Data Driven Approaches to Improve the Drug Discovery Process

a virtual screening quest in drug discovery

JEAN-PAUL EBEJER

St. Anne's College

Department of Statistics

University of Oxford

Hilary 2014

A thesis submitted for the degree of *Doctor of Philosophy*.

Statement of Originality

This is my own work unless where otherwise indicated.

Candidate Jean-Paul Ebejer

Signed _____

Date April 16, 2014

To The Ebejers

To my parents Joseph and Josephine, for their lifelong sacrifices.

To my sister Fleur-Marie, for her support.

To my love Oriana, for hers.

Acknowledgements

I left this page for last, so technically I am one page away from finishing my D.Phil. thesis. The sense of achievement is indescribable, and I almost look forward to the paper-jam printer fights I am going to have in a few minutes. There are a ton of people who have made this work possible and deserve a mention here.

Thanks to my two formidable supervisors; Charlotte and Garrett. You have gently pushed me past my limits, and in doing so I discovered I am more capable than I previously thought. Thanks to Paul Finn, my scientific beacon. I will treasure your anecdotes and quotes ("There are two answers to your question, the first one is *Yes* and the second one is *No*"). Working with you three has been both an honour and a pleasure, in equal measures.

Thanks to Dr. Greg Landrum, for his RDKit toolkit and his out-of-hours support on the mailing lists. Cheminformatics is hard, but RDKit goes a long way to make it easier. His open, responsive and friendly manner are a valuable addition to this already excellent software.

Of course, thanks to past and present OPIG members – *you learn as much from your teachers as you do from your colleagues.* The venerable: Seb, Mir, Anna Vangone, Yoonjoo. My peers: Leila, Jamie, Hannah, James, KK, Henry, Markus. The young padawans: Jin, Claire, Rey, Ali, Anthony, Luis, Malte, Saulo. It is important to keep up the beers (and the pool championships) at the University club after Wednesday group meetings. Eoin please take note and, in my absence, do not let them go astray. Thank-you for this too, Eoin.

In 2007, I was walking with two friends in the tropical island breeze of San Pedro and I told them I wanted to stop coding for the financial industry and do a PhD in an interesting and useful area. I took my time, but I almost did it now. Thanks Ruth and Adrian for being those two friends, and for relentlessly believing in me. Also, thanks to the Ixaris folk, especially to Dr. Mifsud ("mingħajr il-passjoni nsiru *robots*" and "it is a discussion, you write a document and I write another") and Patrick – for my financial industry days, an experience without which I would be incomplete. And also thanks to all my other Maltese supporters: iċ-Ċencu, il-Kros, Gordon, Sergio, Johan, Lukis *et al.*

Thanks to the M.P.G. for those few, but much needed, late nights (or early mornings): David, Michael, Daniel, Tommy, Alex, Shauny, Joyti, and Monsieur Soyer. Special mention to my Oxford family, Daniel and Tommy with whom I lived these past three and a half years. I will miss you and our Sunday meatballs.

Thanks to my parents for all their love, support and sacrifices; where would I be without you? Thanks to my sister, for the fun loving person you are. And last, but definitely not least, to my girlfriend Oriana – three and a half years far away from each other is a long time, but we did it. I look forward to be with you everyday now.

I probably left out a slew of individuals who deserve to be here. Please do not feel wronged, forgive me and *THANK YOU*! A last note to the readers of this thesis. I hope you enjoy reading this as much as I enjoyed this incredible experience in Oxford.

Grazzi ta' kollox ħuti.

Abstract

Drug discovery has witnessed an increase in the application of *in silico* methods to complement existing *in vitro* and *in vivo* experiments, in an attempt to "fail fast" and reduce the high attrition rates of clinical phases. Computer algorithms have been successfully employed for many tasks including biological target selection, hit identification, lead optimization, binding affinity determination, ADME and toxicity prediction, side-effect prediction, drug repurposing, and, in general, to direct experimental work.

This thesis describes a multifaceted approach to virtual screening, to computationally identify small-molecule inhibitors against a biological target of interest. Conformer generation is a critical step in all virtual screening methods that make use of atomic 3D data. We therefore analysed the ability of computational tools to reproduce high quality, experimentally resolved conformations of organic small-molecules. We selected the best performing method (RDKit), and developed a protocol that generates a non-redundant conformer ensemble which tends to contain low-energy structures close to those experimentally observed.

We then outline the steps we took to build a multi-million, small-molecule database (including molecule standardization and efficient exact, substructure and similarity searching capabilities), for use in our virtual screening experiments. We generated conformers and descriptors for the molecules in the database. We tagged a subset of the database as 'drug-like' and clustered this to provide a reduced, diverse set of molecules for use in more computationally-intensive virtual screening protocols.

We next describe a novel virtual screening method we developed, called Ligity, that makes use of known protein-ligand *holo* structures as queries to search the small-molecule database for putative actives. Ligity has been validated against targets from the DUD-E dataset, and has shown, on average, better performance than other 3D methods. We also show that performance improved when we fused the results from multiple input structures. This bodes well for Ligity's future use, especially when considering that protein structure databases such as the Protein Data Bank are growing exponentially every year.

Lastly, we describe the fruitful application of structure-based and ligand-based virtual screening methods to *Plasmodium falciparum* Subtilisin-like Protease 1 (PfSUB1), an important drug target in the human stages of the life-cycle of the malaria parasite. Our ligand-based virtual screening study resulted in the discovery of novel PfSUB1 inhibitors. Further lead optimization of these compounds, to improve binding affinity in the nanomolar range, may promote them as drug candidates.

In this thesis we postulate that the accuracy of computational tools in drug discovery may be enhanced to take advantage of the exponential increase of experimental data and the availability of cheaper computational power such as cloud computing.

Author's Publications

The following research articles have been published in refereed journals by the author of this thesis during the course of his D.Phil. The relation of these articles to the contents of this document is highlighted at the start of every chapter (where relevant).

Ebejer J.P., Morris G.M., Deane C.M., Freely available conformer generation methods: how good are they?, *J. Chem. Inf. Model.* **2012** May 25; 52(5): 1146-58.

Withers-Martinez C., Suarez C., Fulle S., Kher S., Penzo M., Ebejer J.P., Koussis K., Hackett F., Jirgensons A., Finn P.W., Blackman M.J., Plasmodium subtilisin-like protease 1 (SUB1): insights into the active-site structure, specificity and function of a pan-malaria drug target, *Int. J. Parasitol.* **2012** May 15; 42(6): 597-612.

Ebejer J.P., Fulle S., Morris G.M., Finn P.W., The Emerging Role of Cloud Computing in Molecular Modelling, *J. Mol. Graph. Model*, **2013**; 44: 177-187.

Ebejer J.P., Hill J.R., Kelm S., Shi J., Deane C.M., Memoir: template-based structure prediction for membrane proteins, *Nucleic Acids Res.* **2013** Jul 1; 41 (Web Server issue): W379-83.

Kelm S., Vangone A., Choi Y., Ebejer J.P., Shi J., Deane C.M. Fragment-based modelling of membrane protein loops - successes, failures and prospects for the future, *Proteins.* **2013**.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

Productivity in the pharmaceutical industry, particularly in discovering new molecular entities, has decreased over the past two decades [Bunnage, 2011; Pammolli and Magazzini, 2011]. Some have attributed this to the increased costs for the development of new drugs and the rise in attrition rates, especially in late phase clinical trials [David et al., 2009]. Others have pointed to more stringent regulations governing the approval of new medicines, and change in the structure, management and priorities of the industry [Bennani, 2012; Payne et al., 2007]. It has also been suggested that the 'low-hanging fruit' has been picked and we must now venture into less explored target and drug space [Fishman and Porter, 2005; Flemming, 2013; Kraus, 2008; Osmond et al., 2010; Williams, 2011].

Recently, this journey has been supported by the exponential growth of protein structure databases and the increase in availability of small-molecule bioactivity data. Using this unprecedented volume of data promises to allow us to provide more effective, knowledge-based Computer-Aided Drug Design (CADD) methods. The development of these computational models can help us contain the spiralling pharmaceutical costs described above by moving *in vitro* and *in vivo* studies, like evaluating a drug candidate's toxicity, to *in silico* studies which can be performed at a much earlier stage in the drug discovery process. Hopefully, as more experimental data becomes available, and we

gain a better understanding of molecular binding determinants, we will be able to refine CADD methods to enhance their performance in the quest for new drugs. However, the massive amount of data also presents challenges such as efficiently searching large scale small-molecule libraries.

In this thesis we define computational methods and outline challenges of representing, standardizing, storing, and searching millions of small-molecules in a database. We describe a novel virtual screening method we developed, called Ligity, which makes use of the growing repository of publicly-available, cognate protein-ligand structure data to find small-molecule inhibitors against a biological target of interest. Finally, we describe how we used the small-molecule database we developed in a prospective virtual screening exercise on a pan-malarial drug target, which helped discover new inhibitors.

1.1.1 Why is Drug-Discovery Hard?

Drug discovery is a multi-objective optimization problem. In order for a drug to work effectively and be approved by a regulatory body, it has to satisfy a number of criteria. The simultaneous optimization of all these properties is one of the main reasons why drug discovery a complex and difficult problem. Sometimes improving one property has a negative effect on another (*e.g.* adding a chemical moiety to a molecule to make it more potent, may also make it toxic). Table 1.1 describes the most important and desirable drug properties. Some of these properties are highly related to one another (*e.g.* if the body is unable to excrete a drug, its resulting high concentrations may lead to toxicity). CADD may help to fulfil the pharma mantra "fail fast, fail early" by using computational models to predict some of these properties for a given molecule (*e.g.* by not going forward with molecules predicted to bind with multiple targets, thereby causing side-effects in the patient). This thesis describes a data-driven approach to search for small-molecule inhibitors against a specific target of interest. This is just one of the many steps in the complex drug-discovery process.

Table 1.1: Aspects of an ideal drug. Suboptimal properties may prevent a compound from being approved as a drug.

Category	Property	Description
Pharmacokinetics	Liberation	Sometimes active ingredients in a drug are formulated with inactive ingredients (known as excipients) which may provide therapeutic improvements. The active ingredient of the drug should be able to 'liberate' itself from these compounds (*e.g.* from the protective coating).
	Absorption	Absorption is the process of bringing a drug from the site of administration into the circulatory or lymphatic system. The drug should be absorbed well by the body, not being lost before it gets to the circulatory or lymphatic system.
	Distribution	The transport of a drug to its site of action (*e.g.* ability to cross blood-brain barrier). A drug should arrive to the site of action efficiently (*e.g.* fat tissue may act as a storage site for lipid-soluble drugs).
	Metabolism	A drug may be broken down by the body into metabolites, rendering it ineffective. A drug should not be deactivated by the body (and, thus, have its potency reduced).

Category	Property	Description
	Excretion	Elimination of a drug and the metabolite(s) from the body. A drug should not accumulate in the body as this may cause adverse effects.
	Potency	The amount of a drug required to have a therapeutic effect. A drug should have a therapeutic effect with a small dose.
Pharmacodynamics	Specificity	A drug may interact with other unintended biological receptors causing side effects. A drug should affect only the intended target and have no side effects.
	Safety and Toxicity	A drug may act as a poison killing an organism, and induce or increase the frequency of mutation in an organism (mutagenicity). Toxicity depends on a number of different factors such as amount of the drug administered and/or the general condition of health of the patient. A drug should be safe to use.
	Efficacy	The ability of a drug to produce the desired therapeutic effect. A drug should be efficacious and have a beneficial therapeutic effect.

Category	Property	Description
Manufacturing	Large-scale synthesis	Chemical manufacture and mass production of a drug. It should be possible to produce a drug in large quantities and in a cost-effective manner (within reasonable timescales and resources).
Commercialization	Patent and Intellectual Property	A drug may be protected by intellectual property laws (to safeguard the research investment of the company owning the drug patent). The structure and composition of a drug should not infringe any existing patents. The structure of a drug should be patentable.
	Price	The selling price of a drug. A drug should have a reasonable selling price for both the company producing it and the patients who need it.
	Resistance	A pathogen may develop resistance to a drug. A drug should be effective for a long period, in the face of any emergent coevolutionary pathogen responses to the use of the drug.
	Shelf-life and Stability	A drug has to be administered by an 'expiry date'. A drug should have a long shelf-life and should not degrade quickly.
	Storage conditions	The requirements for storing a drug. Ideally, it should be possible to store a drug at room temperature.

Category	Property	Description
	Dosing regimen	The amount and periodicity of drug doses required for optimal therapeutic action. A drug should require a minimum number of doses to be administered.
	Route of administration	Defines how a drug is administered to the patient. For example, oral administration is more desirable than intravenous administration. A drug should be administered to the patient in the most convenient and comfortable way.

1.2 The Drug Discovery Process

In this section we briefly describe the drug discovery process and the role that computation plays in it. We give an overview of some of the main application areas of CADD. Finally, we describe our contributions to this area.

The drug discovery and development process is shown in Figure 1.1. The first step of this pipeline is disease-related genomics, which involves studying the effects of the disease on gene expression, the proteins encoded by affected genes, the synthesis of those specific proteins, and their interactions. This step also aims to uncover how changes that are caused by the disease at the molecular level affect larger-scale cells, tissues and organs.

The second step is target identification and validation. Researchers select a single biological entity (typically proteins, genes or RNA), defined as the target, which is believed to play a critical part in the disease. An example of this is an enzyme that is required for the growth of the infecting pathogen. The target will be modulated by the putative 'drug', hopefully eliciting a positive therapeutic response. Target validation is the process

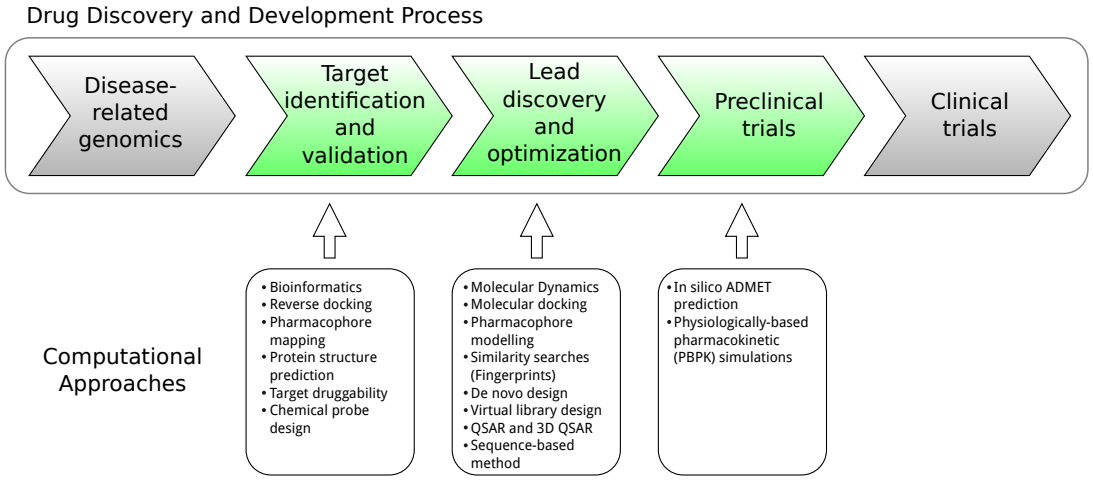


Figure 1.1: Computational approaches in the drug discovery and development process. We are interested in the 'Lead discovery and optimization' stage. This figure is adapted from Tang et al. [2006] and Ou-Yang et al. [2012].

by which the identified target is confirmed experimentally to be crucial to the disease progression (*e.g.* by knocking out a particular gene from the genome of a pathogen, its ability to infect new cells is compromised). A bioassay is developed to be able to measure the biological activity of the target.

After the target is identified and validated, the search for a drug molecule (the 'lead compound') may begin. There are different types of drugs: natural products, steroids, antibiotics, biologicals (*e.g.* antibodies, proteins), peptides, and small-molecules. The latter group is the focus of this thesis. One of the most common methods used to find potential leads is High Throughput Screening (HTS). Using robotics and a biological assay, an HTS experiment is able to screen an entire compound library (hundreds of thousands of compounds) against a target and find molecules, called 'hits', that have a biological effect. Many analogues of these hits are then synthesized and tested in order to explore the active chemical space of the target. The most promising hits are then optimized (for safety, potency *etc.*) and one or more lead compounds are selected from them.

The main aim of preclinical tests, is to determine whether one or more optimized lead compounds are safe for testing in humans. *In vitro* and *in vivo* tests (in cell cultures and animal models) determine the safety profile of the compound(s), and help to shed

some light on how the drug molecules work. The output of this stage is a handful of 'candidate drugs'.

Clinical trials test the candidate drugs in human subjects. There are three phases of clinical trials. In the first phase, candidate drugs are tested on a small group (20-100) of healthy humans to study the pharmacodynamics and pharmacodynamics properties of the compound. In the second phase, the candidate drugs are tested on a small number of patients (100-500). Here researchers test the efficacy of the drug at improving the patients' condition. They also study the most effective dosing regimen for the drug. In the third phase the candidate drugs are tested on a larger group of patients (1000-5000), to generate statistically significant data regarding safety and efficacy.

This whole drug discovery and development process is very costly (close to a billion dollars) and time consuming (typically more than a decade). Computational tools may help lead compounds fail earlier in the pipeline, reducing costs.

1.2.1 The Growing Role of Computation in Drug Discovery

> The "*role of computational models is to increase prediction based on existing knowledge.*" Kapetanovic [2008]

In this section we describe how computational tools contribute to the target identification and validation, lead discovery and optimization, and preclinical trials stages of the drug discovery and development pipeline. CADD is used throughout the drug discovery and development pipeline (some of these application areas are shown in Figure 1.1).

Target Identification and Validation

In silico cheminformatics and bioinformatics tools offer cheap and fast complementary (and sometimes alternative) approaches to the experimental identification and validation of drug targets [Crowther et al., 2010]. These approaches include (amongst others):

- Gene-expression analysis using expressed sequence tags to compare gene expression levels in normal and disease states

- Prediction of gene or protein function using homologous sequence searches to genes or proteins which are already annotated. Function is important in target identification as it allows to determine the role of the target in biochemical or pathophysiological pathways and, therefore, its potential relevance in a disease [Terstappen and Reggiani, 2001]

- Systems biology approaches that allow us to study target associations and responses in networks which vary in scale and complexity (*e.g.* molecular pathways, regulatory networks, cell behaviour,tissues, organs *etc.*) [Butcher et al., 2004; Chan et al., 2010; Cho et al., 2006]

- Use of chemogenomics to identify (possibly multiple) drug targets by mining data describing the modulation of genomic responses using chemical compounds [Bender et al., 2007; Bredel and Jacoby, 2004; Hughes et al., 2011; Wang et al., 2013]

For a detailed review of how CADD may be used for target identification please refer to the work by Koutsoukas et al. [2011]. Since target identification is a key first step to the drug discovery process [Yang et al., 2009], many different methods are used for target validation – including *in silico* ones.

Lead Discovery and Optimization

One of the most prominent areas where CADD has been applied is lead discovery. In virtual screening (VS), libraries of small-molecules are searched to identify inhibitors against some biological target [Melagraki and Afantitis, 2011; Villoutreix et al., 2009]. Virtual Screening is an important theme in this thesis, and is described in detail in Section 1.3.

Another CADD technique used in lead discovery is *de novo* design, where small-molecule inhibitors with novel molecular structures are assembled 'from scratch' either in an atom-by-atom fashion or by using larger building blocks (*e.g.* a benzene ring) [Ou-Yang et al., 2012]. This is in contrast to virtual screening, where a library of 'whole' molecules is searched for actives. There are three main decisions to be made in a *de novo* design experiment [Hartenfeller and Schneider, 2011]. First, the strategy of how to assemble the

compound must be selected (either atom-based or fragment-based). Second, a scoring function is required which evaluates the molecule in its current state. Both structure-based (requires 3D receptor structure) and ligand-based (requires some reference ligand, also known as 'templates') scoring functions have been developed. Lastly, an algorithm which systematically visits the search space for the next molecular modification is needed.

One approach for *de novo* design is 'growing'. Typically, the first step is to anchor a fragment in the pocket and explore the rest of the pocket by adding more fragments which optimize binding interactions (such as electrostatic or van der Waals) [Congreve et al., 2005]. Another approach is 'linking' where multiple fragments docked in distinct parts of the protein pocket are then linked together using a linker or scaffold molecular fragment. These two approaches are analogous to the ones used in Fragment Based Drug Design [Blundell et al., 2002; Murray and Rees, 2009]. Indeed, Loving et al. [2010] argue that there is substantial overlap between the two areas. A common criticism of *de novo* methods is that they do not always produce compounds which are amenable to chemical synthesis [Hartenfeller and Schneider, 2011].

Preclinical Trials

CADD has been used to predict the absorption, distribution, metabolism, excretion and toxicity (ADME-Tox) of molecules with varying degrees of success [Hou and Wang, 2008; Tetko et al., 2006]. Computational models, usually based on experimental data, have been built to quantify drug-likeness [Jorgensen, 2004].

Physiologically based pharmacokinetic (PBPK) modelling or simulation is a mathematical technique that is also used to predict absorption, distribution, metabolism, and excretion. The difference with classical ADME-Tox predicition methods is that PBPK models do not look only at the chemical composition of the molecule but also have parameters for the broader anatomical and physiological processes. They also include specific compartments (organs and tissues) involved in exposure, toxicity, biotransformation and clearance processes connected by blood flow [Reddy et al., 2013]. PBPK models are also useful for cross-species extrapolation, where physiological and biochemical parameters from the animal model may be swapped for human values [Clewell III et al., 2002].

In the next section we describe virtual screening, how computer algorithms are employed to search small-molecule libraries for inhibitors against biological targets of interest. After this we introduce a malarial drug target, PfSUB1, which is the focus of our virtual screening effort.

1.3 Virtual Screening

The practice of virtual screening pre-dates the first use of this term. Still, the earliest appearance of the term 'virtual screening' in the literature is in the 1997 work by Dragos Horvath where a database of 2,500 molecules was screened to find micromolar inhibitors of the enzyme trypanothione reductase in an early grid-based, rigid-body docking experiment [Horvath, 1997]. Virtual screening may be defined as the use of computational models to search for bioactive molecules in a library of compounds.

Molecules in a virtual library may be represented in a large number of ways, from compact two dimensional (2D) Simplified Molecular Input Line Entry Specification (SMILES) [Weininger, 1988] representation to full atomic three dimensional (3D) coordinates. Depending on the virtual screening method employed other descriptors may be used. Molecular descriptors may be scalar quantities in a single dimension (1D), *e.g.* molecular weight or polar surface area. A single 1D descriptor on its own is typically insufficient to compare molecules adequately, so multiple 1D descriptors are used. 2D descriptors describe the atoms and connectivity of a molecular graph, *e.g.* topological indices. 3D descriptors are calculated based on the molecule's 3D shape *e.g.* using inter-atomic distances or molecular surfaces [Willett, 2011]. Higher order descriptor dimensions exist, *e.g.* in ElectroShape searches a molecule is represented as a vector of 15 real numbers representing the statistical moments of distributions of intramolecular distances between all atoms and carefully chosen reference points where the coordinates of the atoms and reference points are in 4D (x,y,z,charge) [Armstrong et al., 2010].

The aim of a virtual screening experiment is to distinguish between actives and inactives (sometimes referred to as decoys). Central to the concept of virtual screening is the quantitative ranking of a list of molecules by some algorithm, and a selected

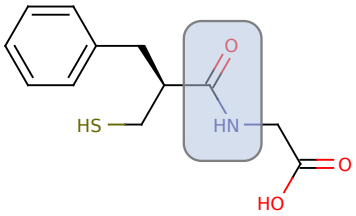
threshold above which all the molecules are deemed actives (compounds which elicit a biological response) and below which they are considered as inactive. The top fraction of the virtual screening ranking list is tested for activity in a biological assay. Hits are compounds that are classified as putative actives and are confirmed by experimental data to show some bioactivity, typically molecules with an $IC_{50} < 25$ μM. These molecules may be optimized (*e.g.* by adding or removing functional groups) in a second virtual screening study, using more detailed and time-consuming protocols. This optimisation is an iterative process, and may happen a number of times complemented by feedback from medicinal chemists who comment on synthetic tractability, toxicity, potency, and other pharmacokinetic and pharmacodynamic properties. Hopefully, activity improves into the nanomolar range (nM) and the most potent compound is selected as a 'lead', which is carried forward to more detailed chemical and biological analysis in an iterative process. The lead compound will eventually progress to a full drug development programme [Valler and Green, 2000].

