual examination of the map may then be conducted such that each neuron is coloured continuously depending on the number of data points that have mapped to that neuron. The population density of a neuron reflects the redundancy of the data points in that neuron. Therefore, if you wished for a diverse subset, you might just select a single data point from each neuron in the SOM. However, if you desired close analogues to a specific data point, you could simply take all data points mapped to that neuron, and potentially neighbouring neurons should there not be sufficient representatives in the winning neuron.

Kohonen maps have been popularised in chemoinformatics by Jure Zupan and Johann Gasteiger, with a seminal volume published on the application of Kohonen maps and other ANNs.[8] One recent application of SOMs has been in the prediction of biological activities, selection of screening candidates (cherry picking), and selected representative subsets from large compound libraries such as those generated by combinatorial chemistry.[9] In the first example, that of predicting biological activity, the training examples also contain a binary variable classifying it either as active or inactive. The classification value of each training data point is not used in the training of the actual map. The resultant map is then coloured by the numbers of actives and inactives represented by each neuron. Compounds for which the biological activity was not known were then mapped to the SOM and a prediction made regarding its activity based on it representative neuron activity. The method was also demonstrated to be effective at separating different chemical series from each other in the map.

8.3.6 Principal Component Analysis

An alternative multidimensional scaling method, or data reduction method, that has been proven to be one of the most popular in chemistry data analysis, is Principal Component Analysis (PCA). PCA was invented by Karl Pearson, one of the most significant statisticians of his generation, in 1901.[10] Hotelling later developed PCA independently in the 1930s and also gave it its name.[11,12] Similarly to SOM, PCA offers an excellent method for exploratory data analysis and is also used in predictive modelling.

The simplest way to think of PCA is to imagine fitting an n-dimensional ellipsoid to your data. The longest axis of the ellipsoid is the first principal component. The second component is given by the second longest axis of the ellipsoid, and so on. By their nature, each principal component is orthogonal to each other. The first principal component explains the most variance of your data since it extends on the longest axis of the ellipsoid. Each additional principal component explains concomitantly less variance up to a cumulative value of one, and where the number of principal components is the same or fewer than the number of descriptors describing your data. The result of a PCA is new and orthogonal co-ordinate system that optimally describes the variance in a single dataset.

The outputs of a PCA, in addition to the variance explained by each principal component, are the *scores* and the *loadings*. The scores are the PCA data for the objects in your dataset, in our case the chemical structures. Plotting the data points of your dataset reveals the underlying structure of these data: the more proximate the points, the more similar those objects are with regard to the molecular descriptors used. The loadings are the PCA data for the descriptors in your dataset, in our case the molecular descriptors. Here, molecular descriptors that lie close together in the loadings plot can be said to be explaining the same variance and therefore exhibit similar behaviour. Taking the scores and loadings plots together, it is possible to understand the descriptors and how they influence the differences between the different chemical structures.

8.4 Supervised Learning

Conversely to unsupervised learning, supervised learning learns using a dependent variable, such as a biological endpoint like pIC_{50}. The concept is to use the known variable to derive a model that optimally separates the interesting from uninteresting data points, or active from inactive molecules. Many different methods have been applied in the field of computational drug discovery, and a few are provided below for further consideration of the methods.

8.4.1 Naïve Bayesian Classification

One of the simplest probabilistic classifiers that has found traction in the field is the naïve Bayesian classifier (NBC). Based on the Bayesian theory of prior probabilities to predict whether a new observation belongs to one class or another based on interpolated posterior probabilities. One of the potential weaknesses of the NBC is that the features used are assumed to be independent and contribute independently to its predicted class. However, regardless of their apparent simplicity, the NBC has been used to great effect in the chemoinformatics community.

The abstract definition of an NBC is a conditional model, such that the dependent class (C) relies on a number of features or descriptors ($F_1...F_n$):

$$p(C|F_1...F_n) \tag{8.2}$$

Assuming a random distribution of 40 red and 20 green objects, it is given that the probability of being red is two thirds (40/(40 + 20)) and of being green, one third (20/(40 + 20)). These are the prior probabilities of belonging to each of these classes.

When a new object is identified, for which we want to predict its class membership, the objects proximate to it (given a radius defined *a priori*) are identified, which gives the likelihood of being red or green. Therefore, the probability of being red is the number of red objects near the new object, divided by the total number of red objects in the larger data set. The posterior probability of being red, therefore, is the product of the prior probability of being red and the likelihood of it being red based on its location in the dataset (Figure 8.2).

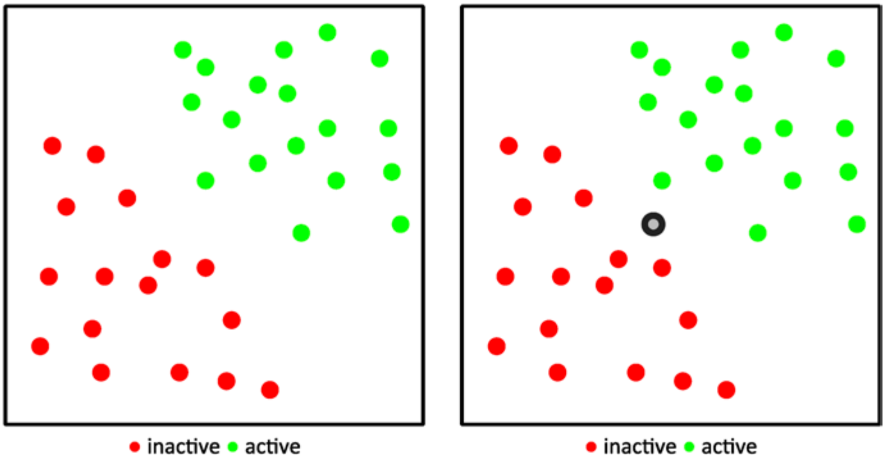


Figure 8.2 A schematic example regarding the decision as to which class a new data point belongs using the naïve Bayesian classifier.

8.4.2 Support Vector Machine

Another often-used classification modelling method used in the field is the Support Vector Machine (SVM), although the approach can also be applied in regression modelling. Dissimilarly to NBCs, SVMs are non-probabilistic binary linear classifiers. Essentially, an SVM training algorithm assigns the dataset into two classes that maximises the partition that separates those two classes. A prediction for a new object can then be made according to which class in the dataset the new object is more proximate. The partitioning takes place in the high-dimensional space equal to the length of the descriptor vectors being used. Therefore, in a 1024-bit fingerprint the partitioning is made in a 1024-dimensional space. The partitioning itself is achieved by the optimisation of a hyperplane (Figure 8.3) that maximally separates the data points in each dimension, referred to as the separation with the maximum margin.

8.4.3 Partial Least Squares

Partial Least Squares (PLS) is a regression method, as opposed to the classification methods discussed above. PLS uses similar principles to PCA, but instead identifies a linear regression model by projecting the predicted variables, molecular descriptors and the observed variables onto a new space.[13] PLS discovers the commonality between two matrices and finds the multidimensional direction in the X-space (independent variable) that maximally explains the variance in the Y-space (dependent variable). The result is a series of Latent Variables (LVs) that cumulatively improve the measure of fit of the regression model.

PLS was developed by the Swedish statistician Herman Wold, and later expanded by his son, Svante Wold, with particular application to challenges

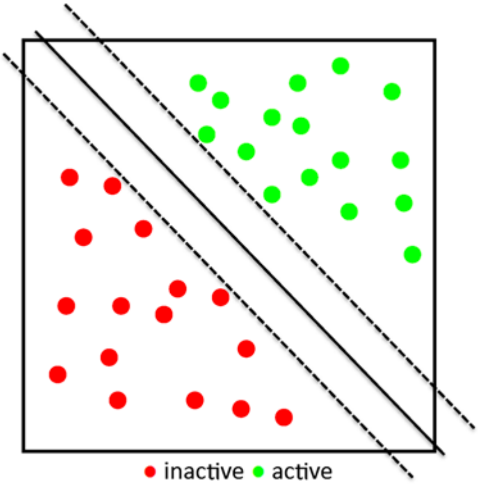


Figure 8.3 A maximum-margin hyperplane separating two classes using a support vector machine.

in chemistry. PLS is defined by Svante Wold, and he argues more correctly, as Projection to Latent Structures, but Partial Least Squares appears to be the most favoured naming.

8.5 Best Modelling Practice

As in all scientific disciplines, it is important to apply appropriate methods and safeguards to ensure that the results of the experiment are valid and relevant. This is especially the case in statistical modelling since it is very simple to develop a model that agrees with hypothesis, regardless as to whether this is valid or not. Therefore, there are number of best practices that are recommended to be applied in all statistical modelling to ensure that the predictions have value and do not lead to incorrect expectations. This chapter explores a number of recommended practices in building useful statistical models given the data available.

There are many approaches to building a statistical model, or indeed making best use of any data set you may have, and therefore it can be very easy to make small mistakes or honest misjudgements that can lead to significant errors in prediction and affect the confidence of those predictions downstream. This can then lead to even more costly mistakes in selecting synthetic targets in medicinal chemistry and subsequent assaying. However, there are a number of accepted best practices that are designed to reduce potential issues in statistical modelling such that one may have an educated confidence in the extent to which the model predictions can be used reliably to make decisions. In this chapter, we will go through accepted best practices that have been published in the literature and understand how these can be used to best effect to build useful and functional statistical models. While this is not an exhaustive consideration of the different statistical modelling best practices that have been reported, it will give a flavour of the types of things for which one should be on the look out.[14]

There is a wide range of statistical measures of confidence that are used in statistical modelling. One of the most commonly reported statistics is R^2. This provides a measure of fit between the experimental data and the modelled predictions, whether for the training set or the test set. The R^2 is a measure of fit between these two datasets as a straight line that minimises the error, or residuals, of each data point in the set. Therefore, the line of best fit can be useful in surmising general model quality, but it does have a number of issues with which it is associated.

One of the most common limitations with only trusting the R^2 value is the *dynamic range* of the data being modelled. If the range is too small, any signal in the model may be lost in the experimental and modelling error. The model may still be useful, but it will be important to look at different measures of quality, such as standard errors (*vide infra*). However, it is important to note that too narrow a range will typically be beyond the experimental limits of the data being modelled. A similar issue can be observed when the data being modelled is of a higher dynamic range, but falls in groups of data that therefore define the line of best fit. For example, considering a situation where there are two groups of data being modelled: one at a high activity and the other indicating low activity. With the line of best fit, these two groups of data in the range are essentially the two points that define the line. If one was to look at the errors of each data point, between the experimental and the modelled values, it is likely that the actual prediction errors are much larger than suggest by the R^2 statistic.

Standard errors (SE) of prediction are useful in this case where the structure of the response data is not evenly distributed in the dynamic range. Using standard errors, root-mean-square error of estimation or prediction, the actual errors from experiment of each data point can be scrutinised and summarised into a single measure of the quality of the predictions being made.

8.6 Summary

Statistical learning methods are of great importance in computational drug discovery. The learning methods summarised here allow for the understanding of the large and multivariate (and even megavariate) spaces in which we work, and also allow for making predictions about the future so that decisions can be made, with caveats, regarding the likelihood of it being successful or not.

Unsupervised learning algorithms are a statistical learning method that is useful when you need to understand the underlying structure of your data and make informed decisions regarding the data points in that space. Structuring the data in clusters (SAHN and *k*-means clustering), or distributed in discretised (SOM) or continuous (PCA) space allows for the visual inspection of the data. The applications of unsupervised learning methods have been discussed in terms of numerous challenges in chemoinformatics, such as: activity prediction, consideration of chemical clusters in data analysis, selection of representative or diverse subsets to reduce the number of chemical structures that must be taken into more computationally intensive algorithms, such as docking or through to synthesis (or purchase) and biological testing.

One positive to unsupervised learning that is very important to consider, yet ostensibly the most obvious, is that the methods offer different views of the same data. Looking at your data in different ways is one of the most important steps in exploratory data analysis and should not be underplayed.

Supervised learning offers potentially even great power than unsupervised learning in that it becomes possible to not only understand the space in which you are working, but make predictions about the spaces that are most likely to bear the most fruit in terms of success for a drug discovery project. With supervised statistical learning it is possible, given sufficient data and data quality, to derive models that can accurately classify or even predict a value for a biological endpoint by using the extant data alone. The chapter on quantitative structure–activity relationships will discuss this in greater detail.

Statistical learning methods are very important in the toolbox of the modeller and their power should not be under-estimated. However, it is important to understand how they work, their limitations and strengths, and also the appropriate measures of success that indicate the degree to which a model can be trusted. There are many parameters to consider and only a few have been presented here. Working from the position that all models are wrong can be useful as it helps to focus efforts on understanding whether predictions that are made are valid and can be used prospectively.

References

1. H. E. Driver and A. L. Kroeber, Quantitative expression of cultural relationships, *University of California Publications in Am. Arch. and Ethn.*, 1932, **31**(4), 211–256.
2. R. Sibson, SLINK: an optimally efficient algorithm for the single-link cluster method, *Comput. J.*, 1973, **16**(1), 30–34.
3. S. P. Lloyd, Least squares quantization in PCM, *IEEE Trans. Inf. Theory*, 1982, **28**(2), 129–137.
4. S. P. Lloyd, Least square quantization in PCM. Bell Telephone Laboratories Paper, 1957.
5. J. B. MacQueen, Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability 1*, University of California Press, 1967, pp. 281–297.

6. T. Kohonen, Self-Organized Formation of Topologically Correct Feature Maps, *Biol. Cybern.*, 1982, **43**(1), 59–69.
7. A. Turing, The chemical basis of morphogenesis, *Philos. Trans. R. Soc.*, London, 1952, vol. 237, pp. 5–72.
8. J. Zupan and J. Gasteiger, *Neural Networks in Chemistry and Drug Design: An Introduction*, Wiley-VCH, Weinheim, Germany, 1999.
9. P. Selzer and P. Ertl, Applications of Self-Organizing Neural Networks in Virtual Screening and Diversity Selection, *J. Chem. Inf. Model.*, 2006, **46**(6), 2319–2323.
10. K. Pearson, On Lines and Planes of Closest Fit to Systems of Points in Space, *Philos. Mag.*, 1901, **2**(11), 559–572.
11. H. Hotelling, Analysis of a complex of statistical variables into principal components, *J. Educ. Psychol.*, 1933, **24**, 417–441, and 498–520.
12. H. Hotelling, Relations between two sets of variates, *Biometrika*, 1936, **27**, 321–377.
13. M. Haenlein and A. M. Kaplan, A Beginner's Guide to Partial Least Squares Analysis, *Understanding Stat.*, 2004, **3**, 283–297.
14. L. Eriksson, J. Jaworska, A. P. Worth, M. T. Cronin, R. M. McDowell and P. Gramatica, Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs, *Environ. Health Perspect.*, 2003, **111**(10), 1361.

Part 5
Modelling Methodologies

CHAPTER 9

Similarity Searching

9.1 Overview

Modern rational drug discovery relies significantly on the concept of molecular similarity, since molecular similarity also suggests similarity in the biological end-point of interest. While empirical and subject to exceptions, the similar-property principle is a highly effective approach to identifying interesting structural analogues that are likely to invoke similar interactions and therefore prioritise compounds for purchase or synthesis based on vast virtual libraries.

Molecular similarity is increasingly used in virtual sets of compounds to prioritise those for testing. The prioritisation approach can be as simple as ranking a list of virtual structures compared with a single reference ligand of interest. The approach can be extended to multiple ligands of interest when these are available, and the resulting structure ranks combined using an approach called, data fusion or consensus scoring.

Another challenge in similarity searching is that structures that are identified as being similar to a known or probe ligand are not necessarily similar to each other. Indeed, as the similarity to the probe decreases, the potential for similarly scored hits to be different to each increases in probability. Therefore, it may also be prudent to consider clustering the identified hits to prioritise specific groups for follow-up, rather than just taking the top one hundred hits, which may not evenly represent the available space.

In the development and enhancement of virtual screening methods, it is important to have, as in all science, appropriate controls. An appropriate positive control in virtual screening is comparison to an extant method that is known to work effectively and is typically used widely. However, this

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

comparison requires metrics to identify whether a newly developed method is actually effective for the intended purpose. There are many ways to quantify and visualise the success or otherwise of a virtual screening method. A number of these approaches that are commonly used will be covered in this chapter, but many are also covered in the supervised statistical learning chapter.

Similarity searching is highly effective when one only has information regarding a single, or perhaps a few, ligands of interest. However, even in situations where one can generate a pharmacophore model from multiple ligands or a structural model, or even when a protein–ligand crystal structure is available, it is important not to forget the power of similarity searching. It is not only an effective approach to identify the most interesting structures to consider from a large list, it is also typically incredibly rapid in calculation compared to other methodologies, such as docking. Therefore, similarity searching is typically represented in some form in any virtual screening cascade. The key is to use the tools available on the available data to benefit a drug discovery programme and also improve the probabilities of success.

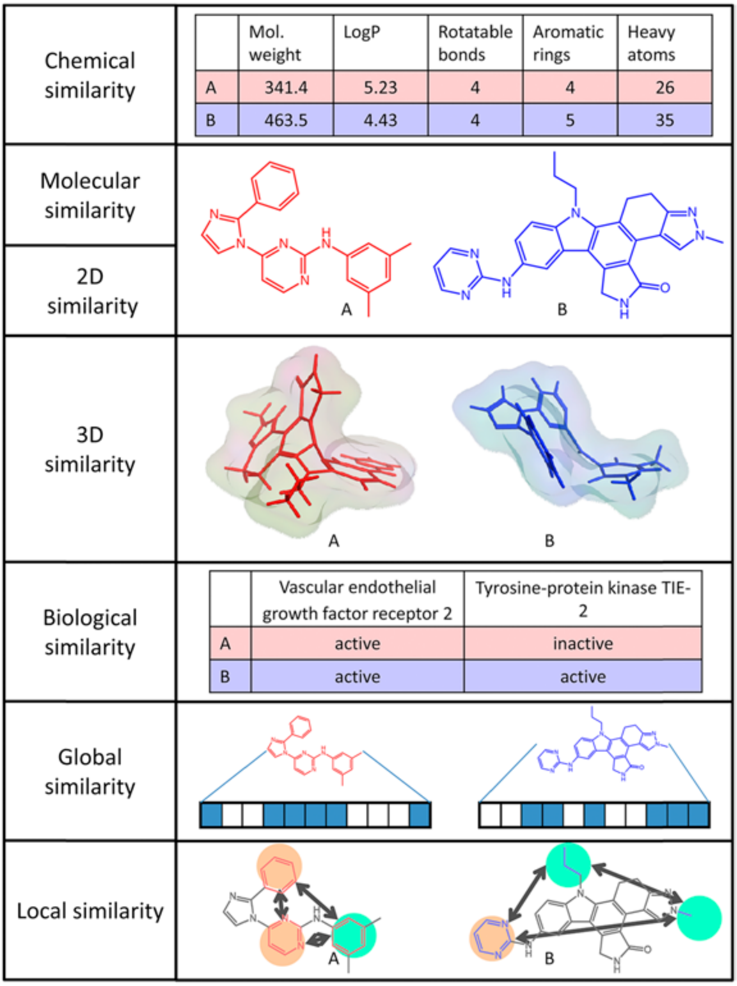
9.2 Similar Property Principle

The similar-property principle has a long history, going back to Alexander Crum Brown and Fraser in 1868,[1] where their work consisted of "performing upon a substance a chemical operation which shall introduce a known change into its constitution, and then examining and comparing the physiological action of the substance before and after the change." This can be seen as an early and empirical investigation of molecular similarity. Alexander Crum Brown later stated clearly that physiological action is a function of chemical constitution, leading to what is arguable the first Quantitative Structure–Activity Relationship (QSAR) model: "It is obvious that there must exist a relation between the chemical constitution and the physiological action of a substance, but as yet scarcely any attempts have been made to discover what this relation is."

Clearly, the concept of the similar-property has been known for well over a century, but it was not formalised until 1990 by Johnson and Maggiora as "similar compounds have similar properties."[2] However, as we have discussed previously, the concept of molecular similarity is quite philosophical, so care must be taken and due consideration given to the appropriate molecular descriptor for a particular similarity search. The similar-property principle is also called neighbourhood behaviour, and is a key concept in the field.[3]

Similarity searching of chemical databases was not commonplace until the pioneering work of Carhart *et al.*[4] and Willett *et al.*[5] in the mid-1980s. Until the advent of these methods, most chemical databases were limited to structure and substructure searches.

Molecular similarity is a contentious area and much discussion has been published in the field with regard to what makes a useful measure of molecular similarity. Two excellent references for additional discussion and conclusions are Kubinyi[6] and Maggiora *et al.*[7] An example of the main classes of molecular similarity is given in Figure 9.1 from Maggiora *et al.*, which covers chemical (or physicochemical) similarity, molecular and two-dimensional (2D or topological) similarity, three-dimensional (3D, topographic, or geometric), biological similarity, global similarity and local (or pharmacophoric) similarity.



	Mol. weight	LogP	Rotatable bonds	Aromatic rings	Heavy atoms
A	341.4	5.23	4	4	26
B	463.5	4.43	4	5	35

	Vascular endothelial growth factor receptor 2	Tyrosine-protein kinase TIE-2
A	active	inactive
B	active	active

Figure 9.1 Similarity perception and concepts. Two exemplary vascular endothelial growth factor receptor 2 ligands are shown, and different ways to assess their similarity are illustrated. Reprinted with permission from G. Maggiora, M. Vogt, D. Stumpfe and J. Bajorath, Molecular Similarity in Medicinal Chemistry: Miniperspective, *J. Med. Chem.*, 2013, **57**(8), 3186–3204. Copyright 2013 American Chemical Society.

9.3 Molecular Similarity and Virtual Screening

The concept of molecular similarity has been discussed already in Chapter 4, but some aspects are worthy of reiteration here. The general concept of molecular similarity is an entirely philosophical one and can depend on the context in which the comparators are compared. Thus, two molecular structures can be seen to be topologically similar if they share the same core structure, or they may be said to be similar if the surface electronics are similar regardless of the underlying topological structure. However, for reasons of pragmatism, and getting the job done, it is important to select a descriptor of relevance, and a similarity coefficient that are deemed to be appropriate for the challenge at hand: a molecular descriptor that fulfils the requirements of the similar-property principle.

In Similarity Searching, which can also be called Virtual Screening (VS), molecular similarity is important since the objective is to identify molecules from a large library of potential hits that are more likely to be those hits. Many different molecular descriptors and similarity measures have been demonstrated to be effective at enriching the number of active molecules recalled in a ranked list, at least retrospectively.

The benefit of conducting a Virtual Screen, which can be achieved using any of the modelling methods discussed in this section, as opposed to performing a full High-Throughput Screen is illustrated schematically in Figure 9.2.

The aim of a virtual screen is to screen as fewer molecules as possible to enrich actives that are then validated in far few real experiments. There is a trade-off to be had in how many compounds need to be screened *in vitro* with regard to the probabilities of success from the virtual screen. However, this potential success can vary substantially from project to project. Another key potential advantage of performing a virtual screen and then *in vitro* experiments is that it is possible to return to the screening library, with some knowledge in hand from the smaller screen guided by virtual screening, and attempt a virtual screen again to identify further compounds.

9.4 Data Fusion

Data fusion, often called Consensus Scoring in docking applications, is a method by which multiple ranked lists are combined from different experiments to generate a new ranked list.[8] The anticipation with data fusion is that the combined ranked list will be superior to conducting a single experiment by the introduction of additional methods and/or chemical probes, in the case of virtual screening like similarity searching. Another driving factor for using data fusion is that particular methods, although typically quite effective, can often under-perform relative to expectations, and it would not be possible to know this *a priori* when the objective is to reduce the predictions to the physical experiments of synthesis and biological testing. This approach is also used with disparate methods, such as shape-screening and

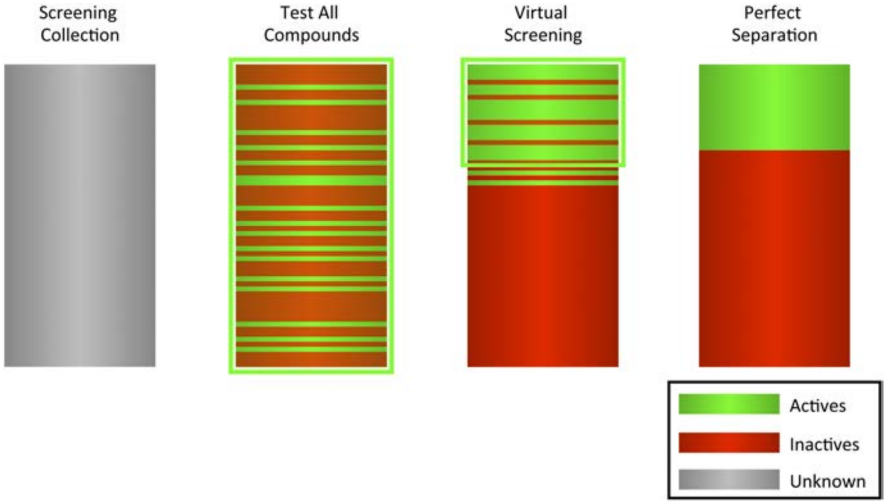


Figure 9.2 The potential benefits of conducing a virtual screen over a full biological screen is illustrated schematically, with the left-most cylinder representing the entire set of compounds in a screening library prior to any testing. The second cylinder represents a full screening campaign where the entire collection has been tested, and the small number of active molecules identified in green. With the knowledge of the full screen, it is now possible to demonstrate what the potential number of active molecules would be in the right-most cylinder, a highly unlikely perfect separation of active and inactive molecules. The third cylinder, where the molecules in the full deck were tested *in silico* and ranked according to the results, represents the virtual screen and only a top slice of the ranked list is tested *in vitro* giving enrichment in active molecules.

Table 9.1 List of common Data Fusion rules applied to challenges in chemoinformatics.

Fusion rule	Formula
MAX	$\max\{S_1(d_j),S_2(d_j)...S_n(d_j)\}$
MIN	$\min\{S_1(d_j),S_2(d_j)...S_n(d_j)\}$
SUM	$\frac{1}{n}\sum_{1}^{n} S_i\left(d_j\right)$
MED	$\mathrm{median}\{S_1(d_j),S_2(d_j)...S_n(d_j)\}$

virtual ligand docking, since it has been observed that when one under-performs, the other method tends to maintain effectiveness (Table 9.1).

There are a large number of ways that a set of ranked or quantitative lists may be combined to generate a single output for prioritisation. The SUM method was described by Ginn *et al.* in which the rank positions of each point in the ranked lists are summed giving a new position.[9] The summed list is then re-ordered by the descending value of the summed ranks and a new ranking given.

9.5 Enrichment

There are many methods to quantify and visualise the output from a similarity search or virtual screen. The main concern when selecting a virtual screening model to apply to *in vitro* experiments is the potential for enrichment in first 1–5% of the ranked list. A number of methods have been reported to evaluate the enrichments of virtual screens. Some of these enrichment evaluation statistics are given below.

9.5.1 Lift Plots

The lift plot, or enrichment curve, is one of the simplest methods by which the ability for a virtual screening method to recall active molecules can be measured. This is particularly important in early recall, and trivial to understand and explain, making it perfect for presentations particular to audiences who may not have seen such kinds of results previously.

The enrichment curve is an *xy*-plot, where the *x*-axis represents the precise number, or percentage, of molecules screened in a database and the *y*-axis represents the precise number, or percentage, of active molecules retrieved from that database.

A schematic lift (or enrichment) plot is given in Figure 9.3. The *x*- and *y*-axes represent the percentage of the total ranked database that have been screened and the percentage of active structures identified from that percentage of the database, respectively. The perfect separation line is highlighted in green, where this would represent that all actives (*a*) were discovered in the

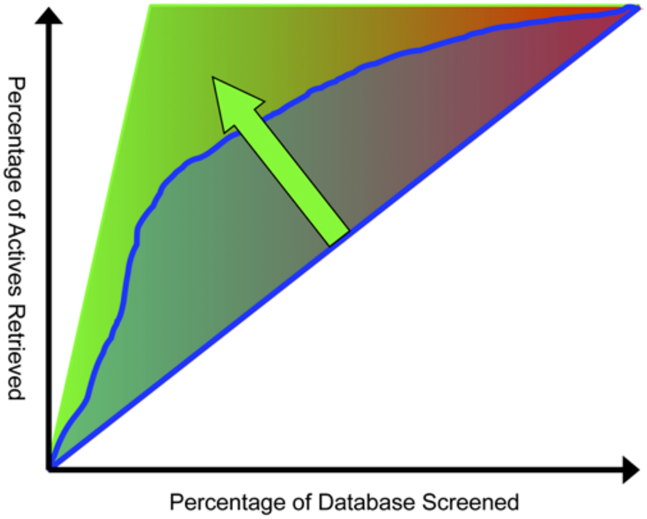


Figure 9.3 An enrichment plot demonstrating perfect separation (green line) and a typical enrichment rate (blue line), above a random enrichment (the diagonal line).

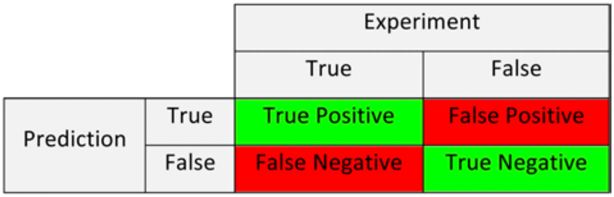
first *a* structures in the screened database. The random recall is specified by the diagonal line, which indicates that after *n*% of the ranked database has been screened, only *n*% of the active structures are recalled. The diagonal specifies a random model that contributes nothing above a random search of the entire library. The third trace is an idealised typical virtual screening result, indicating the objective of moving as far from the diagonal (random recall line) towards the perfect separation line.

9.5.2 Confusion Matrix

The confusion matrix is one of the simplest methods of summarising the classification of objects into their observed classes using a predictive model. The confusion matrix for two categories, active and inactive structures, represents how well the model has correctly classified the objects. The confusion matrix allows for a much more reliable and detailed analysis of the predictive power of a classifier than the accuracy of the model (accuracy simply means the proportion of correct guesses). Accuracy is not reliable in datasets where there is an imbalance in the numbers of objects in the classes. This is important as typically in drug discovery we have far fewer actives than inactives.

The *pro forma* confusion matrix of experimental outcome *versus* predicted class is given in Figure 9.4. Additional statistics can be calculated from the confusion matrix using the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), and these are summarised in Table 9.2.

Two of the most important statistical measures of classifier model quality are sensitivity and specificity. Sensitivity is the measure of the proportion of actual positives that were correctly predicted as such. Similarly, specificity is the measure of actual negatives that were correctly predicted as such. A perfect model would have 100% sensitivity and 100% specificity, which means all actives and inactives are correctly predicted as such, respectively.



		Experiment	
		True	False
Prediction	True	True Positive	False Positive
	False	False Negative	True Negative

Figure 9.4 The generic confusion matrix for the two-class predictive model against experiment (gold standard) and prediction (test outcome). The higher the values in the green cells, the better the separation of the positives and negatives, whereas higher values in the red cells would represent a model of less quality in terms of separating the classes.

Table 9.2 Summary of the terminology of the confusion matrix and the derivation statistical measures of quality and their respective equations.

Terminology	Equation
Positive	P
Negative	N
True positive (hit)	TP
True negative (correct rejection)	TN
False positive (Type I error, false alarm)	FP
False negative (Type II error, miss)	FN
Sensitivity or true positive rate (TPR)	TPR = TP/P = TP/(TP + FN)
Equivalent with hit rate, recall	
Specificity (SPC) or true negative rate (TNR)	SPC = TN/N = TN/(FP + TN)
Precision or positive predictive value (PPV)	PPV = TP/(TP + FP)
Negative predictive value (NPV)	NPV = TN/(TN + FN)
Fall-out or false positive rate (FPR)	FPR = FP/N = FP/(FP + TN)
False discovery rate (FDR)	FDR = FP/(FP + TP) = 1 − PPV
Miss rate or false negative rate (FNR)	FNR = FN/P = FN/(FN + TP)
Accuracy	ACC = (TP + TN)/(P + N)

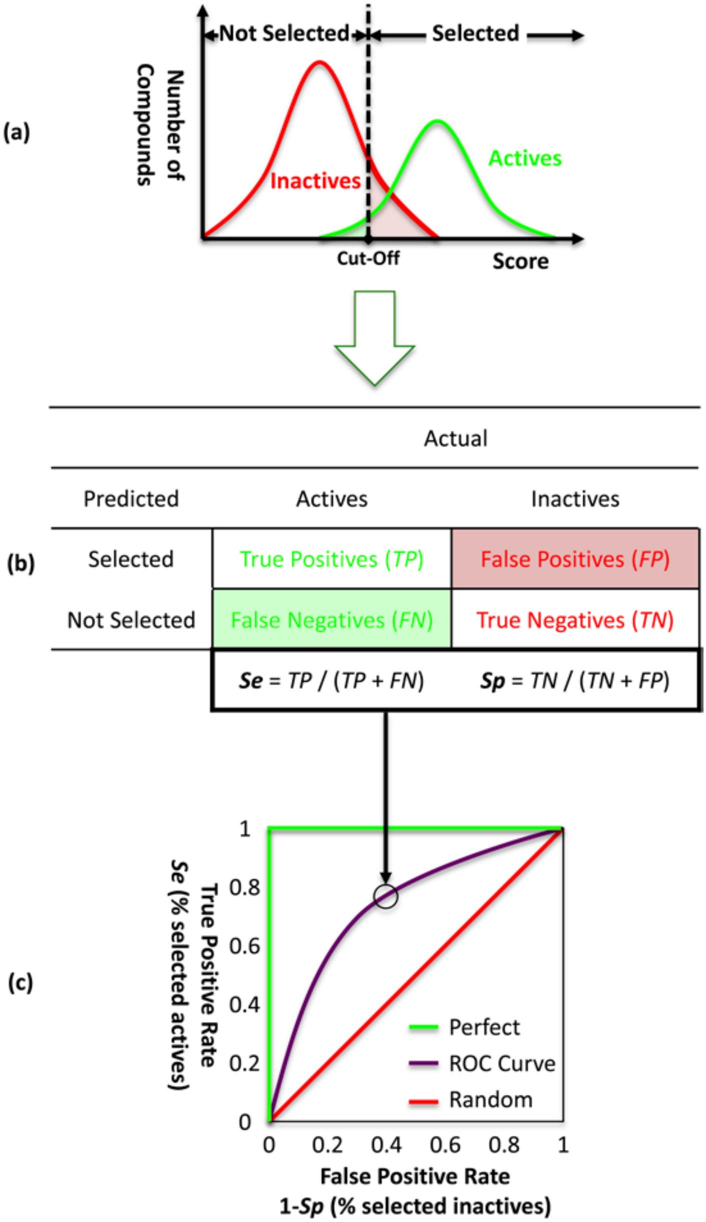
9.5.3 Receiver Operating Characteristic Curves

An alternative to the lift plot, and much more popular in reporting virtual screening results for good reason, is the Receiver Operating Characteristic (ROC) curve. Again, the ROC curve is an *xy*-plot, but this time the *x*-axis represents the false-positive (1 − specificity) rate, and the *y*-axis represents the true-positive (sensitivity or recall) rate. That is, as compounds are removed from the top of the ranked list, the point moves along the *x*-axis one unit if the compound is inactive and one unit up the *y*-axis if the compound is active. The diagonal of the plot represents a random recall of actives and inactives, with a line plotting the extreme upper triangle indicating perfect separation of the data, and a line plotting the extremes of the lower triangle of the plot indicating that all actives were found in the last of the ranked list (Figure 9.5).

The ROC curve was developed during the Second World War by electrical and radar engineers to detect objects in battlefield situations. The methodology was soon introduced to psychological applications to understand how stimuli are perceived. ROC curves are now used widely in many sciences, including: medicine, radiology, biometrics, machine learning and data mining research. One of the early applications of ROC curves in drug discovery was by Triballeau *et al.*, where they emphasised its benefits over the enrichment curve.[10]

9.5.4 Enrichment Factors

Another method to summarise the results of a virtual screen is called the Enrichment Factor (EF). EFs are calculated as the number of experimentally discovered active structures in the top *x*% of the sorted databases of active and inactive structures. Enrichment factors (EF) after *x*% of the prioritised library are calculated according to eqn (9.1), where $N_{\text{experimental}}$ = number of experimentally discovered active structures in the top *x*% of the sorted database,



	Actual	
Predicted	Actives	Inactives
Selected	True Positives (*TP*)	False Positives (*FP*)
Not Selected	False Negatives (*FN*)	True Negatives (*TN*)
	Se = *TP* / (*TP* + *FN*)	***Sp*** = *TN* / (*TN* + *FP*)

Figure 9.5 ROC curves in a nutshell. (a) Theoretical distributions of scores are obtained for both actives (red) and inactives (blue) after processing the sample by a suitable computer test. For intelligibility of the figure, it was hypothesised that the scores for both active and inactive compounds had normal (*i.e.*, Gaussian) distributions, although they are unlikely to be so in a usual case. Generally, these distributions overlap, leading to false predictions (coloured area). Upon threshold modification (dashed line), proportions of such erroneous classifications (reported in a confusion matrix (b)) change dramatically. (c) For all possible score thresholds, the evolution of the deduced sensitivity (Se) and specificity (Sp) is reported on a ROC graph with Se as a function of 1 − Sp. Calculating the area under the ROC curve is a practical way to quantify the overall performance of the computer test.

N_{expected} = number of expected active structures in the top $x\%$, and N_{active} = number of active structures in the whole database:

$$\text{EF} = \frac{N_{\text{experimental}}^{x\%}}{N_{\text{expected}}^{x\%}} = \frac{N_{\text{experimental}}^{x\%}}{N_{\text{active}} \cdot x\%} \tag{9.1}$$

Enrichment factors are particularly useful in comparing virtual screening results on the same datasets, but when applying different virtual screening methodologies. Typically, in virtual screening studies, the EFs would be reported in the top 1%, 5% and 10% to indicate how far down the ranked lists must be travelled to find a suitable number of active structures.

9.6 Summary

Similarity searching is one of the most frequently used methods in computational medicinal chemistry applications. It can be applied in the analogue-by-catalogue approach to identify close analogues that are likely to be active according to the probe molecule used in the search. Analogue-by-catalogue is a useful approach to identify follow-up hit matter in HTS-triage and will be discussed fully in Chapter 15 (*vide infra*).

A similar approach to analogue-by-catalogue, but with a different focus and ambition, is virtual screening by similarity searching. The objective in similarity searching (which will be discussed later in this book) is, given a database of chemical structures and a chemical probe (or multiple probes), structures similar in the database to the probe will enrich on active structures. Using databases, such as those from HTS screens, it is possible to estimate the performance of a virtual screening methodology prospectively by conducting retrospective experiments on this database. Once performed, numerous measures of performance can be reported—such as lift plots, ROC curves and EFs—which will confer a quantitative measure of the performance of each method.

Similarity searching is vitally important in chemical information retrieval systems, where it must be performed very rapidly, often using pre-computed descriptors or calculated indices, and also in virtual screening methods, where computation time is not necessarily as important as the quality of the result. In the latter, it is more important to have reliable enrichments of predicted active structures. The balance to be sought here is between speed of computation for routine similarity searches and quality of prediction for project-critical virtual screens using similarity searching. However, the objective should always be towards improving enrichment regardless of application.

References

1. A. C. Brown and T. R. Fraser, V.—On the connection between chemical constitution and physiological action. Part. I.—On the physiological action of the salts of the ammonium bases, derived from strychnia, brucia, thebaia, codeia, morphia, and nicotia, *Trans. R. Soc. Edinburgh*, 1868, **25**(01), 151–203.

2. A. M. Johnson and G. M. Maggiora, *Concepts and Applications of Molecular Similarity*, John Willey & Sons, New York, 1990, ISBN 0-471-62175-7.
3. D. E. Patterson, R. D. Cramer, A. M. Ferguson, R. D. Clark and L. E. Weinberger, Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors, *J. Med. Chem.*, 1996, **39**(16), 3049–3059.
4. R. E. Carhart, D. H. Smith and R. Venkataraghavan, Atom Pairs as Molecular Features in Structure–Activity Studies: Definition and Applications, *J. Chem. Inf. Comput. Sci.*, 1985, **25**, 64–73.
5. P. Willett, V. Winterman and D. Bawden, Implementation of Nearest Neighbour Searching in an Online Chemical Structure Search System, *J. Chem. Inf. Comput. Sci.*, 1986, **26**, 36–41.
6. H. Kubinyi, Similarity and Dissimilarity: A Medicinal Chemist's View, *Perspect. Drug Discovery Des.*, 1998, **9**(11), 225–252.
7. G. Maggiora, M. Vogt, D. Stumpfe and J. Bajorath, Molecular Similarity in Medicinal Chemistry: Miniperspective, *J. Med. Chem.*, 2013, **57**(8), 3186–3204.
8. P. Willett, Combination of Similarity Rankings Using Data Fusion, *J. Chem. Inf. Model.*, 2013, **53**, 1–10.
9. C. M. R. Ginn, P. Willett and J. Bradshaw, Combination of Molecular Similarity Measures Using Data Fusion, *Perspect. Drug Discovery Des.*, 2000, **20**, 1–16.
10. N. Triballeau, F. Acher, I. Brabet, J. P. Pin and H. O. Bertrand, Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4, *J. Med. Chem.*, 2005, **48**(7), 2534–2547.

CHAPTER 10

Bioisosteres and Scaffolds

10.1 Overview

One of the most important aspects of molecular design is the identification of appropriate substituents and molecular cores that contribute not only to the ligand binding event, but are often also implicated in other properties that are important to optimise in drug design. This chapter introduces the concept of bioisosterism in chemical structures, which enables the identification of functionally equivalent chemical moieties.[1] A brief history of bioisosterism, tracing its routes to the beginning of the 19th century, will be provided. The concept of bioisosteric replacement will be described and how it is used in tuning molecular properties that are important for drug design.[1] A subset of bioisosteric replacement, scaffold hopping, will be introduced and its importance in drug design placed in context.[2] The history of molecular scaffolds in drug discovery will be covered and a selection of popular algorithms for their determination for use in computational analyses is presented. The use of objective and invariant scaffold representations for the analysis of scaffold diversity, or otherwise, of databases, such as marketed drugs, biologically relevant compounds and screening libraries, will be discussed and conclusions drawn on the challenges in designing libraries and how we may use molecular scaffolds to improve these libraries. Lastly, a brief overview of scaffold hopping will be introduced, where the objective is to replace the core functional—either for scaffolding or functional interactions, or both—elements to modulate different parameters that, for example, are important in drug discovery.

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

10.2 A Brief History of Bioisosterism

The concepts of isosterism can be traced back to James Moir in 1909.[3] However, it took a further ten years before the subject was given its name, isosterism, by the famous chemist, Irving Langmuir.[4] Langmuir took the name isostere from the Greek for same (*isos*) and solid shape (*stereos*), literally meaning the same shape. In his seminal work, *Isomorphism, Isosterism and Covalence*, Langmuir identified isosterism according to the different electron configurations in groups of atoms.

Langmuir used the octet rule to identify isosteric groupings of a number of substances and their measured properties. The octet rule is a rule-of-thumb in chemistry that, at least for small molecules (fewer than 20 heavy atoms), eight electrons are preferred in the outer shell valence shell of a particular atom. This relationship was demonstrated by Langmuir to hold true for nitrogen and carbon monoxide with regard to their physical properties (Table 10.1). The same observation was made for nitrous oxide and carbon dioxide through application of data from the Landolt–Börnstein tables and Abegg's handbook (Table 10.2).[5]

The concept of isosterism was extended by Grimm in 1925,[6] extending Langmuir's definition of isosterism through the incorporation of his hydride displacement law, thus:

> *"Atoms anywhere up to four places in the periodic system before an inert gas change their properties by uniting with one to four hydrogen atoms, in such a*

Table 10.1 List of isosteres defined by Langmuir in 1919.[4]

Type	Isosteres
1	H^-, He, Li^+
2	O^{2-}, F^-, Ne, Na^+, Mg^{2+}, Al^{3+}
3	S^{2-}, Cl^-, A, K^+, Ca^{2+}
4	Cu^+, Zn^{2+}
5	Br^-, Kr, Rb^+, Sr^{2+}
6	Ag^+, Cd^{2+}
7	I^-, Xe, Cs^+, Ba^{2+}
8	N_2, CO, CN^-
9	CH_4, NH_4^+
10	CO_2, N_2O, N_3^-, CNO^-
11	NO_3^-, CO_3^{2-}
12	NO_2^-, O_3
13	HF, OH^-
14	ClO_4^-, SO_4^{2-}, PO_4^{3-}
15	ClO_3^-, SO_4^{2-}, PO_4^{3-}
16	SO_3, PO_3^-
17	$S_2O_6^{2-}$, $P_2O_6^{4-}$
18	$S_2O_7^{2-}$, $P_2O_7^{4-}$
19	SiH_4, PH_4^+
20	MnO_4^-, CrO_4^{2-}
21	SeO_4^{2-}, AsO_4^{3-}

> *manner that the resulting combinations behave like pseudo-atoms, which are similar to elements in the groups one to four places respectively, to their right."*

In this work, Grimm, further classified isosteres into two groups: classical and non-classical isosteres. Classical isosteres were those chemical moieties that included monovalent, bivalent, trivalent, tetravalent, and ring equivalences (Table 10.3). The non-classical isosteres included the carbonyl group,

Table 10.2 Experimental data from the Landolt–Börnstein tables and Abegg's handbook for nitrous oxide (N_2O) and carbon dioxide (CO_2).

Property	N_2O	CO_2
Critical pressure (atm)	75	77
Critical temperature (°C)	35.4	31.9
Viscosity at 20 °C	148×10^{-6}	148×10^{-6}
Heat conductivity at 100 °C	0.0506	0.0506
Density of liquid at −20 °C	0.996	1.031
Density of liquid at +10 °C	0.856	0.858
Refractive index of liquid at 16 °C	1.193	1.190
Dielectric constant of liquid at 0 °C	1.598	1.582
Magnetic susceptibility of gas at 40 atm, 16 °C	0.12×10^{-6}	0.12×10^{-6}
Solubility in water 0 °C	1.305	1.780
Solubility in alcohol at 15 °C	3.25	3.13

Table 10.3 Some examples of classical bioisosteres—the groups in each row are equivalent.

Monovalent bioisosteres

OH, NH

OH, NH or CH_3 for H

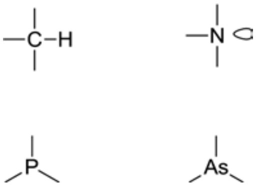
SH, OH

Cl, Br, CF_3

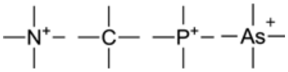
Divalent bioisosteres

C=S, C=O, C=NH, C=C

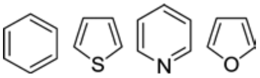
Trivalent atoms or groups



Tetrasubstituted atoms



Ring equivalents



carboxylic acid, hydroxyl, catechol, halogens, amides, esters, thiourea, pyridine and cyclic *versus* acyclic groups.

The term bioisosterism was not itself introduced until 1951 by Friedman,[7] where the term broadened out the concept isosterism such that:

> *"We shall term compounds 'bio-isosteric' if they fit the broadest definition for isosteres and have the same type of biological activity."*

Clearly, this was a much broader concept than that originally proposed and studied by Langmuir, but the introduction of 'bio' into the concept necessarily introduces a degree of fuzziness since we still do not yet have a clear grasp of many pathways and interactions in the human physiology.

10.3 Bioisosteric Replacement Methods

The identification of appropriate bioisosteres for replacement in a particular chemical structure can be achieved using a number of different approaches. Broadly speaking, the methods fall into two camps: knowledge-based and information-based.

10.3.1 Knowledge-Based Bioisosteric Replacements

Knowledge-based bioisosteric replacement methods include those methods where bioisosteric pairings have been identified from experimentally observed phenomena. The first database of bioisosteres to be compiled and distributed is the BIOSTER™ database from Ujvary.[8] In this database, Ujvary has manually curated many thousands of bioisosteric pairings observed in the literature. The database from Ujvary is comprehensive and covers the past 40 years of literature.

The DrugGuru (Drug Generation Using RUles) system was developed at Abbott Laboratories as a tool to help medicinal chemists design their next synthetic targets, a type of *de novo* design system. To this end, DrugGuru uses a well-curated, albeit only available internally, database of molecular transforms for consideration by the chemists in designing their next compounds. Combining the unbiased database of transforms and the intuition of an expert medicinal chemist can lead to much better design ideas than using either approach alone. The inclusion of various predictive models also assists in the generation of new ideas that will progress drug design projects much further.

The ChEMBL database is the largest publicly available database of small-molecule chemical structures and associated biological endpoints. Using this database, one can use the matched molecule pair (MMP) concept to identify all chemical structures that differ in only one position. Using this concept and associated measured of calculated data, relevant MMPs can be identified as potential bioisosteres. One system that uses

ChEMBL and MMPs for the identification of appropriate bioisosteric replacements is *SwissBioisostere*, which has been made available at no cost online.[9]

Moving away from observed bioisosteric pairs from literature data and biological endpoints, the Cambridge Crystallographic Data Centre (CCDC) has used their comprehensive database of small molecule crystal structures, the Cambridge Structural Database (CSD), to mine bioisosteric pairs. Importantly, using the experimentally observed geometric data, the CSD can be used to select potential replacements that exquisitely mimic the geometries required for protein interactions.

10.3.2 Information-Based Bioisosteric Replacements

Complementary approaches to the database approaches to identify bioisosteric pairs are the information-based or descriptor-based methods. Descriptor-based methods, while not necessarily as reliable in practice as knowledge-based methods due to their lack of experimental evidence, can open up the space of available replacements that can be considered extensively and truly look to the future in designing new drugs. The molecular descriptor methods available tend to fall into one of four subsets: physicochemical properties; topological descriptors, such as molecular fingerprints; molecular shape; and protein–ligand environments (although this can be seen as a hybrid between knowledge-based and information-based methods).

The classical descriptors for identifying bioisosteres are the Hammett sigma constants, which play a significant role in describing the electron-donating or electron-accepting power of potential replacements, and the Hansch parameter, which is defined as the difference between the octanol–water partition coefficient (logP) of a substituted molecule and its parent.

One of the most used topological descriptor types for bioisosteric replacement, and indeed scaffold hopping, are the ligand-based topological pharmacophores. This class of descriptor attempts to characterise potential three-dimensional (3D) pharmacophoric representations using atomic abstractions into their potential functionality and using through-graph distances as a surrogate for the through-space distances one would normally consider in a 3D pharmacophore system.

One of the first ligand-based topological pharmacophores to be published, at least for the identification of bioisosteres, was the Similog descriptor. Similog encodes all atom triplets within a molecule and their shortest through-graph distances between each other. The atoms are abstracted according to the DABE scheme: potential hydrogen bond donor; potential hydrogen bond acceptor; bulkiness; and electro positivity. Each atom is therefore represented by a four-bit descriptor describing these three properties, and the triplet encoded with the distance information into Similog fingerprints. Schuffenhauer[10] studied this fingerprint in whole molecule similarity searching to identify structurally dissimilar but functionally similar molecules, so not strictly speaking a bioisosteric replacement tool, but one could see how this could be extended into identifying specific bioisosteric replacements.

A similar ligand-based topological pharmacophore is the Chemically Advanced Template Search (CATS) fingerprint, which was developed explicitly for scaffold hopping. CATS vectors have been discussed in great detail in the earlier chapter on topological descriptors, and the reader is referred to that chapter for more details on their construction. In summary, CATS vectors represent atoms as abstract pharmacophoric types and all pairs of atoms are encoded into the fingerprint with their associated shortest through-graph distance between those atoms. The power of CATS, and indeed Similog, fingerprints is that they disconnect from the underlying connectivity of the molecules under consideration and additionally abstract the atoms into generic pharmacophoric features. These approaches therefore introduce a level of 'controlled fuzziness' to the descriptors to identify molecules that are potentially functionally similar, but not necessarily similar in structure.

10.4 Scaffold Representations

Not long after Langmuir introduced the concept of isosterism, Eugene Markush introduced the concept of molecular scaffolds.[11] The concept that Markush introduced was not truly for reasons of chemistry, but instead to assist in the protection of a chemical series of dyes in one of his patent applications. The Markush structure is now widely used in medicinal chemistry patents to protect a chemical series of compounds rather than just a single compound. However, the definition of a Markush structure as a scaffold of a series is highly subjective and typically the driving factor in the final decision is the ability to patent the structure and series: *i.e.* the Markush structure should be specific enough to be patentable, but also sufficiently generic so as to maximise the region of chemistry space covered. It should also be mentioned, and very importantly, that everything that is patented must be able to be synthesised within reason.

The Markush structure, while used effectively in patents and intellectual property protection, it is important for computational analyses to define objective and invariant approaches to scaffold definitions that can be generated rapidly, unambiguously and consistently.

The earliest reference to a molecular scaffold, as one might define it today, found so far was published in 1969 by Reich and Cram[12] and is defined as: "The ring system is highly rigid, and can act as a scaffold for placing functional groups in set geometric relationships to one another for systematic studies of transannular and multiple functional group effects on physical and chemical properties."

10.5 Scaffold Diversity Analysis

Scaffold diversity is an important characteristic of a screening library to study. The appropriate balance between representing scaffolds and how many representatives are required is a difficult one to obtain. Using molecular scaffold representation methods, it is possible to identify the level of scaffold diversity available. The task here is to identify the scaffolds that are present in a given library using an appropriate scaffold representation method and then calculate the frequency of occurrence of each of these scaffolds.

The Molecular Framework representation from Bemis and Murcko was one of the early scaffold representation methods to be applied to the understanding of scaffold diversity in drug libraries. Bemis and Murcko published an analysis of 5120 drugs in 1996.[13] The Molecular Framework as defined by Bemis and Murcko separates out the ring systems, linkers and side chains in molecular structures as defined fragments. The Molecular Framework is then defined as the substructure represented by all of the rings and linkers in the structure under consideration. A Molecular Framework retains the atom labels and bond orders (Figure 10.1e), whereas the Graph Framework is more abstract, retaining only the nodes and edges and not further information, often represented as a carbon skeleton with single bonds only (Figure 10.1g).

Considering the set of marketed drugs of 5120 unique chemical structures, Bemis and Murcko identified 1179 and 2506 unique graph and molecular frameworks, respectively. Bemis and Murcko then moved on to analyse the distribution of these scaffolds over the set of drugs, finding that 42 of the molecular frameworks are each represented in 10 or more drugs, which represents a total of 1235 drugs, or 24% of all drugs. Furthermore, 1908 of the molecular frameworks are represented in a single compound, representing 76% of the molecular frameworks. This represents an uneven distribution of little scaffold diversity in a vast portion of marketed drugs.

Bemis and Murcko extended their scaffold analysis of marketed drugs using the graph framework, identifying a similar uneven distribution of scaffold representation. This analysis demonstrates that marketed drugs have a small number of highly represented scaffolds and many scaffolds that have only one parent molecule represented.

A similar study to that of Bemis and Murcko was conducted by Lipkus *et al.*,[14] but this time considering the scaffold diversity of the Chemical Abstracts Service (CAS) Registry. The CAS Registry is a far more substantial dataset to work with, containing at the time 24 282 284 unique organic molecules. The Lipkus study again considered the molecular and graph frameworks, but also considered hetero frameworks (Figure 10.1f), where the atom labels are retained, but the bond orders are all set to single bonds. This study found 836 708 unique graph frameworks, 2 594 334 hetero frameworks, and 3 380 334 molecular frameworks.