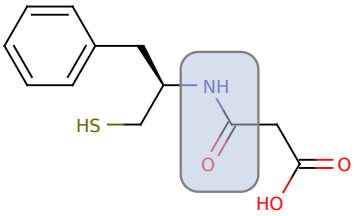
Traditionally, virtual screening is divided into two main categories: ligand-based virtual screening (LBVS) and structure-based virtual screening (SBVS).

1.3.1 Ligand-Based Virtual Screening

Ligand-based virtual screening relies on the molecular Similar Property Principle (SPP); which simply states that similar compounds have similar properties [Johnson and Maggiora, 1990]. If one compound shows activity against a particular target, similar compounds should also show activity to some degree. The open question is the definition of the word 'similar'. More specifically, some studies define similarity as a Tanimoto coefficient (described in Section 1.3.1) which is greater than 0.85 [Matter, 1997], but this is a rather arbitrary definition as it is dependent on the fingerprint algorithm employed, the similarity metric used and the selected similarity threshold. Indeed, some authors argue that similarity has a *context* which defines and limits its use [Bender and Glen, 2004]. It is worth noting that there are exceptions to the widely accepted SPP. Kubinyi [1998] presents a number of examples. Others argue that biological similarity is not as strong as previously assumed. A study by Martin et al. [2002] found that only 30% of molecules

Table 1.2: An exception to the 'similar property principle'. Structurally very similar ligands sometimes show different biological activity depending on the target. In this example two very similar ligands, thiorphan and retro-thiorphan, show similar bioactivities for receptors neutral endopeptidase (NEP) 24.11 and thermolysin but very different inhibition for angiotensin-converting enzyme (ACE).





Receptor	**Thiorphan** (K_i in μM)	**Retro-thiorphan** (K_i in μM)
Thermolysin	$= 1.8$	$= 2.3$
NEP 24.11	$= 0.0019$	$= 0.0023$
ACE	$= 0.14$	> 10

that were similar (defined as Tanimoto similarity of ≥ 0.85 using Daylight fingerprints) to active molecules were active themselves. For a detailed review of shape-based similarity methods the reader is directed to the work by Finn and Morris [2013].

A reported example, based on the work of Roques et al. [1993], shows that two very similar ligands thiorphan and retro-thiorphan which differ only in their amide orientation have similar biological activities in the case of neutral endopeptidase (NEP) 24.11 and thermolysin but have a difference of two orders of magnitude between their inhibition constant for angiotensin-converting enzyme (ACE) [Balkenhohl et al., 1996; Sotriffer et al., 2011]. In this case (Table 1.2) the concept of 'similarity' of biological activity is also dependent on the biological target, and structural similarity of the ligand alone is not enough.

Different types of Ligand-Based Virtual Screening

Some of the most popular methods in LBVS are based on fingerprints, pharmacophores, Quantitative Structure-Activity Relationship models (QSAR) and Ultrafast Shape Recognition (USR). These four methods are explained in detail in the following sections.

Fingerprints. In general, a molecule may be represented as a binary vector (fingerprint) where every element indicates the presence, or absence, of a chemical moiety. Alternatively, non-binary fingerprints record counts of a chemical feature, *e.g.* how many times a carboxyl group is present in the molecule. These fingerprints are then compared to a query (active) molecule fingerprint using a similarity metric (see Section 1.3.1 for more details).

There are four types of 2D fingerprint methods: (i) dictionary-based (ii) topological or path-based fingerprints (iii) circular fingerprints; and (iv) pharmacophores [Riniker and Landrum, 2013].

In dictionary-based methods bits in an array are switched on (and off) to record the presence (or absence) of a predefined list of chemical functional groups. Typically the length of the binary key is the same as the size of the dictionary. Statistical analysis of the most discriminating parts of the compound library is used to build the dictionary. An example of a dictionary-based fingerprint are MACCS keys (subsequently optimized and renamed to MDL keys [Durant et al., 2002]). MACCS keys are 166-bit fingerprints made of 166 structural SMARTS patterns. Each bit indicates the presence or absence of a chemical moiety. Examples of these patterns are: *is there an S-S bond?* or *is there a ring of size 4?*

The first step of a topological or path-based fingerprint is to encode each atom (*e.g.* by a 3-tuple containing the atom's element, number of connected heteroatoms, number of pi electrons). Then, as a second step, distance and connectivity information between the atoms is used to switch bits in the fingerprints on. Two topological (or path-based) algorithms are atom pair [Carhart et al., 1985] and topological torsion [Nilakantan et al., 1987]. In atom pair fingerprints all pairwise atoms of the molecule together with the topological distance (calculated in number of bonds separating the two atoms) are used

as an index to identify which bit to set on. In topological torsions four consecutive bonded non-hydrogen atoms identify which position of the fingerprint to set on.

Circular fingerprints record circular topological information of each atom's neighbourhood [Rogers and Hahn, 2010]. Two types of circular fingerprints are the extended connectivity fingerprint (ECFP) and its variant the functional connectivity fingerprint (FCFP). The main difference between the two is that FCFP is meant to capture more abstract functional role-based groups (*e.g.* acceptor group), rather than the precise atom environment. Initially, in ECFP, each non-hydrogen atom has an integer identifier assigned to it. This integer is generated by hashing six properties: number of immediate non-hydrogen atoms, valence (minus the number of hydrogens), atomic mass, atomic charge, number of attached hydrogens (both implicit and explicit) and whether the atom is contained in at least one ring. After the identifier generation, each atom identifier is updated iteratively. Each iteration uses the previous atom identifiers as input and captures a larger circular neighbour until a specified radius is reached. Atom identifiers of each iteration are captured and hashed into a single value which is added to the fingerprint set and used in the next iteration. This iteration process is based on the Morgan algorithm [Morgan, 1965]. The final step is the removal of multiple identifiers representing equivalent neighbourhoods (*e.g.* a molecule `X-C(=O)N` will have different identifiers if they are centred on the oxygen or nitrogen but these are encoding the same neighbourhood after two or more iterations). These removals are necessary to avoid populating redundant bits in the fingerprint. ECFP (with radius 2) is reported to have better performance than other 2D and 3D shape similarity methods used for virtual screening [Hu et al., 2012]. It is also reported to exhibit the best enrichment when using 2D fingerprints for scaffold hopping [Gardiner et al., 2011].

2D pharmacophore fingerprints record the inter-feature (*e.g.* acceptor) topological distances in a binary or count fingerprint. Distance bins and pharmacophore combinations are used as a key to the fingerprint.

Pharmacophore modelling. Gund's widespread definition of the term *pharmacophore* is "a set of structural features in a molecule that is recognized at a receptor site and is

responsible for that molecule's biological activity" [Gund, 1977]. Another often quoted definition is the official IUPAC (1998) one: "A pharmacophore is the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response" [Wermuth et al., 1998]. The latter definition emphasises that a pharmacophore model does not represent a real molecule, but rather a set of features common across a number of active molecules which determine molecular binding. This feature-set is referred to as 'the largest common denominator' between active molecules.

In essence, pharmacophoric modelling entails finding the hot-spots of query molecules, typically a known small molecule binder, and then using this set of points to match similar molecules in a database. These hot-spots or features include hydrogen bond donors, hydrogen bond acceptors, positive and negative features (charges), hydrophobic features, and customized features (*e.g.* aromatic groups, halogens, and other user-defined feature types) [Leach et al., 2010].

Pharmacophoric elucidation usually involves the alignment of the set of active molecules to find the common features. This may be achieved by aligning a list of predefined anchor-points (features) using least-squares fitting to superimpose them. This is known as point-based alignment. Another method for alignment, known as property-based, is to use molecular fields descriptors for the molecules [Wolber et al., 2008]. Rather then defining the properties of the atoms of the molecule under study these fields define what the receptor 'sees' in terms of charge distribution and shape (on the outside of the molecule). These fields may be represented as a grid of points, or in a more compact way as a set of Gaussian functions [Good et al., 1992]. The alignment is then based on the optimization (maximization) of the intermolecular overlap (similarity) of these Gaussians.

Pharmacophore keys (or fingerprints) have also been reported in literature. Pharmacophoric keys are usually created by enumerating all three or four point pharmacophore combinations of a molecule. Each of these combinations indexes a particular location in a key (binary fingerprint). Pharmacophoric keys have been used as filters to more rigorous protocols by intersecting keys between the query and database molecules [Seidel et al., 2010]. The idea of 'pharmacophoric keys' has been extended to the receptor. Fingerprints

for Ligands and Proteins (FLAP) is a method which generates a four-point combination pharmacophoric fingerprint of the active site based on GRID maps [Baroni et al., 2007]. These are then compared to the pharmacophoric models of ligands. FLAP has been validated for use in virtual screening on the DUD dataset [Cross et al., 2010].

Other pharmacophoric searching methods exist, some of which include information from the protein side. LigandScout is an example of an application which uses protein-ligand complex information to build a single, common pharmacophoric model across different PDB entries using clique detection and 3D alignment [Wolber and Langer, 2005].

For a more detailed review of pharmacophore models, the reader is directed to the work by Leach et al. [2010]. Common criticisms of pharmacophoric models are that they fail to address ligand flexibility, or that choosing anchor points for alignment may be difficult because of dissimilar ligands, or that selection of the proper active molecules to build the pharmacophoric model may be tricky (a different selection of actives may result in a different pharmacophoric model thereby affecting performance) [Yang, 2010].

QSAR. QSAR uses statistical models to predict biological activity (*e.g.* binding or toxicity) from the chemical properties of the molecules [Ekins et al., 2007]. The models are trained on the active molecules' descriptors and associated with a biological phenomenon. Examples of scalar descriptors are lipophilicity or molecular weight. Although successful QSAR models have been published in the literature, there are a number of associated problems such as incorrect data (structures and activities) in the dataset, the size of the modelling set is too small, overfitting, incorrect division of training and testing sets, and use of collinear descriptors [Dearden et al., 2009; Golbraikh and Tropsha, 2002; Kubinyi, 1997; Young et al., 2008].

Ultra-fast Shape Recognition. In ultra-fast shape recognition (USR), all atomic distance distributions are calculated from four reference points in the molecule [Ballester and Richards, 2007]. These four reference points are the molecular centroid (ctd), the closest atom to the centroid (cst), the farthest atom to ctd (fct), and the farthest atom to fct (ftf). This generates four distance distributions, from which the first three statistical

moments of the distributions are calculated (the mean, the variance and the skewness of the distribution). This results in a vector of 12 real numbers representing a 3D conformer of a molecule. To compare these descriptors the inverse of the scaled and translated Manhattan distance is used which gives a similarity of 1 (identical shape) and 0 (minimum similarity). USR creates a very compact yet expressive descriptor for the shape of the molecule. Various extensions exist which increase the dimensionality of the descriptor (*e.g.* adding electrostatics or pharmacophoric features to the shape descriptor) [Armstrong et al., 2010; Schreyer and Blundell, 2012]. Another extension, called Chiral Shape Recognition (CSR), allows to distinguish enantiomers [Armstrong et al., 2009]. The main strength of USR is in its speed, as the comparison only takes the time required for the Manhattan distance calculation. USR may therefore be used to execute ligand-based similarity searches on huge molecular libraries in real time, which previously were not possible. Common criticisms of the method include that the descriptor is conformer dependent (a small change in the conformer might change the original reference point selection) and substructure searches are not possible.

Similarity Metrics

A common task in cheminformatics is the comparison of two homogeneous molecular descriptors [Gregori-Puigjané and Keiser, 2012]. For example, in LBVS a bioactive ligand may be used as a reference (or query) structure to search through a list of database molecules [Willett, 2011]. In this case, both the reference and the database molecules are represented as binary fingerprints (*i.e.* an array of 1s and 0s) which are then compared through the application of a mathematical function. This mathematical function, in this context, is called a 'similarity metric' (Figure 1.2). Extensions to these similarity metrics for count fingerprints (with positive values greater than one) also exist.

Many similarity metrics exist. In this work we make extensive use of the Tversky metric, extended for non-binary fingerprints [Swamidass and Baldi, 2007]. In Equations 1.1 to 1.5 [Willett et al., 1998], $S(A, B)$ is the calculated similarity score (using each of the five different metrics) between non-binary (integer vectors) descriptors A and B. n is the number of elements in the descriptor (*e.g.* in Figure 1.2, $n = 7$). A_i and B_i are the feature

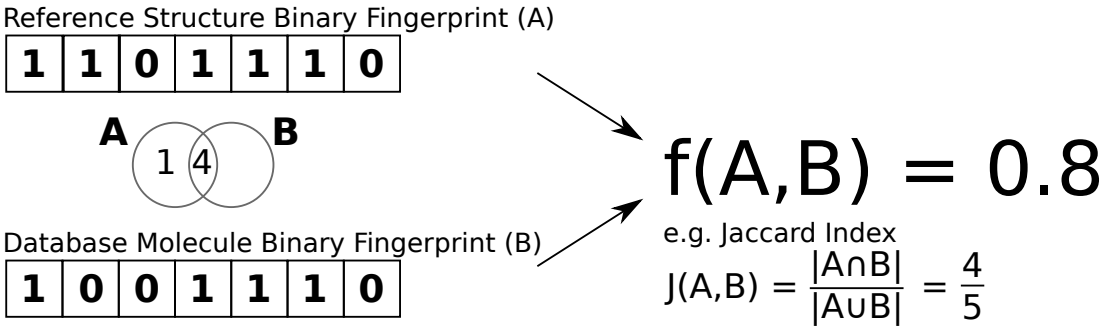


Figure 1.2: A worked example of a similarity metric using the Jaccard index, where $A \cap B = 4$ bits in common and $A \cup B = 5$ bits set by A or B.

counts for the ith element. Some similarity metrics, like Tversky, are asymmetric. The asymmetric nature of Equation 1.1 occurs if unequal values of the α and β parameters are used. For example, by setting $\alpha > \beta$ more emphasis is set on the query descriptor (A).

Tversky

$$S_{\alpha\beta}(A,B) = \frac{\sum_{i=1}^{n} \min(A_i, B_i)}{\alpha \sum_{i=1}^{n} A + \beta \sum_{i=1}^{n} B + (1-\alpha-\beta) \sum_{i=1}^{n} \min(A_i, B_i)} \tag{1.1}$$

Tanimoto

$$S(A,B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sum_{i=1}^{n} {A_i}^2 + \sum_{i=1}^{n} {B_i}^2 - \sum_{i=1}^{n} A_i B_i} \tag{1.2}$$

Dice

$$S(A,B) = \frac{2\sum_{i=1}^{n} A_i B_i}{\sum_{i=1}^{n} {A_i}^2 + \sum_{i=1}^{n} {B_i}^2} \tag{1.3}$$

Cosine

$$S(A,B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} {A_i}^2 \sum_{i=1}^{n} {B_i}^2}} \tag{1.4}$$

Common Counts

$$S(A,B) = \sum_{i=1}^{n} \min(A_i, B_i) \tag{1.5}$$

1.3.2 Structure-Based Virtual Screening

Structure-based virtual screening (SBVS) makes use of the protein (sometimes referred to as receptor) structure to guide the virtual screening exercise. The receptor structure is either determined experimentally or modelled computationally. In the latter case, success of the SBVS study is directly influenced by the quality of the theoretical model [Bordogna et al., 2011] and by the use of multiple models (instead of a single one) [Fan et al., 2009].

A small-molecule is 'docked' to the protein in a typical 'lock and key' fashion, and a score is computed based on how good the fit is. A docking protocol determines the

parameters of the docking experiment. This includes the role of water molecules in the binding, protein side-chain flexibility, ligand flexibility, which scoring function to use and various parameters specific to the docking model (*e.g.* genetic algorithm parameters such as population size, mutation and crossover rates in GOLD [Verdonk et al., 2003]). A scoring function assesses the fit (steric, electrostatic, hydrogen bonding *etc.*) of a particular ligand pose with the receptor. Some of the major critiques of docking are the inability to calculate the free binding energy correctly (possibly because of the additive nature of most scoring functions), protein main-chain flexibility, the correct prediction of water in binding and the intensive computational resources required [Cheng et al., 2012; Kinnings et al., 2011; Leach et al., 2006; Schulz-Gasch and Stahl, 2004].

A popular docking tool is DOCK, which was also the first docking tool available [Kuntz et al., 1982]. Presently, there are two currently maintained versions of DOCK. DOCK version 3.6 defines the negative image of an *a priori* defined binding site by placing spheres in the cavity. These spheres tangentially touch the water accessible surface of the protein and they effectively define the pocket where the small molecule binds. The centres of these spheres are then matched to ligand atoms, which give an orientation of the ligand in the active site based on a geometric matching algorithm (superimposition). This docking pose is scored using a physics-based shape and energy function. DOCK 3.6 is widely used and validated, with more than ten publications of novel binders from the Shoichet laboratory alone [Babaoglu et al., 2008; Brenk et al., 2006; Chen and Shoichet, 2009; Graves et al., 2008; Hermann et al., 2007; Kolb et al., 2009b; Merski and Shoichet, 2013; Powers and Shoichet, 2002; Teotico et al., 2009; Wei et al., 2002, 2004]. DOCK version 6 is richer in features. It supports ligand and receptor desolvation, ligand conformational entropy corrections, AMBER based scoring function which includes receptor flexibility, a full AMBER molecular mechanics scoring function with implicit solvent and molecular dynamics simulation capabilities [Ewing et al., 2001; Lang et al., 2009; Moustakas et al., 2006].

Use of Molecular Dynamics with SBVS

Perhaps the latest anecdote which testifies the 'coming of age' of computational chemistry, and its pivotal role in CADD, is the 2013 Chemistry Nobel Prize award to Martin Karplus, Michael Levitt and Arieh Warshel. One of the main techniques they developed, Molecular Dynamics (MD), has widely varying applications in CADD such as binding free energy estimation [Hansson and Åqvist, 1995], the identification of cryptic and allosteric binding sites [Durrant and McCammon, 2011], and refining virtual screening results [Rastelli et al., 2009].

Molecular Dynamics (MD) simulations are based on the integration of the equations of Newton's laws of motion and a force field that determines the potential energy and, hence, the forces acting on the system. The result is a trajectory that specifies how the position and velocities of the particles in the biomolecule vary with time. MD simulations provide detailed information about the full receptor and ligand flexibility as well as explicit solvent in the binding site [Hansson et al., 2002]. They can be used in conjunction with molecular docking calculations in two ways:

1. to provide different conformations of the protein structure as input of subsequent ensemble based docking calculations

2. to improve the prediction of the binding affinity by rescoring docking solutions

MD simulations can be used to generate different conformations of the receptor structure of interest. A selection of snapshots from the MD trajectories can be used in ensemble based docking, thereby implicitly taking the receptor flexibility into account [Nichols et al., 2012]. This protocol is called relaxed complex scheme (RCS) [Amaro et al., 2008; Lin et al., 2002, 2003].

Another strategy to improve docking calculations is to rescore the top hits of the docking experiment using MD [Alonso et al., 2006]. In a typical workflow, large virtual screening databases are first filtered using fast and inexpensive docking protocols. As a second step, a more accurate and computationally expensive MD simulation is applied to the top hits of the docking experiment to further refine the docking pose and rescore them

subsequently. This rescoring is based on more physically realistic techniques for binding free energy estimations such as thermodynamic integration (TI), free energy perturbation (FEP), linear interaction energy (LIE) and molecular mechanics/Poisson-Boltzmann and surface area (MM/PB-SA). Overall, this provides a more accurate prediction of the binding affinity between the protein and the ligand (compared to the scoring function in docking tools) [Durrant and McCammon, 2011; Okimoto et al., 2009]. TI and FEP are the most rigorous methods, but are very computationally expensive and therefore infeasible for large scale computational screening. LIE is moderately fast, but requires known binding affinities of the molecular system under investigation. MM/PB-SA is reportedly faster (by at least a factor of ten) than TI or FEP and, even if exceptions exist (such as p38 mitogen-activated protein (MAP) kinase complexes investigated by Pearlman [2005]), MM/PB-SA is still able to approximate the binding affinities well [El-Barghouthi et al., 2009; Yang et al., 2011].

1.3.3 Hybrid Methods – Combining Both Ligand-Based and Structure-Based Approaches

The word 'hybrid' in virtual screening is used when multiple approaches are combined together. For example, Cannon et al. [2008] combined MACCS fingerprints and Ultrafast Shape Recognition (USR) into a single descriptor to rank molecules used for doping in sporting competitions. Moro et al. [2007] argue that a "full integration of ligand- and structure-based strategies might sensitively increase the success of VS processes". Output from SBVS experiments may be used as input to LBVS, and *vice versa.* A docked pose from a docking exercise could be used as an input to 3D-QSAR or 3D pharmacophore search. Alternatively, mapping receptor pharmacophores to ligand features has improved the binding pose prediction in protein-ligand docking method Ph4Dock [Goto et al., 2004].

1.3.4 Measuring Virtual Screening Performance

The ideal VS method ranks all active molecules above all inactives. One way to measure the performance of these methods is *via* Receiver Operator Characteristic (ROC) curves [Fawcett, 2004]. ROC measures the performance of a classifier by plotting the fraction of the true positives out of the positives (known as the true positive rate, or sensitivity) versus the fraction of false positives out of the negatives (false positive rate, or 1 - specificity). A perfect classifier would have an area under the curve (AUC) of 1.0, while random performance – where actives are uniformly distributed along the ranking list – would result in an AUC of 0.5.

The problem with AUC scores from ROC curves. A problem with the use of AUC scores (originating from ROC curves) to measure performance of VS methods is that they fail to distinguish early from late enrichment. Early enrichment is an important property in VS. Typically only a small number of the molecules evaluated computationally in a virtual screening experiment are tested *in vitro* or *in vivo*. The practical reasons for this may be limited resources in chemical synthesis, compound purchasing budget, or the cost of bioassay testing. Figure 1.3 shows two ROC curves with similar AUCs (0.5), but the green ROC curve is clearly a preferable outcome for a VS study.

Another commonly used metric to measure VS performance, which addresses the AUC score shortcoming, is enrichment factor (EF). This is the ratio of the fraction of actives found in the top $x\%$ of the database ranked by the VS method over the fraction of known actives in the whole database (Equation 1.6) [Reynolds et al., 2012].

$$\text{Enrichment Factor}_{\%\text{sample of top population}} = \frac{\left(\frac{\text{actives}_{\text{sample}}}{\text{compounds}_{\text{sample}}}\right)}{\left(\frac{\text{actives}_{\text{population}}}{\text{compounds}_{\text{population}}}\right)} \tag{1.6}$$

However popular, there are many critiques of the enrichment factor metric. Arbitrary selection of the sample size (percentage) makes it hard to compare different methods [Nicholls, 2008]. Also, EF is not absolute across different targets but depends on the ratio of actives and inactives for a target [Scior et al., 2012]. A perfect method on two

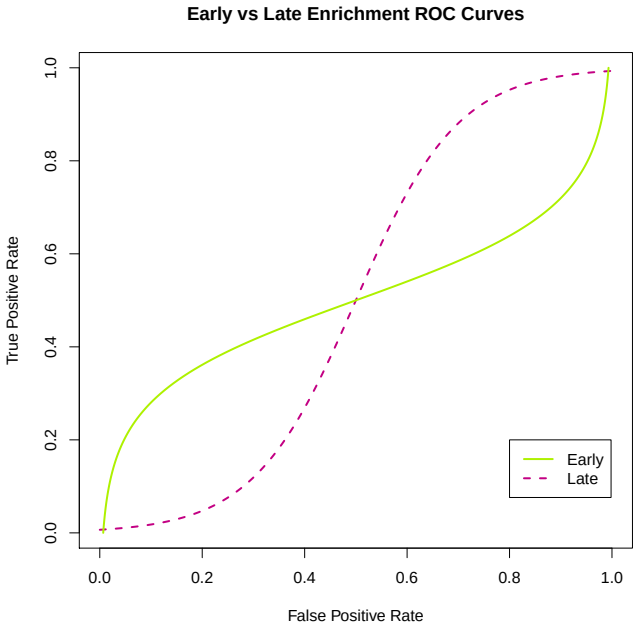


Figure 1.3: Early *versus* late VS enrichment in ROC curves. Early Enrichment (in green) is better because of practical time and money limits on the number of compounds which may be tested in a bioassay.

different targets one with 11 actives and 468 decoys gives a maximum possible EF at 1% (top 5 molecules) = $^{5}/_{5}$ / $^{11}/_{479}$ = **43.5** and another target with 234 actives and 8399 decoys gives a maximum possible EF at 1% (top 86 molecules) = $^{86}/_{86}$ / $^{234}/_{8633}$ = **36.9**.