Considering the hetero frameworks, the Lipkus study considered their frequency of occurrence in the CAS Registry. Overall, 75.5% of the structures from the CAS Registry represented only 5% of the hetero frameworks identified. This result demonstrated that the database is heavily skewed towards

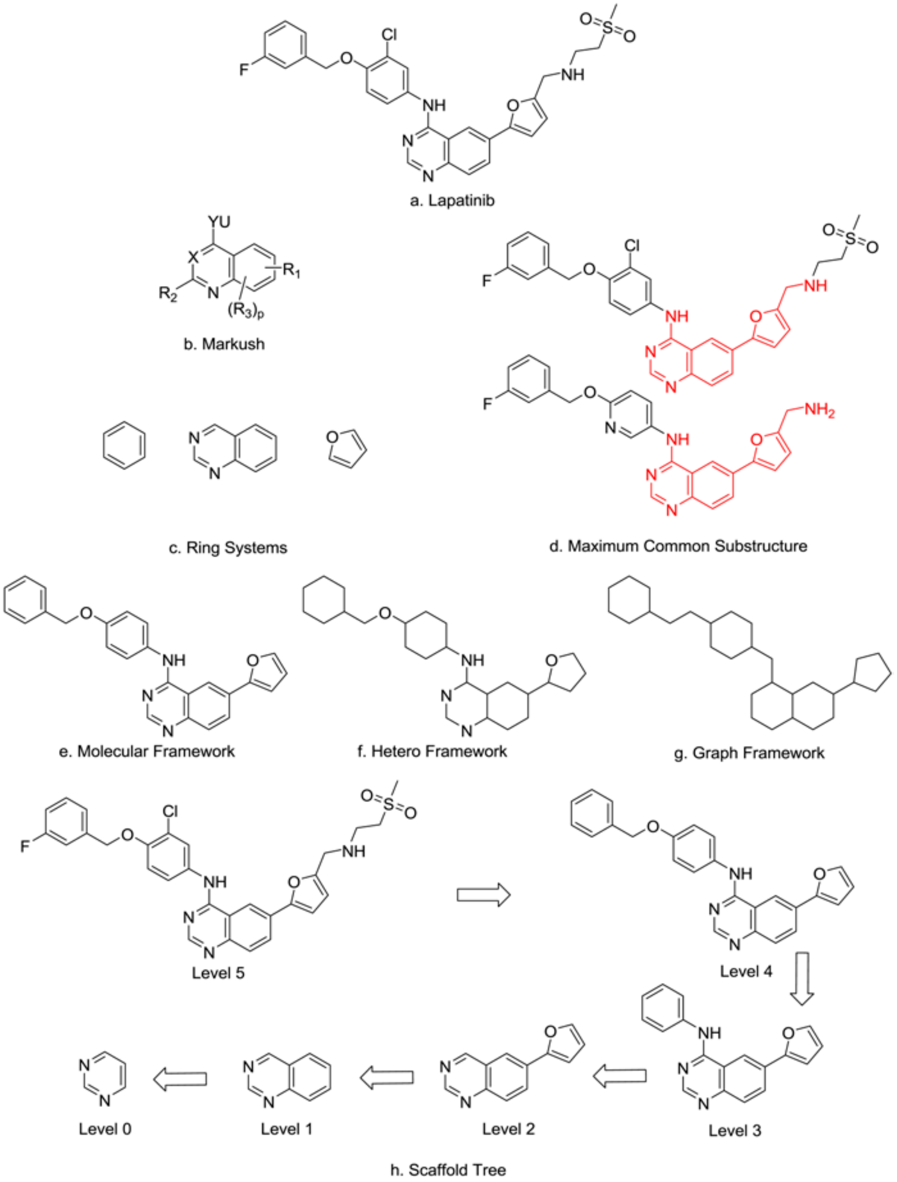


Figure 10.1 The kinase inhibitor lapatinib and exemplar scaffold representations. (a) Lapatinib, (b) Markush structure, (c) Ring Systems, (d) Maximum Common Substructure, (e) Molecular Framework (also known as the Murcko Scaffold, or Bemis and Murcko Scaffold), (f) Hetero Framework, (g) Graph Framework and (h) the Scaffold Tree. Reproduced from S. R. Langdon, N. Brown, and J. Blagg, Scaffold diversity of exemplified medicinal chemistry space, *J. Chem. Inf. Model.*, 2011, **51**(9), 2174–2185.

a small number of scaffolds that represent a vast proportion of the library. Furthermore, taking the top 10 most-frequently occurring scaffolds, it was found that just those few hetero frameworks, representing less than one thousandth of one per cent of the total hetero frameworks, represented 12.7% of the whole Registry.

While the set of marketed drugs and the CAS Registry are clearly good datasets to conduct these scaffold diversity analyses, the typical application of this type of analysis is to assist in designing more representative screening libraries. Therefore, some work has been undertaken to investigate these types of libraries, including the sets from where screening libraries are typically purchased, vendor collections. Analyses of the more appropriate screening-like libraries have also been done using different scaffold diversity methods and the results are quite similar to those above.

A scaffold composition study based on Maximum Common Substructures (MCS) was carried out on 17 screening libraries taken from 12 different suppliers, giving a total of 2.4 million compounds. The MCSs identified in the libraries were categorised as classes if they represented at least two compounds or singletons if they represented only one compound. It was found that for all libraries there were more “singletons” than “classes” and that the distribution of molecules over “classes” is highly skewed with a few very highly populated scaffolds. Several metrics are used to assess the distribution of compounds over scaffolds in the libraries; these will be discussed in more detail in Chapter 11.

One recent analysis of scaffold diversity compared the Molecular Frameworks from Bemis and Murcko with a scaffold representation based on the Scaffold Tree work from Schuffenhauer *et al.*[15] This study, by Langdon *et al.*,[16] considered a library of marketed drugs, a vendor collection, medicinal chemistry structures from the literature (ChEMBL), but also an extant fragment library and a high-throughput screening library that are used routinely at The Institute of Cancer Research, London. The aim of this study was to demonstrate potential issues in using the Molecular Framework representation and a new scaffold representation algorithm that goes some way to reducing certain artefacts of the Molecular Framework representation. The main challenge to Molecular Frameworks in this study was that the resulting scaffolds retained the vast majority of the parent molecule, and this is somewhat distant from what a medicinal chemist would identify as a chemical scaffold. Typically, most substituents or functional groups explored in a drug discovery project will contain ring systems. Therefore, the Molecular Framework would also retain these moieties, which would typically not be retained in a medicinal chemistry analysis.

This study reiterated the previous work with both the Molecular Frameworks and the Level 1 scaffolds. Furthermore, the Level 1 scaffolds demonstrated even more markedly the lack of diversity of coverage of scaffolds in screening libraries. The Level 1 scaffold representation is less granular a representation than the Molecular Framework, since it contains less of the parent molecule. However, the representation is much more aligned to a medicinal chemistry representation of a scaffold.

The studies above have demonstrated that medicinal chemistry relevant libraries contain a very few, but highly represented, scaffolds with a substantial number of singleton scaffolds. Langdon *et al.* proposed potential reasons for the lack of representation of some scaffolds and the significant exploration of others.[16] The challenges are most likely a combination of biological activity being limited to small regions of the chemical space and synthetic, and therefore also commercial availability, accessibility of the scaffolds making them less attractive as medicinal chemistry compounds. The synthetic tractability challenge may also be due to certain scaffolds offering simple syntheses that permit many more analogues to be synthesised during a medicinal chemistry programme.

Much work has gone into the generation, analysis and assessment of molecular scaffolds in the context of drug discovery. The overlap between drugs, bioactive libraries and commercially available screening libraries has been analysed by Shelat and Guy.[17] This work considered the extent to which potential screening libraries have relevance in the biological space. The Shelat and Guy study used the molecular framework as their scaffold representation of choice and identified that commercially available bioactive molecules have a large overlap with scaffolds that are represented in marketed drugs. However, screening libraries generated using the Lipinski rule-of-five or libraries made using diversity-orientated synthesis methodologies represented only a small fraction of scaffolds identified in marketed drugs. Drugs and bioactive compounds will tend to have similar scaffolds represented in the libraries since they are active against the same, well-explored target families. Indeed, many of the bioactives have probably been synthesised and tested within the medicinal chemistry programme or programmes that eventually gave rise to the marketed drug. However, the rule-of-five and diversity-orientated synthesis generated libraries may have activities against those biological targets that have not yet been explored significantly, or simply not tested against the more established targets. One must remember that absence of evidence does not necessarily indicate absence of evidence. That is, one cannot assume inactivity if the compounds simply have not been tested. Therefore, it is important to explore these underrepresented regions of chemical space since they may bear fruit, but this comes at the cost of typically more complex synthesis, further biological testing and the possibility that the endeavour may simply fail due to being over-ambitious in exploring these uncharted territories. However, this should not dissuade scientists from exploring these spaces, but they should to do so with caution and appropriate modelling practices in place to maximise the probabilities of success.

10.6 Summary

The identification and appropriate replacements of functional groups and molecular scaffolds is an important aspect of rational drug design. However, the process by which they are defined and identified is by no means simple, since there is not really a rigorously definable concept applicable in all domains.

The concept of isosterism and bioisosterism stretches back a century, and much research and consideration has been applied to the concept. Many tools have been developed that can be used to identify bioisosteric pairs from molecular descriptors and by mining chemical databases. While the methods are by no means perfect, the approaches do offer an opportunity to narrow down the space of possible replacements to be considered and therefore can provide a much more rational chemistry space to explore through virtual library enumeration.

Molecular scaffold representations are similarly challenging to define since there is no commonly accepted concept to allow definition. However, many different scaffold representation approaches have been published that allow for the objective and invariant generation of molecular scaffolds. These scaffolds can then be compared directly using a variety of methods to understand the diversity of scaffolds covered, which can assist in library design and triaging hit lists from high throughput screening.

Once a scaffold definition can be accepted in the context of a project then possible scaffold replacements can be identified using similar approaches to identifying bioisosteres—since scaffold hopping is simply a subset of bioisosteric replacement.

The methods discussed in this chapter have done a great deal to assist medicinal chemistry thinking and decision making in recent times. The approaches permit rational thought and assist design teams in prioritising the most interesting possible compounds for synthesis. As time progresses and methods undoubtedly improve, bioisosteric replacements and scaffold hopping will become integral components of the medicinal chemistry toolbox. However, we are still some way away from this goal and much research into the fundamental methodologies is needed to make these methods sufficiently generic for application to new challenges.

References

1. *Bioisosteres in Medicinal Chemistry*, ed. N. Brown, Wiley-VCH, Weinheim, Germany, 2012.
2. *Scaffold Hopping in Medicinal Chemistry*, ed. N. Brown, Wiley-VCH, Weinheim, Germany, 2014.
3. A. Burger, Isosterism and bioisosterism in drug design, *Prog. Drug Res.*, 1991, **37**, 288–362.
4. I. Langmuir, Isomorphism, isosterism and covalence, *J. Am. Chem. Soc.*, 1919, **41**, 1543–1559.
5. R. W. H. Abegg and F. Auerbach, *Handbuch der Anorganischen Chemie*, Leipzig, Hirzel, 1909.
6. H. G. Grimm, On construction and sizes of non-metallichydrides, *Z. Elektrochem. Angew. Phys. Chem.*, 1925, **31**, 474–480.
7. H. L. Friedman, Influence of isosteric replacements upon biological activity, in *Symposium on Chemical-Biological correlation*, National Academy of Science-National Research Council, Publication, 1951.

8. I. Ujvary, BIOSTER: a database of structurally analogous compounds, *Pestic. Sci.*, 1997, **51**, 92–95.
9. M. Wirth, V. Zoete, O. Michielin and W. Sauer, SwissBioisostere: a database of molecular replacements for ligand design, *Nucl. Acids Res.*, 2013, **41**, 1137–1143.
10. A. Schuffenhauer, P. Floersheim, P. Acklin and E. Jacoby, Similarity metrics for ligands reflecting the similarity of the target proteins, *J. Chem. Inf. Comput. Sci.*, 2003, **43**(2), 391–405.
11. E. A. Markush, *Pyrazolone dyes and process of making the same*, US Patent Number 1,506,316, 1924.
12. H. J. Reich and D. J. Cram, Macro rings: XXXVII. Multiple electrophilic substitution reactions of [2.2] paracyclophanes and interconversions of polysubstituted derivatives,*J. Am. Chem. Soc.*, 1969, **91**, 3527–3533.
13. G. W. Bemis and M. A. Murcko, The properties of known drugs. 1. Molecular frameworks,*J. Med. Chem.*, 1996, **39**, 2887–2893.
14. A. H. Lipkus, Q. Yuan, K. A. Lucas, S. A. Funk, W. F. Bartelt III, R. J. Schenck and A. J. Trippe, Structural diversity of organic chemistry. A scaffold analysis of the CAS Registry,*J. Org. Chem.*, 2008, **73**(12), 4443–4451.
15. A. Schuffenhauer, P. Ertl, S. Roggo, S. Wetzel, M. A. Koch and H. Waldmann, The scaffold tree-visualization of the scaffold universe by hierarchical scaffold classification,*J. Chem. Inf. Model.*, 2007, **47**, 47–58.
16. S. R. Langdon, P. Ertl and N. Brown, Bioisosteric replacement and scaffold hopping in lead generation and optimization, *Mol. Inf.*, 2010, **29**, 366–385.
17. A. A. Shelat and R. K. Guy, Scaffold composition and biological relevance of screening libraries, *Nat. Chem. Biol.*, 2007, **3**(8), 442–446.

CHAPTER 11

Clustering and Diversity

11.1 Overview

The number of possible molecules that can be purchased or synthesised necessitates the application of methods to rationalise the list of feasible solutions to a number that is pragmatic and cost effective, but still gives a high chance of success with regard to the ongoing challenges of the therapeutic project concerned.

The related approaches of clustering and diversity selection belong to the set of unsupervised learning methods. The approaches seek to understand the general structure of the data sets under consideration and use that structure to pre-select clusters of interest or subsets that are representative of the data set as a whole.

In this chapter, a number of clustering and diversity selection methods will be presented that have been applied in the field of computational drug discovery and chemoinformatics.[1,2] The types of clustering and diversity methods used are illustrated schematically in Figure 11.1: dissimilarity-based compound selection, sphere exclusion, clustering and cell-based selection.[3] The emphasis will not be placed on the appropriate molecular descriptors that can be used with these algorithms, rather the relative advantages and disadvantages of these methods, particularly in terms of the quality of output and computational complexity. For instance, a particular clustering algorithm may be highly desirable in terms of the quality of the resultant clusters, but may be computationally intractable for even modestly sized data sets.

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

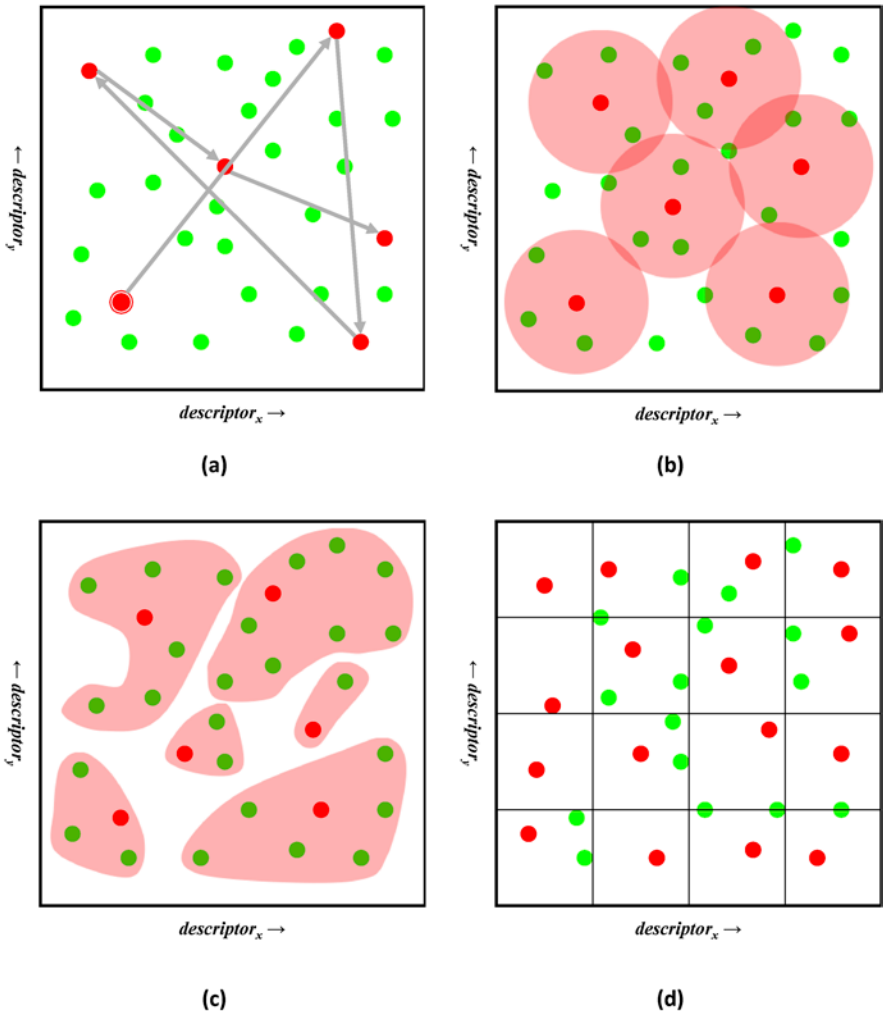


Figure 11.1 Schematic diagrams of clustering and diversity selection methods as a structured approach to exploring the data set under consideration. Qualitative illustration of different diversity selection algorithms. (a) Minimising mean pairwise similarity (MPS) using a dissimilarity-based compounds selection (DBCS) method. (b) Sphere exclusion. (c) Clustering. (d) Cell-based selection.

11.2 Dissimilarity-Based Compound Selection

Sometimes called iterative selection, dissimilarity-based compound selection (DBCS) algorithms work iteratively. An initial seed data point is selected at random or using an heuristic to identify an appropriate starting point, such as the minimum average distance from every other point in the dataset provided or that is closest to the centre of the data set using an appropriate criterion. The algorithm then proceeds to select subsequent points from the data set based on a particular scoring function.[4]

```
subset = []
subset = database[i]
for i = 1 to n
  calculateSimilarity(subset, database)
  append subset maximumDissimilar(subset, database)
```

Algorithm 11.1 Pseudocode of the maximum-dissimilarity selection method using a dissimilarity-based compound selection algorithm.

A number of different algorithms for selecting the next compound in a DBCS algorithm have been proposed, including MaxMin and MaxSum. MaxMin scores each potential new compound to be selected by finding the closest compound that has the highest dissimilarity to it; *i.e.* the next compound to be selected will have its nearest neighbour as the most distant to it when considering all of the other points for selection. MaxSum, however, selects the next compound based on the sum of distances between the compound being considered and the subset so far. The compound with the maximum sum of distances will then be selected. The equations for the scoring functions of MaxMin and MaxSum are given in eqn (11.1) and (11.2), respectively.

$$\text{MaxMin}:\text{score}_i = \text{minimum}(D_{ij}) \tag{11.1}$$

$$\text{MaxSum}:\text{score}_i = \sum_{j=1}^{m} D_{ij} \tag{11.2}$$

One potential limitation of applying the DBCS diversity selection approach is that the computational complexity is $O(n^2N)$, where n is the number of molecular structures to be selected and N is the size of the total dataset from which the subset is to be selected. However, $O(nN)$ algorithms have been reported for both MaxMin and MaxSum, which makes these approaches highly appropriate for clustering and diversity selection. While the application to diversity selection is obvious, the application to clustering is perhaps not so apparent. Once a diverse subset of a certain size has been selected, the diverse subset can be seen as a set of cluster centroids that cover the large space. Once the centroids have been defined, it is trivial to then extend each centroid in turn by incorporating neighbouring molecular structures to be subsumed into the cluster represented by that centroid. This can then be iterated until all molecular structures are contained within a cluster or they are so distal to other clusters that they form their own cluster, a singleton.

MaxMin and MaxSum operate in different ways that can give very different results that may be desirable or undesirable depending on the application of the algorithms. MaxSum tends to pick subsets that represent the extremities of the space being considered, whereas MaxMin, although it begins similarly, will eventually begin to fill the gaps between the points and represent the space in a more balanced distribution. Therefore, MaxSum might be appropriate for sampling the limits of the space under consideration and MaxMin is more appropriate when the objective is to identify a representative subset over the whole space.

11.3 Sphere-Exclusion

While DBCS algorithms were initially designed solely for diversity selection, the algorithm may also be applied as a clustering algorithm. By specifying the number of desired clusters using a heuristic, the diverse points can be used as the cluster centroids. Each cluster can then be expanded from the centroid using a simple neighbourhood metric based on the original descriptors used in the algorithm. Those data points that are closer to one cluster centroid than another will be selected to be absorbed into that cluster. This approach, implemented in PipelinePilot's *Cluster Molecules* component, is very cost effective compared with clustering algorithms that do not scale well, but this speed increase comes at the expense of the quality of the resultant clusters.

Sphere exclusion tends to start from the centre of the space under consideration and expand outwards, making the method more representative of the overall space than the DBCS algorithms above. However, what sphere exclusion gains in representation, it can suffer in diversity.

11.4 Cell-Based Diversity Selection

In contrast to the distance-based (or dissimilarity) methods of the DBCS approaches above, cell-based selection uses a property space discretised into bins (in one-dimension), cells (in two-dimensions), volumes (in three-dimensions), and hyper-volumes (in greater than three-dimensions). The properties selected for cell-based selection, as for all of these methods, should be appropriate for the problem that is being addressed. For example, if you had a large number of 'hit' compounds around a common scaffold then you might consider taking a subset that covers the molecular weight range to effectively sample the space covered by the available structures rather than over-represent those structures closer to the mean of the property being considered.

In the molecular weight example described above, the problem can be addressed by simply discretising the molecular weight into bins of, say, 50 daltons in size. Structures can then be selected from those falling into each bin according to the one, for example, that is closest to the centre of that bin.

An advantage of using cell-based diversity selection methods is that it is not necessary to calculate the distance or dissimilarity between each structure in your dataset. While relatively trivial in computation for low-dimension selections, this still requires $n(n-1)/2$ relative calculations, therefore $O(n^2)$, where n is the number of structures being considered. This permits the consideration of much larger datasets than may be otherwise analysed, such as those coming from commercial vendors. Furthermore, you may also increase the speed of calculation in higher-dimensional spaces, such as molecular fingerprints, by using a dimensionality-reduction method such as Principal Components Analysis (PCA) or Multi-Dimensional Scaling (MDS), with awareness of the concomitant reduction in resolution of data space.

Another advantage of cell-based diversity selection methods is that they represent evenly the entirety of the space under consideration, including those cells that are under-represented. In certain circumstances, you may be interested in both low-occupancy cells and high-occupancy cells. In low-occupancy cells, these compounds (similar to singletons in clustering) may be interesting because they cover underexploited chemistry space. However, one must express caution in these low-occupancy cells since the reason may simply be that the synthesis is challenging and therefore more difficult to conduct an analogue-by-catalogue search and follow-up synthesis. The low-occupancy cells may though offer interesting structures since they are therefore, by definition, under-explored in medicinal chemistry projects. High-occupancy cells may also be of great interest since they offer rapid follow-up in terms of analogue-by-catalogue searches because structures are likely to be readily available from compound vendors. The availability may also reflect the ease of synthesis and therefore facilitate rapid synthetic follow-up to test specific hypotheses: a key element of early stage drug design.

11.5 Hierarchical Clustering

Perhaps the most effective way to both find natural groupings of molecular structures and identify diverse subsets over a whole space is to use a clustering algorithm. However, clustering algorithms are typically very computationally intensive and this must be taken into account when clustering large datasets with many descriptors.

Cluster analysis aims to partition a large number of points into natural groups, where points within the groups are more similar to ones outside the group. Once clustered, the data has additional metadata concerning its cluster membership and this can be used to select individual clusters for specific analysis or select a representative or representatives from each cluster as members of a diverse subset. Clustering can also be useful for identifying singletons that may require specific analyses or perhaps discarded since they do not offer a great deal in terms of information and potential to explore structure–activity relationships.

To conduct a cluster analysis, it is important to first select and calculate the descriptors appropriate for the analysis. For instance, if the objective were to identify structurally similar molecular structures within clusters, it would not make much sense to use physicochemical descriptors. Once the descriptors are calculated, it is then necessary to calculate the pairwise similarity or distance matrix. This step is important since many comparisons will be made and it is much faster to have these calculated and available in a look-up matrix rather than re-calculated every time the comparison is required. Additionally, some consideration of the similarity or distance measure used should be made to ensure artefacts do not appear in the final clustering. Here, only non-overlapping clustering methods will be considered, where each compound is assigned to only one cluster. Typically, one of two clustering methods is applied in chemoinformatics: hierarchical or non-hierarchical.

Agglomerative hierarchical clustering is one of the most common clustering algorithms and begins with all compounds as individual clusters. The two most similar clusters (or singleton compounds) are then merged into a single cluster. The algorithm then iterates, identifying the two most similar clusters at each iteration and merging them. There are a number of different methods of identifying the similarity of clusters. Single linkage, or nearest neighbour, clustering calculates the minimum distance between two compounds, one from each cluster. Conversely, complete linkage, or furthest neighbour, calculates the furthest distance between two compounds, again one from each cluster. By far the most computationally intensive, although arguably more appropriate, is the group average method, which calculates the cluster similarity based on the average of the all-by-all distances between each compound in both clusters. One more approach that is often used, particularly in chemoinformatics, is Ward's method.[5]

Ward's method forms clusters so as to minimise the total variance in the resulting cluster, also known as the minimum variance method. The variance of a given cluster is calculated by the sum of the square of deviations of each compound from the cluster mean. The algorithm proceeds by identifying the two clusters to merge that result in the smallest change in total variance.

Divisive hierarchical clustering takes the opposite approach, by starting with all compounds in one cluster and iteratively partitioning the sets until all clusters contain only one compound. Divisive hierarchical clustering often results in poorer clusters than identified when using agglomerative methods. However, the divisive approach permits the user to terminate the algorithm when the requisite number of clusters is met, which is often typically small. This can make the calculation of the clustering much faster.

Once a hierarchical clustering analysis has been conducted, it is important to decide at which point in the hierarchy to define the optimal set of clusters, a stopping rule. The most common stopping rule applied in chemoinformatics is called the Kelley function. The objective of the Kelley function is to optimise the balance between the number and the spread of the clusters.[6] The optimal number of clusters is the one given by the smallest Kelley value.

11.6 Non-Hierarchical Clustering

In contrast to hierarchical clustering, non-hierarchical methods assign compounds to clusters with no formation of a hierarchical relationship between the clusters. The most common non-hierarchical clustering algorithm used in chemical structure analysis is the Jarvis–Patrick method.[7] The Jarvis–Patrick clustering algorithm requires two integers to be defined to determine cluster placements: *m*, nearest neighbours, and *p*, nearest neighbours in common.

Two compounds are put into the same cluster if they are contained in their respective lists of m nearest neighbours, and they have p nearest neighbours in common. Empirically derived typical values the parameters are $m = 14$, and $p = 8$. Jarvis–Patrick can suffer from disparity in identifying small numbers of large clusters and large numbers of singletons. However, additional parameterisation, such as a similarity cut-off above which to define nearest neighbours, can assist in removing the tendency to identify both large and small clusters.

Another clustering algorithm that is popular in chemistry applications is *k*-means clustering. *k*-Means is a type of relocation clustering method.[8,9] An initial number of clusters is defined, c; typically this is defined as $\sqrt{}(n/2)$ as a rule-of-thumb, where n is the number of objects in the entire dataset being clustered. The initial seeds are typically selected at random or can be selected using a greedy initialisation heuristic. The remaining compounds in the set are then assigned to each of the c clusters, according to the initial seed that is closest to it. Once the first pass has been conducted, the cluster centroids are calculated for each cluster, c. Each point is then reassigned to the closest cluster centroid. This process is repeated until there is no change in cluster allocations or a termination condition is met, such as a maximum number of iterations. It is especially important in *k*-means clustering to identify the level of sensitivity to selecting the initial random cluster seeds, since these seeds can lead to very different clustering. However, in practice, *k*-means offers a computationally inexpensive clustering algorithm that has found widespread use in drug discovery.

11.7 Summary

A wide range of clustering algorithms and diversity selection methods have been investigated and reported in the literature. As with many challenges in the field, there is no single, right answer for every occasion. Arguably, the hierarchical cluster algorithms provide the preferred clustering, not least because the algorithms result in a defined hierarchy of cluster relationships that can be adjusted as required. However, what they offer in refinement of result they lack in the speed at which they can be calculated. Conversely, the non-hierarchical methods tend to be rapid in calculation, but can suffer in quality of output and they do not offer the power of hierarchical methods in selecting different levels of clusters.

The methods reported in this chapter offer solutions to generate appropriate groupings of molecules in a dataset, clusters, and for diverse subset selection. It is important to appropriately identify the molecular descriptors to be used, the measure by which their similarity (or distance) is calculated, the clustering algorithm to apply, and lastly how to apply the results to the challenge at hand. If the requirement is to find delineated clusters that can be considered for further analysis then one of the hierarchical methods is most likely preferred, but can be time consuming. Similarly, to select diverse subsets it is important to understand what is meant by diversity. Diverse sets could be the ones that cover the extremities of the space or that distribute evenly over the entirety of the space. As with all modelling methods, it is important to understand the application prior to selecting the algorithms and other methods since the potential application will affect these decisions. If possible, multiple methods should be used and, importantly, the results visualised to identify whether 'natural' clusters are being identified. While it would be nice to have a generally applicable clustering or diversity selection for all applications, this is wishful thinking and it is still necessary, even given the considerable amount of research in this area, to fully consider the range of approaches and desired outputs.

References

1. J. M. Barnard and G. M. Downs, Clustering of chemical structures on the basis of two-dimensional similarity measures, *J. Chem. Inf. Comput. Sci.*, 1992, **32**, 644–649.
2. G. M. Downs, P. Willett and W. Fisanick, Similarity searching and clustering of chemical-structure databases using molecular property data, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 1094–1102.
3. A. Schuffenhauer and N. Brown, Chemical diversity and biological activity, *Drug Discovery Today: Technol.*, 2006, **3**, 387–395.
4. M. Snarey, N. K. Terrett, P. Willett and D. J. Wilton, Comparison of algorithms for dissimilarity-based compound selection, *J. Mol. Graphics Modell.*, 1997, **15**, 372–385.
5. J. H. Ward Jr., Hierarchical grouping to optimize an objective function, *J. Am. Stat. Assoc.*, 1963, **58**, 236–244.
6. L. A. Kelley, S. P. Gardner and M. J. Sutcliffe, An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies, *Protein Eng.*, 1996, **9**, 1063.
7. R. A. Jarvis and E. A. Patrick, Clustering using a similarity measure based on shared near neighbors, *IEEE Trans. Comput.*, 1973, **100**, 1025–1034.
8. S. P. Lloyd. Least square quantization in PCM. Bell Telephone Laboratories Paper, 1957.
9. E. W. Forgy, Cluster analysis of multivariate data: efficiency versus interpretability of classifications, *Biometrics*, 1965, **21**, 768–769.

CHAPTER 12

Quantitative Structure–Activity Relationships

12.1 Overview

It might seem obvious to us today that the physiological action of a substance is related to its chemical structure; this is the foundation of the similar-property principle after all. However, what might be more surprising is that this observation was first made in 1868 by two leading natural scientists in Scotland, Alexander Crum Brown and Thomas R. Fraser.[1] Essentially, this relationship could be defined as:

$$\Phi = f(C) \tag{12.1}$$

where Φ is the physiological action, or biological activity, and C is the chemical constitution or chemical structure. It is clear that Crum Brown's work in atomistic theory and the chemical structures of compounds would have influenced his thinking in understanding these structural relationships.

Based on the principle from Crum Brown and Fraser, a Quantitative Structure–Activity Relationship (QSAR) is a mathematical model with some associated predictive error.

12.2 Free–Wilson Analysis

Much later than Crum Brown and Fraser, Free and Wilson described their later eponymous methodology (Free–Wilson analysis) for understanding the quantitative contribution that groups or other structural elements could make to a common parent structure.[2] While this was very popular at its time, and is still applied successfully today, the model assumes that group

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

contributions are linear in nature and do not offer any so-called superadditivity, where the groups individually are greater than the sum of their parts, or indeed weaker in many cases. The linear contribution is due to Free–Wilson analysis only considering the presence or absence of groups and not their potential in combination. Free–Wilson analysis uses statistical regression methods to generate models that are fragment-based, or more properly group-contribution generated.

In Free–Wilson analysis, the group contributions are non-overlapping, but the fragment contributions may also be overlapping. This type of group-contribution QSAR can be seen in application in a wide range of physicochemical property QSAR, such as many of those for calculated logP.

12.3 Hansch–Fujita Analysis and Topliss Trees

Corwin Hansch and Toshio Fujita developed one of the first logP calculators that was based on empirical data.[3–5] Hansch analysis subsequently demonstrated the importance of the logP partition coefficient in drug discovery.

An alternative, and arguably more interpretable, modelling method was introduced by John Topliss based on Hansch analysis.[6] Here, the decision on what to synthesise next could be made by deconvoluting what is most likely to allow the biological activity to increase based on the probability of what modification will lead to an improvement in the property under optimisation. Topliss used the Hansch analysis approach to understanding the relative potencies of R groups according to three properties: electronic, hydrophobic and steric.

The Topliss tree has recently been extended using the ChEMBL dataset to *Matched Molecular Series*, where series of modifications have been extracted from published data and the equivalent of Topliss trees generated to suggest modifications that may be of benefit.[7]

12.4 QSAR Model Generation

Many ways have been published by which one can generate an appropriate QSAR model and assess its suitability to be used prospectively. After all, the point of a model is to inform on the information that can be garnered from your dataset, a diagnostic model, or generate appropriate predictions that help make decisions on a project, and understand the extent to which those predictions can be trusted.

A general process to generate a QSAR model in practice is given in Figure 12.1. The workflow starts from the original dataset of chemical structures and the experimental readout of interest, often the biological activity, called the dependent variable, the one variable that is being modelled. The descriptors calculated from the chemical structures are referred to as the independent variables.

One of the most important steps in generating QSAR models is the process by which the dataset is split into modelling (both training and internal

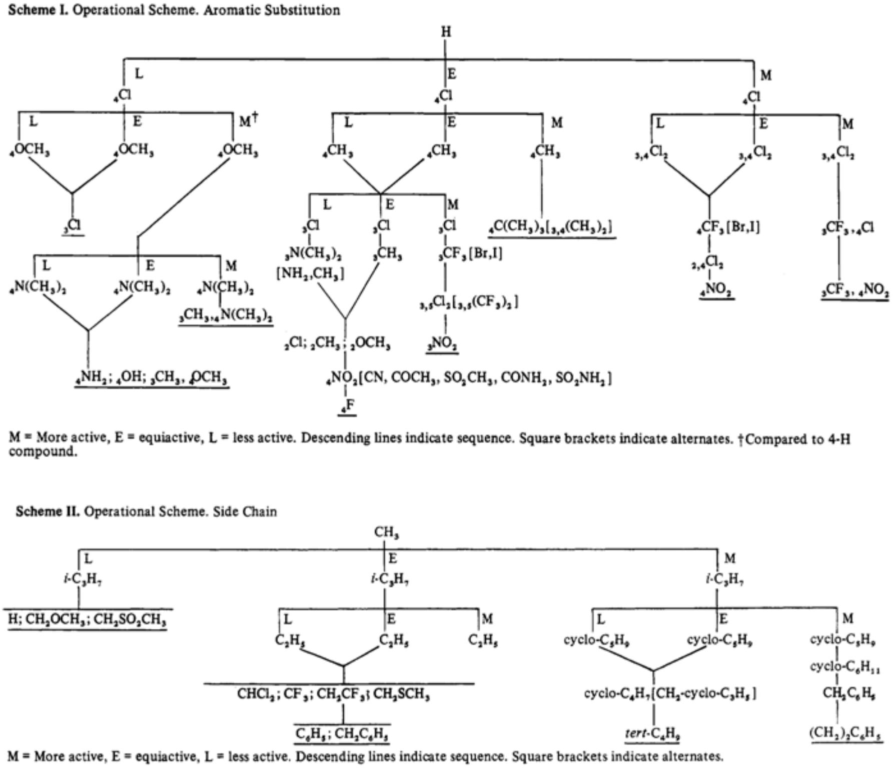

Scheme I. Operational Scheme. Aromatic Substitution
H
L
E
M
4Cl
OCH3
3Cl
N(CH3)2
3CH3,4N(CH3)2
4NH2; 4OH; 3CH3, 4OCH3
CH3
C(CH3)3[3,4(CH3)2]
3N(CH3)2 [NH2,CH3]
3CF3[Br,I]
3,5Cl2[3,5(CF3)2]
3NO2
2Cl; 2CH3; 2OCH3
4NO2[CN, COCH3, SO2CH3, CONH2, SO2NH2]
4F
3,4Cl2
4CF3[Br,I]
2,4Cl2
4NO2
3CF3,4Cl
3CF3,4NO2
M = More active, E = equiactive, L = less active. Descending lines indicate sequence. Square brackets indicate alternates. †Compared to 4-H compound.
Scheme II. Operational Scheme. Side Chain
CH3
i-C3H7
H; CH2OCH3; CH2SO2CH3
C2H5
CHCl2; CF3; CH2CF3; CH2SCH3
C6H5; CH2C6H5
cyclo-C5H9
cyclo-C4H7[CH2-cyclo-C3H5]
tert-C4H9
cyclo-C6H11
CH2C6H5
(CH2)2C6H5
M = More active, E = equiactive, L = less active. Descending lines indicate sequence. Square brackets indicate alternates.

Figure 12.1 Examples of the Topliss trees from his original paper. Reprinted with permission from J. G. Topliss, Utilization of operational schemes for analogue synthesis in drug design, *J. Med. Chem.*, 1972, **15**(10), 1006–1011. Copyright 1972 American Chemical Society.

modelling) sets and external test sets. It is imperative that an appropriate modelling methodology is applied and clearly articulated in any scientific communication allowing not only for reproducibility, but also to understand what potential limitations there may be in the approach used.[8] An example of the QSAR modelling workflow, including the separation of training set, internal and external test sets, and associated tasks is given in Figure 12.2.

Qualities that are desirable in partitioning training and test sets are as follows:

1. The distribution of activities in training and tests should be similar.
2. The training set itself should be distributed within the chemical space of the dataset distribution.
3. All points in the test set should be contained within the applicability domain defined by the training set, at least in the entire descriptor space.
4. Ideally, each point of the training set should be close to at least one point of the test set.

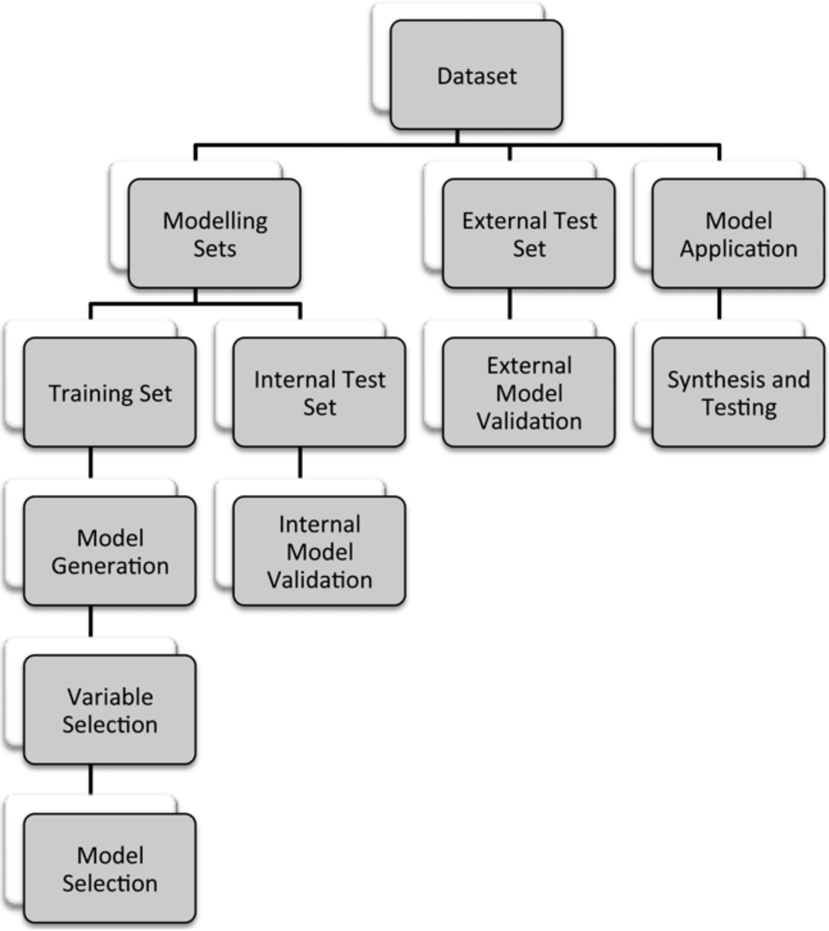


Figure 12.2 Typical QSAR model generation, internal validation, external validation, and experimental validation workflow.

When dealing with small datasets, as is often the case, it can be difficult to partition sufficient data into the respective training and test sets, and here is where pragmatism may come to the fore. One approach may be to remove the requirement to have an internal test set altogether, although not necessarily desirable sometimes the data do not allow for this additional complexity. In situations where the dataset is very small, it is often desirable to utilise multiple train and test set partitions and generate multiple models. This is due to the sensitivity of the modelling methods in terms of their respective partitioning and therefore leading to vast discrepancies in statistical parameters representing the quality of your models. Indeed, you may also want to include multiple representatives of these models in your final model set due to these same reasons.

An extension of investigating multiple training and test set partitions is to use both multiple descriptors representations and multiple supervised statistical learning algorithms, and also combinations of these.

The number of ways of partitioning a set of objects into two or three differently sized sets is fraught with risk. Activity-based training and test set partitioning would take the activity data being modelled, order the data by those activity measurements and select every *n*th activity value to go into a particular. For example, you might require a training set of two-thirds of the overall dataset and the remaining third in the external test set. Therefore, the first and second compounds would be selected for the training set, the third for the external test set, and iteration over this procedure until the end of the set is reached. This is often used when the *dynamic range* of the activity data is quite wide and therefore prone to issues in the quality of predictions on the external test set due to an imbalance of the activities, and therefore likely to not be modelled well due to this imbalance also being represented in the training set. In other words, the model that has been generated may contain a lot of data about inactive compounds, but little about active compounds. Therefore, the model may be prone to inappropriately predicting the entire external test set as inactive because it does not have sufficient data to predict for actives.

As an alternative to using the activity, or measured data above, the selection of the training and test partitions can be achieved by considering the chemical structures in the dataset. The hypothesis here is that the coverage of the structural space will be more representative in the resulting partitions, and it is therefore more likely that the predictions will be more reliable since the entirety of the domain under consideration will be represented in both sets. However, one issue that may arise with this approach is that islands of chemistry space may not be represented in the external test set and these will reveal poor predictions.

Coverage of the chemistry space distribution can be achieved through using many of the unsupervised learning methods described previously, such as clustering. Further methods have been reported in the literature and the interested reader is especially directed to some of the many recent articles on good modelling practice.

12.5 Feature Selection

It is important to identify the appropriate set of molecular descriptors to use in model generation. Typically, a number of camps exist that prefer one method to another. However, one should investigate the statistical learning community to understand where certain approaches can be used to benefit. The statistical learning community refer to feature selection to mean what we may intend by variable or descriptor selection, and there are very many methods by which one can select some number of *n* descriptors from a set of *N* as it suffers from combinatorial explosion in most practical cases.

Feature selection can be applied for a number of reasons in the generation of a QSAR model. One of the most compelling reasons reported in the QSAR community is that appropriate feature selection leads to more interpretable models, which we will discuss more fully below. Another key point is that feature selection may simply be necessary to make model training or learning, and typically to a lesser extent application, faster in execution. However, in the age of Big Data and Deep Learning, and their associated algorithms, renders most of the modelling experiments you would want to perform achievable with a typical desktop computer. However, simply because one can do something, it does not follow that one must or even should—the foundation of hypothesis-driven science. There are many variable selection methods used in chemoinformatics, and even more in the field of statistical learning, but a few are provided below that are commonly applied.

One of the commonly reportedly methods for variable selection in the beginnings of QSAR was to select descriptors whereby they mean something experimentally and can be measured or modelled, as in one of the first uses of ClogP. This variable selection method is highly intuitive, but can lead to models that are not very predictive, especially when the model and the structures being predicted are unsuitable for the calculated properties. However, the interpretability of a QSAR model should not be underestimated and can be highly effective in certain cases. Having said that, though the calculated physicochemical properties can offer additional limitations in this, since it is not actually that simple to deconvolute what the, *e.g.* group contributions in ClogP actually contribute and if this is chemically meaningful (Figure 12.3). This can be seen as simply introducing another level of indirection from reality, but with the comfort of 'knowing' what the descriptors on which the new model is generated mean.

Another popular approach to variable selection, the Greedy Forward Selection approach, is also common amongst scientists from the philosophy of less complex and, assumedly, more interpretable models. Here, the algorithm will often begin with a one or a few 'orthogonal' descriptors from the set available and evaluate the quality of the resultant model. Therefore, the likely high number of model evaluation steps can often lead to very substantial calculations and subsequent runtimes, unless appropriate pragmatic heuristics are used. Similarly, Greedy Backward Elimination uses a greedy algorithm by removing potential descriptors from the larger set, which could lead to even longer runtimes if model evaluation is strongly dependent on the number of descriptors used.

Common heuristics used in the greedy variable selection methods are also manifold, but oftentimes, simple correlation values are used between descriptors. For example, many implementations of calculated logP may be available in a given software package and it is unlikely to add much value to include all of them. Here, one could simply discard all but one of the logP calculators. This would tend to be a more manual step, as reported by Paul Labute, where a handcrafted set of descriptors was optimised to be applicable

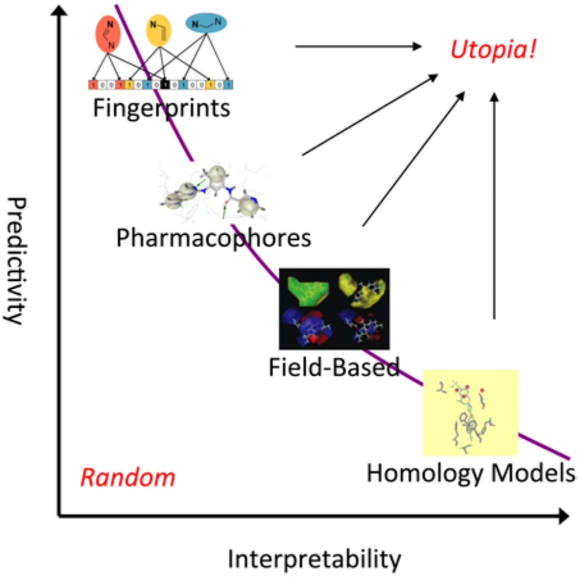


Figure 12.3 The trade-off surface between interpretability and predictivity in predictive modelling. As the models become simpler—simple modelling methods and descriptors—the models themselves become easier to understand, but less predictive: models in the diagnostic mode. As the models become more complex—complex modelling methods and descriptors—the models themselves become better at predictions, but it is less easy to interpret what the predictions mean: models in the predictive mode.

to different challenges, including only descriptors that could be readily calculated from connection table representations, including: atomic contributions to van der Waals surface area, logP (octanol/water), molar refractivity and partial charge.[9]

The last variable selection to be considered here that is widely used in QSAR is using a *Genetic Algorithm* (GA), a population-based natural heuristic optimisation algorithm developed as an analogue of Darwinian evolution in nature. Here, subsets can be selected from the entire set, and each subset represented as a *chromosome*. A population of chromosomes can be generated, evaluated in terms of the selected optimality criterion, or *fitness function*, and sampled according to how optimal each chromosome is. Once sampled according to the fitness function, analogues of *recombination* (called *crossover*) and *mutation* are applied, theoretically to take advantage of the genetic material present in the population and introduce limited new genetic information, respectively. The GA approach has been used widely in this field, and many others, and has been shown to be competitive in rapidly identifying globally optimal solutions.

As mentioned, many other feature selection algorithms are available, and likely yet to be designed, and the challenge is not a solved problem.

12.6 Methods for Estimating Model Validity, Predictive Power, and Applicability Domains

One of the key advances made in more recent times is the realisation that models do not necessarily help you if you apply them blindly and without consideration for the dataset, the molecular descriptors used, and the statistical learning method used to generate the classification or regression models.[8]

The dataset is important to consider, as it will have many different variables. The sheer size of the dataset will indicate from the start the ability of any model you may be able to generate and whether it is worth trying to build a model at all. One has to consider that for good modelling practice, you must have at least an external test set partition, and preferably an internal one too, to develop an understanding of whether the model is of any quality. Relying on internal model statistics of predictive power is not sufficient in and of itself—*e.g.* the Kubinyi Paradox.

There are many measures of both internal and external model suitability that have been reported in the field, far too many to cover in any great detail here. The essential measures are introduced below with brief descriptions as to how they may be applied appropriately to critically analyse QSAR statistical models.

The coefficient of determination, R^2 or r^2, pronounced *R*-squared, is the workhorse of QSARs and statistical learning in general. The R^2 parameter indicates how well correlated, or otherwise, a particular model is, and it is based on the observed *versus* predicted values, those as predicted by the model. There are a number of definitions that are used in the field, but to maintain simplicity, we will refer to the R^2 here as equivalent to the square of the Pearson correlation coefficient between observed and modelled data of the dependent variable, in our case most often the biological activity. In linear least squares regression this is with an estimated intercept term. The R^2 takes the range −1 to 1, representing anti-correlated and correlated, with 0 indicating decorrelated points.

The cross-validation version of R^2 is called q^2 and is calculated in the same way, but this time it represents the predicted values for each one of the predicted *versus* measured points when cross-validation is used. Cross-validation is a simple, and often flawed, measure of predictive power that iteratively removes some portion of the dataset being modelled, adjusting the model appropriately, and predicting for the removed points. Leave-one-out (LOO) cross-validation was used most often in the past, where only one point is removed at each iteration of the cross-validation and the model updated, but this has been identified to not reflect the predictive power of the model since the removed subset size of one has very little effect and is often supported by nearest neighbours. Extensions to LOO cross-validation consider *n*-fold cross-validation or leave-many-out (LMO). In *n*-fold cross-validation, a proportion of the dataset is removed as before, but this portion is more of the order of 10–15% of the dataset. With *n*-fold cross-validation, the effect on the model of the removed points is more marked and can highlight if the model is over-trained to the training set. However, this is still an estimate of the correlation between measurement and prediction. The next test is to calculate the same statistical measure on the internal and external test sets to get a truer consideration of the predictive power. However, this is still limited to the planned utility of the QSAR model.[10–13]

R^2 is highly used, perhaps over-used, in QSAR models, but it should be clear that it offers a very global statistic that can lead to misinterpretation of the model quality. For instance, many distributions can offer the same indication of quality that may not be reflected uniformly over the data. Furthermore, the dynamic range of the data can affect the quality of the model. If the data range is too wide then the R^2 may appear high, and indeed is high, but the R^2 at the area of interest, often at the submicromolar level, may not allow a decision to be made rationally because of limited information regarding the R^2 of the model at that level, and therefore any point may be selected with undefined confidence.

The highlighted issues with just using R^2 in estimates of model quality require consideration of the error of the model in terms of the differences in each of the predictions and the measured data. Therefore, the root-mean-square errors of estimation (RMSEE) and prediction (RMSEP) can be used to inform more on the quality of the residual errors between measured and predicted data.

12.7 Automated Model Generation, Validation and Application

Much effort has also been expended on automatic generation and validation of models: AutoQSAR[14] and DiscoveryBus.[15] Typically, these were intended as temporal models, being updated over time and as new data became available. These systems are now quite commonplace in large pharmaceutical and some smaller biotech companies, but as always caution should be taken when applying and interpreting the models.[16]

These automated model generation and validation systems have been shown to be of great importance to the field, but regular checks of data quality, changes in assay conditions, *etc.* can all lead to issues when applying them. Additionally, the models generated may start occupying different regions of chemical space, or more densely representing the same spaces considered previously, a form of positive reinforcement. Therefore, careful consideration of the applicability domains of the models and where your predictions lie is important.

The field has now reached a point where the automated model generation, validation and application stages can be achieved using a range of off-the-shelf Application Program Interfaces (APIs) available for many languages, either implemented directly in that language or with a wrapper implemented to facilitate the use of the functionality in different languages to the ones in which they are natively implemented. A combination of RDKit (http://www.rdkit.org), a chemoinformatics API made open source by Greg Landrum at the Novartis Institutes for BioMedical Research, and scikit-learn (http://scikit-learn.org), now readily permits the implementation of relatively simple, and vastly more complex, workflows in Python. RDKit can be used, as in many other applications discussed in this book, to calculate molecular descriptors, such as Morgan fingerprints, to enable the generation of appropriate variables for machine learning applications. The scikit-learn API, a combination of NumPy, SciPy and matplotlib, can take care of the remainder of the machine learning, such as partitioning of the datasets, treatment of the data, generation of models, and calculation of the appropriate statistical measures that permit for an appropriate leave of validation: internal, external and prospective. In addition, scikit-learn offers not only the supervised statistical learning methods discussed in this chapter and previously, classification and regression, but also unsupervised methods, such as clustering algorithms.

12.8 Summary

QSARs and other predictive models have a long history in drug discovery and design, going back a century and a half to the pioneering work of Crum Brown and Fraser. With the advent of computers that slowly became available to scientists in the 1950s and 1960s, more analyses could be conducted and models generated. This work extended with much pioneering work on QSAR equations, model generation and validity testing, which is still ongoing to the present day. The main inflection point for QSAR and its use in active research was with the work of Hansch and Fujita, which led to a wide range of applications and developments over the past fifty years.

One of the main challenges in QSAR methods is the understanding of when and where you can use a model reliably and with appropriate measures of the reliability. The challenge of model reliability only became apparent relatively recently in the significant sense and has had a profound impact on the use of QSAR in drug design over the past twenty years or so. It was once thought that the internal prediction quality measures were sufficient and many models were published that offered these as the only improvements. Indeed, models began to be more predictive and with less error than the experimental data on which they were trained. Clearly, this was a watershed moment in QSAR, and all statistical learning methods, considering how we use these models. Thankfully, many scientists have tackled these challenges and much advice has been offered to the community. This advice will undoubtedly continue, but already we can see much more reliable models being generated and applied, and it is now often a requirement for a journal publication to include experimental validation of the model predictions—truly the acid test.