Another measure used to analyse early enrichment, which addresses the limitations of EF, is Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic (BEDROC) [Truchon and Bayly, 2007]. BEDROC has the advantages of ROC (the limiting values [0-1] are independent of number of actives) and tackles the early recognition problem. BEDROC gives early rankings more weight than late rankings, by applying an exponential function to the ranking list. This exponential function is governed by a parameter α, which controls the degree of early recognition required. It is important to note that "there is no absolute, interpretable meaning to a BEDROC number, only a relative meaning when ranking methods" [Nicholls, 2008].

1.3.5 Data Fusion of Results in Virtual Screening

Data fusion refers to the practice of making separate VS runs, and aggregating the results in one 'fused' ranking. Data Fusion of virtual screening results (also referred to as consensus scoring) has been reported to give better enrichments in both docking and ligand virtual screening experiments [Feher, 2006; Hert et al., 2006; Wang and Wang, 2001]. There is debate on the reasons why this occurs [Willett, 2006], some of the suggestions are: (i) using multiple scoring functions is akin to repeated sampling, and results are less biased by a single run, (ii) using multiple methods allows for better clustering of actives, which results in more actives being recovered, and (iii) different methods seem to agree more on the ranking of actives than inactives [Baber et al., 2006]. There is a large number of ways how to fuse results together (Nasr et al. [2009] list 14 methods), and choosing the best performing method is not straightforward.

1.3.6 Virtual High-Throughput Screening

Virtual high-throughput screening (vHTS) rapidly screens a huge number of compounds *in silico*. The exact number, in terms of screened compounds per second, is increasing over time due to advances in computer hardware and VS software. Multi-core processors, GPUs, higher communication bandwidths and cloud/high performance computing environments all contribute to the speeding up of virtual screening methods. On the opposing side, compound databases are growing larger. A few years back virtual libraries consisted of hundreds of thousands to a few million molecules [Seifert et al., 2003]. The current release of the freely available ZINC (version 12) database contains approximately 20 million commercially available molecules [Irwin and Shoichet, 2005]. GBP-17 which is a virtual enumeration of all possible molecules (up to 17 atoms) containing only H, C, N, O and S atoms has 166.4 billion molecules [Ruddigkeit et al., 2013].

While, even by the most conservative of estimates, ZINC and other commercially available databases are a fraction of chemical space (the lower bound is thought to be around 3.4×10^9 and the upper bound is set to 10^{60} molecules [Drew et al., 2012; Reymond and Awale, 2012]), these databases are at the high-end of what can be pragmatically

screened today in a VS study.

In order to deal with this amount of data smarter algorithms are required which break down data in a hierarchical fashion, where little time is spent on a structure that is unlikely to be active, and more fine-grained analysis is made on the more plausible compounds. For example, in SBVS this could be a pre-computed volume filter, which removes any molecule that is too large to fit in a binding site [Alonso et al., 2006].

1.3.7 Evaluating the Success of Virtual Screening

Most unsuccessful VS experiments are never published, making it hard to estimate the percentage success based on a literature review alone. While there are reported successes, the false positive rate remains high [Lyne, 2002]. Even within the field's reported successes, reviews only include a dozen or so cases (*e.g.* 12 different receptors by Lyne [2002], 8 different targets by Oprea and Matter [2004], 12 different target classes by Ghosh et al. [2006], and 17 different target classes by Villoutreix et al. [2009]). Even if computational studies have identified new ligands for over 50 targets [Shoichet, 2004], it is hard to keep track of success for a multitude of reasons. First, there are varying definitions of the term 'success' (*e.g.* is an IC_{50} of 50 μM to be considered a success?). Second, the computational role in each study may vary *e.g.* some studies employ computational methods as a pre-filter to a high throughput screen, while others use solely a computational method with minimal human intervention. Third, some of these studies are proprietary and carried out in industry and may never be published because of intellectual property and/or patenting concerns. Lastly, some rigorous computational experiments are not followed up by a biological, physical experiment to confirm the computational findings. Perhaps one of the most detailed reviews of CADD success stories to date is from Kubinyi [2006].

1.4 Virtual Screening Application to a Malarial Target: PfSUB1

Malaria is an endemic disease which, in 2010, affected 219 million people and was responsible for killing approximately 660,000 people – over 2% of all deaths worldwide that year [WHO World Malaria Report, 2010]. A recent study by Murray et al. [2012] suggests that this mortality rate is greatly underestimated, and the number of deaths caused by malaria is nearly twice this figure.

Malaria is caused by parasites from several of the Plasmodium genus. *Plasmodium falciparum* is the most common parasite and causes the most severe form of the disease in humans. *Plasmodium vivax*, *Plasmodium ovale*, *Plasmodium knowlesi* and *Plasmodium malariae* cause mild forms of the disease in humans. Recent alarming reports indicate that infections with *Plasmodium vivax* and *Plasmodium knowlesi* may also be fatal [Price et al., 2009; William et al., 2011]. Malaria is widespread in tropical and sub-tropical regions, and while most deaths occur in Sub-Saharan Africa – other regions in Asia, Latin America and to a lesser extent the Middle East are also affected [Sachs and Malaney, 2002].

There are seven drug classes which are currently used for the treatment of malaria: 4-aminoquinolines, arylaminoalcohols, 8-aminoquinolines, artemisinines, antifolates, inhibitors of the respiratory chain, and antibiotics [Schlitzer, 2008]. Due to rapid mutations in the parasite's genome combined with incomplete compliance with prescribed drug regimens, resistance against monotherapies is widespread [Biagini et al., 2009; White, 2004]. Treatments where resistance has developed at a slower pace (such as quinine) cause substantial adverse side effects [Achan et al., 2011]. The current strategy adopted to cure malaria is to use multiple combinations of drugs, the most popular of which is the artemisinin-combination therapy (ACT). Resistance against multiple therapies develops more slowly than single-drug therapies [White, 1999]. However, reports have emerged in recent years of resistance to ACT [Dondorp et al., 2009; Wongsrichanalai and Meshnick, 2008]. Currently, there is no adequate alternative to this malarial treatment.

Presently, there is no clinically effective vaccine against malaria. There are two vaccines that are showing promise in clinical trials. The first, RTS,S/AS01, is designed for infants in the areas hit by the disease [Abdulla et al., 2008; Ballou, 2009; Lell et al., 2009] and has shown moderate success in Phase III trials [Agnandji et al., 2012]. The second, FMP2.1/AS02A (in phase II trials) has been found to be safe and highly immunogenic in malaria-exposed adults [Polhemus et al., 2007; Thera et al., 2008]. The development of these, if successful, together with simple preventive measures such as mosquito nets will help reduce the numbers of new cases of malaria in the affected regions. Still, the need for novel drugs to counteract the parasite's resistance to the current medications continues.

1.4.1 Malaria Life Cycle

The malaria parasite has a complex, multi-stage life cycle which takes place in two different hosts (Figure 1.4). The primary host and transmission vector for *Plasmodium* is the female mosquito of the *Anopheles* genus. The infected mosquito transmits the parasite by biting the secondary, or intermediate, vertebrate host (*e.g.* humans). The parasite, in the form of sporozoites (this is the cell form that infects new hosts), passes from the infected saliva of the mosquito into the blood stream of the human host, where it rapidly transfers to the liver and invades liver cells called hepatocytes. The sporozoites reproduce asexually within the hepatocytes to form merozoites, which rupture the liver cell and are released in the blood stream. Some parasites, *Plasmodium vivax* and *Plasmodium ovale*, may lie dormant in the liver stage in what is called a hypnozoite form. These hypnozoites will activate after a period of time (ranging from days to years) and cause clinical relapse. The re-activation mechanism is not well understood, and in *Plasmodium falciparum* and *Plasmodium malariae* the dormant form of the parasite does not occur. When the merozoites are released into the blood stream, they invade the red blood cells (erythrocytes) where they asexually replicate synchronously. The infected red blood cells rupture and a new wave of parasites is released, which go on and infect other blood cells. This repeating cycle depletes the body's oxygen supply and coincides with the fever and chills symptoms in humans [Cowman and Crabb, 2008]. Within a red blood cell, some

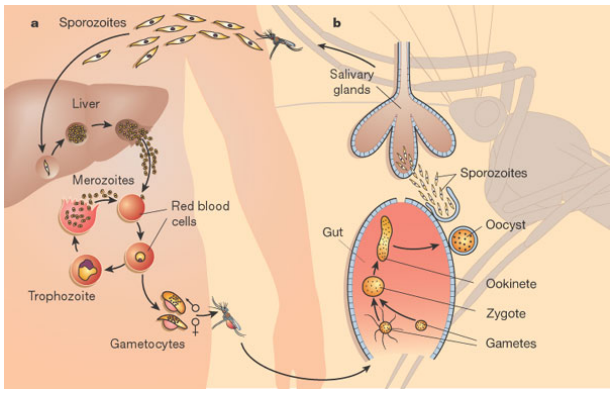


Figure 1.4: The malarial life cycle. Part of the parasite's life cycle takes place in humans (shown in [a]) while part takes place in the mosquito (shown in [b]). Reprinted by permission from Macmillan Publishers Ltd: *Nature* [Wirth, 2002], http://www.nature.com/nature, ©2002.

of the merozoites differentiate into male and female gametocytes. These gametocytes are ingested by the mosquito during feeding. The male and female gametocytes develop into mature sex cells called gametes and then fuse to form a zygote in the mosquito gut. The zygote develops into a slow moving, elongated ookinete which invades the midgut wall of the mosquito and becomes a stationary oocyst. The oocysts grow, rupture and release sporozoites which travel to the mosquito's salivary glands. The mosquito's next feeding will give rise to the whole infection cycle once again.

Throughout its lifetime the malaria parasite is relatively protected from the human immune system because for most of the human stage it lies hidden in the liver or inside red blood cells.

1.4.2 *Plasmodium falciparum* Subtilisin-like Protease 1 (PfSUB1)

Plasmodium falciparum subtilisin-like protease 1 (PfSUB1) is a serine protease that is critical in the process of *Plasmodium falciparum*'s egress from the red blood cell in humans and, hence, in maintaining the asexual erythrocytic life cycle of the parasite. Its importance is also highlighted by the fact that its sequence is highly conserved amongst the *Plasmodium* genus [Yeoh et al., 2007]. The parasite's egress is poorly understood, but it appears to happen in a quick multi-step process that is probably tightly regulated [Roiko and Carruthers, 2009].

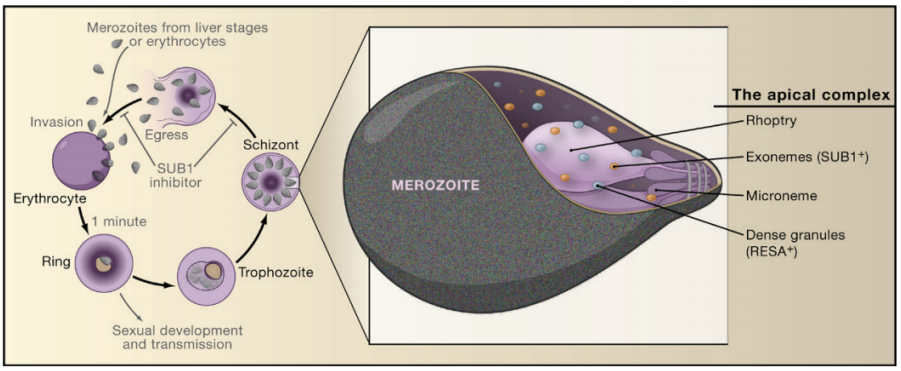


Figure 1.5: The specifics of the blood stage in the malarial life cycle. Reprinted from *Cell* [Janse and Waters, 2007], `http://www.cell.com/cell`, ©2007, with permission from Elsevier.

During the human blood stage of the disease, the parasite enters the red blood cell (RBC) as an individual merozoite which forms and develops within a parasitophorous vacuole inside the RBC. The merozoite grows further during what is known as a trophic period. The early (immature) trophozoites are often referred to as the "ring form" because of their morphology. Trophozoite enlargement is accompanied by an active metabolism including the ingestion of host cytoplasm and the proteolysis of hemoglobin into amino acids. The end of the trophic period is manifested by multiple rounds of nuclear division resulting in a schizont. PfSUB1 activity is required for the rupture of the mature schizont [Arastu-Kapur et al., 2008; Yeoh et al., 2007] and erythrocyte re-invasion. The cellular and molecular mechanism of the parasite's invasion of the human RBC has recently been reviewed by Cowman et al. [2012].

PfSUB1 is not the only protein involved in this egress from the erythrocyte. PfSUB1 is initially housed in the exoneme (Figure 1.5) and later secreted in the parasitophorous vacuole in the final stages of schizont maturation. Here it digests and activates the papain-like SERA family of proteins, which are abundantly expressed in the late stage blood form of the parasite (*i.e.* just before egress) [Blackman, 2008]. SERA proteases play a key role in the subsequent destruction of the parasitophorous vacuole and, possibly, in the destruction of the erythrocytic cell wall itself.

PfSUB1 as a Malarial Drug Target

Choosing PfSUB1 as a malarial drug target is not devoid of challenges. At the time of writing, there is no published experimental structure of PfSUB1. A homology model of PfSUB1 was published in the literature some time ago [Withers-Martinez et al., 2002]. We have complemented this by building new homology models using more recent sequences and higher resolution template structures from the PDB. These models are used in our virtual screening experiments and are described in detail in Section 5.2.1.

Another possible criticism for PfSUB1 as a malarial target is that there are 178 serine proteases encoded in the human genome [Stolze et al., 2013], making specificity an issue. For a genomic overview of serine proteases, the reader is directed to the work by Yousef et al. [2003]. This specificity issue is mitigated by the fact that there are no obvious human homologues of PfSUB1. The closest related human enzymes (*e.g.* tripeptidyl-peptidase II) have different substrate specificity and are structurally distinct [Withers-Martinez et al., 2012]. Most similar sequences (and structures) are from non-humans proteases, such as bacterial subtilisins or thermophilic enzymes (which are only present in micro-organisms). PfSUB1 exhibits some unique features (*i.e.* a highly polar S1 pocket and a hydrophobic S4 pocket) which may be exploited to find selective inhibitors [Withers-Martinez et al., 2004].

In a recent *in silico* study by Ludin et al. [2012] 40 candidate drug targets against malaria were identified. Their selected criteria to identify these candidate targets, based on the whole *Plasmodium falciparum* proteome were: (i) has conserved orthologues in all the mammalian-pathogenic *Plasmodium* genus (ii) has no other (sequence) match in *Plasmodium falciparum* (iii) does not have a match in the human proteome (iv) is expressed in the asexual and gametocyte stages (v) is predicted to function as an enzyme, receptor, or transporter. PfSUB1 matches criteria (i), (ii), (iv) and (v). The sequence identity cut-off used in the study for (iii) was 10%. The sequence identity of PfSUB1 to its closest human orthologue is 22%, which explains why PfSUB1 does not feature in the final 40 candidate drug targets. We do not consider this an issue because, as described previously, the human orthologue has a different substrate specificity and is structurally different.

Inhibiting PfSUB1, and therefore stopping the parasite from exiting the red blood cell, would effectively reduce the infection rate and the host sickness bouts. It is generally accepted, though never formally proven, that the infected red blood cells would then be destroyed (together with the parasites) in the spleen [Ho and White, 1999; Krucken et al., 2005].

Recent developments in the understanding of parasite egress from the red blood cell make PfSUB1 a novel and interesting drug target, especially as serine proteases have long been studied [Hedstrom, 2002] and we can therefore make use of a large, existing knowledge-base. The serine protease class is known to be druggable, *e.g.* thrombin and factor Xa. Still, for a drug to act on the PfSUB1 active site the compound must cross two membranes; the parasitophorous vacuole membrane and the erythrocytic wall.

An overview of all reported PfSUB1 inhibitors is given in Section 5.3.2. Most of these molecules are not drug-like and present a number of challenging drug design issues (*e.g.* they are peptidic, so are easily metabolized by the body).

Reaction Mechanism

The function of the PfSUB1 protease is to cleave a polypeptide at a specific location called the scissile bond, in order to digest and activate the papain-like SERA family of proteins (as mentioned earlier). This happens in PfSUB1's active site, and the reaction is specifically carried out by three residues, ASP372, HIS428 and SER606, known as the catalytic triad. The catalytic triad is characteristic of serine proteases in general and the common reaction mechanism is shown in Figure 1.6, using chymotrypsin as an example.

The polypeptide substrate binds to the active site of PfSUB1 (Figure 1.6a), with the carbonyl carbon of the scissile bond positioned close to the hydroxyl oxygen atom of the nucleophilic serine (Figure 1.6b). The epsilon nitrogen on the histidine's imidazole ring accepts the proton from the serine, while the oxygen atom (on the serine) becomes a potent nucleophile, attacking the carbonyl carbon on the peptide bond. This serine attack on the peptide forms a covalent bond and, hence, a tetrahedral intermediate (Figure 1.6c). This tetrahedral intermediate is stabilized by hydrogen bonds from an oxyanion hole [Kraut, 1977], specifically formed by the PfSUB1 residue ASN520 [Jean et al., 2005]. The

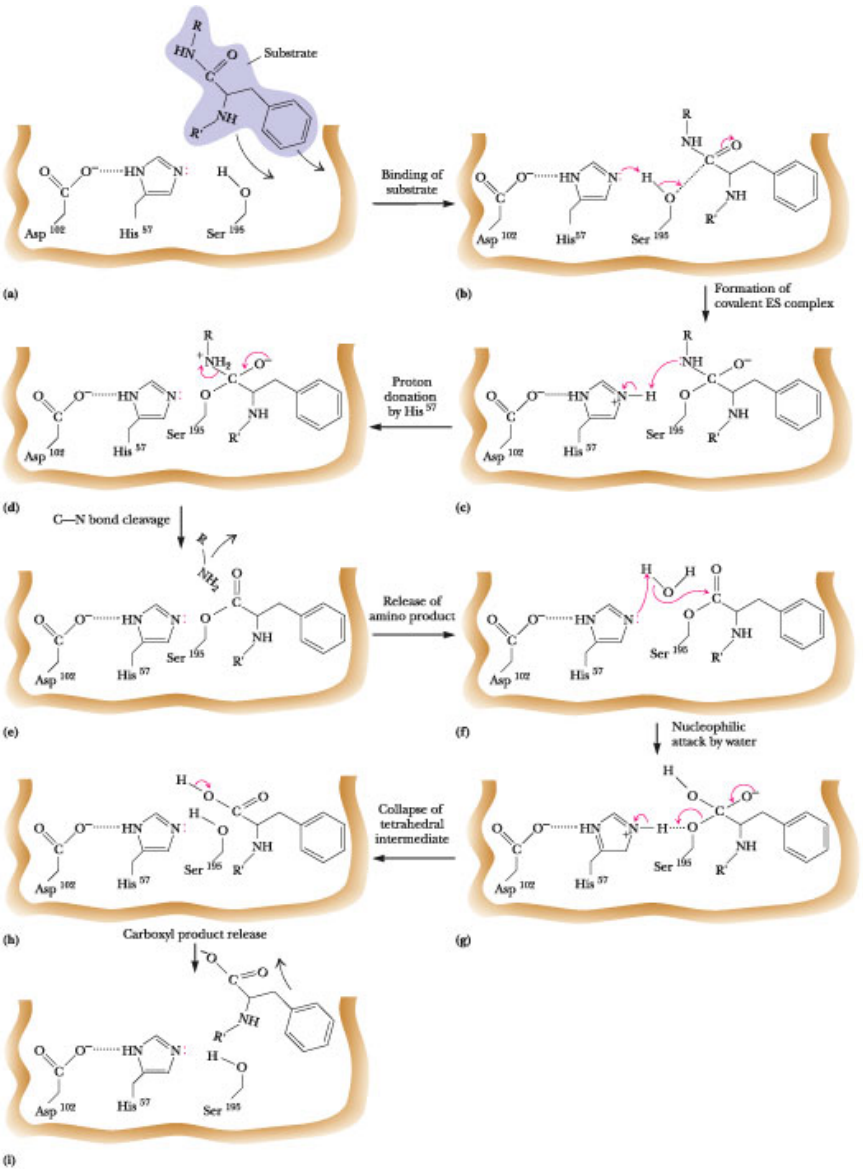


Figure 1.6: The reaction mechanism of another serine protease, chymotrypsin. This diagram is taken from Figure 14.21 in Garrett and Grisham [2008][1], and is explained in detail in the text.

[1]From GARRETT/GRISHAM. Biochemistry, 4E. ©2008 Brooks/Cole, a part of Cengage Learning, Inc. Reproduced by permission. http://www.cengage.com/permissions.

histidine protonation and the serine attack are thought to be concerted.

The peptide bond between the carbon and the nitrogen is cleaved to form an acyl-enzyme intermediate. This happens because the histidine nitrogen donates the proton to the nitrogen attached to the serine, and this covalent bond is lost in this attack (Figures 1.6d and 1.6e). Once this bond is cleaved, the amino product can leave the pocket and is replaced by a water molecule (Figure 1.6f). This water helps form a second tetrahedral intermediate, by attacking the carbonyl carbon in the acyl-enzyme. Electrons from the carbonyl double bond move onto the oxygen making it negative and a bond between the oxygen of the water and the carbonyl carbon is formed. This, together with the histidine epsilon nitrogen accepting a hydrogen from the water, produces another tetrahedral intermediate which is also stabilised by the oxyanion hole (Figure 1.6g).

In the final step of the reaction, the histidine epsilon nitrogen donates its hydrogen back to the serine oxygen breaking the bond between the serine and the carbonyl carbon which was originally on the substrate. The carbonyl carbon forms a double bond with the oxygen, generating a carboxylic-acid from the acyl-enzyme (Figure 1.6h). This product is then free to leave the active site. This restores the active site to its original form, ready to cleave another substrate (Figure 1.6i).

The role of the aspartate in the active site is two-fold: to orient the histidine to make it a better proton acceptor and to facilitate the proton transfer by electrostatic stabilization. Even if it does not actively feature in the mechanism described above it is shown to be critical for the efficient functioning of serine proteases [Craik et al., 1987].

1.4.3 Collaborative Framework

This project is part of a larger collaboration within the European Union Seventh Framework Programme. The three stakeholders in this study are InhibOx Ltd. (in collaboration with the University of Oxford), the Latvian Institute of Organic Synthesis (LIOS) and the Medical Research Council (MRC). InhibOx was responsible for the computational discovery of potential drug hits through virtual screening methods. These hit compounds were then synthesized by LIOS and later delivered to the MRC where they were biologically tested on the malarial pathogen. Any indications (in terms of biological activity)

stemming from the bioassay testing were then used to optimise the hit compounds and develop them into leads and repeat the drug discovery, synthesis and testing cycle. The MRC group is led by Michael J. Blackman, who is the principal investigator who first characterized the PfSUB1 gene [Blackman et al., 1998].

The work presented in this thesis is funded by an EU Marie Curie Fellowship – Initial Training Network, grant agreement 238490.

1.5 Main Contributions and Thesis Structure

In this introduction we have given a general overview of CADD methods with particular attention to virtual screening, or how to use computational methods to find bioactive, small-molecule inhibitors. We discussed the different types of virtual screening techniques, LBVS and SBVS. This is an active field of research, which has gained a lot of momentum in recent years – partially driven by the need to reduce costs and find new drugs in the pharmaceutical industry. We have also described a malarial target, PfSUB1, which will be the focus of virtual screening experiments presented in the last part of this thesis.

Throughout this thesis we have developed methods that are central to virtual screening. We review in detail a number of conformer generation methods (Chapter 2). This is a ubiquitous process in computational drug discovery, both in rigid docking and ligand 3D similarity searches. We suggest a protocol for generating conformers, based on the number of rotatable bonds, which strikes a balance between the diversity of the conformer ensemble and the number of conformers generated. We use this protocol to build a multi-million small-molecule database for use in virtual screening exercises (Chapter 3). Several challenges are outlined, such as standardizing molecules, providing fast and easy access to the database, and creating a diverse subset of the database for use in more exhaustive and rigorous virtual screening studies. We describe the development of Ligity (Chapter 4) – a fully-automated pharmacophore-based virtual screening method making use of a novel descriptor. When tested retrospectively on current benchmark datasets, this method is shown to perform better than existing 3D methods on average. We then describe two prospective virtual screening studies using a pan-malarial target,

PfSUB1 (Chapter 5). We highlight the promising results and challenges of both LBVS and SBVS methodologies on this target. Our LBVS study resulted in the discovery of novel inhibitors which have expanded the structure activity relationship. Finally, we describe future work which may be undertaken based on this thesis (Chapter 6).