References

1. A. C. Brown and T. R. Fraser, On the connection between chemical constitution and physiological action; with special reference to the physiological action of the salts of the ammonium bases derived from strychnia, brucia, thebaia, codeia, morphia, and nicotia,*J. Anat. Physiol.*, 1868, **2**(2), 224.
2. S. M. Free and J. W. Wilson, A mathematical contribution to structure-activity studies,*J. Med. Chem.*, 1964, **7**(4), 395–399.
3. C. Hansch, P. P. Maloney, T. Fujita and R. M. Muir, Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients, *Nature*, 1962, **194**, 178–180.
4. C. Hansch and T. Fujita, p-σ-π Analysis. A method for the correlation of biological activity and chemical structure,*J. Am. Chem. Soc.*, 1964, **86**(8), 1616–1626.
5. T. Fujita, J. Iwasa and C. Hansch, A new substituent constant, π, derived from partition coefficients,*J. Am. Chem. Soc.*, 1964, **86**(23), 5175–5180.
6. J. G. Topliss, Utilization of operational schemes for analog synthesis in drug design,*J. Med. Chem.*, 1972, **15**(10), 1006–1011.
7. N. M. O'Boyle, J. Boström, R. A. Sayle and A. Gill, Using matched molecular series as a predictive tool to optimize biological activity,*J. Med. Chem.*, 2014, **57**(6), 2704–2713.
8. L. Eriksson, J. Jaworska, A. P. Worth, M. T. Cronin, R. M. McDowell and P. Gramatica, Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs, *Environ. Health Perspect.*, 2003, **111**(10), 1361.
9. P. Labute, A widely applicable set of descriptors,*J. Mol. Graphics Modell.*, 2000, **18**(4), 464–477.
10. A. Golbraikh and A. Tropsha, Beware of q^2!, *J. Mol. Graphics Modell.*, 2002, **20**(4), 269–276.
11. A. Tropsha, P. Gramatica and V. K. Gombar, The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models, *QSAR Comb. Sci.*, 2003, **22**(1), 69–77.
12. A. Golbraikh, M. Shen, Z. Xiao, Y. D. Xiao, K. H. Lee and A. Tropsha, Rational selection of training and test sets for the development of validated QSAR models,*J. Comput.–Aided Mol. Des.*, 2003, **17**(2–4), 241–253.
13. A. Tropsha, Best practices for QSAR model development, validation, and exploitation, *Mol. Inf.*, 2010, **29**(6–7), 476–488.
14. S. L. Rodgers, A. M. Davis, N. P. Tomkinson and H. van de Waterbeemd, Predictivity of simulated ADME AutoQSAR models over time, *Mol. Inf.*, 2011, **30**(2–3), 256–266.
15. D. E. Leahy and D. Krstajic, Automating QSAR expertise, *Chem. Cent. J.*, 2008, **2**(suppl. 1), 1.
16. P. Gedeck, B. Rohde and C. Bartels, QSAR-how good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets,*J. Chem. Inf. Comput. Sci.*, 2006, **46**(5), 1924–1936.

CHAPTER 13

Protein–Ligand Docking

13.1 Overview

Protein–ligand docking is a commonly used method to predict the likely binding mode of a ligand in a protein-binding site. Given a particular ligand and an extracted binding site, the algorithm explores potential binding modes through rotation and translation in three-dimensional space, and then scores each pose identified for suitability. Various docking algorithms can additionally consider ligand and even protein flexibility.

There is a wide range of methods to search the space of potential binding modes in ligand docking and some of these approaches are introduced in brief. The second aspect of protein–ligand docking to be introduced, and the most challenging to perfect, is the scoring function, the method by which binding poses are scored. More in-depth information is given on one protein–ligand docking algorithm that has been applied and validated extensively in retrospective and prospective studies.

13.2 Search Algorithms

Docking methods can be classified into three distinct types of varying complexity. The first is rigid docking where both the ligand to be docked and the binding site are rigid. Although this is limited, since no account of flexibility of the system is considered, typically multiple pre-generated conformers are used to overcome this problem. The second type of docking algorithm includes optimisation of both the rotation and translation in three-dimensional space for the ligand, as with rigid docking, but also explores multiple conformers on-the-fly as part of the search and optimisation process.

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

The last type of docking approach introduces protein flexibility into the method. One approach, called flexible side-chain docking, as expected allows the side-chains of the amino acid residues to move during optimisation, while maintaining the rigidity of the alpha carbon backbone of the protein. Alternatively, implicit flexibility can be introduced with the addition of multiple crystal structure conformations, where available, in a process called ensemble docking. The results of ensemble docking can then be combined using consensus scoring (or data fusion) to permit the prioritisation of results.

Docking has been applied successfully in virtual screening to identify potential hits that can then be tested. These hits can come from a large screening library of real samples, or from the virtual space of potential ligands. In the case of virtual ligands, oftentimes the library will be a focussed one around the scaffold of interest and docking is used frequently in lead optimisation.

13.3 Scoring Functions

Many different scoring functions have been developed to improve the accuracy of docking algorithms.[1,2] Although none works perfectly, judicious selection of an appropriate scoring function can improve the quality of results in particular domains. The scoring functions that have been developed typically fall within one of four specific categories: force field; empirical; knowledge-based; and consensus scoring. Many more recent docking algorithms also allow the user to define their own scoring function.

Force field scoring functions are a composition of the ligand binding energy and the interactions with the protein. GOLDScore is the default scoring function in GOLD and is made up of four components: protein–ligand hydrogen bond energy (external H-bond); protein–ligand van der Waals (vdw) energy (external vdw); ligand internal vdw energy (internal vdw); and ligand torsional strain energy (internal torsion).

ChemScore, also available in GOLD, is a scoring function using hydrogen bond energies, atomic radii and polarisabilities, torsion potentials, hydrogen bond directionalities, *etc.* However, ChemScore is derived empirically based on a regression model of these parameters with binding affinities from 82 protein–ligand complexes. It is not clear whether ChemScore is superior to GOLDScore, but since its objective is to model measured binding affinity, it is anticipated that the values will be more directly comparable.

For knowledge-based scoring functions, it is typical to extract structural information from the complex of the protein and the docked ligand and then use the Boltzmann law and calculate pairwise atom potentials that are distance dependent. However, this approach typically omits the directionality of the interactions, although efforts have been made to improve this type of scoring function.

The last general type of scoring function is not itself a scoring function but a method by which other scoring functions may be combined to mitigate limitations in some scoring functions that may not be known *a priori*. The combination scoring approach is commonly known as consensus scoring (or data fusion, particularly in similarity search). As one can imagine, there are a large number of ways of combining multiple scores, from simple sums of scores, or weighted sums, to taking the maximum (or minimum) value discovered with each method.

13.4 GOLD: Genetic Optimisation for Ligand Docking

GOLD[3–6] (Genetic Optimisation for Ligand Docking) was developed in 1995 and uses a Genetic Algorithm (GA) as the means by which the space of solutions is explored.[7] GAs have been used widely in computational chemistry, with a book published reviewing a wide range of different applications by Clark, including for protein–ligand docking applications.[8] GAs are search and optimisation heuristic algorithms that are modelled on Darwinian evolution, and they are often used for typically intractable optimisations in a reasonable time frame. The problem space is encoded as a chromosome representation, typically a binary string, which defines the genotype. Initially, a population of these chromosomes is randomly generated, often using greedy heuristics to encode any domain knowledge that may be known. Each chromosome is then mapped from this genotypic space into the phenotypic space and a score given for each individual chromosome by means of a fitness function.

Once scored, the entire population is sampled using a fitness-based sampling approach, where each individual will be represented in some proportion to their overall suitability based on their fitness scores. One of the most common fitness proportionate sampling methods is the roulette wheel sampling scheme, with each chromosome being given a portion of the roulette wheel proportionate to its fitness score in the population. The roulette wheel is then spun *n* times, where *n* is the number of chromosomes in the population, resulting in a stochastically populated fitness proportionate sample of the population.

On completion of sampling, the population undergoes computational analogues of recombination, called crossover in GAs, and mutation. Crossover takes two chromosomes from the sampled population in turn, randomly defines a crossover point, or locus, and exchanges the genetic material either side of this locus. Crossover attempts to exchange desirable genetic material information between increasingly highly scoring chromosomes in an attempt to capitalise on the genetic information in both and tends not to destroy the genetic information, but this can depend on the encoding strategy. Mutation, however, introduces some noise into the population that is crossed-over by randomly inverting bits in the population according to some probability, typically low. The recommended probabilities of the crossover (chromosome-based) and mutation (gene-based) operators in a simple GA is typically around 0.7 and 0.01, respectively, but these tend to require a significant degree of optimisation for specific applications and problem domains.

In the published implementation of GOLD, a similarly structured algorithm is applied, but with some key alterations. The chromosome representation for each binding pose consisted of a chromosome each for the ligand pose and the protein conformation. Each byte (8-bits) in the chromosome encoded an angle of rotation around a defined rotatable bond in the ligand or protein, or torsion angle. This angle was encoded in step-sizes of approximately 1.4° from −180° to +180°, with each byte representing 256 of these steps. When considering a rigid protein, only one binary chromosome is required to represent the ligand. In addition to the binary chromosomes, there are also two integer strings that represent the mapping of possible hydrogen bonds between the protein and ligand. A least-squares fitting procedure is applied on decoding the chromosome for evaluation to optimise the number of hydrogen bonds.

The original fitness function considered six aspects of the chromosomal representation of each binding pose:

1. Conformation of both the ligand and protein was generated;
2. Ligand was in the active site applying a least-squares fitting procedure;
3. Hydrogen bonding energy was generated for the complex (H_Bond_Energy);
4. Steric interaction energy between ligand and protein (Complex_Energy);
5. Internal energy of the ligand obtained using molecular mechanics expressions (Internal_Energy);
6. Summation of all energy terms, if present, to give a fitness score for each pose.

A number of alternative docking algorithms have been applied successfully in drug discovery. Glide is one such docking algorithm from Schrodinger.[9] In brief, Glide considers a systematic yet approximate search of the entire space of all potential solutions based on potential conformations, orientations, and positioning of the ligand under examination. Through a series of refinement filters applying ever more complex calculations, Glide finalises a pose using a Monte Carlo sampling of the pose conformation and a model energy function that is a combination of empirical and force-field based approaches.

Another very popular docking algorithm is FlexX from BioSolveIT.[10] FlexX implements an alternative method to modelling protein–ligand binding and scoring. The algorithm proceeds by extracting the preferred sets of torsion angles for acyclic single bonds and ring conformations. The torsion angles for multiple bonds and the bond lengths and angles are applied as given to the software. After additional preprocessing to generate conformers using structures from the Cambridge Structure Database (CSD) and ring conformation exploration using CORINA from Molecular Networks GmbH, the algorithm proceeds to a fragmentation and incremental growth method. The ligands are fragmented into subunits by cleaving all acyclic single bonds. Next, base fragments are placed in the binding site and an incremental growth procedure begins where the original ligand is gradually rebuilt with scoring at each stage until ligand poses are generated and a final score given.

While the docking algorithms reported above operate with different search and scoring algorithms, it is difficult to identify a preferred algorithm in all cases. However, much research is being conducted to improve on the algorithms available and it is expected that further improvements will be published over time.

13.5 Model Validation

As with all modelling approaches, it is important to sufficiently validate the model system to ensure that the results from prospective studies may be trusted, or at least to understand the extent to which they may be trusted. A common approach in validating a docking protocol for a particular protein structure is to dock ligands for which the bound pose or biochemical readouts are known.

For protein–ligand complexes (*holo* structures), readily available as public data from the Protein Data Bank (PDB) but often also available in-house, the objective in model validation is to recapitulate the bound conformation observed experimentally. One approach is to define the protein-binding site based on residues that are proximate to the ligand in the co-crystal structure and then extract the bound ligand. An important next step is to take the bound ligand conformation and remove any geometry bias that would not necessarily be present in a prospective study. It is often sufficient to render the ligand as a two-dimensional molecule and then generate a low-energy conformation for docking. The docked pose, or poses, can then be evaluated with reference to the bound conformation using the root-mean-square distance (RMSD) in ångströms (Å) of the heavy (non-hydrogen) atoms of the molecule. The lower the RMSD value, the closer the modelled pose is to the experimentally observed pose.

For protein structures where protein–ligand complexes are not available (*apo* structures), common in the early stages of drug discovery, it is necessary to identify the most likely binding site using druggability analysis to find the 'druggable' pocket. In this situation, since no bound conformation for a ligand is available, it may be possible to use extant ligands and biological data to investigate whether the ranking is similar using the docking scoring function as a surrogate for biological activity. This approach reflects how the model system would be applied prospectively in virtual screening. However, caution must be shown since the conformation of the protein binding site may not be similar to that in the eventual protein–ligand complex, due to induced fit.

13.6 Docking in Prospective Studies

Docking has promised much over the thirty or so years since its inception and it has certainly contributed positively when applied very carefully. Indeed, it is possible to recapitulate the bound conformation of a ligand in its own binding site with a high degree of accuracy. Unfortunately, when the similarity of the ligands that are being docked decreases away from that of the bound ligand, the docking algorithms begin to reduce in this ability significantly and it becomes much more difficult to model these ligands. Therefore, it should be clear that docking of close analogues to the ligand found in the bound crystal structure will be relatively successful, but this does not really inform much over and above what is already known from the original protein–ligand crystal structure. Indeed, this does not offer much above three-dimensional ligand similarity re-scored in the context of the binding sites.

Clearly, docking does not live up to the anticipations and expectations of its early days and often this stymied its growth as a method since its results could be seen to be unreliable, at least in high-throughput docking studies, or requiring a substantial degree of manual inspection of the docked poses that would take a great deal of time to process.

It is evident that docking works sometimes, and that it tends to work more reliably when the ligands that are being docked are more similar to the ligand in the protein–ligand complex. It is likely that this is to some extent due to the induced fit problem; ensemble docking is one approach to overcoming this by introducing multiple protein conformations that have already been crystallised and using these as an ensemble of structures against the ligands being docked and the best-scored poses taken or somehow post-processed to combine the top poses. However, this relies on additional crystal structures that may not be available, particularly early on in a drug discovery project. Furthermore, the computational run times will increase linearly with the number of crystal structures available, thereby increasing further the overall docking experiment, which is not insubstantial compared with other methods.

Recently, a study by Broccatelli and Brown[11] examined probabilities of success in prospective docking studies using cross docking as a surrogate for the prospective data. Cross docking is when multiple protein–ligand crystal structures are available for the same protein but different ligands, in this case CDK2. Each ligand from each of the protein–ligand crystal structures is then docked into every other protein structure in turn and the best poses retained. The native docking, or self-docking, was compared to the cross docking experiments and the results are shown in Figure 13.1. The results from the cross docking experiment demonstrate that the docking success rate drops significantly across all variants of the Glide algorithm used. This reduction in docking performance when moving from native docking to cross docking highlights the kind of success that could be expected in a prospective study.

The study continued to investigate this challenge by moving away from native docking, which works well, to use an external test set of CDK2 ligands that had been published but for which no crystal structures have been published. The CDK2 ligands had been tested and included both actives and inactives. From Figure 13.2, it can be seen that, when ligand similarity is high, the docking results tend to be reliable, but as has been discussed this is expected and demonstrates that ligand similarity is a major driver in docking experiments. Typically, the more interesting experiments are at the lower ligand similarity levels and this study demonstrated a ‘sweet spot’ of similarity in

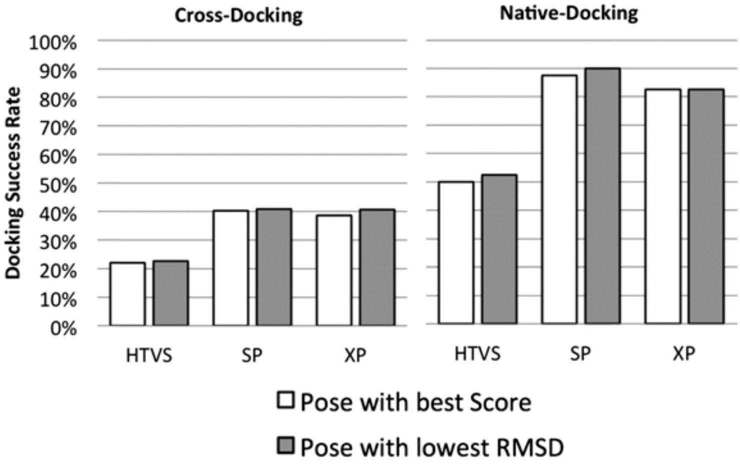


Figure 13.1 Docking success rate for three Glide protocols: HTVS, SP and XP. Docking pose prediction is considered correct if the RMSD from the crystallographic ligand is below 2 Å. Reprinted from F. Broccatelli, N. Brown, Best of both worlds: on the complementarity of ligand-based and structure-based virtual screening,*J. Chem. Inf. Model.*, 2014, **54**(6), 1634–1641.

the mid-ranges, Tanimoto similarity on Extended Connectivity Fingerprint with a diameter of four (ECFP_4) of 0.2–0.4, where it might be best to focus more considerable docking efforts. This study also demonstrated the likely probabilities of success at different ligand similarity levels, which can also assist in understanding how many ligands that have been docked should be put through screening to expect a significant signal in terms of activity enrichment.

The study clearly demonstrated, as with many of the methods applied in computational modelling, that using all of the available data is important, as in ensemble docking, and also that data fusion should be used where possible to make use of the benefits of both ligand-based and structure-based methods. Lastly, the study demonstrated that the probability of success of docking decreases significantly as the ligand similarities to the native ligands drop.

13.7 Summary

Docking has been part of the computational chemistry and molecular modelling toolbox for more than three decades. Many different search algorithms have been developed that are capable of exploring the conformation of the ligands being docked, their rotations in three-dimensional space, and the positioning in that same space. These search algorithms are important in being able to explore the potential search space. Equally importantly is the scoring function that scores each one of those poses and the potential interactions the pose may make.

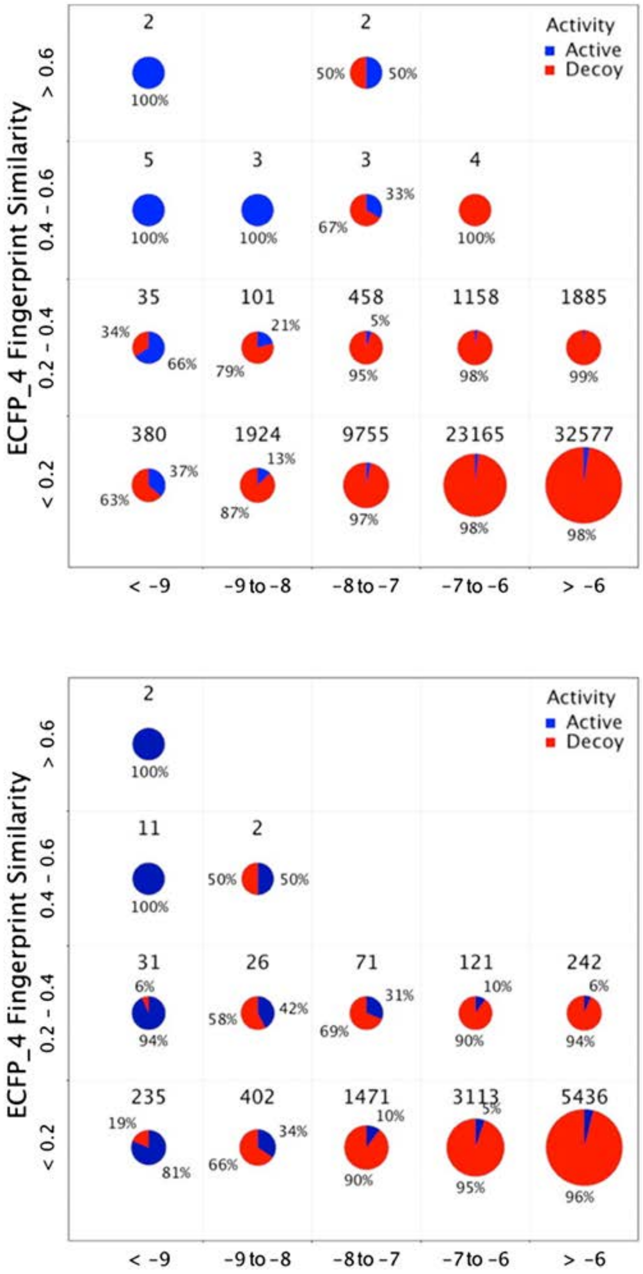


Figure 13.2 Relative percentage of actives and decoys (Directory of Useful Decoys (DUD) data set, upper plot), and actives and inactives (GlaxoSmithKline data set, lower plot) in different Extended Connectivity Fingerprint with a diameter of four (ECFP_4) and High-Throughput Virtual Screening (HTVS) Glide docking score ranges. Reprinted from F. Broccatelli, N. Brown, Best of both worlds: on the complementarity of ligand-based and structure-based virtual screening, *J. Chem. Inf. Model.*, 2014, **54**(6), 1634–1641.

Many different improvements have been made in docking research since its inception. It is inevitable that the methods will improve over time, but a number of studies have demonstrated that docking still has a number of limitations to overcome. Not least of these is the improvement of scoring functions. The reduction in docking success between native docking, an experiment where the answer is known already, and cross docking, where it is expected we know the answer, has shown that docking is very good at recapitulating its own pose, but not effective at performing in an effectively prospective experimental study.

Using ligand similarity, a recent study has furthermore demonstrated that expectations of success can be estimated in docking. Highly similar ligands will tend to succeed, but there is a point at which docking, in the mid-range ligand similarity measure, will have a measurable probability of success. This type of study demonstrates that all methods are useful in computational medicinal chemistry, but it is important to find ways to understand when and where certain methods will and will not work, at least with some level of probability of success.

Docking has come a long way, and still has some way to go, but this does not mean it should not be used. As with all modelling, careful and considered construction of hypotheses and expectations will permit new methods, often hybrid methods, to work positively for modelling in the prospective use-case scenario.

References

1. 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.
2. C. A. Baxter, C. W. Murray, D. E. Clark, D. R. Westhead and M. D. Eldridge, *Proteins*, 1998, **33**, 367–382.
3. G. Jones and P. Willett, Docking small molecule ligands into binding sites, *Curr. Opin. Biotechnol.*, 1995, **6**, 652–656.
4. G. Jones, P. Willett and R. C. Glen, Molecular recognition of receptor sites using a genetic algorithm with a description of solvation, *J. Mol. Biol.*, 1995, **254**, 43–53.
5. G. Jones, Willett and R. C. Glen, A Genetic algorithm for flexible molecular overlay and Pharmacophore elucidation, *J. Comput.–Aided Mol. Des.*, 1995, **9**, 532–549.
6. G. Jones, P. Willett, R. C. Glen, A. R. Leach and R. Taylor, Development and Validation of a Genetic Algorithm for Flexible Docking, *J. Mol. Biol.*, 1997, **267**, 727–748.
7. D. E. Goldberg, Genetic Algorithms in Search, *Optimization and Machine Learning*, Addison-Wesley Longman Co., Inc., Boston, MA, USA, 1989.
8. *Evolutionary Algorithms in Molecular Design*, ed. D. E. Clark, Wiley-VCH, Weinheim, Germany, 2000.
9. 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, Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy, *J. Med. Chem.*, 2004, **47**, 1739–1749.

10. M. Rarey, B. Kramer, T. Lengauer and G. Klebe, A Fast Flexible Docking Method using an Incremental Construction Algorithm, *J. Mol. Biol.*, 1996, **261**, 470–489.
11. F. Broccatelli and N. Brown, Best of both worlds: on the complementarity of ligand-based and structure-based virtual screening,*J. Chem. Inf. Model.*, 2014, **54**(6), 1634–1641.

CHAPTER 14

De Novo Molecular Design

14.1 Overview

Combinatorial explosion is a significant challenge in all modern science, but nowhere can this be realised more clearly in this field than that of the sheer number of possible *drug-like* molecular objects that can be realised from just a couple of dozen carbon atoms, some oxygens and nitrogens, and perhaps a halogen or two. However, the estimated number of unique drug-like molecular structures has been conservatively estimated at 10^{60}.

Considering just a single scaffold of interest in a medicinal chemistry project, with two or three points of variation and a relatively conservative 1000 common medicinal chemistry relevant substituents, a single virtual library for that scaffold contains one million to one billion unique molecules that could be considered for synthesis. However, a typical medicinal chemistry project considers only a few thousand in terms of synthesised compounds.

Clearly, computers can consider vastly more space than could realistically be synthesised with the same amount of laboratory work. However, even computers, with the vast amount of computing resource that could be dedicated to the challenge, could still only consider a few tens or hundreds of millions with sufficient scientific effort and hypothesis-driven science. This would result in useful output that has been validated in more routine algorithms, but also the more complex and computationally intensive methods, such as docking and shape-based searching. Therefore, a ‘simple’ enumeration of the space would be too vast to consider.

Here, the field of *de novo* design comes to the assistance of the drug design project team. *De novo* design is essentially an umbrella term for everything that is trying to be achieved in the field of computational drug discovery. All of our modelling methodologies are used to appropriately sample the

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

drug-like space for those molecules that are of interest, from molecular descriptors and similarity searching to QSARs and ligand docking, to identify the few that should be synthesised that have a high probability of success against the multitude of read-outs we need to satisfy in modern drug design.

From the age of atomistic theory came the idea that chemists could summon up matter to design molecules that fulfil distinct purposes and we are now beginning to realise that dream routinely in drug design reaping real benefits in saving time, but more importantly not wasting time on molecules that are unlikely to be of interest. However, it has been a long road with many pitfalls in that journey. Many new methods arose that promised so much, a common problem in scientific computation, over the past thirty or forty years or more with peaks and troughs in the success and perceptions of *de novo* design. However, only recently have the systems been developed and the will to test the outputs, the synthesis of machine-designed structures, has the field really borne fruit. Indeed, the first volume dedicated to *de novo* molecular design was published only recently by Schneider.[1] The interested reader is certainly directed to the book by Schneider for a full and comprehensive treatment of the state-of-the-art from many of the key leaders and champions of the field. Here, an overview of some of the different approaches used in *de novo* design is provided, from atom-based designs through to fragment-based and reaction-based approaches.

The methods by which candidate solutions have been scored in the past and new methods for optimising across multiple read-outs will be discussed with reference to recently published algorithms.[2]

14.2 Receptor-Based Methods

As with all medicinal chemistry design efforts, some information is needed to be able to optimise away from and towards the ultimate goal. Therefore, some information is required: ligands of interest that exhibit a biological response and/or receptors of interest in a known drug binding site or a new desirable target. *De novo*-based design originated with the advent of structure-based methods in drug design. The first methods were published at a time when receptors were becoming increasingly available and the methods designed were mimics of what a molecular modeller would do if a receptor structure were available. Furthermore, at that time QSAR and other modelling methods were not sufficient to enumerate all of the ligands of interest that would be required. Instead, the receptor-based *de novo* design methods focussed on optimising interactions in the binding site.

Regions in the binding sites of interest would be identified through structure visualisation and pharmacophore maps generated to optimise interactions with complementary sterics and electronics on the designed ligand or ligands. The receptor-based methods were not overly successful in the early days of *de novo* design due in part to the difficulty in optimising appropriately in protein binding sites due to lack of algorithms, computing power and appropriate scoring functions. In addition, there was a tendency for purely structure-based methods to over-optimise to the target, which is itself a dynamic system captured as a snapshot. With these limitations, *de novo* design was limited in predictive power and optimisation of appropriate drug-like ligands.

However, this does not indicate that structure-based *de novo* design methods were not successful, but instead they were highly beneficial in conducting rational design experiments that permitted scientists to ask questions of structural data that were otherwise quite difficult to answer. This period opened up the space of potential ligands for synthesis that otherwise would not necessarily have been considered.

One of the challenges of the time was to select appropriate building blocks for the design of new ligands that contained common structure moieties seen in the medicinal chemistry. Through this requirement a number of retrosynthetic fragmentation schemes were developed, such as Retrosynthetic Combinatorial Analysis Procedure (RECAP), which applied a set of rational retrosynthetic rules that fragment given parent molecular structures into small building blocks that can be used for ligand generation.[3]

Another key method used at this time was skeleton graphs that were molecular templates where the atoms were labelled only with the hybridisation type and bond orders. By placing the skeleton graphs in the binding site of interest, it was possible to explore more of the chemical space around those skeleton templates by focusing on geometry and optimising molecular interactions afterwards, thereby focusing on one aspect of the search space to reduce the search space of the atom mutations required later.[4,5]

A number of different approaches have been applied to generate new molecular structures: fragment linking, fragment growing, and sampling approaches, including additional atom-based and fragment-based methods. Lastly, there are reaction-based methods. The next section will consider each of the approaches in turn with the successes and relative drawbacks to each.

14.3 Fragment-Linking Methods

Some of the first methods of exploring ligand space were the fragment-linking methods, where appropriate structural groups, such as those described above from RECAP, are positioned in a receptor-binding site and their positions optimised for their direct interactions. The fragments, once positioned, were then linked using skeletons, such as those from a dictionary of aliphatic and acyclic linkers, and optimised on a lattice of possible atom positions.

Once two or more fragments were linked by one of the methods described above, it was possible for the new structure to be optimised *in situ* to identify any potential undesirable steric clashes that may have been caused by a change in bond angles. The fragment-linking methods were successful at the time and are still used today, but they tend to have issues in small local changes, particularly towards the centre of the molecules, which led to undesirable global conformational changes that were not able to be predicted through the design stage.

14.4 Fragment-Growing Methods

A second, arguably more successful, approach is a fragment-growth strategy. While the fragment-linking strategies typically used two linkers and combined them geometrically, in fragment growing a base fragment (or seed point) is positioned and optimised for individual interactions. The second step of this algorithm proceeded by connecting additional fragments or individual atoms. The ligands would then grow to optimise the sterics and electronics within the protein-binding site.

Fragment-growing methods were somewhat more successful than fragment-linking, finding particular favour with Astex Pharmaceuticals in their Pyramid™ platform for fragment-based drug discovery. The benefits of the technology available at Astex, namely biophysical techniques, X-ray crystallography and nuclear magnetic resonance, made it possible to detect bound fragments at a much higher sensitivity than otherwise possible with other methods. These technologies, in concert with fragment-based computational design, have been highly successful in identifying and optimising fragments using a combination of experimentally assisted design strategies.

14.5 Sampling Chemistry Space

The methods discussed so far tend to be based on local searches from seed fragments that are linked or grown. While these approaches are now being used to generate ligands in this way, they still tend towards the stepwise linking or growing of the ligands. This is a pragmatic decision to perform a local search around these seed fragments; it would not be possible to simply explore the vast space available and predict positively in these binding sites without globally sampling the space as opposed to the local search from fragments described above.

Sampling chemistry space typically works in a population-based optimisation approach, where multiple candidate solutions are generated and evaluated at once. These solutions are then scored and ranked, often using some form of Genetic Algorithm (GA) or Evolutionary Algorithm (EA). The best aspects of each of the candidate ligands are then reassembled using a crossover operator or other genetic exchange method and small amounts of mutation to introduce new genetic material. The *de novo* design system then iterates over generations, constantly scoring and perturbing the structural elements in each generation. Eventually, after some termination condition is met, a set of final candidate solutions is output.

The scoring functions used in the sampling strategies tend, at least, to not directly link the sampling and the scoring, as is the case in the fragment-linking and fragment-growth methods described above. Instead, there is a distinct evaluation step, called a fitness function, which can combine a multitude of methods, including docking, but also ligand-based approaches, such as molecular similarity and Quantitative Structure–Activity Relationships (QSARs).

14.6 Atom-Based *De Novo* Design

One such *de novo* design system, the Compound Generator (CoG), which is a dual atom-based and fragment-based population-sampling algorithm, was developed and implemented by Brown *et al.*[6] CoG used bespoke genetic operators that were adapted from binary chromosome representations typically used in GAs into analogue algorithms that worked not on strings—although they could—but also on graph structures, including the typical cycles found in many organic molecules. The crossover operators were able to break and form ring systems, while also attempting to not disrupt the genetic material too much, which would otherwise have the side effect of a pseudo-mutation operator. CoG also included new graph-based mutation operators, which were able to add, prune, insert and delete atoms and bonds where these changes conformed to the standard valence bond model.

CoG was designed as a modular *de novo* design engine, or cog in the system, allowing multiple optimisation procedures to be undertaken. Brown *et al.* first demonstrated the use of CoG to design molecules that were maximally similar to two probe molecules of interest, using Fingal fingerprints[7] and Pareto ranking to score the molecules (*vide infra*), and called them median molecules. The median molecules concept was applied successfully to generate many different candidate solutions that were maximally similar to two extant molecules, typically in the hundreds or sometimes thousands. The authors proposed that this would be an appropriate and controlled method to design focussed analogues around certain molecules of interest that could then be prioritised for synthesis using other methods, or perhaps for the identification of novel scaffolds.

The second approach proposed by Brown *et al.* was to use Quantitative Structure–Property Relationships (QSPRs) in the inverse mode, essentially *de novo* design but with statistical models.[8,9] The authors realised that the predictions against two different models became untrustworthy due to the *de novo* design system optimising molecules that exploited the models. Therefore, they introduced additional parameters to the multiobjective optimisation procedure to include information not only regarding the actual prediction, but also the quality of those predictions in terms of their domains of applicability. The inclusion of these parameters, *Residual Standard Deviation* (RSD) and *Leverage*, optimised ligands that fit within the domain of the model much more appropriately.

14.7 Fragment-Based *De Novo* Design

More recently, Firth *et al.*[10] developed a new multiobjective *de novo* design system, Multiobjective Automated Replacements of Fragments (MOARF), which is an example of a fragment-based *de novo* design system. The system is highly extensible to new challenges with their exemplar being the optimisation of new ligands for a drug discovery project, CDK2, using ligand-based shape similarity, QSAR models (Random Forests), and maintaining the ligands to within a physicochemical property of interest.

MOARF introduced two new methods for dealing with fragments in this publication: Synthetic Disconnection Rules (SynDiR) and Rapid Alignment of Topological Fragments (RATS). SynDiR is new retrosynthetic fragmentation scheme that has fewer rules than RECAP and generates few fragments of lower molecular weight. These fragments were also shown to be more relevant to medicinal chemistry based on a comparison to physical samples available from Sigma-Aldrich.

The RATS system is a new fragment replacement algorithm that is able to align replacement fragments based on a simple topological fingerprint. This is an important step in fragment-based *de novo* design, particularly as shown above in receptor-based methods. The RATS method was demonstrated to work comparably with the much more computationally intensive and complex BROOD algorithm, one of the leading methods for bioisosteric replacements. In this comparison, RATS recapitulated BROOD results far above what would be expected randomly, and reaching above 80% correspondence the majority of the time.

Using MOARF, SynDiR and RATS, Firth *et al.* considered a virtual space of *ca.* 200 million unique structures while only sampling fewer than 50 000 of those molecules *in silico*. More importantly, 14 of the top 25 resulting solutions were synthesised and tested against Cyclin-Dependent Kinase 2 (CDK2) and Human Liver Microsome (HLM) exhibiting a good maintenance of potency against CDK2 while significantly reducing HLM turnover. Maintaining metabolic stability is a key objective in drug design that is often overlooked until late in a programme after potency has been optimised. The optimisation to target potency is a common trope in medicinal chemistry and this paper demonstrated that it was possible to design in additional and necessary parameters to increase the probability that these compounds could be considered further as part of a drug discovery programme.

14.8 Reaction-Based *De Novo* Design

The last *de novo* design method to be considered in this chapter is the reaction-based scheme, which is perhaps most familiar to the synthetic organic chemist. Synthesis can be seen mathematically as a context-dependent molecular graph transform, but this over-simplifies the art and science of making new compounds that have never been made before. One of the main challenges to other *de novo* design methods described above is that, although the compounds are predicted to be of relevance to the objective under optimisation, their synthesis may not be trivial or even possible with current chemistry techniques.

It is important to consider in *de novo* design not just the optimisation of the desired property, but also the fact that someone has to make the compound. Furthermore, if the synthesis is likely to take 5–10 times longer than a more trivial synthesis that still has the potential to answer valid hypotheses, then the latter is likely to win in the decision-making priorities of a project. Here, reaction-based methods can be used to optimise not only the molecular structures, but at least make some headway towards the most likely synthetic route to make that compound with common medicinal chemistry synthetic methods and building blocks.

The Design of Genuine Structures (DOGS) system from Schneider *et al.*[11,12] allows for the reaction-based *de novo* design of synthetically accessible drug-like molecules. Using DOGS, it was shown that from 25 144 available synthetic building blocks and 58 established reactions, the algorithm first selects the reaction to be conducted *in silico*, and then which of the identified reactants to be taken forward to the next step. DOGS can introduce additional reactions with the decision to terminate the synthetic simulation taken based on whether the current product is not desirable according to a similarity-based kernel mode, or the size of the molecule becomes too large to be considered drug-like. Once a termination condition is met, the final product is stored, and the process iterated to identify a shortlist of possible products for subsequent consideration.

The DOGS system has been demonstrated to identify novel scaffold hops and has more recently designed a novel and selective inhibitor for Polio-Like Kinase 1 (PLK-1). The Schneider laboratory is now developing a system that could theoretically lead to automated design, synthesis and testing for the identification of novel leads, “Leads on Demand”. While this may seem far-fetched, the tools and techniques being developed for this system are clearly highly effective at identifying novel scaffolds and also novel inhibitors, which are clearly valuable to a medicinal chemistry programme, whether automated or not.

Reaction-based *de novo* methods have been shown to be effective at identifying potential new ligands that come readymade with a potential synthetic route that at the very least will give a medicinal chemist a head start in synthesising that compound. However, one challenge to this approach of *de novo* design using virtual reactions is that the space covered by the products may be limited compared to the other *de novo* design methods discussed previously. This may be the case, but going a long way to solving the challenge of synthetic tractability in *de novo* design is a great step forward and might be the method by which *de novo* design becomes a routine tool actively sought out by medicinal chemists, not just as an ideas generator, but also a compound maker.

14.9 Multiobjective Optimisation

A key aspect of *de novo* design that is increasingly becoming a focus is multiobjective optimisation, as opposed to single-objective optimisation. Historically, as we have seen in the work from Firth *et al.*, drug design programmes typically focus on target potency early on in a project and can often get "locked in" to undesirable regions of chemical space, or not even allow for the possibility of exploring those spaces that may or may not be of interest (Figure 14.1). Here, multiobjective methods can be highly beneficial and are increasingly being used in *de novo* design systems.

Since for Food and Drug Administration (FDA) approval it is necessary for a small molecule to be safe and efficacious at its administrated dose, it is clear that drug design is an inherently multiobjective challenge. In fact, safety and efficacy are convolutions of multiple properties of small molecules, not just target potency. Additional properties that must be considered and optimised in a drug design project are: improved selectivity; few side effects; decreased toxicity; improved pharmacokinetics; and increased metabolic stability. More prosaically, but also vastly important, are the additional challenges of simplified synthetic routes and patented lead compounds.

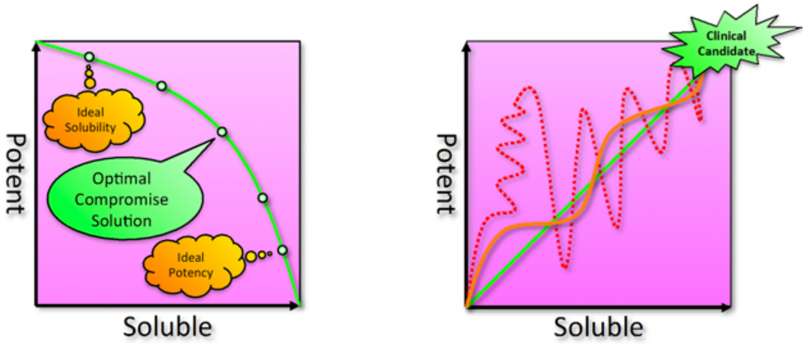


Figure 14.1 Schematics of the multiobjective optimisation challenges in drug discovery. On the left there is a trade-off surface identified between solubility and potency. The over-riding objective for this project is to identify the preferred window on that trade-off surface and design small molecules that satisfy those objectives. On the right is a stylised illustration of the trajectory taken through a drug discovery programme. The green line shows the ideal, where both solubility and potency are being optimised simultaneously, which is highly unlikely. The red-dotted line illustrates the typical trajectory through a drug discovery programme, where potency is typically optimised early, with random fluctuations in solubility. It is not until later in the programme that solubility is necessarily considered and the late stage optimisation can lead to wild fluctuations in the properties being optimised. What might be preferred is the orange line that approaches the optimal green line, but exhibits occasional changes in priority with regard to the property being optimised, but sights are always set on the ultimate outcome.

Therefore, it is key that these types of properties, or appropriate surrogates, are considered as early as possible in a drug design project and appropriately included in the design, make and test cycle, and one way of achieving this can be through the use of multiobjective *de novo* design methodologies.

The first approach to incorporating multiobjective optimisation methods into *de novo* design was by Brown *et al.* in 2004,[6] for the optimisation of median molecules (*vide supra*). The multiobjective optimisation was achieved by means of Pareto optimality. Pareto optimality is a method of identifying non-dominated solutions in a given co-ordinate space of competing objectives to identify the optimal trade-off surface of solutions that are optimal in, for example two dimensions, with no other solution more optimal. This leads to a family of equally optimal points defining the trade-off surface. In the case of *de novo* design, each of the points is a potential molecular structure that could be prioritised for synthesis.

Many more recent multiobjective *de novo* design systems have been developed using different approaches. Two reviews have been published recently covering the current state-of-the-art in multiobjective *de novo* design.[13,14]

14.10 Summary

De novo design is a relatively recent advance in drug design and even more recently being used routinely in drug design projects. However, it has had a tremendous effect on medicinal chemistry thinking and projects.

In the early receptor-based *de novo* design projects, the emphasis was towards optimising ligands against a target of interest, with a relative disregard for other properties. Furthermore, the receptor-based methods were limited in their search to a relatively local search algorithm.

The main advances in *de novo* design really came to the fore when population-sampling methods were developed to more effectively explore and exploit the search space compared to receptor-based *de novo* design. However, the approach still required further refinement, particularly in methods for synthetic accessibility and effective modelling methods.

Recently, approaches have been introduced that make multiobjective *de novo* design a more effective and potentially routine tool that can be used in medicinal chemistry projects of the future. It is possible now, and already premature projects from the Schneider laboratory at ETH-Zurich and Cyclofluidic are demonstrating large parts of the entire *de novo* design workflow, including the synthesis of compounds and their subsequent testing, being operated as an automated iterative design paradigm with little human input but maximising human benefit.

References

1. *De novo Molecular Design*, ed. G. Schneider, John Wiley & Sons, 2013.
2. G. Schneider and U. Fechner, Computer-based de novo design of drug-like molecules, *Nat. Rev. Drug Discovery*, 2005, **4**(8), 649–663.

3. X. Q. Lewell, D. B. Judd, S. P. Watson and M. M. Hann, Recap retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry,*J. Chem. Inf. Comput. Sci.*, 1998, **38**(3), 511–522.
4. R. A. Lewis and P. M. Dean, Automated site-directed drug design: the formation of molecular templates in primary structure generation, *Proc. R. Soc. London, Ser. B*, 1989, **236**(1283), 141–162.
5. R. A. Lewis and P. M. Dean, Automated site-directed drug design: the concept of spacer skeletons for primary structure generation, *Proc. R. Soc. London, Ser. B*, 1989, **236**(1283), 125–140.
6. N. Brown, F. Gilardoni, B. McKay and J. Gasteiger, A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules,*J. Chem. Inf. Comput. Sci.*, 2004, **44**(3), 1079–1087.
7. N. Brown, B. McKay and J. Gasteiger, Fingal: a novel algorithm for the generation of geometric fingerprints, *QSAR Comb. Sci.*, 2005, **24**(4), 480–484.
8. N. Brown, B. McKay and J. Gasteiger, The *de novo* design of median molecules within a property range of interest, *J. Comput.–Aided Mol. Des.*, 2004, **18**(12), 761–771.
9. N. Brown, B. McKay and J. Gasteiger, A novel workflow for the inverse QSPR problem using multiobjective optimization,*J. Comput.–Aided Mol. Des.*, 2006, **20**(5), 333–341.
10. N. C. Firth, B. Atrash, N. Brown and J. Blagg, MOARF an Integrated Workflow for Multiobjective Optimisation: Implementation, Synthesis and Biological Evaluation,*J. Chem. Inf. Model.*, 2014, **55**, 1169–1180.
11. M. Hartenfeller, H. Zettl, M. Walter, M. Rupp, F. Reisen, E. Proschak and G. Schneider, DOGS: reaction-driven *de novo* design of bioactive compounds, *PLoS Comput. Biol.*, 2012, **8**(2), e1002380.
12. B. Spänkuch, S. Keppner, L. Lange, T. Rodrigues, H. Zettl, C. P. Koch and G. Schneider, Drugs by Numbers: Reaction-Driven *De Novo* Design of Potent and Selective Anticancer Leads, *Angew. Chem., Int. Ed.*, 2013, **52**(17), 4676–4681.
13. C. A. Nicolaou and N. Brown, Multi-objective optimization methods in drug design, *Drug Discovery Today: Technol.*, 2013, **10**(3), e427–e435.
14. C. A. Nicolaou, N. Brown and C. K. Pattichis, Molecular Optimization Using Multi-Objective Methods, *Curr. Opin. Drug Discovery Dev.*, 2007, **10**(3), 316–324.

Part 6
Applications in Medicinal Chemistry

CHAPTER 15

Applications in Medicinal Chemistry

15.1 Overview

As has been shown throughout this volume, there are many computational approaches that have been developed to address challenges in medicinal chemistry, and drug discovery and development. But how are these methods applied in practice? This chapter aims to place some of the methods discussed, and derivatives of those discussed, in context and demonstrate their actual impact in dug design. The chapter will address four of the main stages of drug discovery, from the identification of new drug targets, through to the identification of new hit matter that are relevant to those targets, followed by exploration and development of the hits in the hits-to-leads phase. Lastly, lead optimisation will be considered to bring to bear the wealth of medicinal chemistry data, information and knowledge to optimise the final compound or compounds to be taken through to the clinical context. Lead optimisation necessarily includes consideration of ADME/Tox (Absorption, Distribution, Metabolism, Excretion, and Toxicology), and this will also be discussed. A schematic of the major processes involved in drug discovery is provided in Figure 15.1.

A drug discovery programme and the biology and medicinal chemistry effort involved, including high-throughput screening, X-ray crystallography, cell biology, and more besides, is a vastly complex process, and it is impossible to cover all of the challenges in one section alone. However, it is important to provide a flavour of the work done to support the scientists, and more importantly the overall project progression, using computational drug discovery methods.

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

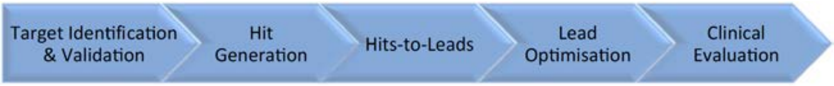


Figure 15.1 A Schematic overview of the general drug discovery and development workflow.

15.2 Early Stage Targets

Without targets of clinical relevance, it would be impossible to progress drug discovery projects in the modern world of therapeutic development. It is important to understand, as best we can, the cellular and molecular processes that have been implicated in disease.

A protein structure may have been identified that has been implicated in a disease indication, but this does not necessarily render that protein structure as a drug target. There are potential targets and there are druggable, or ligandable, targets, where the latter suggests that it is possible to somehow interact with a protein target to bring about a desirable biological effect that is beneficial as a potential therapeutic indication. For example, in a kinase structure one of the aims is to design ATP-competitive (adenosine triphosphate) ligands that block the ATP binding site so that the enzyme cannot catalyse its intended reaction. Therefore, a suitable ATP-competitive ligand will block the mode of action of an enzyme, which has often been shown to be important in killing a pathogen or correcting a metabolic imbalance. Protein kinases are clearly a well-established target class with over 500 identified protein kinases, many of which have been investigated as potential drug targets, and the protein family appears to have eminent druggability potential. The high druggability potential of protein kinases has been demonstrated both by modelling, but also, arguably mainly, by experiments. However, the understanding gained in exploring such druggable protein families can be incorporated into druggability analyses for new and exciting protein families that are only just being discovered or have yet to be discovered. It is here where computational methods can offer an opportunity to identify where a given protein structure is druggable.

It is important here to take a moment to consider the terms druggability and ligandability. As with lead-like and drug-like, there are not really formal definitions for what we mean by these terms. For example, a protein may be identified as druggable by a modelling tool. However, the only proteins that can truly be said to be druggable are those for which we have drugs. The same challenge to these neologisms occurs with ligandability. For a protein to be liganded it is sufficient to say that a ligand is bound to it at its endogenous binding site or another site on the protein. However, this does not suggest that that protein is druggable. Furthermore, it does not suggest it is not. While the terms ligandable and druggable are useful shorthand to describe briefly what we are trying to do, no small blessing when sitting in a conference on the subject, it is important to understand what is implied by these definitions and ensure that these implications are valid for the work being done and the conclusions drawn.

With the caveats and assumptions described above in place, it is however possible to gain great insights into new protein target classes using druggability analyses. Recently, an article appeared on the application of one druggability analysis to consider the novel protein target class of bromodomains.[1] Bromodomains are a new and exciting protein target class that have been implicated in cancer and therefore are well worth consideration of study. However, what is the most effective way to use the wealth of structural data to provide an objective assessment of which bromodomains are potentially druggable and therefore more likely that a chemical probe can be designed?

In the study, the authors identified that not all bromodomains were as druggable as each other, but this was based on a subjective assessment through visualising overlays of the protein binding sites. Therefore, they realised that, rather than manually inspect all binding sites available, perhaps it was possible to use a computational method to assess the druggability of these sites. After some exploration of available software tools and their published applications to druggability analyses, the authors decided to use the SiteMap software from Schrödinger.

SiteMap uses a combination of measurable parameters that can be observed from a crystal structure: volume of the pocket; enclosure of said pocket, or buriedness; and the degree of hydrophobicity in the identified pocket. From these parameters, it is possible to calculate a druggability score, or DScore, which provides an estimate of how druggable each structure may be. This analysis provided a prioritised list of bromodomains where the top-ranked may be considered more druggable than the lower ranked ones. Figure 15.2 illustrates the range and distribution of each of the bromodomains analysed in this study as a box-and-whisker plot.

The authors continued their study by analysing the identified common binding site features and this allowed a grouping of bromodomains to be determined (Figure 15.2). This allowed for a new classification of bromodomains and significant differences to the whole-sequence alignment similarity cluster performed previously. Using the structural information, and focussing on the acetyl-lysine binding site, the authors posit that this bromodomain grouping may be more appropriate for consideration in drug discovery (Figure 15.3).

This druggability study demonstrates one way in which computational methods can contribute positively to a drug discovery project. The authors went to expand on this prospectively with the identification of novel small-molecule inhibitors of the BRD4 bromodomain through structure-based design[2] and the design of a chemical probe for the BAZ2A and BAZ2B bromodomains.[3]

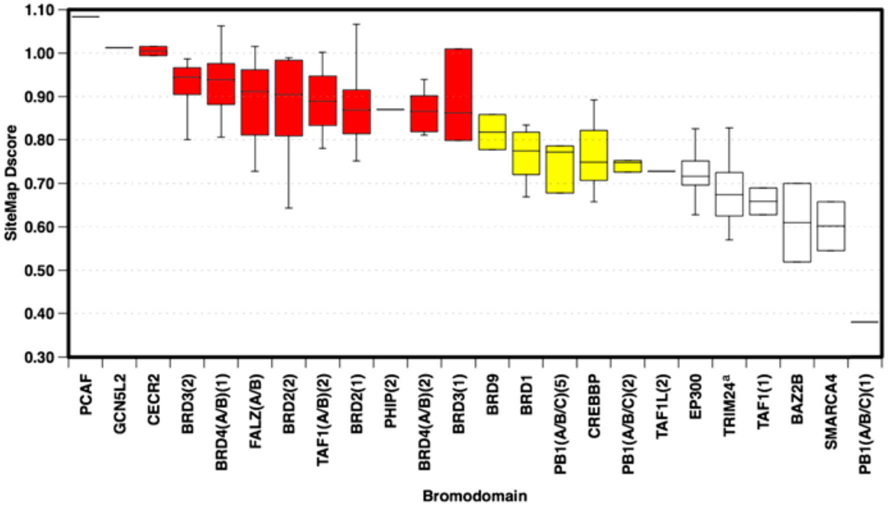


Figure 15.2 Box-and-whisker plot showing the range and distribution of druggability for each bromodomain across available structures passing imposed filters (including presence of binding site water molecules). Ranked by median Dscore. Colours indicate druggability classification: red, druggable; yellow, intermediate; white, difficult. Reprinted from L. R. Vidler, N. Brown, S. Knapp and S. Hoelder, Druggability analysis and structural classification of bromodomain acetyl-lysine binding sites, *J. Med. Chem.*, 2012, **55**(17), 7346–7359.

15.3 Hit Generation

One of the first tasks of any drug discovery programme is to identify validated hit matter. Sometimes this is achieved through serendipity and an understanding that we are lucky, and capitalising on that. Sometimes good quality hit matter already exists from published data in the literature. More often than not, many cycles of trial and error are required to identify and validate good quality hit matter. One of the most common, and effective, methods for hit generation is High-Throughput Screening, where vast libraries of hundreds of thousands of diverse lead-like small molecules are screened in different assays to identify anything that can be identified as a hit, in other words something that brings about the desired biological response. However, a HTS often generates such vast amounts of data that it is difficult to understand and prioritise the most important compounds to take through and investigate further.

Computational methodologies often play a key role in what is called HTS triage: the process by which we identify the compounds to take forward and those to drop. There are many processes by which computational methods can benefit a HTS triage. Property and structural filters may be applied that remove unwanted lipophilic compounds or undesirable structural moieties, respectively. One may argue that if those compounds are undesirable then

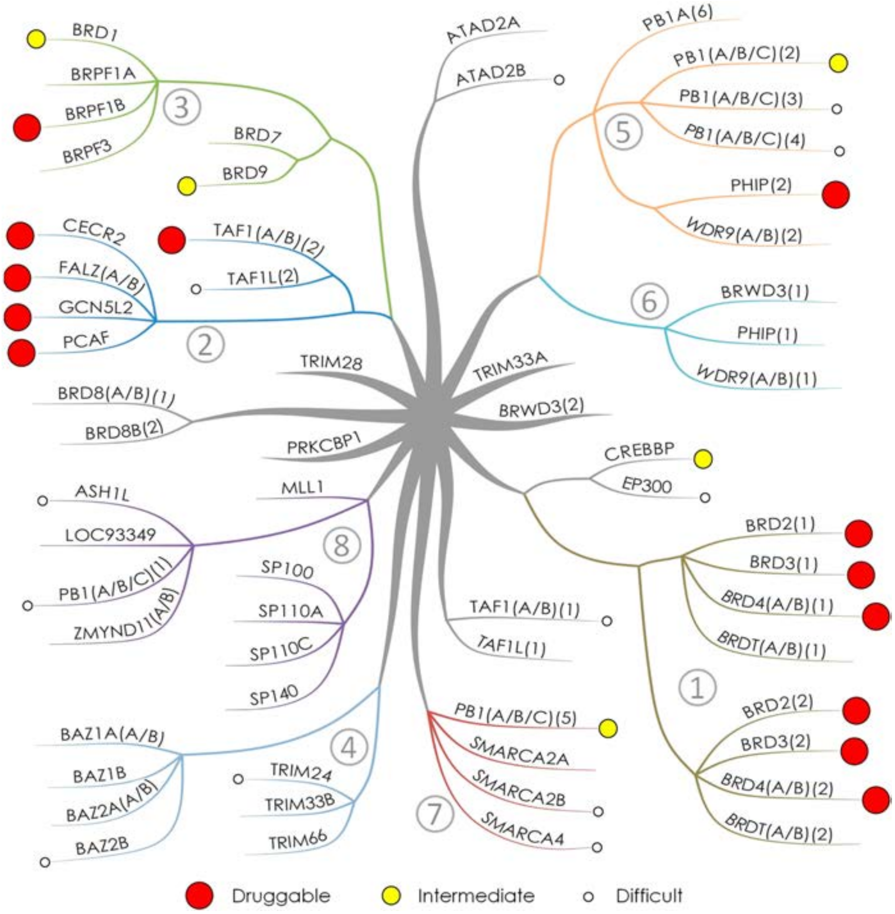


Figure 15.3 Bromodomain classification tree generated on the basis of eight binding site amino acid signatures showing bromodomain druggability. Reprinted from L. R. Vidler, N. Brown, S. Knapp, and S. Hoelder, Druggability analysis and structural classification of bromodomain acetyl-lysine binding sites, *J. Med. Chem.*, 2012, **55**(17), 7346–7359.

they should not have been screened in the first place. However, when working with vast chemical libraries, the logistics of frequently removing compounds that should not be in the library are non-trivial and will most likely lead to even more errors creeping into the chemical compound database.