1.5.1 Chronology of Events

The chronology of the work presented in this thesis has not followed the order of the chapters herein. First, the PfSUB1 SBVS experiment (including the PfSUB1 homology modelling) was carried out. The conformer generation study was done in parallel to this. After the initial SBVS results, we decided to build a more rigorous and high-quality small-molecule database for use in our virtual screening studies. We later used this improved database in the PfSUB1 LBVS experiment. Finally, based on these improvements, we developed a novel virtual screening method called Ligity.

2 Conformer Generation

Most of the work in this chapter has been reproduced *verbatim* from **Freely available conformer generation methods: how good are they?**; Jean-Paul Ebejer, Garrett M. Morris, Charlotte M. Deane; *Journal of Chemical Information and Modeling*, 52(5):1146–58, 2012.

2.1 Background

The vast majority of small molecule drugs work through physical interaction with specific biological macromolecules, usually proteins. The principal determinants of molecular recognition are complementarity of shape and properties between the two molecular entities. Thus the biological function of a drug is intimately related to its three dimensional structure. Furthermore, most drug molecules are flexible and can adopt a variety of shapes (conformations) in aqueous solution, existing as an ensemble of low-energy conformations in equilibrium with one another. The "biologically active" conformation (that which binds to the target protein) may be similar to one of the solution conformations or it may be a new conformation induced by protein binding [Perola and Charifson, 2004]. Different proteins may induce different conformations of the same ligand. For example, two arbitrarily selected conformations of the same ligand *cellotriose*, with the Protein Data Bank (PDB) [Berman et al., 2000] chemical component identifier `CTR`, bound to two different cellulases from different micro-organisms (PDB entries `2rfz` and `2xqo`), have a root mean square deviation (RMSD) of 3.25 Å from each other. This also serves

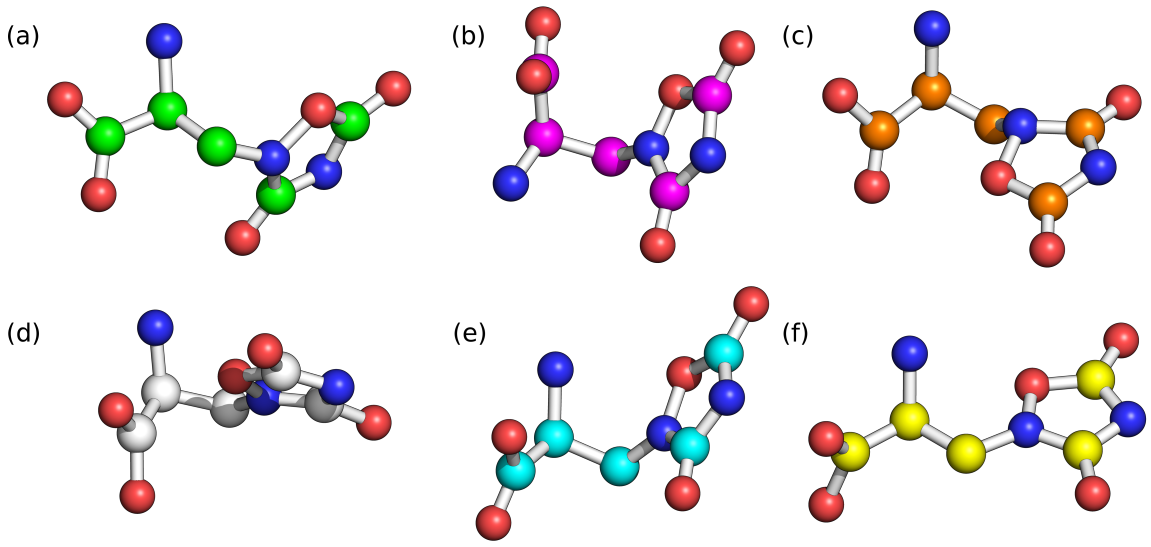


Figure 2.1: Examples of randomly selected conformations for *quisqualate* (PDB chemical component identifier `QUS`) with four rotatable bonds generated by the various tools in our review: (a) the X-ray crystal structure (taken from PDB entry `1p1o`), (b) Balloon, (c) Confab, (d) Frog2, (e) MOE, and (f) RDKit. Note that the conformer generated by Balloon has inverted stereochemistry, *i.e. (R)* instead of *(S)*. More details about this may be found in Section 2.2.4.

to highlight that a molecule's "shape" may be more accurately described by an ensemble of low-energy conformations that it may adopt, and thus the diversity of conformations generated by a method is an important consideration [Agrafiotis et al., 2007].

Therefore, for Computer-Aided Drug Design (CADD) studies, an effective method is required for generating conformational models that captures the bioactive conformation as one of a set of diverse, energetically accessible conformations (because this bioactive conformation will generally not be known in advance). Figure 2.1 shows a comparison of the crystal structure with the various conformations obtained for each of the methods reviewed here, when only a single conformation is generated. CADD methods that use such models are widespread and include shape-based similarity searches [Hahn, 1997], pharmacophore modelling [Kristam et al., 2005; Schwab, 2010], 3D QSAR [Verma et al., 2010], structure-based [Lorber and Shoichet, 1998; Lyne, 2002; Makino and Kuntz, 1997] and ligand-based [Stahura and Bajorath, 2005] methods in virtual screening.

The aim of the conformer generation process is to build a set of representative conformers that covers the conformational space of a given molecule. The generated

conformers should reasonably sample the energy landscape and not produce highly unlikely structures. In applications such as virtual screening, conformational models must be generated for very large number of compounds and as a result, methods that are very computationally intensive are impractical. All the conformer generation methods we consider are therefore based on a molecular mechanics approach.

While reviews of conformer generation software have been published, most of these do not reflect recent developments [Sadowski et al., 1994] and/or review only commercial software [Agrafiotis et al., 2007; Boström, 2001; Chen and Foloppe, 2008; Kirchmair et al., 2006]. Test set sizes in these publications vary from 32 to 778 molecules.

In this study we benchmark the performance of four freely available conformer generation methods and one implemented in the commercial package MOE. Key metrics are the ability to reproduce the experimentally determined conformation, the coverage of conformational space and the computational efficiency of the methods.

2.1.1 Conformer Generation

Conformer generation algorithms may be broadly classified as either systematic or stochastic [Gu and Bourne, 2009]. In the systematic approach, a regular sampling of each of the dimensions of the search space, and hence of all possible conformers, is conducted. This is achieved by incrementing the torsion angles of all rotatable bonds by some pre-defined amount. Each conformer is therefore enumerated. This is impractical for molecules with a large number of rotatable bonds, due to the combinatorial explosion in the number of rotameric states. In the stochastic approach, the conformational space is randomly sampled using techniques such as Monte Carlo simulated annealing [Chang et al., 1989; Wilson et al., 1991], genetic algorithms [Liu et al., 2009; Mekenyan et al., 1999; Vainio and Johnson, 2007] and distance geometry [Havel et al., 1983; Spellmeyer et al., 1997].

Knowledge-based methods [Feuston et al., 2001; Klebe and Mietzner, 1994] use pre-defined libraries for torsion angles and ring conformations [Brameld et al., 2008; Sadowski and Boström, 2006]. These libraries are created by considering known, experimentally determined structures in databases such as the Cambridge Structural Database

(CSD) [Allen, 2002] and the PDB. A molecule is decomposed into its constituent fragments and libraries of the possible conformations of these fragments are used to reassemble the whole molecule, in either a stochastic or systematic manner. Energetically favourable conformers are subjected to energy optimization in torsion angle space. The output conformer ensemble is typically generated based on energetic and geometric criteria.

2.1.2 Tools Compared

The conformer generation tools reviewed and compared in this study are Balloon [Vainio and Johnson, 2007], Confab [O'Boyle et al., 2011c], Frog2 [Leite et al., 2007; Miteva et al., 2010], RDKit [RDKit, online] and MOE [MOE, online]. All of these tools are free and publicly available, with the exception of MOE. Confab, Frog2 and RDKit are open source. Frog2 is also accessible through a web interface. We give a brief description of each tool and refer the interested reader to the cited literature and the user manuals of these software packages for more details.

RDKit

RDKit uses the distance geometry approach described by Blaney and Dixon [2007]. In this approach, a matrix representing the lower and upper bound of all pairwise distances in a molecule is created. This matrix effectively describes the whole structural space for a molecule. The triangle inequality rule is applied to smooth and further refine this matrix (where the three vertices of the triangle are different atoms, so for atoms A, B and C the distance $AC \leq AB + BC$). To generate multiple conformers, random distance matrices which satisfy the bounds matrix are generated. These conformers are then typically cleaned up using a force field. RDKit does not guarantee that the generated structures are low energy but it is possible to discriminate and keep only conformers which are a certain RMSD threshold apart. The documentation [RDKit manual, online] suggests optimizing the generated conformers using the Universal Force Field (UFF) [Rappe et al., 1992]. It also states RDKit's conformer generation is designed to supply 3D structures

quickly.

RDKit is an open source cheminformatics toolkit made available under the permissive Berkeley Software Distribution (BSD) licence.

Balloon

Balloon uses distance geometry to generate an initial 3D structure for a ligand, followed by a multi-objective genetic algorithm approach which modifies torsion angles, the stereochemistry of double bonds, tetrahedral chiral centres and ring conformations. The aim is to generate conformers which are near the global energy minimum, whilst also being diverse and distinct. Generated conformers are optimized using a parametrized force field. Conformers which are closer than an RMSD threshold to a lower energy conformer are discarded, ensuring that the conformers generated are different from each other.

Balloon is free and supports a number of different operating systems (Linux, Mac OS X and Windows). The source code is not available and its use is governed by a proprietary licence.

Confab

Confab is a knowledge-based conformer generation tool. It uses a systematic approach to generate and test all conformers described by a set of torsion rules. Conformers are generated by varying torsion angles, therefore no conformers are generated for molecules with zero rotatable bonds. The number of conformers to test may be specified by a cut-off (default 10^6), in which case the conformational space is visited randomly in order to ensure adequate sampling. Only the conformers within a certain energy threshold of the lowest energy conformation are kept. Conformers are also discarded if they are similar in shape to other selected structures (*i.e.* their RMSD falls within a user-selected value; the default is 0.5 Å). Note that it is not possible to generate a user-specified number of conformers using Confab. Furthermore, and unlike the other tools reviewed here, it only accepts 3D structures as an input. Confab does not explore ring conformations.

Confab is an open source project available under the GNU General Public License (GPL) version 2 licence.

Frog2

Given a one- or two-dimensional description of a molecule, Frog2 will break it down into a graph of rings and acyclic elements. This graph is then used to generate conformers. For the ring nodes, a conformation is selected from a library using a knowledge-based approach. The acyclic elements are built using literature-based canonical bond lengths and valence angles. Various combinations of dihedral angles are considered, supplemented by using a Monte Carlo search to vary these angles by a small amount. Conformations are then energy minimized using the AMMOS force field [Pencheva et al., 2008].

Frog2 has a web interface [Frog2, online] and the source code is available under the GNU General Public License (GPL) version 3 licence.

MOE

Molecular Operating Environment (MOE) is a fully integrated commercial drug discovery software package. It offers a wealth of functionality covering structure-based design, pharmacophore discovery, protein modelling, molecular simulations, cheminformatics, QSAR and medicinal chemistry applications. MOE offers three methods for conformer generation: systematic search, stochastic search and low mode molecular dynamics. The systematic search method generates conformers by rotating the dihedral angles of the molecule by a discrete pre-defined amount (this type of generation is only suitable for molecules with a small number of rotatable bonds). In the stochastic search method all the rotatable bonds in the molecule are randomly rotated (including ring bonds) and stereochemistry may also be randomly inverted. The low mode molecular dynamics simulation generates conformers by running a brief molecular dynamics simulation, with velocities initialized to low-frequency vibrational modes. Although this method is efficient for small molecules, chiral centres are rarely inverted. Irrespective of the method chosen the output is subjected to energy minimization (by default, a modified version of the MMFF94 [Halgren, 1996] force field).

MOE is commercially available from Chemical Computing Group.

Other Tools

There are several other popular, commercial conformer generation software tools not featured in this review.

Corina is a commercial product available from Molecular Networks [Molecular Networks, online] which generates a single, high-quality conformer. It takes a rule and data-based approach to generate a low-energy conformer [Gasteiger et al., 1990]. First, bond lengths and bond angles are set to standard values based on a table. Bond lengths are specific to atom types, hybridization states and bond order of a particular atom pair. Bond angles depend on the atom type and hybridization state of the central atom. Secondly, the molecule is fragmented into ring systems and acyclic parts. Corina can handle small-ring, rigid polymacrocyclic systems and flexible macrocyclic systems in different ways and is able to produce a list of conformations for ring systems. The resulting geometries are optimized using a reduced force field [Sadowski and Gasteiger, 1993]. Rotate is a complementary program also available from Molecular Networks which generates diverse conformational ensembles by applying a set of rules that resulted from a statistical analysis on the conformational preferences of experimentally determined molecular structures of small molecules.

Omega is another commercial conformer generation tool available from OpenEye Scientific Software [OpenEye Scientific, online] and uses a systematic, knowledge-based approach. It works by first assembling the initial 3D structure from a library of fragments. Secondly, it exhaustively enumerates all rotatable torsions using predefined libraries and finally samples this large conformational space using geometric and energy criteria [Hawkins et al., 2010].

For a more detailed and comprehensive review of these (and other) tools we direct the interested reader to a review by Schwab [2010].

2.1.3 Test Set

Our validation of freely available conformer generation methods is based on a test set of 708 distinct small molecules. The selection of these molecules is derived from the work

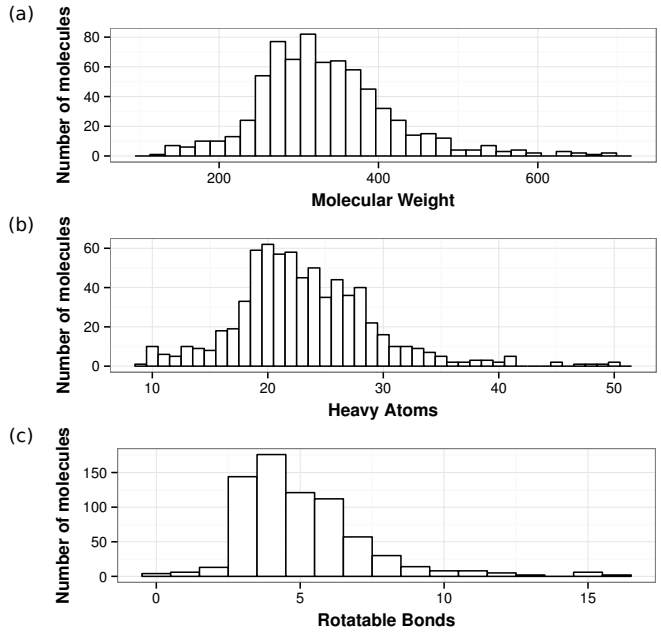


Figure 2.2: Test set distributions for (a) molecular weight, (b) number of heavy atoms, and (c) number of rotatable bonds.

of Hawkins et al. [2010] as well as the ligands found in the Astex Diverse Set [Hartshorn et al., 2007]. These molecules come from high quality X-ray crystal structures in the PDB and CSD. A list of PDB and CSD identifiers and Simplified Molecular Input Line Entry Specification (SMILES) [Weininger, 1988] representations of the ligands used in this study have been made publicly available. For the PDB structures we also supply the reference file used. The molecular weight, heavy atom counts and rotatable bond distributions for the molecules in the test set are shown in Figure 2.2. Note that these distributions are similar to the distributions for drug compounds published in the literature [Feher and Schmidt, 2003; Oprea, 2000] and comparable to other datasets used for conformer validation [Chen and Foloppe, 2008; Kirchmair et al., 2006].

2.2 Methods And Materials

In this section we describe the conformer generation tools we used in this study (and the parameters used to run each of these). Here, we also explain how we assembled the test set.

2.2.1 Conformer Generation Tools

The program installations used in this review are described in Table 2.1.

Table 2.1: Tools used for conformer generation.

Tool	Version	Platform	Cost/Licence	Distribution
MOE	2010.10	Linux (Ubuntu 10.10 64-bit)	Commercial	Binary
Balloon	1.2.0.915	Linux (Ubuntu 10.10 64-bit)	Free/Proprietary	Binary
Confab	1.0.1	Linux (Ubuntu 10.10 64-bit)	Free/GNU GPL	Source
Frog2	2.13 (patched from authors)	Linux (Ubuntu 10.10 64-bit)	Free/GNU GPL	Source
RDKit	2011.09.1	Linux (Ubuntu 10.10 64-bit)	Free/BSD	Source

Default parameters recommended by the creators of the respective programs have been used throughout when running the conformer generation applications, but in order to make a fair comparison we try to replicate the same behaviour across all the different tools. For each tool, input and output molecule file parameters were specified as command line arguments. Wherever present the option to generate conformers which are at least 0.5 Å apart was set (before energy minimization). The following points of interest are relevant to each application:

- MOE is very rich in terms of the options for conformer generation. In this work a MOE Scientific Vector Language (SVL) script was written that uses the Conformation Import functionality (*i.e.* `conf_Import`) with all filtering options turned off. The `conf_Import` function breaks a large molecule into smaller fragments. For common fragments it uses conformations from a standard library. If a fragment is not found in the standard library, these fragments are generated using a stochastic conformation search. The number of conformers was specified by

setting the `outputConformationLimit` parameter. Note that the generated structures undergo energy minimization using MOE's MMFF94 modified force field (the `outputRefine` parameter was set to 1). The output was converted to an MDL SD file to keep the same workflow and input/output requirements as the other tools. The `mmSuperposeRMSD` parameter which sets the criterion for conformer equality was set to 0.5 Å RMSD.

- Balloon was run using the settings listed on the usage page on the website [Balloon, online], with the MMFF94 force field (option `f`) and the number of conformers to generate (option `nconfs`) specified. Note that the latter parameter does not specify the number of final conformations, which may be slightly more or less depending on the flexibility of the molecule, but rather the initial ensemble size. The default value for `RMSDtol`, which specifies the inter-conformer RMSD for the final pruning of conformers, is 0.5 Å.

- For Frog2, apart from specifying the work path (option `wrkPath`) and log file (option `logFile`), the `multi` option was also set to 10, 50, or 100 depending on the number of conformers being generated. This setting, together with the `gnb` option, limits the number of conformations for each compound. We also set the `rmsd` parameter to generate conformers which are at least 0.5 Å different from each other.

- We generated RDKit conformers by writing a simple Python script as shown in Section 3.5 in the RDKit User Manual [RDKit manual, online]. As a first step, molecules are loaded from a SMILES file and hydrogens are added to each using RDKit. The only difference from the code snippet provided in the manual was that a call to `AllChem.EmbedMultipleConfs` (instead of `AllChem.EmbedMolecule`) was made. This function generates a user-defined number of conformers rather than just one. Also, the parameter `pruneRmsThresh` in this function call is set to 0.5 Å to ensure that the generated conformers are at least this far apart from each other. The script then loops over the conformers to energy minimize them using RDKit's implementation of the UFF force field, as suggested in the manual.

- The Confab conformers were generated using default parameters, which include an RMSD diversity cut-off of 0.5 Å. The conformers outputted from Confab were subjected to energy minimization using the MMFF94 force field. This was done using the `Obminimize` program available in OpenBabel version 2.3.1 [Guha et al., 2006; O'Boyle et al., 2011a; OpenBabel, online].

2.2.2 Test Set Selection

One of the main challenges was to acquire a diverse test set of high quality structures of small molecules on which to base our conformer tools comparison.

Our starting test set was originally built by combining the 85 bound small molecules from the Astex Diverse Set and a further 677 molecules used in the testing of commercially available conformer generation tool Omega [Hawkins et al., 2010].

The Astex Diverse Set is made up of high quality, high resolution crystal structures of complexes containing drug-like ligands from the PDB. It is typically used for protein-ligand docking performance validation. From this set of 85 small molecules, we removed the ligand for PDB entry `1x8x`, since this was a duplicate ligand also seen in entry `1of6`. This duplication may be an oversight in the Astex Diverse Set as the PDB entries label the ligands as D- and L-isomers of tyrosine for `1of6` and `1x8x` respectively, but in fact these structures are both L-tyrosine.

The molecules from the Omega test set included the PDB codes of 197 drug-like, high quality structures from well resolved structures in the PDB and 480 molecules from the CSD originating from a previous publication [Brameld et al., 2008]. Unfortunately, there is often more than one ligand associated with each PDB code provided, but the ligand identifier is not supplied. In order to determine which ligand to use for a particular PDB code, the following procedure was applied:

1. apply similar filtering criteria used in the Omega validation publication (*i.e.* $0 \leq$ rotatable bonds ≤ 16, and $8 \leq$ heavy atom count ≤ 50);

2. use the CoFactor database [Fischer et al., 2010] to remove any cofactors present in the structure file;

3. review the literature describing the PDB deposition to determine which was the inhibitor molecule;

4. search the ligands in DrugBank [Knox et al., 2011] to find a match to one of the ligands listed (and manually select the match).

If after this rigorous manual process there were still multiple potential ligands for a particular structure, the PDB entry was removed from the test set. For each ligand in our test set which has more than one instance in its corresponding PDB entry, we selected an arbitrary structure.

Each identifier for the CSD molecules corresponds to exactly one 3D structure. The selected molecules from the Omega test set were validated using the test set distributions published in the Omega paper.

2.2.3 Test Set Preparation

A one-dimensional SMILES representation with stereochemical information was produced from the reference 3D structures of the molecules in the test set using both OpenBabel 2.3.1 and RDKit 2011.09.1. From these two equivalent SMILES representations we generated the corresponding InChI keys (with the stereochemistry layer). Where there were mismatches between these keys, the molecule's SMILES representation was inspected manually and, where possible, replaced with the correct SMILES. Otherwise the molecule was removed from the test set. This gives us confidence that the SMILES representations with stereochemistry information (which we made publicly available[1]) have been correctly generated and are true representations of the 3D reference molecule.

Also, the canonical SMILES and InChI keys were generated for each ligand (also using OpenBabel 2.3.1) to make sure that there were no duplicates across the whole set. The SMILES representation of the molecule is the starting point of our conformer generation process, ensuring that the conformers are not geometrically biased by the coordinates of the original X-ray structures.

[1] http://pubs.acs.org/doi/suppl/10.1021/ci2004658/suppl_file/ci2004658_si_003.zip

Not all conformer generation tools accept a SMILES string as an input: Confab accepts only a 3D structure file. In order to get around this limitation, we generated 3D coordinates from the SMILES representation using OpenBabel 2.3.1 (with the `gen3d` option). We then used this as the input file to Confab. Once again, we generated InChI keys (with the stereochemistry layer) on the 3D files generated by OpenBabel and compared them to the InChI keys of the 3D reference molecule file from the test set. If these did not match, we removed the molecule from the test set so as not to negatively bias the results for Confab.

The above processing removed 37 out of the initial 197 Omega PDB entries, 11 out of the initial 480 Omega CSD entries and 5 out of the 84 Astex Diverse Set entries (duplicate ligand entry `1x8x` had already been removed). The remaining 629 molecules from the Omega test set, were added to the 79 small molecules from the Astex Diverse Set to give a combined set of 708 molecules.

2.2.4 A Note on Stereochemistry

Stereochemistry for the molecules in the test set is defined in the SMILES used to generate conformers. Most of the tools featured in this review are constrained by the defined stereochemistry and therefore generate the correct stereoisomer, but on occasion some tools incorrectly invert stereocentres. An example of this may be seen for the conformer generated by Balloon in Figure 2.1 (this incorrect behaviour seems to have now been fixed in Balloon version 1.4.1.1068, released in July 2013). This stereocentre inversion leads to higher RMSDs between the reference structure in the test set and the generated conformers. If stereochemistry is not defined in a molecule's SMILES representation, tools like RDKit will sample randomly between the different stereoisomers.

2.2.5 Determining Molecular Descriptors and RMSD between Molecules

The molecular weight, the number of heavy atoms and the number of rotatable bonds were calculated using the Pybel API [O'Boyle et al., 2008] (available when building

OpenBabel 2.3.1 with Python language bindings). The rotatable bonds definition used does not include ring bonds as rotatable.

The RMSD between the experimental structures and conformers generated by each tool was calculated using `Obfit` (a program in the OpenBabel 2.3.1 suite). The alignment generated by this program ignores hydrogen atoms and takes in account symmetry.

2.2.6 Number of Generated Conformers

In a small number of cases the software tested failed to generate any conformers (Confab: 20 cases, Frog2: 4 cases for the 50 generation run and Balloon: 2 cases for the 50 conformer generation run). In some cases this may be attributed to software bugs (*e.g.* segmentation fault). Confab does not generate conformers for molecules with zero rotatable bonds as its approach is based on changing torsion angles of rotatable bonds.