The primary hit list from an HTS is often defined as the top two or three standard deviations above the mean end-points, typically a single-point percentage inhibition measurement. This would often give more compounds to contend with than is practically possible to consider in a confirmation screen, or other follow-up. Therefore, other computational analyses must be brought in to consider the remaining data. The property filters, or preferably flags, above can be used to reduce the dataset, but it must be remembered that it is much easier to take away rather than re-introduce in these types of analyses; therefore it is always better to flag these data points rather than remove.

At the point of attempting to triage an HTS hit list, a multitude of chemoinformatics methods are often introduced to tackle the challenge. One of the most common approaches is to first cluster or group the hit list based on the chemical structures. Descriptor-based clustering can be very powerful, but is prone to structural mismatches where particular molecular structures fall into inappropriate clusters. The mismatches can occur for a number of reasons, but typically it is either an issue with the molecular descriptor or the clustering method used, or a combination of the two, that renders it difficult to deconvolute the actual problem. Another issue with clustering is that it can be very computationally intensive, and heuristics applied in attempting to reduce the computational complexity can often lead to increased mismatches. Rather the priority in triage should be towards understanding what the desired output from the triage is: reliable hit series that demonstrate, or may have potential to demonstrate, some Structure–Activity Relationship (SAR). Therefore, we must return to considering molecular scaffold representations.

Molecular scaffolds are a method by which compounds can be grouped on common scaffolds or cores. One hypothesis being that those structures that naturally group together will tend to be of medicinal chemistry relevance, that is a medicinal chemist will consider them similar and also be able to identify the potential for chemistry expansion around those common cores, which can lead to much scope for more exploration and exploitation in the lead optimisation phase. However, as have seen in Chapter 10, the ways by which we define scaffolds can be limited since oftentimes the scaffold can over-represent the determined scaffold in the parent molecule.

The Scaffold Tree method was developed to improve the scaffold representations, not only in terms of how scaffolds are defined objectively and invariantly from parent molecules, but also to provide a hierarchy of molecules and their pruned fragments.[4] The limitation of Murcko scaffolds is that the resulting scaffolds are over-specified and can therefore be reduced, since if the scaffold is over-represented in the scaffold tree, it will have very few sub-ordinate nodes in the tree. Using Scaffold Tree and the Scaffold-Hunter software, therefore, can permit a multi-level scaffold analysis, where different levels of the tree may be taken for different scaffolds. This can be a very important consideration when dealing with a large HTS with many random screening compounds, but also some focussed libraries around a common core. An illustration of the Scaffold Tree results from applying the method to the results from a pyruvate kinase assay is provided in Figure 15.4. The Scaffold Tree method can therefore permit a greater control over the scaffold analysis of the available space than may otherwise be the case with alternative scaffold groupings and from cluster analysis. Scaffold-Hunter was published by Wetzel *et al.*[5] and is available as open source from http://www.scaffoldhunter.sourceforge.net/.

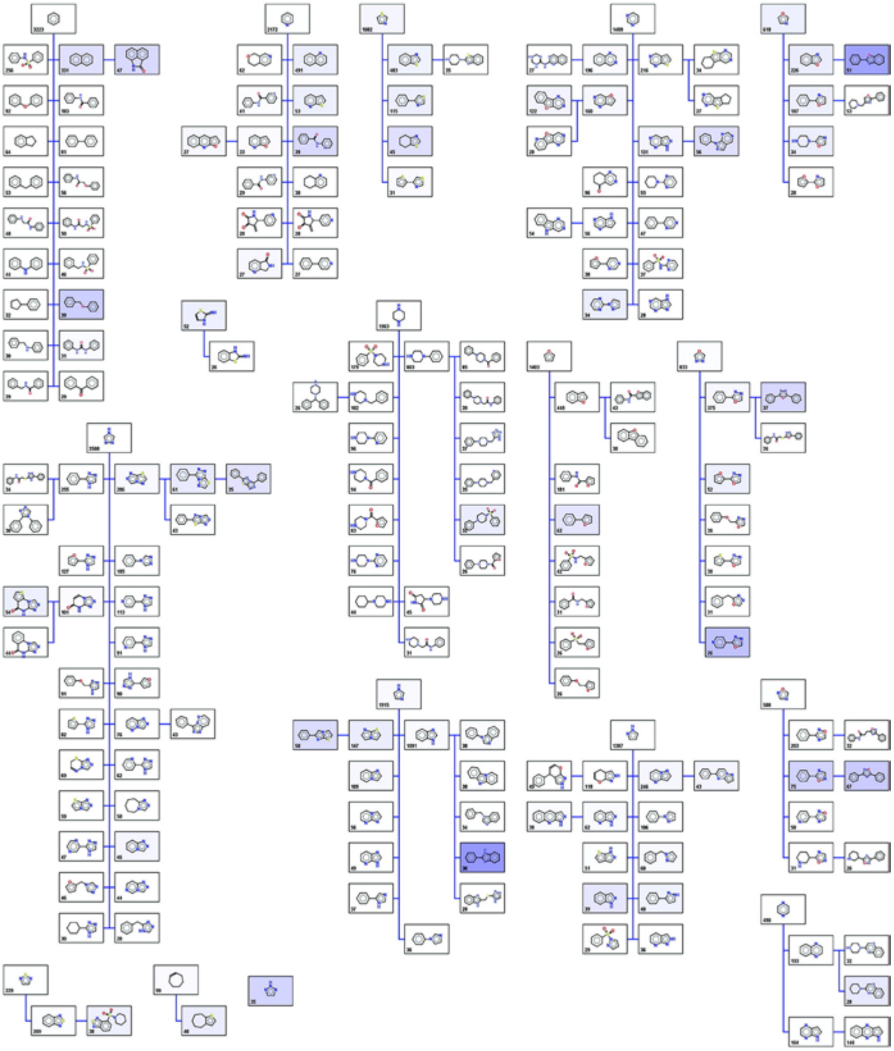


Figure 15.4 Scaffold tree for the results of a pyruvate kinase assay. Colour intensity represents the ratio of active and inactive molecules with these scaffolds. The 2-phenyl-benzoaxazole scaffold can be found at the top right corner. Reprinted with permission from A. Schuffenhauer, P. Ertl, S. Roggo, S. Wetzel, M. A. Koch, and H. Waldmann, The scaffold tree-visualization of the scaffold universe by hierarchical scaffold classification, *J. Chem. Inf. Model.*, 2007, **47**(1), 47–58. Copyright 2007 American Chemical Society.

Another very powerful technique for triaging HTS data is through interactive data exploration or data mining. Interactive data exploration allows a user to explore vast quantities of data across chemical structure and molecular properties. Chemical structure analysis typically is delivered through substructure and similarity searches, right through to cluster and scaffold analysis. The interactivity of these methods tends to arise from the interactive dynamic queries where it is possible to dynamically use sliders to make changes in, for example, property ranges, such as molecular weights between 100 and 350 Da, and see those points appear alone, filtering out anything that does not conform to that query. Another aspect of explorative data analysis is the use of close-coupled data visualisation, where each of the plots is dynamically linked. If a selection of points or a dynamic query is introduced, then all views of the data, scatter plots, histograms, *etc.* are automatically updated.

It is difficult to convey the power of explorative data analysis in static pictures (Figure 15.5), but thankfully Actelion Pharmaceuticals has recently released a version of the in-house explorative data analysis tool as open source that is available at no cost from http://www.openmolecules.org/datawarrior/.[6] The software contains many features that you would otherwise find in commercial software, such as Spotfire, Gigawiz Aabel and Miner3D. Some of the features include:

- Interactive data visualisation and analysis
- Built-in chemical intelligence
- Real-time data filtering on alphanumerical and chemical criteria
- Prediction of molecular properties from the chemical structure
- Dedicated chemoinformatics modules to support drug discovery
- Table view with columns containing alphanumerical or chemical information
- Versatile graphical 2D-view for scatter plots, bar and pie charts, box plots, ...
- Graphical freely rotatable 3D-view for scatter plots and bar charts
- Dedicated chemical structure view with optional alphanumerical data
- Form based view with form designer and form based data editing
- Multiple views are shown side by side or are stacked on top of each other
- Views can be highly customised to reveal multiple dimensions of the data

DataWarrior is a very powerful piece of software that every modeller should have in their toolkit, and the developers are congratulated on contributing this excellent resource to the wider field.

15.4 Hits-to-leads

Once the hit series have been identified from the previous hit discovery phase, the next stage is to explore those hit series. Historically, the process in hit-to-lead would have involved much trial and error in exploring the potential SAR in each hit series. The exploration would have taken the form of adding groups, replacing groups with bioisosteres, and removing groups. While this can often be a little haphazard in its exploration of the chemistry space around each series, it should not be underestimated how powerful

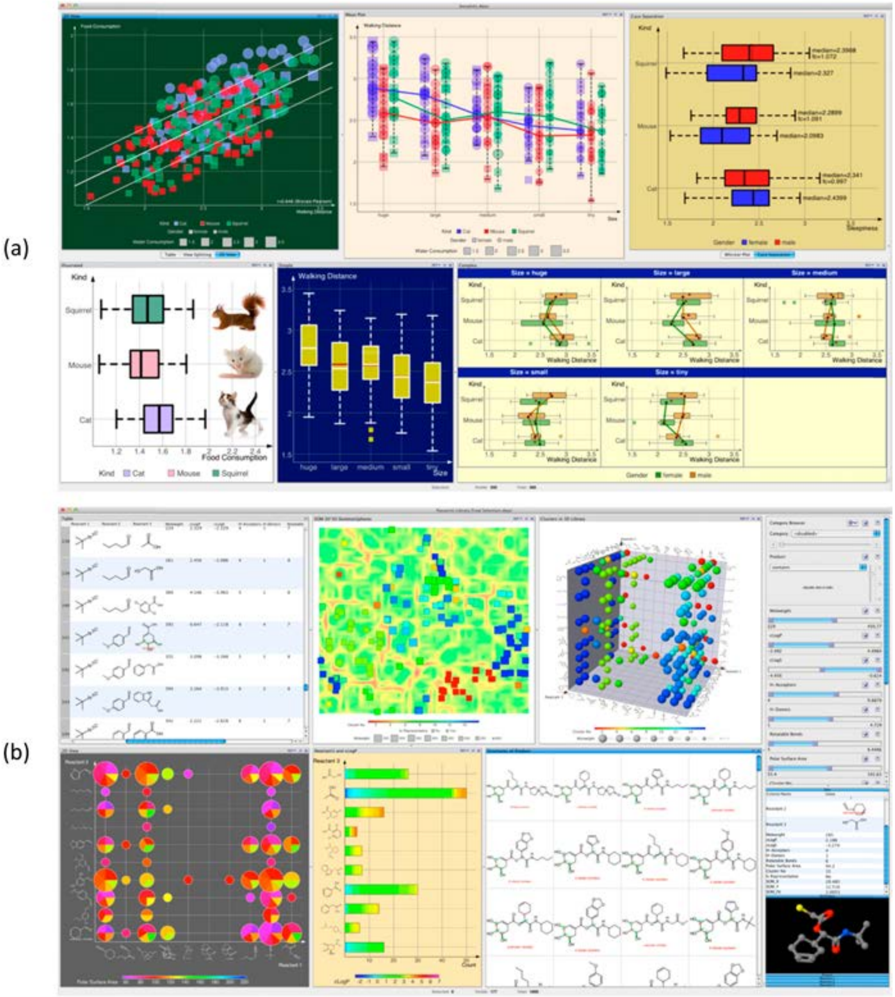


Figure 15.5 Screenshots from the DataWarrior explorative data analysis software released by Actelion Pharmaceuticals. (a) Correlation plot, box plots, whisker plot and statistical parameters. (b) Combinatorial library analysis. From a generic three-component reaction and provided reactant structures, DataWarrior calculated the product structures (bottom half-right) and their physicochemical properties. Products were clustered and arranged on a 2-dimensional self-organised map (top centre). Filters were adjusted (top right) and views created to show various aspects of the selected library subset. Reprinted with permission from T. Sander, J. Freyss, M. von Korff and C. Rufener, DataWarrior, An Open-Source Program For Chemistry Aware Data Visualization And Analysis, *J. Chem. Inf. Model.*, 2015, **55**(2), 460–473. Copyright 2015 American Chemical Society.

this technique can be with expert medicinal chemists. However, as with all humans, there can be a tendency to focus on exploitation rather than exploration, and towards being blinded by local searches and a conservative nature in understanding the SAR around their series of interest. More recently, however, two particular methods have been used to great effect to assist in exploring these series: analogue-by-catalogue and virtual combinatorial libraries.

Analogue-by-catalogue is a relatively simple process and takes advantage of the vast number of compounds that are commercially available from compound vendors and brokers. Typically, most of the big vendors offer many millions of compounds that have been synthesised or could be synthesised rapidly if the client requests. With this vast resource of available chemicals, there is a significant likelihood that chemical series of interest will also be present in these catalogues and that have not yet been tested for the particular end-point in which you are interested.

An analogue-by-catalogue search can be as simple as analysing your extant series and defining a single substructure as your series of interest. With the advent of improved online structure searching, the substructure of interest can be drawn directly into the preferred compound vendor website and within seconds a list of available compounds in that series can be obtained, together with quantity availabilities, estimated costs, and projected delivery dates. The ease of use of this system and reliable feedback of estimated costs and delivery dates can allow the computational and chemistry teams to liaise with the screening teams to expedite biological testing and feedback to the project team for further analysis and discussion on new chemical targets for synthesis.

The substructure search described above can be seen as quite primitive, particularly when the project may already have identified a specific SAR that suggests exploration of particular vectors would not contribute much in addition to what is already available. Here, the team can define SMARTS queries that can more explicitly represent the expected decoration on the scaffold of interest. Some of the simple rules that can be encoded are: blocking exit vectors; specifying a preference for particular groups; and whether aromatic or aliphatic atoms should be present or not at specific positions. Using this rich chemistry pattern search, it is possible to refine and finesse the substructure search with great control. However, caution must be used when applying these filters since it can be quite simple to accidentally ‘freeze’ out your search space and miss important virtual hits.

Another approach that is complementary to the analogue-by-catalogue search above is the use of combinatorial virtual libraries. As the name suggests, a combinatorial virtual library takes the scaffold as defined through analysis of the SAR of the chemical series identified by the project team with exit vectors at which exploration may be made. Next, a set of potential substituents at each exit vector can be defined, or simply mined from datasets using fragmentation schemes, with the potential for filtering down on simple parameters, such as heavy atom count. Given the scaffold, the potential exit vectors, and substituents available at each of those exit vectors, a virtual library can be defined combinatorially by enumerating every combination in this set.

One of the main challenges in applying virtual combinatorial libraries in practice is that it is very easy to define a vast library of potential targets for chemical synthesis that far outweighs the available resource assigned to that project. Here, there are two options that are not mutually exclusive: reduce the space or prioritise the space. Reducing the space would typically take the form of applying filters to remove synthetic targets that were desirable prior to enumeration, but when combined may be, for example, too lipophilic, or too homogeneous (cluster analysis). However, the space may be too vast simply because too many options for substituents have been provided. Reducing the number of substituents considered can drastically reduce the size of the virtual library. Prioritising the space requires statistical models of interest by which the virtual molecules can be considered. These models may be simple models of physicochemical properties applied as surrogates for other properties that cannot be modelled appropriately. However, care should be taken with using property models as surrogates for other properties as the correlation may not be sufficiently strong to warrant that application. Alternatively, it may be possible to design bespoke statistical models, even at this relatively early stage, based on the data generated in hit discovery and hit-to-lead until that point. Simple naïve Bayesian models have been demonstrated to be very effective at modelling HTS readouts and these could be used here. Lastly, it is possible to combine models using multiobjective optimisation to prioritise the virtual molecules for synthesis that satisfy multiple properties of interest, which will tend to reduce the space to a very focussed library, but care should be taken that the screen-out is appropriate.

15.5 Lead Optimisation

The final stage of a medicinal chemistry drug discovery programme to be considered in this section is lead optimisation. Lead optimisation (LO) is a critical stage of a project, if any stage can be said to be non-critical. Here, compounds do not need to only optimise, or maintain, potency against the enzyme target, but many other parameters need to be balanced. Cellular and toxicity assays remain vastly important, but by this stage it may be possible to define sufficiently reliable models for guiding predictions for use in synthesis. In addition, consideration also needs to be made regarding the properties that govern good oral absorption, slow metabolic clearance *in vivo*, and displaying activity in animal models of the disease.

It is clear that LO is a critical and difficult challenge and fraught with multiple competing readouts. At this stage, the data regarding the chemical series that remain in the project, typically one priority series and at least one back-up series, may now be sufficient to build more reliable models than previously possible in the project to predict for enzyme and cellular potency, and also potentially some off-target (selectivity) and anti-targets (such as hERG liabilities), but this may not be the case. However, regardless, the mantra in

LO should be to make best use of the data but do not become beholden to it. The data can certainly inform the decision making in the team, but care must be taken to ensure critical and significant changes are reduced to experiment where possible to test the assumptions made on the models.

Given all of the challenges in LO, and the critical role that data plays at this stage in the process, it is clear that the Matched Molecular Pair Analysis (MMPA) concept could play a pivotal role. The Matched Molecular Pair (MMP) concept, introduced by Kenny and Sadowski, is one of the most recent advances in chemoinformatics and could offer a great deal to LO projects.[7] MMPA takes a pair of molecules that differ in only one minor way, but also have been measured using the same assay, to identify small structural changes that may have a desired effect. Given that the same matched molecular pair will often be seen in different pairs, confidence can be increased in the change in readouts as the number of pairs with the same modifications increases.

It is often claimed that QSARs are not as interpretable as would be desired, which is sometimes the case, but they can still be useful. It is true to say it can be difficult to understand precisely what modification had an effect and how these modifications can be designed. Regardless, the benefit of conducting multiobjective *de novo* design in a late stage project was demonstrated in the previous chapter. However, regardless of opinions regarding QSARs, it is clear that MMPA offers an alternative and interpretable approach to making changes to molecules such that the modification can be rationalised structurally and it is an ideal approach, perhaps complementary to or in combination with other methods.

MMPA is clearly reliant on the data available and this may also be a reason for reports of its benefits to projects in pharmaceutical companies, since they tend to have substantially more, and more relevant and reliable, data than is otherwise publicly available. However, more recently, an online bioisosteric replacement tool has been developed by Wirth *et al.*[8] using the MMPA concept. The SwissBioisostere tool is available online at http://www.swissbioisostere.ch/.

15.6 Summary

In this chapter, methods that have had a demonstrative impact on the various stages of drug discovery have been reported. It is often thought that computational methods in medicinal chemistry can tend to contribute only at certain stages within the drug discovery process, but it is clear for the examples above, and the many more examples that exist, that computational methods have been applied to a wide range of diverse challenges in drug discovery and medicinal chemistry.

The evaluation of targets early in a potential drug discovery project is essential to understand whether the target has potential liabilities that may stymie its progress downstream. Using druggability analyses, it is possible to help prioritise potential drug targets for consideration to initiate a project. The analysis of bromodomains using just such a tool discovered that some bromodomains had a higher chance of success in terms of exploiting the endogenous binding site than others and that those protein targets could therefore be progressed to further analysis, and even the identification of chemical tools as demonstrated in that project.

It is always important to look at the available data, as it becomes available, in different ways and using different approaches and statistical modelling methods to help reveal any hidden relationships and trends. Two such visualisation tools that have wide application within drug discovery projects beyond the hit generation stage were presented in this chapter: ScaffoldHunter and DataWarrior. ScaffoldHunter allows the organisation of structural data through iteratively pruning back the rings in the molecular structures in a given set and organising the structures hierarchically. By mapping the biological assay data, it is possible to highlight scaffolds that could be flagged for follow-up since they are represented by many structures that exhibit activity. Similarly, DataWarrior offers the ability to crosslink data in different data views that allows for the interactive exploration of data. Furthermore, the dynamic filter queries enable the data scientist to drill down to the data of most interest and, if necessary, return to the data and ask different questions of the data.

The last approach covered in this chapter is one that is highly relevant to lead optimisation, but can also be applied with benefit in the earlier stages, bioisosteric replacement. With the large amounts of public data that have been collated and (re)digitised, it is possible to data mine structure trends in big data. Using ChEMBL and MMPA permits just such an analysis and can recover molecular substructures that have been demonstrated, by experiment, to help maintain potency while ameliorating other properties, such as solubility. SwissBioisostere now offers this resource online so that anyone is able to mine these data in a quick and simple way and re-adjust their hypotheses as the data reveals different possibilities.

It is important not to get carried away with all of the methods available today, but it is clearly difficult when they offer so much promise. Therefore, it is essential to appropriately formulate the question being asked of the method or the data, and understand what success would be and how it would be measured. By maintaining this clear focus on research questions and expectations from the data and the methods, these approaches are incredibly valuable.

References

1. L. R. Vidler, N. Brown, S. Knapp and S. Hoelder, Druggability analysis and structural classification of bromodomain acetyl-lysine binding sites, *J. Med. Chem.*, 2012, **55**(17), 7346–7359.
2. L. R. Vidler, P. Filippakopoulos, O. Fedorov, S. Picaud, S. Martin, M. Tomsett, H. Woodward, N. Brown, S. Knapp and S. Hoelder, Discovery of novel small-molecule inhibitors of BRD4 using structure-based virtual screening,*J. Med. Chem.*, 2013, **56**(20), 8073–8088.

3. L. Drouin, S. McGrath, L. R. Vidler, A. Chaikuad, O. Monteiro, C. Tallant, M. Philpott, C. Rogers, O. Fedorov, M. Liu, W. Akhtar, A. Hayes, F. Raynaud, S. Müller, S. Knapp and S. Hoelder, Structure enabled design of BAZ2-ICR, a chemical probe targeting the bromodomains of BAZ2A and BAZ2B, *J. Med. Chem.*, 2015, **58**(5), 2553–2559.
4. A. Schuffenhauer, P. Ertl, S. Roggo, S. Wetzel, M. A. Koch and H. Waldmann, The scaffold tree-visualization of the scaffold universe by hierarchical scaffold classification, *J. Chem. Inf. Model.*, 2007, **47**(1), 47–58.
5. S. Wetzel, K. Klein, S. Renner, D. Rauh, T. I. Oprea, P. Mutzel and H. Waldmann, Interactive exploration of chemical space with Scaffold Hunter, *Nat. Chem. Biol.*, 2009, **5**(8), 581–583.
6. T. Sander, J. Freyss, M. von Korff and C. Rufener, DataWarrior, An Open-Source Program For Chemistry Aware Data Visualization And Analysis, *J. Chem. Inf. Model.*, 2015, **55**(2), 460–473.
7. P. W. Kenny and J. Sadowski, *Chemoinformatics in Drug Discovery*, Wiley-VCH Verlag GmbH & Co. KGaA, 2005, pp. 271–285.
8. M. Wirth, V. Zoete, O. Michielin and W. Sauer, SwissBioisostere: a database of molecular replacements for ligand design, *Nucleic Acids Res.*, 2013, **41**(D1), D1137–D1143.

Part 7
Summary and Outlook

CHAPTER 16

Summary and Outlook

16.1 The Past

Arguably, computers and chemistry have gone hand-in-hand since even before computers as we know them existed. Mathematics and chemistry have certainly been closely linked for many centuries. It is important for any scientist to understand the context of what they are learning and this is why a history was given regarding the close links between graph theory and chemistry, going back to the advent of atomistic theory.

The work of Alexander Crum Brown and others gave rise to what we now see as the chemistry *lingua franca* almost two hundred years ago. Crum Brown's further work on what we now call structure–activity relationships was pioneering at its time and it is amazing to think now that this mathematical insight as to how molecules work physiologically is down to a simple, but oh-so-complicated, function of chemical constitution.

Moving forward to Langmuir's work on isosteres, which founded a whole new field that only in recent times has begun to resonate with importance as to its impact in drug design and medicinal chemistry. Langmuir's work led to the foundation of bioisosteric replacement, which is fast becoming one of the most important facilitators of new ideas for some time.

Markush introduced the protection of ideas to the chemistry world in 1924, but not just ideas, also families of ideas. Now every medicinal chemistry patent has generic Markush structures to represent the space to be covered by the patent, with the necessary caveats of the need for reduction to practice.

The history has shown us much and I think still has much to show us. There are most likely methods to be rediscovered, ideas from minds thinking well ahead of their time and awaiting the development of algorithms that can finally release those answers that were considered so long ago. What is clear

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

is that history has shown us that we have been doing a lot of what we think is new for a very long time. We may not have realised that we have already been doing it, but looking back to the history of the field and related fields demonstrates that many of the theories and contexts are the same.

16.2 The Present

Now, everyday computers are used in every aspect of drug discovery, from their use as Electronic Laboratory Notebooks (ELNs), to central repositories or enterprise information systems that store all of our metadata about our compounds, to online search engines that permit you to find compounds you want to test using the Internet, or calculate properties, or perform a substructure or similarity search to identify other interesting compounds.

All of the systems mentioned above use the foundations of everything that has been covered in this book. The concepts of molecular similarity are used every day, whether it is simply in finding something similar to what was wanted, through to analogue-by-catalogue searches where a range of similar compounds might be needed to follow up on a high-throughput screen.

Calculated properties are all-pervasive to the point that sometimes one needs to be reminded that they are calculations after all. Data analysis techniques now easily allow us to conduct vast analyses on tens of thousands of chemical structures, across dozens of descriptors. Models can be generated on an *ad hoc* basis, and derivatives used to generate ideas that might be otherwise outside where the project was planning to go.

Other methods from areas of theoretical chemistry are also beginning to find their feet in computational drug discovery. Molecular dynamics simulations are now frequently being used in drug discovery programmes to understand where loops may move and how the overall carbon backbone may flex allowing the identification of potential binding sites we otherwise would not have recognised.

Over the past few years, the increase in routine calculations using molecular modelling approaches has been rapid. However, it is of utmost importance for us as scientists to question and formulate our hypotheses appropriately and have a realistic expectation of what to expect from the results. It is possible to generate vast swathes of data, but they cannot be considered results until they have been analysed and the conclusions interpreted and justified.

16.3 The Future

It is already possible to see some of the advances that could be seen as futuristic only a couple of years ago. The ready access to data will only become more pronounced as search and retrieval techniques will allow us to not only release more and more data as it is generated, but also to go back into the archives and retrieve ‘lost’ data and chemical structures from chemical papers and archives.

The new data that will be generated and released will enable many new analyses, and even the development of methods that hitherto were merely pipedreams due to the paucity of available data. Large-scale databases, such as those from ChEMBL, have already revealed interesting trends that otherwise would not have been identified. This will only increase in the future as we map more of the chemistry spaces in which we are working.

Future successes will rely on data and new algorithms, but also on the scientists not losing sight of the goal and using the different methods appropriately, based on hypothesis-driven science. It is important to understand how the algorithms work and what their potential failings may be. Without this understanding, the methods are black boxes that cannot reveal whether the conclusions are valid and can therefore be very costly downstream when these ideas leave the computer and enter the lab, with the relatively much more expensive processes of synthesising these new compounds and evaluating their utility in the biology labs. It is important as a modeller to ensure that every aspect has been considered appropriately and to understand limitations so that the project team, including the modellers, can be involved in the decision as to whether this idea or that is worthy of further consideration.

Structure-based methods continue to improve and will most likely continue to do so in the future. Fundamental challenges remain, but these are being addressed as time passes. For instance, understanding when and where docking methods can be applied, and an understanding of probabilities of success, will inform on the ligand structures that we are investigating with these methods. Additional up-stream filtering and consideration will likely become more flexible to deal with new challenges as they are discovered. As more structural data becomes available, it will be possible to make more accurate predictions routinely using multiple models and consensus or ensemble docking.

It is now becoming ever more apparent that no single method will outperform another in all circumstances, but the understanding of where these different methods can be applied is becoming clearer. Methods of combining different methods—molecular descriptors, statistical learning, shape-based and pharmacophore analyses, and protein–ligand docking—will be refined with more methods used in concert to optimise not only the predictions, but also the quality of those predictions.

Perhaps the most exciting aspect of current and on-going research is the prospects that *de novo* design offer in terms of designing new molecular entities *in silico*. These methods will allow the modeller and the chemist to reveal hitherto uncharted regions of chemical space, and dig down to exploit those areas to maximum effect. It must be made clear that these approaches are unlikely to replace the expertise of the modeller or the chemist since these skillsets are still essential in using the tools appropriately as well as critically evaluating the results. However, it is clear, from a number of recent studies that *in silico* drug design is coming of age in terms of a demonstrable impact in drug discovery.

16.4 Summary

This book covers a wealth of research, both historical and still active, in the field of computational drug discovery. Much had to be omitted due to clarity and brevity to appropriately cover the field for the beginners, but the book still also offer items of interest to those who have some more experience in the field. Where space was limited, appropriate review articles have been referenced as pointers to further and more in-depth information.

It is clear from the breadth of methods and algorithms discussed in this volume that the field covers a wide range of methods and therefore requires scientists from all relevant disciplines to get involved and help make the future discoveries that will ultimately result in benefits to humanity as a whole.

Appendices

APPENDIX A

Glossary of Terms

Summary

Many words and terms used in the field of *in silico* medicinal chemistry are non-obvious and potentially confused with their use in other fields. Here, a glossary of terms used in the context of *in silico* medicinal chemistry is provided that seeks to act as a ready reckoner of these terms and for clarification in reading both this text and other papers and books published in the field.

Term	Definition
Ångström (Å)	The standard unit of length in chemistry where 1 ångström is equal to 10^{-10} metres, or one ten-billionth of a metre. As an example, a typical hydrogen-bond interaction is around 1.5 to 2.5 ångströms
Area under the curve (AUC)	The calculation of the area under the curve of a receiver operator characteristic or enrichment plot to quantify the relative enrichment of active molecules at a particular point or points in a screening experiment
ClogP	Calculated descriptor of the octanol/water partition coefficient. Not to be confused with the measured property, logP
Cluster analysis	Grouping objects according to similarity whereby the same groups or clusters are more similar to each other than they are to objects in other clusters
Conformer	A specific spatial arrangement of atoms in a molecular structure in *xyz* space
Confusion matrix	A simple method of summarising the quality of a binary statistical classifier by comparing the known classes and the predicted classes. This results in a two-dimensional matrix that summarises how many true positives, true negatives, false positives and false negatives have been predicted

(*continued*)

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

Term	Definition
DA	Discriminant analysis
Dependent variable	Represents the output or effect, often a biological end-point, such as percentage inhibition or inhibitory concentration
Descriptor	A calculated, or sometimes measured, value that describes an aspect of the molecular structure. These range from calculated physicochemical properties and those determined from the connectivity of a structure to the geometries or conformers of the molecules
Diversity	In the context of molecules, diversity is a measure of the evenness of distribution of the molecular structures over a defined descriptor space. Diversity is important in designing diverse libraries and analysing hit matter from high-throughput screens
Docking	The process by which, given a protein binding site, a small molecule (or ligand) is placed into the binding site such that internal and external strains and interactions are minimised and maximised, respectively. The search algorithms typically explore conformations, rotations, and positioning of the ligand in three-dimensional space
EC_{50}	Half maximal effective concentration
Enrichment	The measurement of the quality of a ranked list of structures from a virtual (or any) screen in terms of the discrimination of active structures from inactive structures
Fingerprint	A vector of binary, categorical, integer or real-valued data that can be applied as a mask screen, a complexity measure, or to calculated similarities between objects
*f*sp$_3$	Fraction of sp$_3$ centres in a given molecular structure as the number of heavy atoms contained in the whole molecule. Often presented as a descriptor of three-dimensionality, it is actually a descriptor of molecular complexity that could give rise to increased three-dimensionality. It is rapid to calculate, from topological structures, and therefore structural conformers are not required
Genetic algorithm (GA)	A natural heuristic optimiser that is a computational analogue of Darwinian evolution. GAs are used widely in computational drug discovery where they are accepted as rapid and effective global optimisation algorithms
H-bond	Hydrogen bond
HA	Heavy atom
HBA	Hydrogen bond acceptor
HBD	Hydrogen bond donor
High-throughput screening (HTS)	The physical measurements of a, typically, biological end-point of a large library of compounds in a suitably short timeframe, typically a few weeks
IC_{50}	Half maximal inhibitory concentration
Independent variable	Represents the variables that can be tested to see if they are the cause
Intellectual property (IP)	The means by which novel molecular entities are protected to enable revenue generation as compensation for the expense of drug discovery and development

K_d	Dissociation constant
K_i	Inhibition constant
Ligand	In biochemistry or pharmacology, a ligand is usually a small molecule that complexes with a biological macromolecule to trigger or inhibit a signal. In drug discovery, the ligand is typically the small molecule being optimised and will eventually lead to a small-molecule drug
logD	Logarithm of distribution coefficient
logP	Logarithm of partition coefficient
Ligand efficiency (LE)	A parameter calculated as the ratio of the biological activity and size of the molecular structure, typically in heavy atoms. The higher the ligand efficiency, the more potency is generated without a concomitant increase in size
Lipophilic ligand efficiency (LLE)	The equivalent of ligand efficiency, but simply subtracts the ClogP from the potency in pIC_{50} (negative logarithm of the IC_{50} data)
Matched molecular pairs (MMPs)	Two molecular structures differing in only one substituent or group and having the same measured property. The difference in the measured property can then be inferred from the change in the matched molecular pair
Molecular weight (MW)	The mass of a molecular structure in terms of its atoms' contributions to its mass measured in daltons (u or Da)
Multiobjective optimisation	The process by which an optimiser takes into account multiple objectives or parameters when optimising. Typically in drug discovery this is used to mean optimising molecules that satisfy multiple properties, such as: enzyme potency, cell potency, solubility, toxicity, metabolism, *etc.* This is sometimes called multiparameter or multiparametric optimisation
Murcko scaffolds	The Murcko scaffolds (proposed by Bemis and Murcko) is an objective and invariant scaffold representation of a molecular structure such that only the cyclic groups are retained and their interconnecting acyclic groups. The scaffold is applied in scaffold similarity and diversity analyses
Naïve Bayesian classifier (NBC)	A simple supervised statistical learning method that uses Bayesian probabilities priors on extant objects to predict for unknown new objects. Essentially, if a property has been observed before and was identified as 'good', it will be seen as good in any new object that exhibits that same property
Pareto ranking of optimality	The calculation of a ranking of objects in greater than two-dimensions, such that the non-dominated objects are said to be those in which no other object out-performs them in all dimensions. The non-dominated front is the family of all non-dominated solutions
Partial least squares (PLS)	A supervised statistical learning method to build a regression model. PLS can often work well when there are many more variables than observations
Principal components analysis (PCA)	A statistical projection method used in multivariate and megavariate analysis to reduce the number of dimensions, but retain the explanation of as much statistical variance of the original data matrix

(*continued*)

Term	Definition
Pharmacophore	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
Physicochemical properties	Calculated properties that correlate with measured phenomena. Examples are molecular weight, ClogP, *etc.*
pIC_{50}	The negative logarithm base-10 of the IC_{50} value: • IC_{50} of 1 μM = 10^{-6} M: pIC_{50} = 6.0 • IC_{50} of 100 nM = 10^{-7} M: pIC_{50} = 7.0 • IC_{50} of 10 nM = 10^{-8} M: pIC_{50} = 8.0 • IC_{50} of 1 nM = 10^{-9} M: pIC_{50} = 9.0 • IC_{50} of 30 nM = 3×10^{-7} M or $10^{-7.5}$ M: pIC_{50} = 7.5
Plane of best fit (PBF)	A molecular descriptor as a measure of three-dimensionality that calculates the best-fit plane between each of the heavy atoms and the average of the distances of those atom to the fit plane
Principal moments of inertia (PMI)	A molecular descriptor that quantifies the degree to which a molecular structure is rod-like, disc-like, and sphere-like. The calculation can therefore be applied to quantify the three-dimensionality of a given ligand structure
Probe	In the context of similarity searching and virtual screening, a probe is a ligand of interest that is used, through ligand similarity methods (sometimes using metadata), to identify similar ligands that may be of interest
Protein–protein interaction (PPI)	The interface between two proteins that can be usefully disrupted by designed small molecules to therapeutic effect
Quantitative structure–activity relationship (QSAR)	Quantitative structure–activity relationship, a supervised model that attempts to correlated an activity endpoint with molecular structure
Receiver-operating-characteristic (ROC) curve	A plot that illustrates the performance of a binary classifier model
Similarity	Normally used to denote molecular similarity. A quantification of the level of similarity between two chemical structures based on descriptors, such as molecular fingerprints, structural overlap or physicochemical properties
Similarity searching	Given a probe molecule and a database of chemical structures, the probe is compared to each structure in the database in turn using some similarity measure. Once the probe has screened the database, the database can be ranked in descending order of similarity. The most similar molecules to the probe will now appear at the top of the list for focussed analysis
SMILES arbitrary target specification (SMARTS)	Related to SMILES strings, but an extension that optimally permits the encoding of substructure search queries with varying levels of specificity
Simplified molecular-input line-entry specification (SMILES)	A line notation to denote the two-dimensional (or topological) structure of a molecular graph

Supervised learning	Model that attempts to correlate an independent variable, typically biological activity, with dependent variables, typically calculated molecular descriptors
Unsupervised learning	Model that reveals the structure of a given data matrix without any external training or correlating property
Virtual screening	The process by which a large virtual library of chemical structures is prioritised using similarity searching, statistical model and simulations (such as docking) such that the most relevant structures will appear at the top of the list

APPENDIX B

Professional Societies

Summary

This appendix provides an overview of some of the most important professional societies in computational drug design worldwide. The list is not exhaustive and is limited to those with direct relevance to drug design.

Chemical Information and Computer Applications Group (CICAG)

Parent: The Royal Society of Chemistry, London

Website: http://www.rsc.org/Membership/Networking/InterestGroups/CICAG/

Overview: The Chemical Information and Computer Applications Group (CICAG) emerged in 2007 from the unification of the Chemical Information Group (CIG) and Computer Applications Subject Group (CASG), both from within The Royal Society of Chemistry, London. The objective of the committee is to raise awareness of chemical information and applications, including services and developments in this field of research that change very quickly. The CICAG organise a range of conferences of relevance to its membership.

The UK-QSAR and Chemoinformatics Group (UK-QSAR)

Website: http://www.ukqsar.org/

Overview: The UK-QSAR and Chemoinformatics group was formed in Yugoslavia at the European QSAR Meeting held in 1986. It was identified by a group of British scientists that there was an opportunity to offer a similar

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

meeting to EuroQSAR to serve research in the UK. Initially formed as the UK QSAR Discussion Group, the name later integrated chemoinformatics into its name to reflect the close links between these areas of endeavour and to co-ordinate activities. The UK-QSAR and Chemoinformatics Group now organises two one-day events per year around the United Kingdom at no charge to delegates interested in these fields.

Chemical Structure Association Trust (CSAT)

Website: http://www.csa-trust.org/

Overview: The Chemical Structure Association Trust (CSAT) is an internationally known society that promotes and supports education, research and development in the field of storage, processing and retrieval of information concerning chemical structures. The CSAT enables their support through a wide range of awards and travel grants and has supported many scientists during its existence. The CSAT publishes a regular newsletter that is available online.

Molecular Graphics and Modelling Society (MGMS)

Website: http://www.mgms.org/

Overview: The Molecular Graphics and Modelling Society (MGMS) was formed in 1981 to represent and support all scientists conducting research at the interface between different fields of study, such as chemistry, physics, biology, mathematics and computer science, who have a unified interest in molecular modelling and graphics. The MGMS is a charity that is voluntarily funded from both academia and industry. The society organises many conferences under its umbrella and in collaboration with other learned societies, such as CSAT, and also organises lecture series of interest to members. The society also organises the long-running and highly popular Young Modellers' Forum, at which a number of early-stage scientists are invited on abstract submission to present either orally or by poster presentation at a dedicated meeting for early-stage scientists. The society also supports its official journal, the Journal of Molecular Graphics and Modelling.

Division of Chemical Information (CINF)

American Chemical Society
Website: http://www.acscinf.org/

Overview: The ACS Division of Chemical Information (CINF) has as its primary objective the aim to promote the generation of, access to, and use of all the information and knowledge generated worldwide. As such, CINF balances its focus between expertise in science informatics, information technology, and librarianship to ensure all of its membership is covered and represented.

Division of Computers in Chemistry (COMP)

American Chemical Society
Website: http://www.acscomp.org/

Overview: The ACS Division of Computers in Chemistry has a highly diverse membership and seeks to support the application of the latest innovations in theoretical chemistry to the experimental, physical and biological sciences. The ACS COMP organise a twice-yearly session at the ACS National Meetings, which are held in a variety and ever-changing selection of American cities. In addition, the ACS COMP awards a number of prizes for excellence in computational chemistry at all levels of scientific endeavour from graduate students to established scientists.

The Cheminformatics and QSAR Society

Website: http://www.qsar.org/

Overview: The International QSAR Society was founded in 1989 at a Gordon Conference on QSAR. At the 1995 QSAR Gordon Conference, the title of the society was changed to The QSAR and Modelling Society. In spring 2007, the Board decided to change the title of the society to The Cheminformatics and QSAR Society. The change in name reflects the changing emphasis of computational research in chemistry to increasingly include chemoinformatics. All scientists who are involved in chemoinformatics and/or investigate quantitative structure–activity relationships in medicinal, agricultural or environmental chemistry are encouraged and invited to join the Cheminformatics and QSAR Society.

Chemistry-Information-Computer (CIC)

The German Chemical Society (Gesellschaft Deutscher Chemiker e.V.)
Website: https://www.gdch.de/netzwerk-strukturen/fachstrukturen/chemie-information-computer-cic.html

Overview: The CIC organises an annual conference on chemoinformatics, held in Germany. The conference has been held for a number of years; the official langue of the conference is English and invitations to speak are accepted from all scientists active in the field.

APPENDIX C

Journals

Summary

What follows is a list of some of the most popular journals in the field of molecular design and computational chemistry modelling. This list is by no means meant to be comprehensive, but represents a substantial subset of the most read journals in the field that would be well worth considering reading for someone starting out in the field.

Journal of Chemical Information and Modeling

J. Chem. Inf. Model.
American Chemical Society
2004–Present

Previously:
Journal of Chemical Documentation (1961–1974)
J. Chem. Doc.
Journal of Chemical Information and Computer Sciences (1975–2004)
J. Chem. Inf. Comput. Sci.

Overview: The Journal of Chemical Information and Modeling has long been the main journal in the field. The journal began in 1961 as the Journal of Chemical Documentation. The remit of this journal is to publish new methodologies and applications in the fields of chemical information and molecular modeling. Specific areas of interest include: computer-based searching of chemical databases; molecular modeling; and computer-aided molecular design of new materials, catalysts or ligands. It is one of the most highly rated journals in modeling and also in computer science itself. This journal

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

is an excellent place to start investigating the long history of chemical information right up to the present day, represented by a relatively high impact factor.

Website: http://pubs.acs.org/journal/jcisd8

Journal of Computer-Aided Molecular Design

J. Comput.–Aided Mol. Des.
Springer-Verlag GmbH
1987–Present

Overview: This journal covers the theory and application of computer-based methods in the analysis and design of molecules. The journal covers the following topics, but this is not exhaustive: theoretical chemistry; computational chemistry; computer and molecular graphics; molecular modelling; protein engineering; drug design; expert systems; general structure–property relationships; molecular dynamics; and chemical database development and usage. The journal also incorporates the journal Perspectives in Drug Discovery and Design, and each volume contains issues dedicated to the remit of that journal. The journal is the official journal of The Cheminformatics and QSAR Society (www.qsar.org).

Website: http://www.springer.com/chemistry/physical+chemistry/journal/10822

Molecular Informatics

Mol. Inf.
Wiley-VCH GmbH
2009–Present

Previously:
QSAR and Combinatorial Science (1982–2009)
QSAR Comb. Sci.

Overview: This journal covers all aspects of molecular informatics, including biology, chemistry and computer-assisted molecular design. In particular, the journal seeks to enhance the understanding of ligand–receptor interactions, macromolecular complexes, molecular networks, design concepts and processes. The journal also includes the unique 'Methods Corner' review-type articles that feature important technological concepts and advances within the scope of this journal.

Website: http://www.onlinelibrary.wiley.com/journal/10.1002/(ISSN)1868-1751

Journal of Molecular Graphics and Modelling

J. Mol. Graphics Modell.
Elsevier B.V.
1996–Present

Previously:
Journal of Molecular Graphics (1983–1996)
J. Mol. Graph.

Overview: This journal is dedicated to the publication of papers on the application of computers in theoretical investigations of molecular structure, function, interaction and design. The scope covers all aspects of molecular modelling and computational chemistry. The journal publishes in association with two highly active professional societies in the field: Molecular Graphics and Modelling Society (MGMS) (www.mgms.org) and the American Chemical Society Division of Computers in Chemistry (COMP) (www.acscomp.org).

Website: http://www.sciencedirect.com/science/journal/10933263

Journal of Chemical Theory and Computation

J. Chem. Theor. Comp.
American Chemical Society
2004–Present

Previously:
Journal of Chemical Documentation (1961–1974)
J. Chem. Doc.
Journal of Chemical Information and Computer Sciences (1975–2004)
J. Chem. Inf. Comput. Sci.

Overview: This journal covers more of the theoretical and computational chemistry aspects of the field, but tends to focus more on new theories and methodologies related to quantum electronic structure, molecular dynamics and statistical mechanics. Specific topics of interest to the journal include: *ab initio* quantum mechanics, density functional theory, design and properties of new materials, surface science, Monte Carlo simulations, solvation models, QM/MM calculations, biomolecular structure prediction, and molecular dynamics in the broadest sense, including gas phase dynamics, *ab initio* dynamics, biomolecular dynamics and protein folding. The journal explicitly does not accept papers describing straightforward applications of known methods, including DFT and molecular dynamics, instead preferring to focus on fundamental breakthroughs and advances in theory or methodology with applications to compelling problems.

Website: http://pubs.acs.org/journal/jctcce

Journal of Cheminformatics

J. Cheminf.
Chemistry Central—Open Access chemistry platform of Springer Science+ Business Media
2009–Present

Overview: This journal is entirely open access, but all papers are still peer-reviewed, and it covers all aspects of molecular modelling and cheminformatics. Coverage includes: chemical information systems, software and databases; molecular modelling; chemical structure representations and their application in structure, substructure and similarity searching of chemical substance databases and reaction databases. The journal also covers molecular graphics, computer-aided molecular design, expert systems, QSAR and data mining techniques.

Website: http://www.jcheminf.com/

Wiley Interdisciplinary Reviews: Computational Molecular Science

WIREs Comp. Mol. Sci.
John Wiley & Sons Ltd.
2011–Present

Overview: This is a journal of essentially review articles that cover a wide variety of methods in the field of computational molecular science. The journal covers the following top-level disciplines as part of its remit: electronic structure theory, molecular and statistical mechanics, computer and information science, computational chemistry, and theoretical and physical chemistry.

Website: http://wires.wiley.com/WileyCDA/WiresJournal/wisId-WCMS.html.

APPENDIX D

Resources for Computational Drug Discovery

Summary

It was not that long ago that both data and the software with which to analyse them were unavailable unless large licence fees were paid to various companies. While it is still the case that commercial vendors offer databases and software for licence fees, and often those resources can be worth it, much more data and software are becoming available to all as free downloads as both closed and open source offerings.

This appendix covers a range of the leading resources on offer in the field, from compound datasets with biological activities and software to analyse those data to software application program interfaces (APIs) to enable software developers to design and implement new algorithms and protocols to conduct novel research. The list is not exhaustive, but covers a representative set of tools that are discussed frequently at relevant conferences. Apologies to any tools that have been left off the list, this was entirely by accident.

RDKit

Resource: Chemoinformatics API for C++, C#, Java and Python

Website: http://www.rdkit.org/

Overview: RDKit was first developed at Rational Discovery to provide an API platform for building supervised statistical models, but then in June 2006 Rational Discovery was closed down and the RDKit was released as open

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

source under a BSD licence. Since 2006, RDKit has remained open source and has been support by scientists at the Novartis Institutes for BioMedical Research and an increasingly sizeable user base in academia and industry.

The core functionality of RDKit is implemented in C++, but Python, Java and C# wrappers are available, with the Python API appearing to be the most popular with users. The RDKit offers a substantial amount of functionality for both two-dimensional and three-dimensional molecular operations, and reads and writes most common file formats, including SMILES/SMARTS and SDF. The chemoinformatics functionality includes substructure searching, generation of canonical SMILES, support for chirality, chemical transformations and reactions, and serialisation of molecules into text.

In addition to the functionality above, RDKit offers a wide range of 2D fingerprints (including ligand-based topological pharmacophores), similarity and diversity selection, generation of 2D and 3D co-ordinates, and a variety of common molecular and physicochemical descriptors. RDKit also has very close integration with IPython allowing for interactive and explorative scripting in IPython Notebooks, which has been proposed as a method by which methods can be published with the actual implementations in journals in the future.

Regular updates to RDKit are released at six month intervals and the community is very active over email on the discussions list for asking questions and, if you can get there before Greg, answering them!

Use of RDKit has increased substantially in recent years, with the first RDKit UGM being held in London in 2012, and has been held annually every year since.

Scikit-Learn

Resource: Statistical Learning and Visualisation Python API

Website: http://www.scikit-learn.org/

Overview: While not strictly a domain-dependent API, scikit-learn offers many of the unsupervised and supervised statistical learning methods mentioned in this book and those that are used widely in the community. Integration with RDKit permits IPython Notebooks to be written that can handle the chemoinformatics processes first, followed by statistical learning methods, and subsequent interrogation of numerical and structural data.

ChEMBL

Resource: Chemical database of bioactive molecules

Website: https://www.ebi.ac.uk/chembl/

Overview: ChEMBL is one of the most outstanding resources of chemical biology data to be released in recent years. Originally, the database was developed by a biotechnology company called Inpharmatica and was called StARlite. In 2008, thanks to an award from the Wellcome Trust, ChEMBL was created and a chemogenomics group at the Wellcome Trust in Cambridge created, led by John Overington, to support its maintenance and pursue research activities using the data.

The ChEMBL database currently contains 10 744 targets, 1 715 667 compound records, 1 463 270 distinct compounds and 13 520 737 activities manually curated from 59 610 publications.

SureChEMBL

Resource: Chemistry Patent Database

Website: https://www.surechembl.org/

Overview: From the guys who brought you ChEMBL! SureChEMBL is a huge online patent database that not only has access to the original patent documents, but it has also abstracted out the text and structural information. The system offers online structure searches using substructures and molecular similarity.

myChEMBL

Resource: Chemoinformatics Virtual Machine

Website: ftp://ftp.ebi.ac.uk/pub/databases/chembl/VM/myChEMBL/current/

Overview: Again from the ChEMBL guys comes a virtual machine that is ready-to-run with many different computational drug discovery and chemoinformatics functionalities. myChEMBL comes preinstalled with PostgreSQL, which is preloaded with the latest ChEMBL version, including additional tables to enable RDKit similarity searching. Therefore, as you would expect, RDKit is also installed. The system also comes with IPython preinstalled and ready-to-run, allowing interactive chemoinformatics scripting in RDKit using ChEMBL.

DataWarrior

Resource: Visual explorative data analysis and interactive data mining

Website: http://www.openmolecules.org/datawarrior/

Overview: DataWarrior combines dynamic graphical views and interactive row filtering with chemical intelligence. Scatter plots, box plots, bar charts and pie charts not only visualise numerical or category data, but also show trends of multiple scaffolds or compound substitution patterns. Chemical descriptors encode various aspects of chemical structures, *e.g.* the chemical graph, chemical functionality from a synthetic chemist's point of view or 3-dimensional pharmacophore features. These allow for fundamentally different types of molecular similarity measures, which can be applied for many purposes including row filtering and the customisation of graphical views. DataWarrior supports the enumeration of combinatorial libraries for the creation of evolutionary libraries. Compounds can be clustered and diverse subsets can be picked. Calculated compound similarities can be used for multidimensional scaling methods, *e.g.* Kohonen nets. Physicochemical properties can be calculated, structure–activity relationship tables can be created and activity cliffs visualised.

KNIME

Resource: Interactive and Visual Workflow Tool

Website: https://www.knime.org/

Overview: KNIME is a free, yet closed source, tool to develop visual data pipelines using prewritten components or bespoke developed components. Many tools plug into KNIME, including RDKit and scikit-learn. The platform also has nodes available for commercial software tools, such as MOE from Chemical Computing Group and the Schrödinger suite. The company does offer enterprise editions of their software under commercial licences, but the standalone version is free.

PyMOL

Resource: Interactive Protein Structure Viewer

Website: https://www.pymol.org/

Overview: PyMOL is a user-sponsored molecular visualisation system on an open-source foundation. PyMOL offers protein structure visualisation that is interactive and has extensive functionality for adapting the visualisations and annotation of protein structures. PyMOL is excellent for high-resolution, publication-quality images.

SwissBioisostere

Resource: Online Bioisosteric Replacement Tool

Website: http://www.swissbioisostere.ch/

Overview: SwissBioisostere is a web-based bioisosteric replacement suggestion tool using the Matched Molecular Pair Analysis (MMPA) concept, and it mines its data from the ChEMBL database. The tool has an easy-to-use interface and suggests replacements with associated target class annotations based on the potential change in activity. The system is very simple to use and ideal as an ideas generator for medicinal chemists.

OpenBabel

Resource: The Open Source Chemistry Toolbox

Website: http://openbabel.org/

Overview: Open Babel is a chemical toolbox designed to primarily convert between the multitudes of chemical structure file formats that have been designed for different software packages; currently, this stands at 100 file formats. In addition, OpenBabel has a wide range of ready-to-use software programs to convert between file formats, generate conformers, generate fingerprints, calculate molecular descriptors, and many more.

canSAR

Resource: Integrated multidisciplinary knowledge-base

Website: https://cansar.icr.ac.uk/

Overview: From The Institute of Cancer Research, London, canSAR is a vast knowledge base of linked data coming from biology, chemistry, pharmacology, structural biology, cellular networks and clinical annotations, and applies machine learning approaches to provide drug-discovery useful predictions.

DrugBank

Resource: Drug database with associated metadata

Website: http://www.drugbank.ca/

Overview: The DrugBank database is a unique bioinformatics and chemoinformatics resource that combines detailed drug (*i.e.* chemical, pharmacological and pharmaceutical) data with comprehensive drug target (*i.e.* sequence, structure and pathway) information. The database contains 7759 drug entries including 1600 FDA-approved small molecule drugs, 160 FDA-approved biotech (protein/peptide) drugs, 89 nutraceuticals and over 6000 experimental drugs. Additionally, 4282 non-redundant protein (*i.e.* drug target/enzyme/transporter/carrier) sequences are linked to these drug entries. Each DrugCard entry contains more than 200 data fields with half of the information being devoted to drug/chemical data and the other half devoted to drug target or protein data.

ScaffoldHunter

Resource: Visual Exploration of Scaffold Trees

Website: http://scaffoldhunter.sourceforge.net/

Overview: Scaffold Hunter is a Java-based open source tool for the visual analysis of data sets with a focus on data from the life sciences, aiming at intuitive access to large and complex data sets. The tool offers a variety of views, *e.g.* graph, dendrogram and plot view, as well as analysis methods, *e.g.* for clustering and classification. Scaffold Hunter has its origin in drug discovery, which is still one of the main application areas, and has evolved into a reusable open source platform for a wider range of applications. The tool offers flexible plugin and data integration mechanisms to allow adaption to new fields and data sets, *e.g.* from medical image retrieval.

CheS-Mapper

Resource: Chemical Space Mapping and Visualization in 3D

Website: http://ches-mapper.org/

Overview: CheS-Mapper (Chemical Space Mapper) is a 3D-viewer for chemical datasets with small compounds. The tool can be used to analyse the relationship between the structure of chemical compounds, their physicochemical properties, and their biological or toxic effects. CheS-Mapper embeds a dataset into 3D space, such that compounds that have similar feature values are close to each other. It can compute a range of descriptors and supports clustering and 3D alignment.

Subject Index