RDKit was set to generate 10, 50 and 100 conformers for each molecule. Similarly, Frog2 has a command line argument to generate a set number of conformers per stereoisomer. Balloon has a parameter for the number of conformers to generate but this is used as an indication only and the number may slightly vary depending on the flexibility of the structure. MOE has an option to limit the number of conformations output. Confab does not have any options in this regard so the number of conformers varies greatly *e.g.* the ligand `UN6` in PDB entry `2f70` has 10 rotatable bonds and generated 53,340 conformers. The histogram of the number of conformers generated by Confab is shown in Figure 2.3. Moreover, sometimes Confab generates just one conformer for flexible molecules. This is a known issue because if a molecule has a very large conformer space and a systematic search is carried out within that space only a small fraction of conformers will fall within 50 kcal/mol of the lowest energy conformer. Apart from this, in general and as expected, fewer conformations are generated for less flexible structures (smaller number of rotatable bonds). For tools which generate more than the required number of conformers (*i.e.* mostly Confab but occasionally also Balloon) we sample the set randomly.

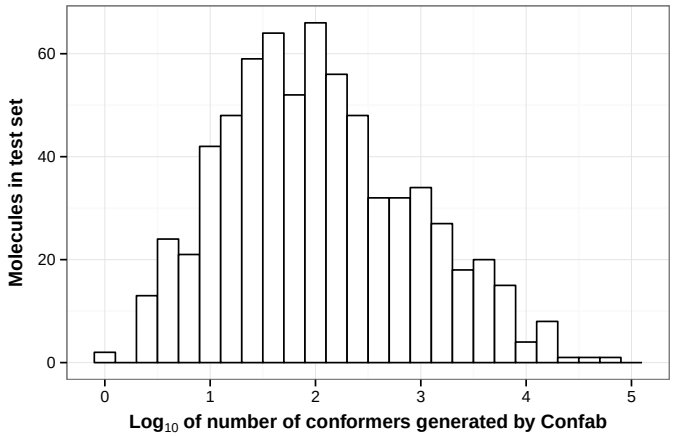


Figure 2.3: Histogram of number of molecules in the test set versus $\log_{10}$ of the number of conformers generated by Confab for each molecule in the test set. The median of the number of conformers generated for our test set is 92.5, and the mean is 928.615. The mean is biased by a few molecules which generate a huge number of conformers, *i.e.* 11 molecules in the dataset generate more than 10,000 conformers.

2.2.7 Statistical Tests

The following tests were performed to find whether the distributions of minimum RMSD values from the theoretical conformer to the experimentally determined structure were statistically significantly different for each method.

Firstly, a Kruskal-Wallis rank sum test was performed for each toolkit's RMSD distribution when generating 50 conformers. This test was selected as there are five groups (tools) and we cannot assume a normal distribution in the underlying minimum crystallographic RMSD values. Secondly, since the Kruskal-Wallis test only indicates if there is an effect, a *post-hoc* test using the pairwise Wilcoxon signed-rank test (paired and with Bonferroni correction) was carried out on toolkit pairs to find which conformer generation methods are statistically different from each other.

2.3 Results and Discussion

We consider the following three criteria in our review:

1. **accuracy**: how close in terms of positional RMSD is a generated conformer to one

of the experimentally observed X-ray crystallographic structures?

2. **diversity**: how different or similar are the generated conformations?

3. **speed**: how much computational time is required to generate the conformers?

Our findings show that on average RDKit and Confab are the best conformer generators amongst the tools we considered. Statistically there is no difference between selecting one of these two tools. However, RDKit is much faster.

Even if the conformer diversity threshold parameter is set, RDKit generates many similar conformers after energy minimization (with RMSD < 0.5 Å) for molecules with few rotatable bonds. We later describe an algorithm which corrects this.

2.3.1 Quality of Generated Conformers

In order to determine which of the methods most frequently produces conformers closest to the crystallographic conformation, we carried out the following test. Taking each molecule in the test set in turn, we generated 50 conformers using each of the five methods, and for each method we computed the RMSD from each of the 50 conformers to the same arbitrarily selected X-ray structure for that ligand. We then selected the minimum RMSD value, which gives us the closest conformer to the X-ray structure. For tools such as Confab that do not have the option of generating a user-specified number of conformers, we sampled randomly from the generated set (for more details see Methods and Materials). The effects of selecting randomly, selecting the minimum energy conformers, and selecting the minimum RMSD structures out of the whole conformer ensemble for Confab are shown in Figure 2.4.

The pairwise comparison of methods can be performed using these minimum crystallographic RMSD values, as shown in Figure 2.5. In each panel two methods are compared: each point corresponds to a molecule from the test set with its minimum RMSD for each of the two methods providing the x- and y-coordinates. Note that the points on each graph are smoothed using a kernel density function (*i.e.* the default function, `bkde2D`, in R's `smoothScatter` routine) to avoid overplotting. The diagonal line indicates the

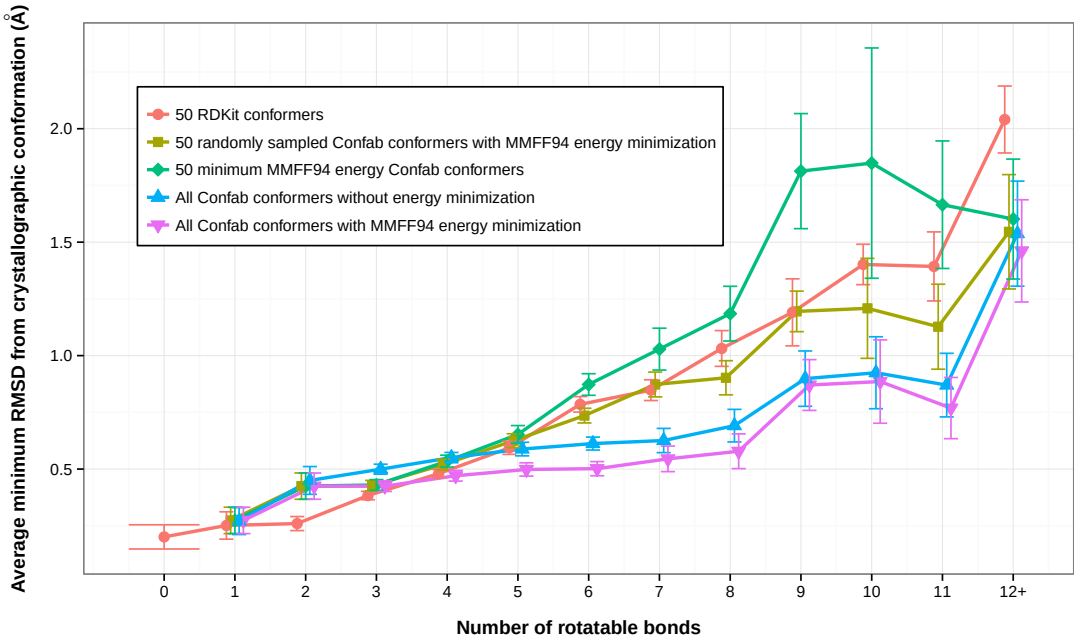


Figure 2.4: Analysis of choosing a random sample of conformers for Confab and its effect on the ability to reproduce the experimental structure. In this figure we show the minimum RSMD value from each reference structure to the (i) set of 50 conformers generated by RDKit (red), (ii) the set of 50 conformers sampled out of all conformers generated for Confab (yellow), (iii) the set of 50 lowest energy conformers using MMFF94 for Confab (green), (iv) the set of all Confab conformers without energy minimization (blue) and (v) the set of all Confab conformers with MMFF94 energy minimization (violet). With fewer rotatable bonds, the set of 50 sampled Confab conformers does as well as the other Confab conformer sets as at this point few structures generate more than 50 conformers (so the whole set is sampled and performance does not vary). With a larger number of rotatable bonds many more conformers are generated by Confab, so its ability to reproduce the X-ray crystallographic structure is increased.

theoretical position if both methods performed identically for every molecule in the test set.

Taking the plot of RDKit (y-axis) versus Frog2 (x-axis) in Figure 2.5 as an example, it can be seen that more points fall below the diagonal than above it. In practice, this means that generally the RMSD between a theoretical conformer and its corresponding experimentally determined structure is greater for Frog2 than RDKit. Thus, RDKit tends to be more accurate in terms of generating experimentally observed structures of drug-like molecules than Frog2.

The distribution of the minimum crystallographic RMSD values for all 708 small

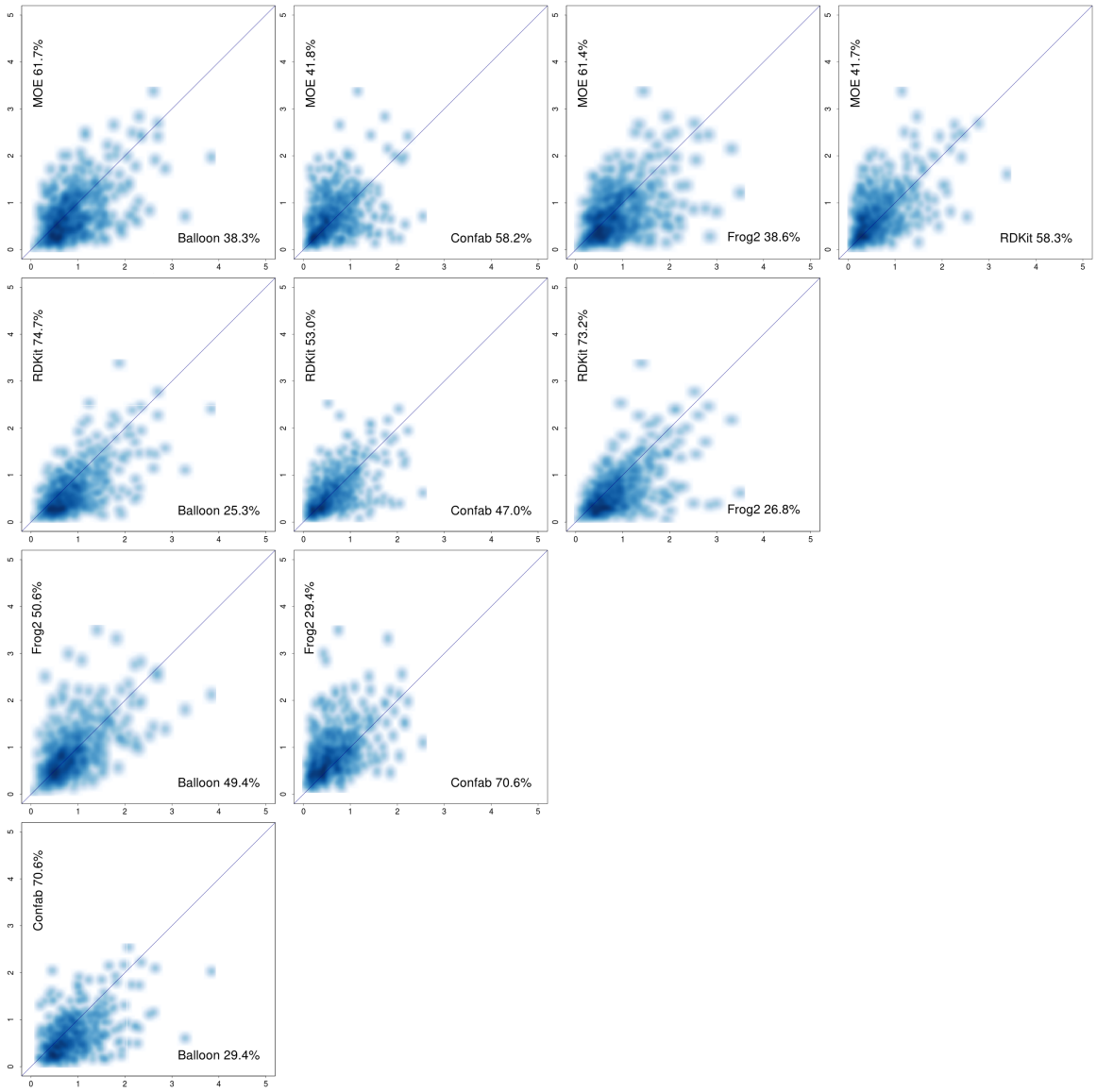


Figure 2.5: Pairwise comparison of minimum RMSD values between all the conformers generated by the method and the X-ray structure for all toolkits. These plots represent the densities of the 708 molecules in the test set. It can be seen that RDKit performs better than the other toolkits, because it tends to generate conformations with RMSD values from the crystallographic structure that are generally lower than the second method it is compared to, *e.g.* RDKit versus Frog2, there is a higher density of points below the diagonal than above. The percentages next to the method labels show the proportion of all ligands in the dataset (708) for which that particular method is closer to crystallographic structure than the other method plotted on the same panel. All units shown are in Å.

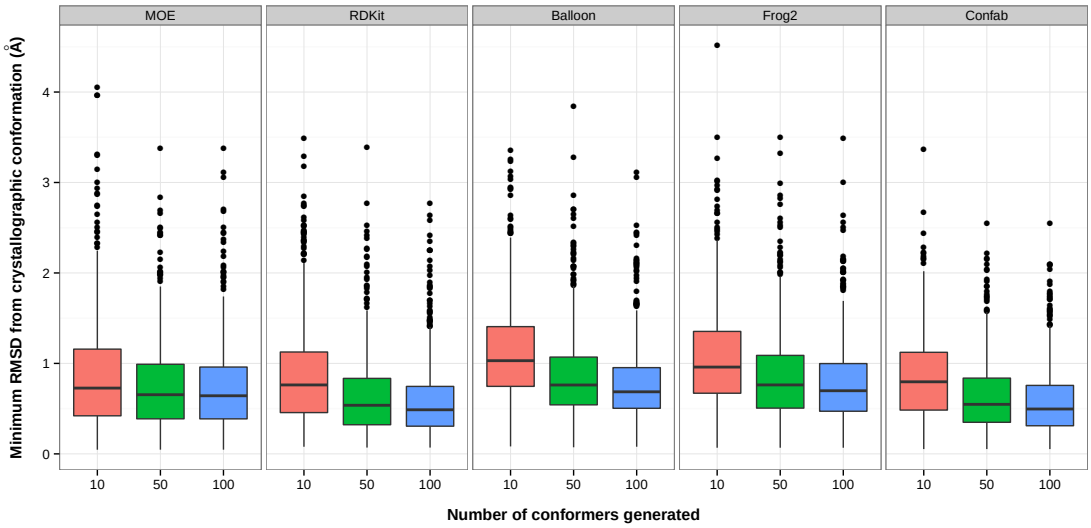
molecules for each method based on 10, 50 and 100 conformer generation runs is shown in Figure 2.6 (a). Each boxplot shows the minimum RMSD value (from the X-ray crystal

structure) for a set of conformers for each molecule in the test set. Thus each boxplot is made up of 708 data points. Figure 2.6 (a) shows that RDKit and Confab have better (lower) average minimum RMSD values than the other tools and for more than 75% of the ligands (when generating 50 and 100 conformers) they generate a structure with an RMSD that is less than 0.9 Å from the experimentally observed conformation of the molecule.

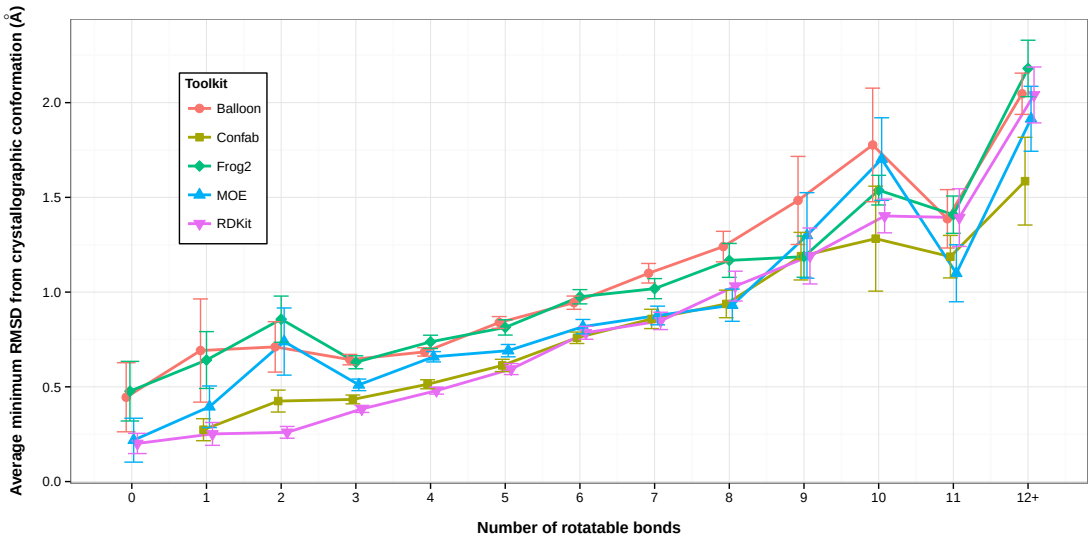
Using the Omega tool, it is reported that approximately 83% of the PDB structures and 84% of the CSD structures are reproduced within a 1 Å RMSD of the experimental structure (generating a maximum of 200 conformers and using default settings) [Hawkins et al., 2010]. In order to make our test set comparable to the validation set used in the Omega study, we removed the ligands from our test set originating from the Astex set. For the 50 conformer run, RDKit generates 70.0% of the 160 PDB structures and 90.4% of the 469 CSD structures in our set within a 1 Å RMSD of the experimental structure. All the other tools show inferior performance to RDKit.

Plots of the average minimum RMSD from the experimentally determined structure (with the standard error indicated by error bars) versus the number of rotatable bonds in the test set for the 50 conformer generation run are shown in Figure 2.6 (b). Unsurprisingly, as the number of rotatable bonds in the molecule increases it becomes more difficult to generate the crystallographic form. With a couple of exceptions, both RDKit and Confab achieve better results than the other methods, with RDKit doing better or similar to Confab when the number of rotatable bonds is less than ten, and Confab showing better performance for molecules with ten or more rotatable bonds. Only a few molecules in our test set have ten or more rotatable bonds (see rotatable bond distribution in Figure 2.2).

When considering all 50 conformers generated by each tool across all 708 molecules (rather than just the best conformer), the percentage of the generated structures with crystallographic RMSD values less than 2 Å is as follows: RDKit 76.7%; Confab 73.2%; Balloon 70.9%; Frog2 70.1% and MOE 63.4%. These percentages at this frequently used threshold [Boström, 2001; Liu et al., 2009] give an indication of how plausible the generated conformations tend to be. On the other hand, a lower percentage might



(a) Box plots showing the minimum RMSD from the crystal structure for each of the 708 molecules in the test set, when generating 10, 50, and 100 conformers. In general, as the number of conformers generated increases, the mean RMSD from the crystal structure decreases.



(b) The variation in the ability of each method to reproduce crystallographic conformations as the number of rotatable bonds increases. Note that the data points in this graph are jittered horizontally to avoid overplotting. Also, the molecules with 12 or more rotatable bonds are grouped together. It can be seen that in general, amongst all methods, RDKit and Confab are best at finding the lowest crystallographic RMSD values for the molecules in the test set.

Figure 2.6: Minimum RMSD distances from generated conformers to X-ray crystal structures.

indicate more diversity in the conformational model.

The distributions for the 50 conformer runs presented in Figure 2.6 are all statistically significantly different from each other with the exception of the Balloon-Frog2 and Confab-RDKit pairs ($\alpha = 0.05$). This means that there is a statistical difference in selecting one conformer tool over another.

A Kruskal-Wallis test revealed a significant effect of the toolkit on the minimum RMSD distance to the experimentally determined structures ($\chi^2(4) = 194.961, p < 0.01$). The Wilcoxon signed-rank test (paired) with Bonferroni correction showed significant differences between the tools shown in Table 2.2.

Table 2.2: Statistical tests for generated conformers show that there is a difference between the generated conformer sets for the below pairs.

Toolkits	Median (Å)	Z score	p-value
Balloon vs Confab	0.663	12.357	< 0.01
Balloon vs MOE	0.697	6.496	< 0.01
Balloon vs RDKit	0.646	13.706	< 0.01
Confab vs Frog2	0.651	−11.844	< 0.01
Confab vs MOE	0.598	−5.357	< 0.01
Frog2 vs MOE	0.696	5.730	< 0.01
Frog2 vs RDKit	0.637	13.052	< 0.01
MOE vs RDKit	0.581	6.003	< 0.01

2.3.2 Difficult Cases

As discussed earlier, the larger the number of rotatable bonds the more difficult it is for these tools to generate the X-ray crystallographic structure. The reason for this is that the conformer space grows exponentially with the number of rotatable bonds.

There are a small number of molecules which consistently score badly across all tools. These molecules typically have a central core with flexible, large parts of the molecule stemming from that core as shown in Figure 2.7 (a). The RMSD calculation is

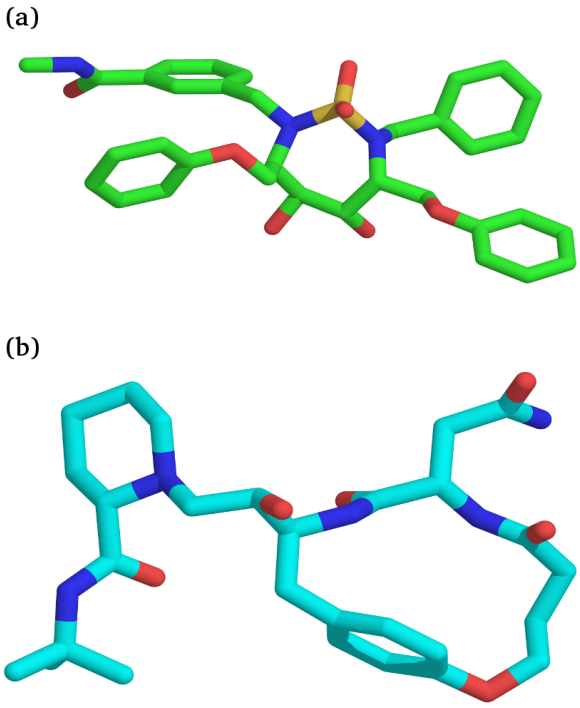


Figure 2.7: (a) small molecule with identifier `NM1` taken from PDB entry `1g2k` in the test set produces conformers across all toolkits which are, on average, 3.459 Å from the experimental structure. (b) small molecule with identifier `PI4` taken from PDB entry `1b6l` in the test set produces conformers across all toolkits which are, on average, 2.790 Å from the experimental structure. In this case, the inability to reproduce the experimental structure may be attributed to the macrocycle.

based on the position of these heavy atoms, and even a small torsional rotation of the bonds emanating from the "core" will offset many of the heavy atoms. The presence of macrocycles in the molecule, such as the one in Figure 2.7 (b), also makes the conformer generation tools perform badly as most of them are unable to sample the ring conformer space correctly. In most cases, the tools tested are unable to reproduce the experimentally determined ring conformation for rings with seven or more atoms (refer to Figure 2.8 for more details).

2.3.3 Diversity of Generated Conformers

Diversity is an important consideration in conformer generation. Running a docking experiment using rigid ligands where these have very similar structures is clearly sub-

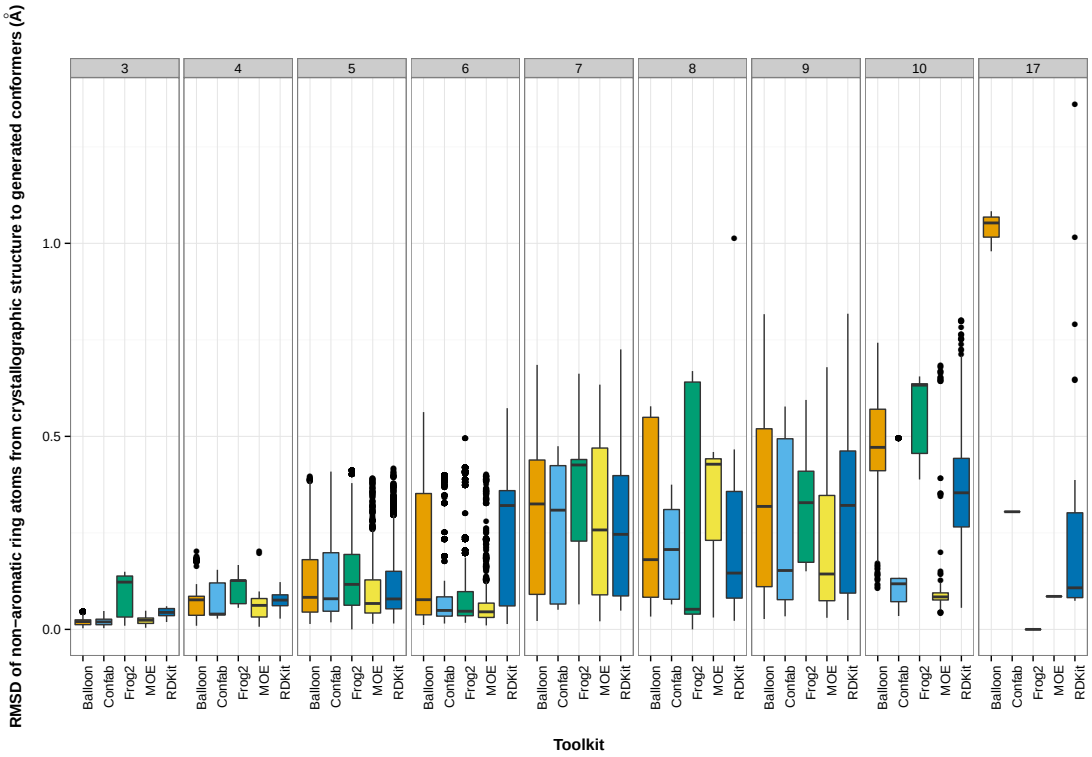


Figure 2.8: Analysis of how well the conformer generation tools do at generating ring systems. This figure shows the RMSD (in Å) between the X-ray crystallographic determined co-ordinates of all the non-aromatic ring systems in our test set and the co-orindates of these rings in the 50 conformer generation runs for each tool. This figure is grouped by number of atoms in the ring system. We consider fused rings as one ring system; *e.g.* the 17 atom ring system on the right originates from *dexamethasone* (PDB chemical component identifier `DEX` in structure `1M2Z`) which consists of three six-membered rings and one five-membered ring fused together. This figure shows that all tools have difficulty in reproducing the experimentally determined conformation for larger ring systems (≥ 7 atoms). Of note is that on average RDKit does worst than the rest for six-membered rings.

optimal both in terms of the experiment's running time and the space needed to store these conformers (assuming they are stored on the file system or in a database rather than generated on the fly). On the other hand, structures that are very different from the experimentally determined conformation may be unlikely in practice because of energetic and conformational constraints. A representative sample that covers all the low energy structures in each molecule's conformational space is therefore required.

For every molecule in the test set and then for each conformer generation method, we have calculated the pairwise RMSD distances between every pair of generated conformers

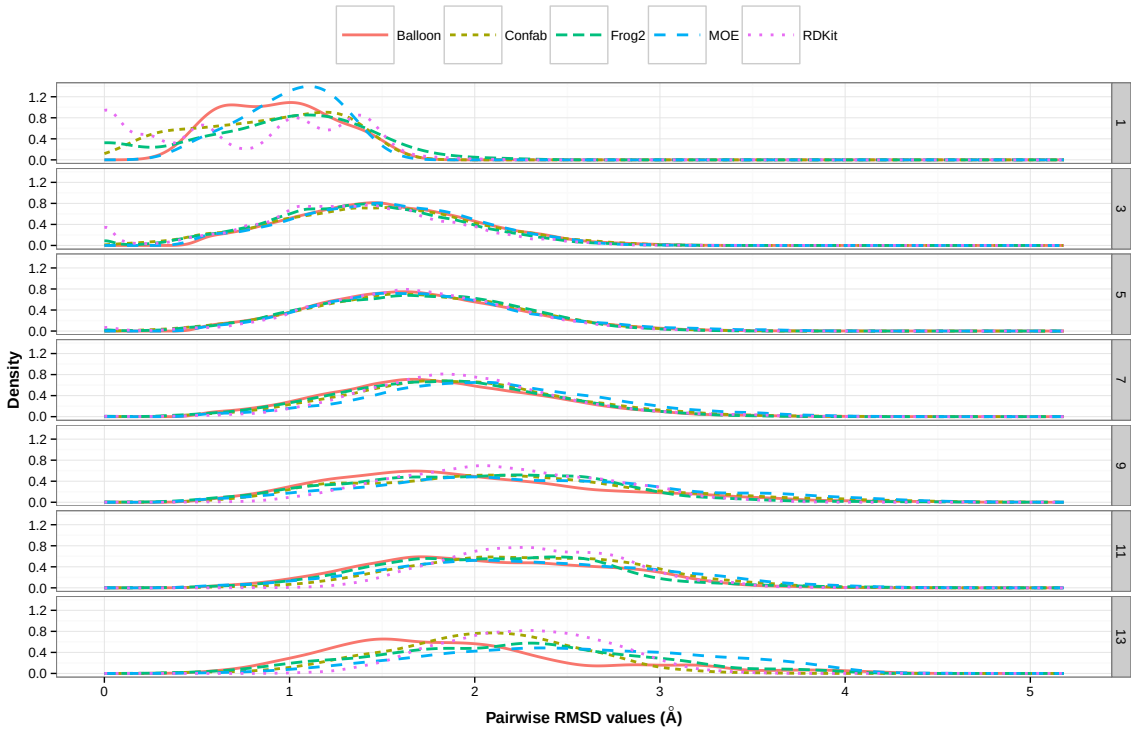


Figure 2.9: Density-smoothed distribution showing the variation in pairwise RMSD values for conformers per molecule for each toolkit, for sets of molecules with an odd number of rotatable bonds.

(for the 50 conformer generation run) and partitioned the results by sets of molecules with an odd number of rotatable bonds, and the toolkit used. The resulting graph in Figure 2.9 shows that as expected, the average of the minimum crystallographic RMSD increases with the number of rotatable bonds. This is because as the flexibility of the molecules increases, the size of the conformer search space increases and this allows for more diverse conformers to be generated. Note that even if the conformer diversity threshold for these tools was set to 0.5 Å, this is applied before the energy minimization step.

RDKit appears to produce many very similar conformations for molecules with low numbers of rotatable bonds (Figure 2.9). The multimodal peaks can be attributed to the fact that when there are few rotatable bonds, the conformer space is limited and energy minimization will make some of the conformers converge to similar geometries. This highlights the problem with checking for diversity before running the selected energy

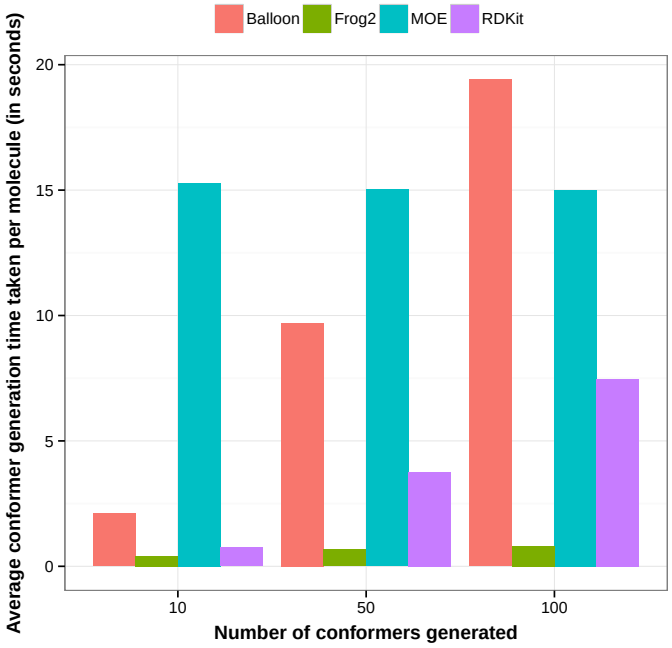


Figure 2.10: The average time (in seconds) for each molecule for each of 10, 50 and 100 conformer generation runs. Confab is not shown due to the inability to generate a specific number of conformers.

minimization protocol (as happens with most of these tools).

2.3.4 Conformer Generation Speed

All the conformer generation runs were executed in isolation and benchmarked using the same hardware (using an Intel Core i7-2600K CPU running at 3.40GHz with 4 GB RAM) and the same operating system (Ubuntu Linux 10.10 64-bit) to get comparable results. Also, the same system tool (*i.e.* the `time` command line utility in Linux) was used for benchmarking the execution speed of each conformer generation tool.

We measured the CPU time taken by the conformer generation runs. Our dataset consists of 708 separate SMILES files for each molecule in our test set. Each file is run as a separate process. The average CPU time (in seconds) per molecule in the 708 molecule test set for 10, 50 and 100 generated conformers is shown in Figure 2.10.

Frog2 is an order of magnitude faster than any other tool. Speed is of primary importance for Frog2 considering its web application nature. MOE shows constant time

regardless of the number of conformers generated. When running a different experiment, we noticed that MOE was considerably faster when all the molecules were in the same file and a single MOE process launched as opposed to launching a new process for every molecule. The reason for this is two fold; first there is a setup cost involved in running a conformer generation job (such as licence acquisition or database creation) and, secondly, the common fragments file is cleared at every single molecule run. RDKit and Balloon both show an increase in the time taken with respect to the number of conformers generated, with Balloon having a much steeper gradient (*i.e.* slower).

Confab is not included in this comparison because it is not possible to generate a specific number of conformers and make a side-by-side comparison to the other conformer generation programs. The Confab run (without energy minimization) took 254 minutes of CPU time for a total of 638,887 conformers generated for the 708 molecules in the test set. The MMFF94 energy minimization for this whole conformer set took approximately 53 hours. This measurement is biased by the huge number of conformers ($> 10,000$) generated for a few molecules (refer to Figure 2.3 for more details).

2.3.5 A Note on Energy Minimization

While an in-depth comparison of energy minimization on small molecules using a variety of force fields is beyond the scope of this chapter, we comment on the effect of energy minimization on the generated conformers. Taking the set of 10 conformers for each molecule in the dataset generated by RDKit, we investigate the effect of the force field on the ability to reproduce the conformation of the experimentally determined structure.

Figure 2.11 shows that minimization is important to increase the quality of the conformational models. Moreover, energy minimization is a critical step after RDKit's distance geometry approach as it is required to produce "clean" structures. The MMFF94 force field seems to be better suited for molecules with zero or one rotatable bonds, while UFF does better for the rest of the molecules. The mean RMSD value for all conformers from their respective experimental structures is of 1.113 Å when not using any optimisation, 0.878 Å when minimizing using UFF and 0.945 Å when minimizing using MMFF94.

Another point of interest is the difference between the energy of the experimental

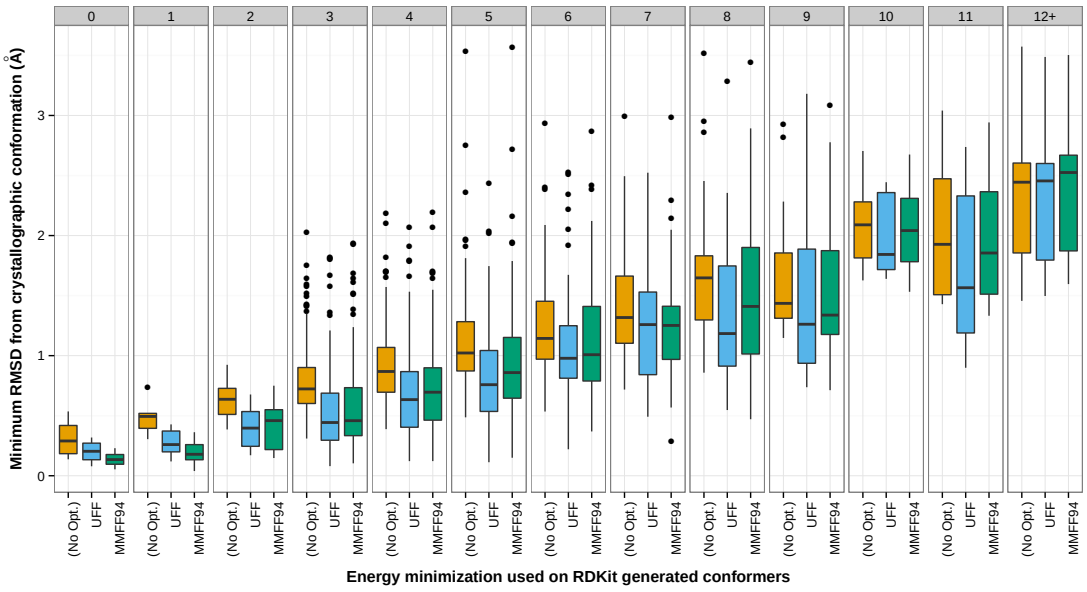


Figure 2.11: Boxplots showing the minimum RMSD from the experimental structure to the RDKit set of 10 generated conformers for every molecule in the test set, partitioned by number of rotatable bonds. We show the results for conformers with no optimisation (yellow), and with UFF (blue) and MMFF94 minimization (green).

structure compared to the energies of the conformers generated (shown in Figure 2.12). The prevalence of positive values on this histogram shows that the experimental structures have higher energies than the conformers generated. This may indicate an inherent deficiency in these force fields, that the force field used when refining the crystal structure is different, or could be due to crystal packing effects or constraints placed by the bound protein when the ligand crystal structure is taken from a complex.

Generating ten conformers for every molecule in the data set using RDKit and without any energy minimization took 1 minute 46 seconds on the machine described in the previous section. UFF minimization on these molecules took 8 minutes 2 seconds while MMFF94 minimization took 113 minutes 10 seconds. Considering energy minimization makes up the most computationally expensive operation in the conformer generation workflow it may be beneficial to optimise the energy minimization parameters (*e.g.* maximum iterations or force tolerance).

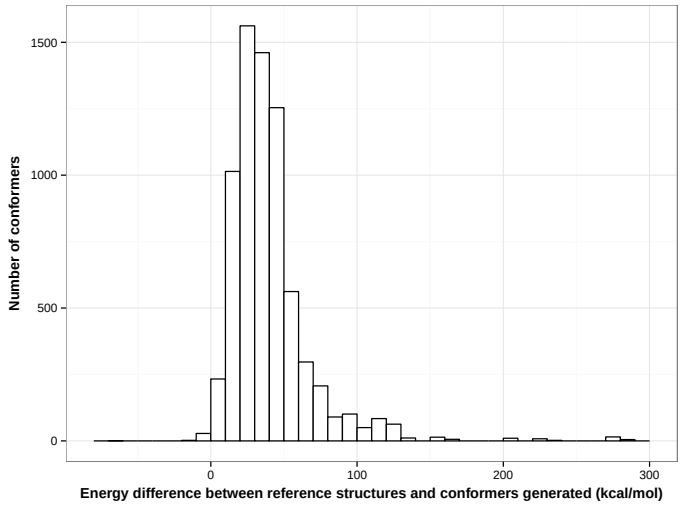


Figure 2.12: The histogram showing the number of conformers versus the UFF-calculated energy difference between the crystal structure and the ten RDKit generated conformers for each molecule in the dataset (7,080 data points). The positive values show that the energy of the crystal structures is typically higher than that of the energy minimized conformers.

2.3.6 RDKit – Conformer Generation Post-Processing

The positive results for RDKit in terms of accuracy and speed make it a viable candidate for inclusion in a computational drug discovery project.

One of the main issues with RDKit is that the number of conformers generated must be specified. This means that for less flexible molecules with a small number of (or even no) rotatable bonds, most conformers generated will be similar to each other. This has important repercussions for both the space needed to store the conformers as well as processing time.

RDKit has an option to keep only conformations that are at least a particular RMSD threshold apart from one another (`pruneRmsThresh`) but this gives distinct conformers before the force field is applied. Theoretically, performing a force field-based energy minimization after the filtering might cause two different conformers close to one another in conformational space to fall into the same local energy minimum and become structurally very similar after energy minimization.

We therefore developed the following alternative approach for filtering the conformers

generated by RDKit to resolve this conformational diversity problem:

1. Using RDKit, generate n conformers in set C_{gen}.
2. Energy minimization (using the UFF force field) is performed on every conformer. The conformer list is sorted by increasing energy value and the lowest energy conformer (the first conformer in the list), c_{low}, is recorded.
3. Remove c_{low} from C_{gen} and add it to C_{keep}
4. For each conformer, c, in C_{gen}, compute the RMSD between c and each conformer in C_{keep}
   (a) If any RMSD value is smaller than a fixed threshold, d_{min}, discard c as we already have a representative of that point in conformational space.
   (b) Otherwise add c to C_{keep}

There is ongoing debate as to whether bioactive conformations lie close to an energy minimum [Butler et al., 2009] or are significantly above it [Nicklaus et al., 1995; Perola and Charifson, 2004]. We use the lowest energy conformer as an initial starting point for the sampling of conformational space.

The advantage of sorting the list by increasing energy values is that this will give us the lowest energy conformer in the d_{min} partition.

At the end of the above procedure the set C_{keep} will contain the lowest energy conformer and all its elements will be at least d_{min} Å RMSD apart.

We have performed several experiments to determine the optimal values for the parameters specifying the initial number of conformers to be generated (n) and the minimum threshold distance between each conformer kept (d_{min}). The value of n is a function of the number of rotatable bonds. As expected, more flexible molecules require a larger value of n to cover sufficiently the conformational space.

A good starting value of n is one which still generates a conformer that is similar to the X-ray crystallographically determined structure. Figure 2.13 shows the minimum RMSD to the experimentally determined structure versus multiple RDKit runs that generate

10, 50, 100, 200, 300, 400, 500 and 1000 conformers for each molecule in our test set (partitioned by number of rotatable bonds). Based on Figure 2.13, plausible values for n could be for $n_{rot} \leq 7$, use $n = 50$; for $8 \leq n_{rot} \leq 12$, use $n = 200$; and for $n_{rot} \geq 13$, use $n = 300$.

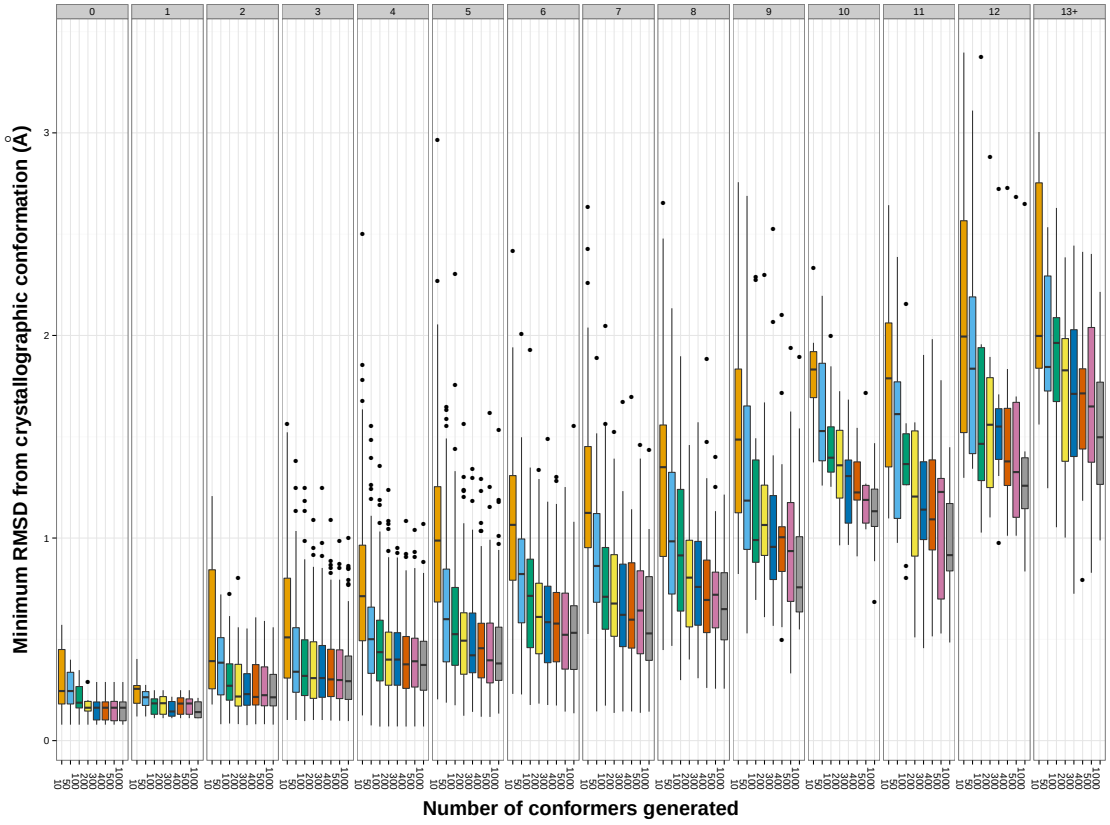
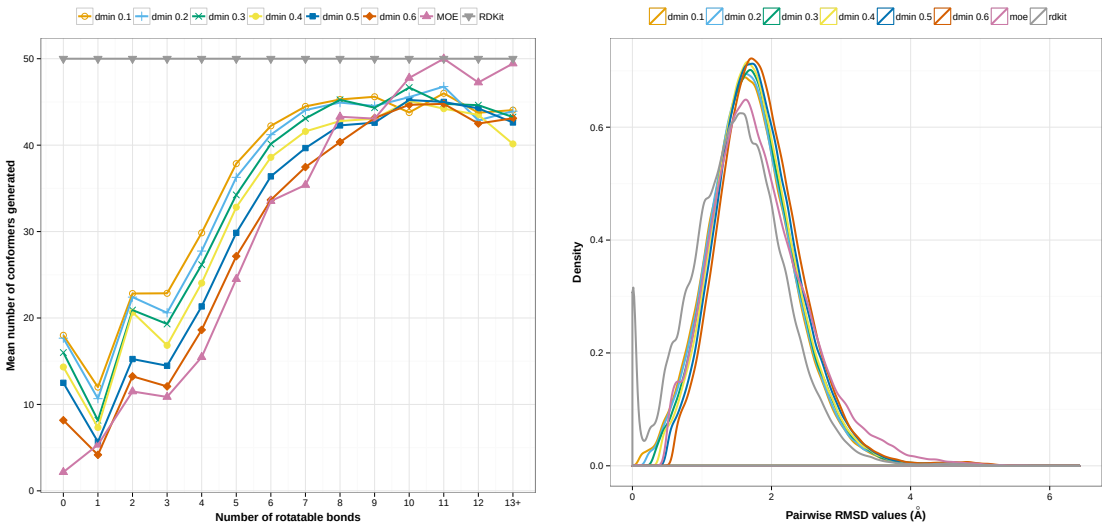


Figure 2.13: Variation in minimum crystallographic RMSD versus the number of rotatable bonds in the ligand, and how these values change with number of conformers generated using RDKit.

We also estimated the value of the clustering threshold, d_{min}, and studied its effects on the number of filtered conformers. Figure 2.14 shows that as expected the larger the value of d_{min}, the more conformers are filtered.

When applying our post-processing algorithm on RDKit, using $n = 50$ and $d_{min} = 0.35$ Å we obtain distinct conformers and the initial peak which can be seen in Figure 2.14(b) for RDKit disappears. This is because similar molecular geometries are filtered out from the generated conformer set. It can also be seen that RDKit with filtering behaves similarly to MOE in terms of number of conformers generated and conformer



(a) This graph shows the average number of conformers generated for each rotatable bond subset using the RDKit-based algorithm presented here. Note that increasing the d_{min} threshold filters out more conformers (if d_{min} was large enough only one conformer would be left after post-processing).

(b) This graph shows the density distribution of the pairwise RMSD values between conformers of a molecule. Note also that the initial peak of very similar conformers for RDKit is removed using the post processing algorithm.

Figure 2.14: d_{min} parameter investigation

diversity.

So, using the number of conformers to generate (n) as:

$$n = \begin{cases} 50 & \text{if } n_{rot} \leq 7 \\ 200 & \text{if } n_{rot} \geq 8 \text{ and } n_{rot} \leq 12 \\ 300 & \text{otherwise} \end{cases}$$

and a d_{min} value of 0.35 Å, the total CPU running time for the whole test set was 102 minutes. This is significantly faster than using the one-size-fits-all value of 300 conformers per molecule without compromising the quality of the results. The minimum crystallographic RMSD distributions were similar to those shown for the 50, 200 and 300 conformer generation runs in Figure 2.13.

2.4 Conclusions

We have reviewed the performance of four free and/or open source conformer generation software packages: Balloon, Confab, Frog2 and RDKit, and compared them to the Conformation Import method implemented in the commercially-available package MOE.

We are interested in three measures, specifically the ability of these tools to generate a conformation close to the experimentally observed structure; the coverage of the conformational space of a molecule; and the performance of these tools in terms of speed. These are critical aspects of the computational drug discovery process.

For our benchmarks, we rebuilt the dataset used to validate another popular commercially available conformer generation toolkit, Omega, and augmented it with the ligands present in the Astex Diverse Set. The resultant dataset consists of 708 molecules from the PDB and CSD with high resolution X-ray crystal structures and which are mostly drug-like in their properties.

When considering the ability to generate conformers which are structurally similar to the experimentally determined structures, we have found that both RDKit and Confab do better than the other toolkits, with the latter performing better with more flexible molecules (*i.e.* ≥ 10 rotatable bonds). This can be attributed to the systematic exploration of the conformer space as opposed RDKit's stochastic approach.

When analysing the ability of a method to explore or "cover" conformational space (by measuring the pairwise RMSD between each conformer generated) RDKit tends to generate more similar conformers than the other methods. We presented a post-processing algorithm we developed to filter out similar structures from the RDKit output, using the lowest energy conformer as the starting point of the conformational space sampling.

In terms of speed, Frog2 was the fastest conformer generator by an order of magnitude and is only slightly affected by the number of conformers generated. After that, RDKit is significantly faster than the other toolkits.

Finally, the choice of a conformer generation tool depends on a number of factors other than the primary ones considered above, *e.g.* ability to explore the energetic land-

scape, the ability to integrate with other software (either through source code or in a workflow), licensing model and pricing. Even in conformer generation, open source tools offer a viable alternative to commercial, closed source, proprietary software.

3 Building a Small-Molecule Database

In this chapter we discuss the development of a small-molecule database, Scopius-CSpace, for use in prospective virtual screening (VS) experiments. First, we describe the motivation for this work and we give a brief overview of existing compound databases. Second, we outline the steps to build such a database. Last, we describe the database structure and the tools used to access it.

This work has been partially carried out using cloud computing. In addition to this chapter, the interested reader is directed to a complementary description of how we built a virtual library of ~28 million molecules in the cloud in our topical perspective: **The emerging role of cloud computing in molecular modelling**; Jean-Paul Ebejer, Simone Fulle, Garrett M. Morris, Paul W. Finn; *Journal of Molecular Graphics and Modelling*, 44C:177–187, 2013.

3.1 Background

Small-molecule databases have wide-ranging applications in both organic chemistry and biology [Chen et al., 2005]. In chemistry, geometric data from experimentally determined structural databases is used to validate newly resolved molecules with similar structures [Spek, 2009]. Also, conformational analysis can make use of torsion angle distributions found in these databases to increase the confidence in the generated structures [Bruno

et al., 2004]. Small-molecule databases may also be used to supply building blocks in combinatorial libraries for use in chemogenomics [Agrafiotis et al., 2002]. From a biological perspective, small-molecules are used as modulators to systematically explore 'biological-activity space' [Stockwell, 2004]. In biomedical and pharmaceutical research, small-molecule databases are used *in silico* for: (i) virtual screening, (ii) drug target identification, (iii) drug design, (iv) drug repurposing, and (v) the prediction of drug-drug interactions, pharmacokinetics and pharmacodynamics [Wishart et al., 2006].

3.1.1 Current Small-Molecule Databases

A number of currently available small-molecule databases are presented in Table 3.1. Several other small-molecule databases exist, but almost all of these suffer from one or more of the following limitations.

A chemical supplier makes available a compound catalogue (database) which lists only the molecules that the vendor is able to synthesize (or has available in stock). These catalogues are usually chemically biased depending on the supplier's experience, synthetic chemistry abilities, chemical manufacturing process (including range of equipment) and customer base. Compound brokers, like eMolecules or MolPort, alleviate this bias by collecting and collating catalogues from a large number of vendors.

Some of the current small-molecules databases available are specific to only a few biological targets and are, therefore, limited in general applicability (*e.g.* ChemDB HIV, Opportunistic Infection and Tuberculosis Therapeutics Database [NIAID, online]). Some databases have specific applications (*e.g.* NIST Chemical Kinetics Database contains small molecules from reactants and products in gas-phase reactions [Manion et al.] or BRENDA which is an enzyme information system which also contains enzyme-ligand interactions [Schomburg et al., 2013]). Others only hold information on a specific class of small molecules (*e.g.* lipids or metabolites).

Some small-molecule databases suffer from ill-maintenance; some are rarely updated or out-dated, others are poorly maintained and/or undocumented. Some have missing or incomplete information which limits their use in a VS setting (*e.g.* 3D data). In others, the chemical data is not curated and of low quality (*e.g.* pentavalent carbons,

missing formal charges, using the Hill system to define chemical formulas, improper use of stereochemistry or not addressing stereochemistry at all, different ionization states and not removing duplicate structures thereby inflating database sizes). Some databases lack proper querying and chemistry-aware functionality. Some chemical information is stored in repositories that lack the necessary format to query and/or retrieve the data (an example of this may be seen in Wikipedia's description of compound 12-Crown-4[1], which contains all the information pertaining to this compound in a human-readable format but has no obvious way how to query it electronically).

Some databases have limited or no information on the compound's commercial availability, a critical aspect of VS if an organic synthesis resource is not available.

We developed a small-molecule database of commercially available compounds, named Scopius-CSpace version 6, which addresses some of the above limitations. Our goals when building this database were threefold:

1. **consistency**: to represent all molecules in the database in a uniform way (*e.g.* same ionization state) to make it easier to search the database and to make the search results more consistent.

2. **validation**: all molecules stored in the database are of high quality and error-free (*e.g.* correct valency, low-energy conformers, *etc.*).

3. **semantic mark-up**: allow molecules in the database to be annotated or 'tagged' to highlight an important property (*e.g.* 'drug-like' or 'commercially-available').

This database is a collection of non-redundant, standardized and high-quality chemical structures, including stereochemistry and 3D conformer ensembles. It has pre-calculated molecular properties, fingerprints and ElectroShape descriptors which are used for fast molecular similarity, substructure and exact searches. The molecules are stored in a relational database and we have developed a web-based front-end to allow easy access for non-technical users.

[1] http://en.wikipedia.org/wiki/12-Crown-4

Table 3.1: Some current small-molecule databases.

Small-Molecule Database	Approx. No. of Molecules	Description	Reference
ChEBI	30,000	Database and ontology of chemical entities of biological interest.	Degtyarenko et al. [2008] Hastings et al. [2013]
ChemBank	1,200,000	Stores raw screening data using a rigorous definition of screening experiments in terms of statistical hypothesis testing. Also, related assays are stored into screening projects.	Seiler et al. [2008]
ChEMBL	1,300,000	Bioactive, drug-like small molecules and the related binding data. High quality data is manually abstracted from published literature and is then curated and standardised.	Gaulton et al. [2012]
ChemDB	5,000,000	A database of commercially available compounds from 150 chemical vendors. Machine-learning predictors are available for octanol/water partition coefficient, aqueous solubility and melting point. Is able to search virtual chemical space based on a set of reactions and the readily-available compounds in the database.	Chen et al. [2007]
DrugBank	6,811	Combines drug data (chemical, pharmacological, pharmaceutical) with target data (sequence, structure, pathway)	Wishart et al. [2006]
EDULISS	5,000,000	Structural, physicochemical and pharmacophoric properties of commercially available small molecules with USR searching capabilities.	Hsin et al. [2011]
PDBeChem (formerly MSDChem)	17,015	A database of chemical components (small molecules and monomers) found in the PDB.	Dimitropoulos et al. [2006]
Pubchem Compound	47,000,000	A database of small molecules and their biological activities. Also contains 3D conformer data, ontology classification, literature and patent information.	Bolton et al. [2008]
SMPDB	1,195	Small molecule database of human metabolic pathways.	Frolkis et al. [2010]
STITCH	300,000	Database of publicly available knowledge on protein-chemical interactions (with a confidence score for each interaction).	Kuhn et al. [2012]
ZINC[12]	21,000,000	Database of commercially-available compounds for virtual screening	Irwin and Shoichet [2005] Irwin et al. [2012]

3.1.2 Relational Databases

Relational databases, which are based on set theory, have long been studied in the field of computer science [Codd, 1983]. Relational databases enable the persistent storage of a collection of structured data. This is in stark contrast to the emerging NoSQL database movement which focuses on the storage of unstructured (schema-less) data. Relational databases model entities and their relationships. Each entity (*e.g.* `Molecule`) is represented as a table in the database. The entity's attributes (*e.g.* molecular mass) are stored in the columns of a table. Specific rules, called constraints, define the domain of an attribute and describe the values it may take (*e.g.* a check constraint may be implemented that ensures that the number of rotatable bonds is between zero and ten, inclusive). A tuple represents a particular entity which is being modelled in the database. A tuple is stored as a row in a table.

The collection of tables and their relationships which implement a data model is called a database schema. Figure 3.1 gives an example of a simple database schema with two entities, `Molecule` and `Tag`.

A primary key is used to uniquely identify each row in a table (an entity contains a set of tuples and a set cannot have any duplicates). Any column (or columns) which is designated as the primary key must be unique and not null. Relationships between different entities (*e.g.* a molecule has multiple tags) are modelled via foreign keys. Foreign keys are columns in one table which reference the primary key in another table (*e.g.* in Figure 3.1, the column *MoleculeIdentifier* in table `Tag` is a foreign key to column *Identifier* in table `Molecule`). Foreign keys define a parent-child relationship between two tables. What happens to the child (*e.g.* a tag) when a parent record (*e.g.* a molecule) is deleted or updated depends on propagation constraints, which control the creation of orphan rows. There are three types of propagation constraints: (i) restrict (ii) set null; and (iii) cascade. In the first case, restrict, the database management system will give an error if one tries to delete a molecule which has tags associated to it (via the foreign key). In the second case, set null, the foreign key attribute in the child row is set to null if the parent row is deleted. In the last case, cascade, if the molecule is deleted, the associated tag rows are

Molecule

Identifier	SMILES	Molecular Mass	HBA	HBD	Rotatable Bonds	TPSA
AAAABN	Nc1cc(CC(=O)[O-])ccc1Cl	184.602	3	2	2	66.15
AAAADF	O=Cc1ccc(-c2ccccc2)[nH]1	171.199	2	1	2	32.86
AAAAEP	CC(C)(C)OC(=O)N1CCC(C(N)=S)C1	230.333	4	2	4	55.56
AAAAHZ	[NH3+][C@@H](C(=O)[O-])c1ccncc1	152.153	4	3	2	80.66
...	...	...	...	...	...	...

Tag

Molecule Identifier	Tag
AAAABN	AVL
AAAADF	AVL
AAAAEP	AVL
AAAAHZ	AVL
AAAABN	SCR
AAAADF	SCR
AAAAEP	SCR
AAAAHZ	SCR
AAAAEP	DRG
AAAABN	DRG
...	...

Figure 3.1: Example of a relational database schema. Table `Molecule` has attributes: *Identifier*, *SMILES*, *MolecularMass*, *HBA*, *HBD*, *RotatableBonds*, and *TPSA*. Table `Tag` has attributes: *MoleculeIdentifier* and *Tag*. Primary keys for every table are highlighted with an underline. The relationship between the two tables is through the foreign key *MoleculeIdentifier* in table `Tag` (links to *Identifier* in `Molecule`).

also deleted leading to a cascading effect.

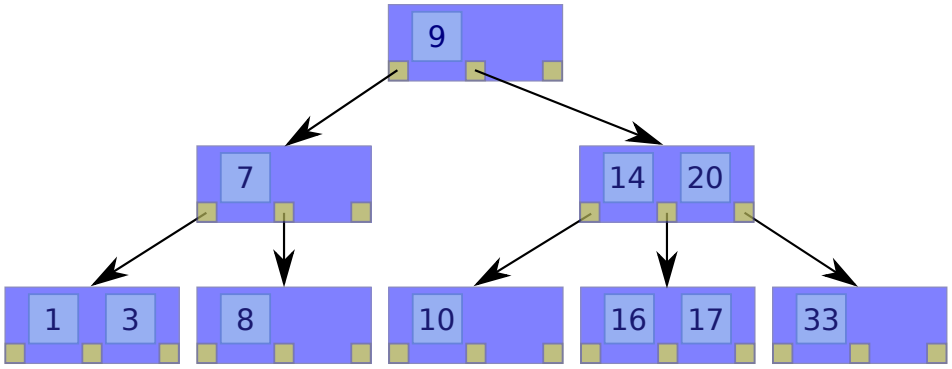
Cardinality (and optionality) of a relationship defines the degree of participation of an entity in a relationship *e.g.* one-to-one, one-to-many, zero-to-many *etc.* Molecules and conformers may have a cardinality of one-to-many, *e.g.* a single molecule may have many conformers. Conversely, a conformer must always have only one molecule (no optionality) associated with it. The cardinality between the molecule and tag entities is many-to-many. One tag, *e.g.* the 'drug-like' tag, is assigned to many molecules and one molecule may have many tags, *e.g.* molecule `AAAABN` has both 'available' and 'drug-like' tags.

Data Access

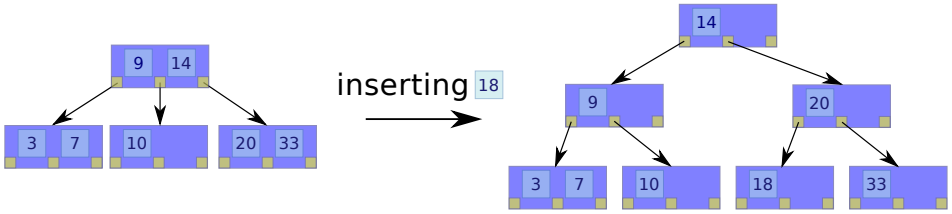
In relational databases, data is accessed through a special-purpose language called Structured Query Language (SQL). SQL's main application is for data manipulation, *i.e.* to create, read, update and delete rows in a table. It also allows for data definition such as the creation of tables and the implementation of check constraints. SQL is used to create indices to speed up database searches. An index is defined on one or more column attributes of a table. B-trees are a common implementation of a database index (this is the default used in our database of choice, PostgreSQL). The defining property of a B-tree is that it is balanced, *i.e.* all leaf nodes have equal depth. An example of B-trees is shown in Figure 3.2a. In B-trees, keys are stored in a sorted manner. Each key has an associated pointer that points to the physical location of the row represented by the key. B-trees have a fixed constant (sometimes defined as the order), n, which defines the maximum number of children for a non-leaf node. A non-leaf node can hold at most $n-1$ keys and n pointers (to other nodes). B-trees have an average search, insert and delete complexity of $\mathcal{O}(log\, n)$. When the index is updated (insertion or deletion of a key), some rearrangement may be required to maintain the balanced property (see Figure 3.2b for an example of an index insertion that triggers a rearrangement).

3.1.3 Clustering of Molecular Structures

Molecular cluster analysis is the process of partitioning a set of molecules (*i.e.* creating subgroups, classes or clusters) based on the similarity of some descriptor. There are four main steps to any clustering process in a molecular context: (i) select and generate a descriptor for every molecule in the dataset (descriptors used in clustering may be based on structural features or a number of molecular 1D properties), (ii) select a similarity measure, (iii) apply clustering method on the similarity scores based on the descriptors, and (iv) analyze the results [Lipkowitz and Boyd, 2003]. Cluster analysis should be distinguished from discriminant analysis in which known classes (*e.g.* active and inactive molecules resulting from biological screening) are used to classify other unknown data points.



(a) An example of a B-tree of order 3. When searching, the query key is compared to the key stored in the node. A decision is then made on whether to traverse the left or right subtree depending on if the query key is smaller or greater than the node key. When the query is equal to the node key or a leaf node has been visited, the search stops.



(b) Inserting key 18 in this B-tree requires the tree to be rearranged. The height of the tree changes from 2 to 3 after the insertion.

Figure 3.2: Examples of a B-tree.

Fraley and Raftery [1998] divide clustering methods into two main types: hierarchical and relocation methods (sometimes referred to as partitional [Jain, 2010]). Many cluster analysis methods exist, and there is divergence on the classification of clustering methods [Rokach and Maimon, 2005]. Han and Kamber [2000] suggest three further categories: density-based methods, grid-based methods and model-based methods. Barnard and Downs [1992] attribute relocation methods as a subset of non-hierarchical methods.

As the name implies, hierarchical methods build a hierarchy out of the set of observations. Two different ways to build this hierarchy are agglomerative and divisive. Agglomerative hierarchical clustering is a bottom-up approach where each observation starts in its own cluster and the most similar clusters are joined together. This happens iteratively until all the clusters are merged together as one, at the root of the hierarchy. Divisive hierarchical clustering is a top-down approach where all observations are

contained in a single cluster and cluster splits occur recursively the deeper the hierarchy is traversed. At the bottom of the hierarchy each observation is in a cluster of its own, *i.e.* the leaf nodes each contain a cluster with one single element. Hierarchical methods have the disadvantage of having (at least) a square running time and memory requirements, which is too onerous for large datasets [Böcker et al., 2005].

Relocation methods start from a given set of initial clusters and move observations iteratively from one cluster to another based on some criterion. Typically, the end objective is to increase the similarity of the objects in one cluster (intra-cluster) while increasing the dissimilarity of the objects between different partitions (inter-cluster). The initial number of clusters does not change. Each cluster must contain at least one observation and each observation must belong to only one cluster (although the latter constraint may be relaxed in fuzzy clustering). A drawback of these algorithms is that the number of clusters must be known in advance.

The main application of molecular clustering is the rational selection (or sampling) of chemical space of compounds from high-throughput screening, combinatorial libraries and external supplier chemical inventories [Lipkowitz and Boyd, 2003; Olah et al., 2004]. Ward's clustering (hierarchical) and Jarvis-Patrick (nearest neighbour approach, non-hierarchical) are two of the most popular clustering algorithms used in cheminformatics [Böcker et al., 2005; Khanna and Ranganathan, 2011]. A comparative study by Brown and Martin [1996] found that Ward's hierarchical agglomerative method was best at separating active and inactive structures based on MACCS key descriptors. However, Ward's methods was the second slowest in this study and the datasets used only consisted of a few thousand molecules raising questions about the application of this approach in a large-scale study.

We cluster the set of drug-like molecules in the database using Stochastic Cluster Analysis (SCA) [Reynolds et al., 1998]. This algorithm has three main advantages over 'traditional' clustering methods. First, some clustering algorithms (*e.g.* k-means) require the number of partitions to be specified. This is difficult to estimate for large molecular datasets. If the number of partitions is too large, a lot of clusters will have a size of one. Conversely, if the number is too small, clusters will contain very diverse structures in the

same cluster. Second, SCA allows addition of molecules to clusters at a later time without restarting the clustering algorithm (a piecemeal approach). Third, SCA is significantly faster than most other clustering algorithms.

SCA works in two steps. First, a diversity step is carried out on the list of compounds to find a diverse set of probes. A compound is selected at random from the list. If it is the first compound, it becomes the first probe. Otherwise, it is compared with the current list of probes. If the compound being tested has a similarity, S, less than an arbitrary similarity threshold, S_C, to all of the current probes the compound is added to the probe list, otherwise it is skipped. This implies that each compound in the probe list will be sufficiently diverse from any other probe. The procedure is repeated until all compounds in the starting list of compounds have been considered. This first step results in a reduced list of diverse probes.

Second, a similarity search is executed. Each non-probe compound is tested for similarity with each probe compound. If $S \geq S_C$ the non-probe compound is clustered with the probe compound. Note that this may lead to compounds being allocated to multiple probe clusters. In that case, compounds are assigned to the cluster of the probe to which the exhibit the greatest similarity (S). This way all compounds in the list are assigned to only one cluster. Note that all non-probe compounds have to be similar to, at least, one probe because of the diversity search in the first step. Some clusters are singletons and contain only the original probe molecule. No attempt is made to aggregate these clusters.

A possible critique of SCA is its stochastic nature – each time the clustering algorithm is run, different results are returned because of the different probe selection. The initial probe selection is dependant on the order in which the molecules are evaluated. Also the probes are not in the centre of the cluster and as such, better representatives of each cluster may be selected.

3.2 Methods and Materials

In this section we describe how we built the new version of Scopius-CSpace. This process is now automated in a generic pipeline which may be applied to any library of molecules. Figure 3.3 describes the main steps of the database assembly.

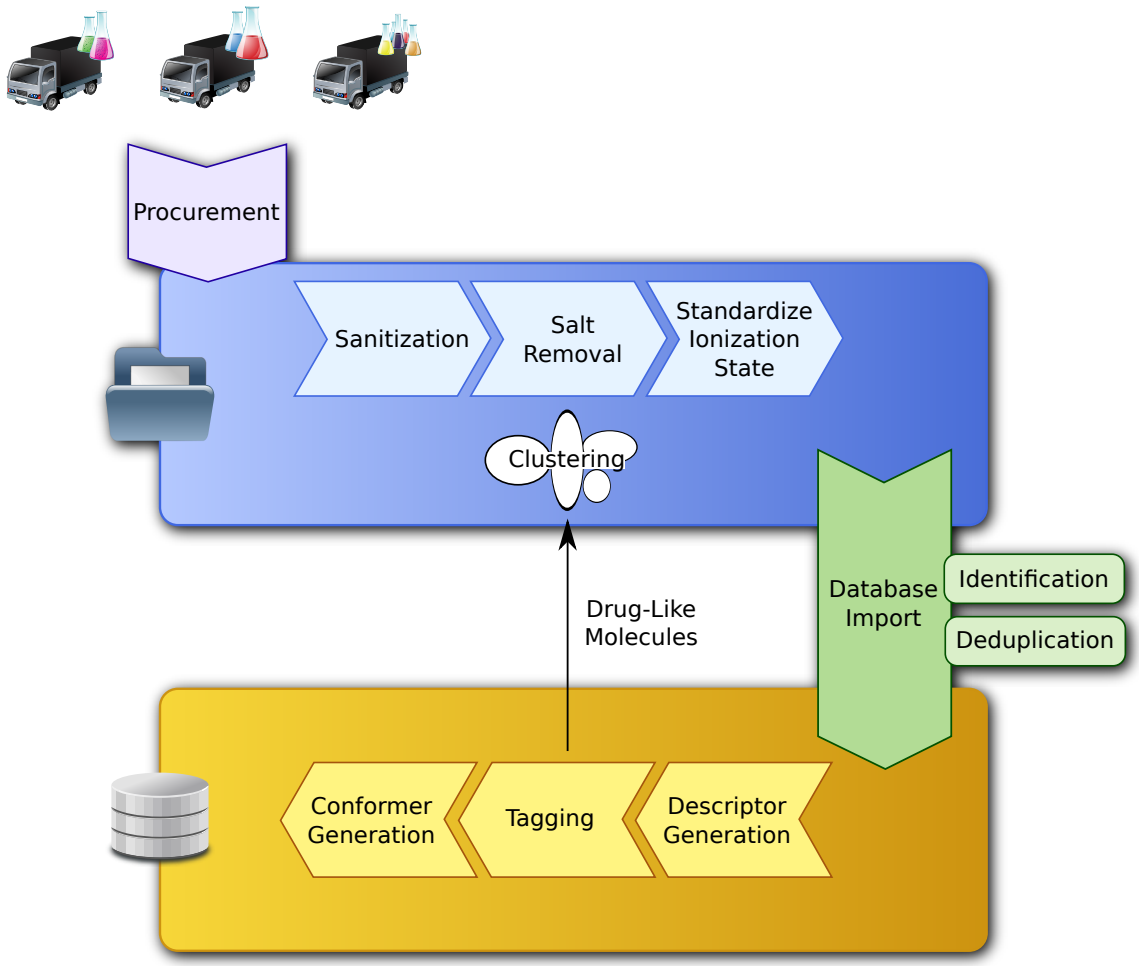


Figure 3.3: Steps in the creation of our multi-million chemical compound database.

All the following steps have been implemented in Python (version 2.7.3) and using RDKit (version 2012.12.1). RDKit is a cheminfomatics toolkit which also provides a SQL database cartridge offering 'molecule aware' functionality in a relational database. The end-product we developed is a relational database which uses PostgreSQL (version 9.1.2). The use of a relational database allows for quick searches using indices on molecular properties, substructure and similarity, and is a noteworthy improvement over

the previous file system storage of molecules (see Section 3.3.6 for more details).

Throughout the implementation of this database we focus on chemical correctness. The main application of this database is drug discovery, and its use in virtual screening studies. Commercial availability is also an important consideration, as chemical synthesis may be a limited resource in small to medium-sized biotech companies.

3.2.1 Molecular Data Procurement

The first step of building a database is to acquire (or procure) the data. The main source of commercially available molecules is, of course, the suppliers who sell them. Molecular files in different file formats are typically available for download from each supplier's website. Each supplier has their own molecule identification and classification scheme. Also, the update frequency and regime (*e.g.* incremental versus full updates) of their catalogues may differ considerably. Suppliers may also have a set of different properties defined in the molecule files. The molecular file formats available for download may also differ greatly; some suppliers offer comma separated value files while others offer 3D molecular formats.

For example, one of the suppliers, MolPort, classifies its 'Available' compounds as 'Shop' or 'Request-for-Quote' compounds. 'Shop' compounds are defined as compounds for quick ordering ("90% of them can be shipped within 10 days of ordering"). The 'Shop' compounds are further divided into 'Building Blocks' and 'Screening Compounds'. 'Request-for-Quote' compounds originate from suppliers that have not integrated with the MolPort electronic procurement platform and therefore require MolPort to request a quote for prices and availability. In 70% of cases this happens within one business day. Additionally, apart from the 'Available' classification, MolPort has two other categories we are not interested in, 'Sold out' and 'Made-To-Order' compounds. The molecules in these categories are not immediately available for off-the-shelf purchase.

Although some of these compound classifications, *e.g. what constitutes a 'Building Block' ?*, are useful and relevant, they vary between suppliers. Therefore, it is more consistent to import all the molecules and apply our own uniform classification across all suppliers. We do this at the tagging stage of our pipeline (Section 3.2.7).

To start with, we download all the molecules available from MolPort. MolPort is a compound supplier broker, and it aggregates compounds from 246 suppliers worldwide. The advantage of using a broker is that we do not have to get the (possibly heterogeneous) molecular data from each individual supplier. We downloaded 13,302,368 molecules in 20 structure data files (SDF) from the 'All available compounds' category via MolPort's FTP site (release 2012-01). Note that although we used only molecules from MolPort to build this database, all the steps described here are vendor-agnostic.

Our pipeline accepts the following file formats: SMILES, SDF, MDL MOL and Tripos Mol2. These are some of the most popular chemical file formats in use today.

A Note on Commercially Available Compounds

Sometimes compounds marked as in-stock or immediately available (off-the-shelf) by some suppliers are still unavailable for purchase. This may happen because the supplier runs out of the compound, or because some suppliers inflate their catalogues with molecules they *could* produce if the demand was high enough. This is especially true for brokers, like eMolecules or MolPort, which may not have updated stock levels for some compound suppliers. In our experience, when ordering hits from virtual screens there are always a handful of compounds that fail to be delivered by the suppliers. This situation is improving as suppliers offer real-time access via web services to their compound stock levels.

3.2.2 Sanitization

We pass all the downloaded molecules through RDKit's sanitization check, and remove the ones which fail. This sanitization step internally includes the following checks and standardizations:

1. The clean-up of substructures to their standardized form (*e.g.* `N(=O)=O` → `[N+](=O)[O-]`). Selecting one 'standard form' over another, allows consistency in the database and confidence that all the relevant compounds will be found in searches.

2. The calculation of implicit and explicit valence of all atoms. Fails on atoms with illegal valency (*e.g.* pentavalent carbon atom)
3. The generation of a symmetrized set of smallest rings.
4. The generation of a Kekulé structure for the molecule. This fails if the Kekulé form cannot be computed or non-ring atoms are marked as aromatic.
5. The assignment of radical counts to each atom.
6. The setting of the aromaticity of the molecule (finds and sets bonds to aromatic).
7. The identification of conjugated bond systems.
8. The assignment of the hybridization state of each atom.
9. The clean-up of atom stereochemistry from non-sp3 centres.
10. The assignment of explicit hydrogen atoms to aromatic heteroatoms (where needed).

This is an important step as most of the applications we developed, including Ligity (described in Chapter 4), use RDKit as their cheminformatics back-end. Unlike other toolkits, RDKit is very rigorous with its input and will fail rather than try to 'guess' a structure. It is better to remove incorrect molecules than let some molecular modelling software interpret a 'broken' molecule (which may have a negative impact on the results of a study). Out of the 13,302,368 downloaded molecules, 10,792 were found to be invalid (0.081%).

3.2.3 Salt Removal

Some of the molecules contain salts which have to be removed from the molecule description. These salt containing molecules have more than one component. Components are defined in SMILES using the period symbol (.), *e.g.* `CCO.O` corresponds to ethanol and water. RDKit has a list of 15 SMARTS patterns which define salts. We complemented this with an additional 412 salt-defining SMARTS patterns. This list was compiled by analysing the most common components found in the downloaded molecules. The top

Table 3.2: Top 20 most common components in the downloaded molecules.

SMARTS Pattern	Occurrences	SMARTS Pattern	Occurrences
`Cl`	112,026	`[Cl-]`	33,907
`OC(=O)C(O)=O`	23,544	`[Br-]`	12,844
`Br`	9,488	`[Na+]`	8,898
`[I-]`	6,837	`O`	6,091
`[K+]`	3,360	`[O-][Cl+3]([O-])([O-])[O-]`	3,106
`F[B-](F)(F)F`	2,067	`CC(O)=O`	1,306
`OS(O)(=O)=O`	1,023	`I`	880
`[F-]`	851	`OC(=O)C=CC(O)=O`	672
`[Li+]`	650	`CN(C)C=O`	610

20 most common components and their number of occurrences are shown in Table 3.2. The output of the salt removal stage is a list of molecules with only one component. If a molecule still contains multiple components after the salt removal, only the largest component is kept for that molecule. In 4,330 cases, the salt removal step removes all the components in a molecule. Some examples of this are magnesium difluoride (`[F-].[F-].[Mg++]`) and lithium chloride (`[Li+].[Cl-]`). These molecules were removed from the dataset, leaving 13,287,246 molecules.

3.2.4 Standardization of the Ionization State

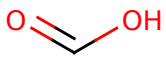
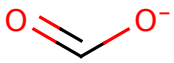
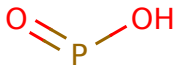
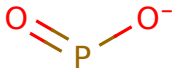
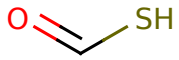
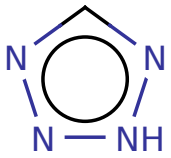
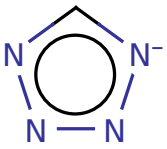
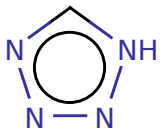
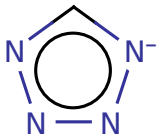
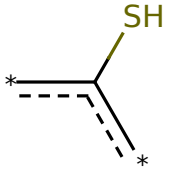
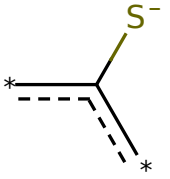
The predominant pH of biological tissues is 7.4, although there are some exceptions (*e.g.* human skin has a pH of 5.5). The ionization (charge) state of a molecule depends on the pH and therefore we standardize the ionization state of all molecules to the species we expect to be most prevalent at pH 7.4. This makes searching the database consistent, because all of the molecules are 'standardized' to biological pH, irrespective of the original ionization state found in the suppliers' catalogue.

Ideally, for virtual screening and other modelling studies, the ionization state of a molecule would be calculated based on the location (and therefore local pH) of the biological target. In principle it is possible to do this by generating the major species of

each molecule at additional pH values, but this would create a large number of redundant structures in the database and is further complicated by the fact that current methods for calculating pK_a are not very accurate. Therefore we do not currently support pH dependent ionization states for a molecule in the database, although this is a possibility for future work.

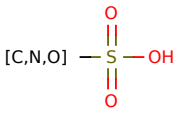
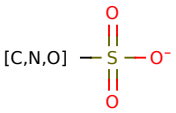
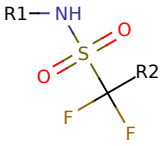
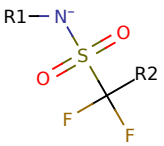
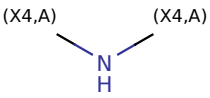
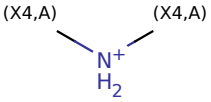
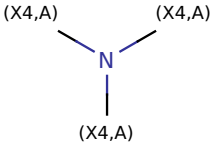
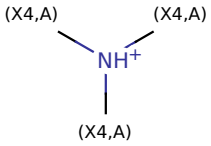
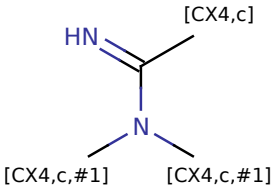
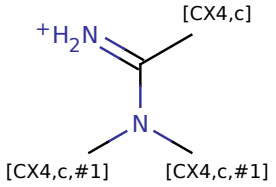
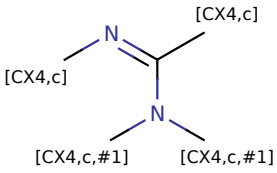
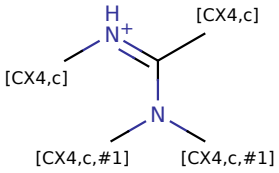
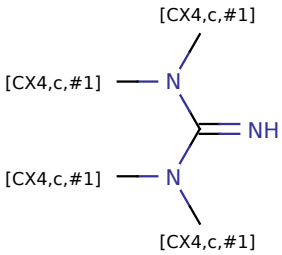
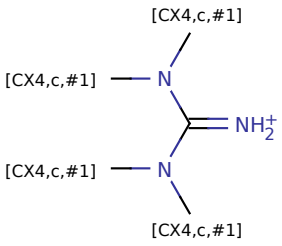
We apply the rules in Table 3.3 to every molecule in the database, generating a standard ionization state for the molecules which we temporarily hold in the file system. Although the ionization rules reported here are not exhaustive, they cover the most common chemical moieties found in drugs.

Table 3.3: Ionization rules. Note that this list was compiled by Paul Finn at InhibOx (private correspondence).

Rule Name	Match	Results
Carboxylic Acid		
Phosphoric Acid		
Thiocarboxylic Acid		
Tetrazole (Tautomer 1)		
Tetrazole (Tautomer 2)		
Aromatic Thiol		

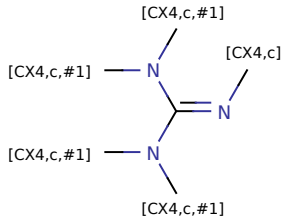
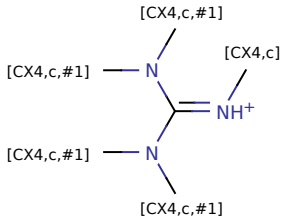
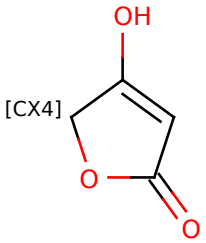
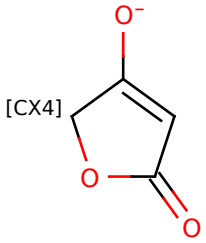
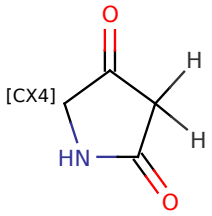
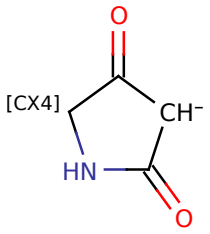
Continued on next page...

Table 3.3 – continued from previous page

Rule Name	Match	Results
Sulphate		
Activated Sulphonamides		
Primary Amines		
Secondary Amines		
Tertiary Amines		
Amidines (1)		
Amidines (2)		
Guanidines (1)		

Continued on next page...

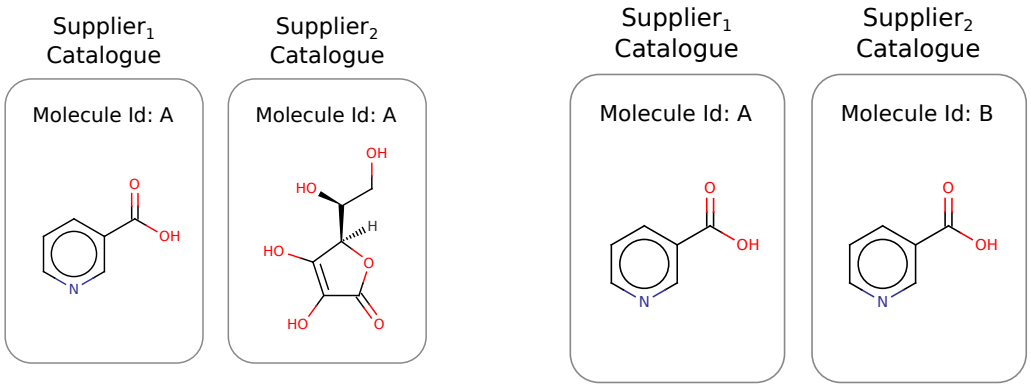
Table 3.3 – continued from previous page

Rule Name	Match	Results
Guanidines (2)		
Ascorbic acid-like		
Tetramic acid-like		

3.2.5 Database Import

The sanitization, salt removal and assignment of ionization state steps described above all process text files as input/output. In the next step, we imported the molecules contained in these files into the PostgreSQL database. The database, together with the RDKit database cartridge, enables fast exact, substructure and similarity searches. It also allows random access and hence fast retrieval of molecules using database indices (*e.g.* B-trees). Chemical data is well-defined and structured and therefore fits well in a relational data model as used by PostgreSQL. This chemical data includes physicochemical properties, molecular structures (both 2D and 3D), supplier information, and meta-data (such as tags).

The state of the molecule becomes immutable once it is imported in the database. We imported the molecules from the file system into the SQL database. This presents us



(a) Different suppliers using the same molecule identifier for different molecules.

(b) Different suppliers using different molecule identifiers for the same molecule.

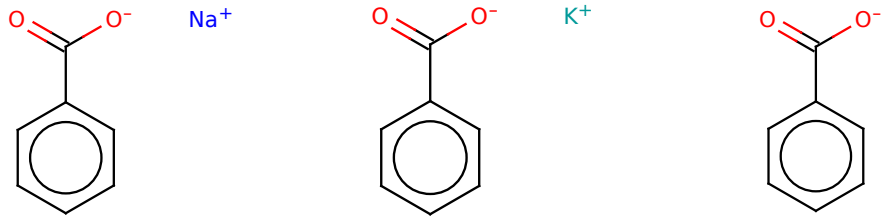
Figure 3.4: Issues with molecules identification across different suppliers. The need for our own identification scheme arises from these two issues.

with two main issues: identification and deduplication of molecules.

Identification

Molecules each have their own identifiers originating from the supplier's compound catalogues (*e.g.* MolPort-002-015-711). The database is meant to support multiple suppliers, where each supplier will have its own molecule naming scheme (identifier). This could lead to two problems, as shown in Figure 3.4. Different suppliers may use the same identifier for different molecules and different suppliers may use different identifiers for the same molecule. The former is a much rarer occurrence than the latter. In order to circumvent these problems we set up our own molecule identification scheme.

Every 'parent' form of a molecule will have a six letter identifiers ($26^6 = 308,915,776$ possible combinations). The definition of a 'parent' form is that of a molecule without a counter-ion or salt (more details on this in the next section). Identifiers are case insensitive (*e.g.* `abcdef` and `ABCDEF` are the same molecule). Identifiers are generated and assigned to molecules in sequence, *i.e.* starting from `AAAAAA`, `AAAAAB` *etc.* Our naming convention also postfixes the conformer identifier to the molecule identifier (*e.g.* `AAAAAA_1` is the first conformer of molecule `AAAAAA`).



(a) Sodium Benzoate **(b)** Potassium Benzoate **(c)** Benzoate

Figure 3.5: An example of two molecules that, after salt removal, will result in the identical 'parent' form. This shows how duplicate entries may be created in the database.

Deduplication

The sanitization, salt removal and assignment of ionization state steps may produce duplicates in the molecule dataset. This may happen when different reagents are used during chemical syntheses which produce the same end-product but different disconnected salts. An example of this are sodium benzoate and potassium benzoate (Figure 3.5). These are two separate molecules in MolPort (with identifiers `MolPort-002-317-245` and `MolPort-005-935-811` respectively). When salts (`Na+` and `K+`) are removed from these molecules, the end product (*i.e.* benzoate) is the same for both cases. In this case benzoic acid is the parent form of these two molecules, and it is the structure we are interested in and which is saved in the database. Deduplication removes the extra copies of the same molecule.

Duplicates may arise from a single supplier's catalogue (*e.g.* a compound broker who gets the same molecule from more than one supplier), multiple suppliers synthesizing the same compound, and also during the execution of pipeline steps described previously. Duplicates could be filtered after each step in the pipeline, and this would avoid, for example, running ionization on identical copies of a molecule. However the deduplication process is computationally intensive and removes only a few molecules at this stage, so it is acceptable to run it once at the database import stage. Duplicate molecules are removed because this redundancy makes all subsequent processing steps slower (*e.g.* running conformer generation a number of times on duplicate molecules), and increases the storage requirements of the database (*e.g.* multiple conformer ensembles and multiple descriptors for the same molecule).

Each molecule is read from the file system and a check is made if it already exists in the database. This check is made by comparing two different representations of the molecule to be imported (a stereochemistry-aware, isomeric canonical SMILES and an InChI key without hydrogen layer) to these same descriptors generated for the molecules currently in the database. If a molecule does not yet exist in the database, its descriptors are computed, a new identifier is generated and the molecule entry is inserted in the database. If a molecule already exists in the database, a database record is created to link the database identifier to the supplier identifier. This is useful when ordering compounds, *e.g.* to find which supplier sells the compound at the cheapest price. Note, that the link between the new identifiers and the original supplier's identifiers is also stored in the database. As a performance optimization, the database indices are created after the initial database import has completed.

The end result of all these steps is an easily accessible chemical database of 12,812,373 commercially-available, non-redundant compounds.

3.2.6 Descriptor Generation

For every molecule in the database we generate fingerprints to use during exact, substructure and similarity searches. Fingerprints are stored in the database and, together with fast indexing technologies on these attributes, allow the database to be easily and readily searched. We generate five different types of fingerprints: (i) atom-pair (binary) (ii) topological-torsion (binary) (iii) Morgan fingerprint using atom neighbourhood chemical-feature invariants (binary, radius 2, FCFP-like) (iv) Morgan fingerprint using atom neighbourhood connectivity invariants (binary, radius 2, ECFP-like); and (v) Morgan fingerprint using atom neighbourhood connectivity invariants (counts, radius 2, ECFP-like).

The following physicochemical descriptors are also generated and stored in the database: (i) molecular mass, (ii) lipophilicity ($\log P$), (iii) number of hydrogen bond donors, (iv) number of hydrogen bond acceptors, (v) number of atoms, (vi) number of heavy atoms (non-hydrogen), (vii) number of heteroatoms (non-carbon), (viii) number of rotatable bonds, (ix) number of rings, and (x) topological polar surface area. These

descriptors allow to filter molecules according to common VS requirements (*e.g.* 'lead-like' or 'fragment-like').

3.2.7 Tagging

The molecules in the database may be tagged based on physicochemical properties and SMARTS patterns. Tagging enables us to work with subsets of molecules using set theory and logical operators, *e.g.* for a virtual screening study, we only want to consider molecules that are tagged as both lead-like and not toxic, *i.e.* {`lead-like` ∩ $\overline{\texttt{toxic}}$}. These tags may be added on a project-by-project basis. All tags are represented as three letter codes, *e.g.* 'drug-like' is represented with the tag `DRG`.

Initially we define only one tag in our database, 'drug-like'. The 'drug-like' tag is defined by eight physicochemical filters shown in Table 3.4 and using 58 SMARTS patterns.

The eight physicochemical filters are based on the ubiquitous Lipinski's 'Rule of 5' (Ro5) [Lipinski et al., 1997]. These rules of thumb are based on 90-percentile values of 2,000 drugs and candidate drugs (in clinical trials) and are meant to be a good indicator of a small molecule's oral bioavailability (*i.e.* the small molecule's pharmacokinetic properties including absorption). Many refinements and permutations of the Ro5, and of measuring drug-likeness, exist and these have been reviewed by Ursu et al. [2011].

The physicochemical filters (Table 3.4) are complemented with 58 SMARTS patterns which define undesired functional groups that may be toxic (*e.g.* cyanamide) or reactive (*e.g.* epoxide) and are therefore ill-suited as drugs (any molecules containing these unwanted functional groups are not tagged as 'drug-like'). Molecules that only contain atoms in the set {`H,C,N,O,F,S,Cl,Br,I`} are tagged as 'drug-like'. This excludes 'drug-like' tagging for any compounds with heavy metals (*e.g.* `Hg`).

Tagging allows to binary classify a molecule in some dimension. Recently, new quantitative measures for drug-likeness have been developed. One of these measures, called 'Quantitative Estimate of Druglikeness' (QED), gives a quantitative drug-likeness score [0..1] based on multiple desirable properties [Bickerton et al., 2012]. These desirable properties are weighted according to underlying property distributions of a well curated

Table 3.4: Physicochemical filters defining 'Drug-Likeness' in the database.

Property	Limits
Molecular Mass	≤ 500 Da
Number of Rotatable Bonds	≤ 7
Lipophilicity	-2.0 $\leq \log P \leq 5.0$
Number of Heavy Atoms	≤ 36
Number of Hydrogen Bond Donors	≤ 5
Number of Hydrogen Bond Acceptors	≤ 10
Number of Carbon atoms	≥ 3
Number of Isotopes	$= 0$

collection of 771 orally approved drugs.

The continuous nature of QED is richer than the binary ('drug-like' or 'not drug-like') classification of our tagging system. Top molecules resulting for a virtual screening study may be sorted (or filtered) in terms of drug-likeness giving a potentially better ranking. Also, promising molecules that fail just one drug-like filter (*e.g.* eight rotatable bonds) would still be removed from the drug-like set. Finally, all of our rules have equal weighting in defining what constitutes a drug-like molecule. This is perhaps unrealistic. Adding quantitative tags to our database molecules is scope for future work.

3.2.8 Clustering

A reduced set of diverse molecules is useful in an initial, rapid, hit identification virtual screening study. Newly identified, promising hits could then be followed by a more thorough investigation of other molecules in clusters that contain the original hits. Clusters with no originating hits could be discarded. This approach is suggested by Willett et al. [1998].

We clustered the drug-like compound set using SCA. We represent each molecule as a Morgan fingerprint using atom neighbourhood connectivity invariants (counts, radius 2, ECFP-like) and we use this as an input to the SCA clustering. We use an SCA

implementation called SUBSET (version 1.0) [Bienfait, 2001]. The Tanimoto coefficient was used as the distance metric and the similarity threshold (S_C) was set at 0.55. This resulted in 600,531 clusters of the drug-like compounds. We select the probe used from every cluster and use this as a reduced database set.

3.2.9 Conformer Generation

Conformers are generated for the drug-like molecules in the database. These are generated using the protocol described in Chapter 2 [Ebejer et al., 2012]. For molecules with stereochemical properties, if the dominant stereoisomer of a molecule is known by the supplier, it is specified in the molecule's SMILES representation in the suppliers' catalogues. In this case the conformer generation protocol is constrained by the defined stereochemistry and generates the correct (dominant) stereoisomer. If this stereochemical information is not defined in the suppliers' catalogues, then the possible chiral and *cis*/*trans* states are sampled randomly by the conformer generation protocol.

189,638,700 conformers were generated for 6,604,127 drug-like molecules, with an average of approximately 29 conformers/mol. The conformers were stored in the PostgreSQL database, which allows to rapidly extract a 3D conformer set of interest. For the 3D conformer structures in the database we also generate ElectroShape USR, CSR, 4D and 5D binary descriptors [Armstrong et al., 2009, 2010]. These descriptors are stored in the file system (~44 GB of disk space required) and are used for molecular similarity searching.

3.3 Results and Discussion

In this section we discuss the final small-molecule database. We discuss how to access the database, and the pipeline tools we developed. We also highlight possible applications for this new version of Scopius-CSpace.

3.3.1 Database Statistics

The Scopius-CSpace database we developed contains 12,812,373 commercially available, distinct compounds. These molecules have been sanitized, assigned standard ionization states, had their salts removed, and deduplicated. They have been assigned identifiers for consistent access across different suppliers. The molecules, including stereochemical and isomeric information, are stored in the database as 2D canonical SMILES and 3D MDL Molfiles. Out of these ~13 million molecules, 6,604,127 have been tagged as drug-like using filters on physicochemical properties and 58 SMARTS patterns to avoid undesirable functional groups (*e.g.* toxic or reactive). We have generated 189,638,700 3D conformers for the drug-like subset.

The distributions of all the computed physicochemical properties of the database are shown in Figure 3.6. The number of heavy atoms and molecular mass distributions have a bimodal nature. This may be attributed to the underlying imported molecule datasets. MolPort has a 'building blocks' set and a 'screening' set available for download, which both have close-to normally distributed molecular mass with different means (~275 Da for building blocks and ~400 Da for screening set compounds). This causes the resulting bimodal distributions.

3.3.2 Database Schema

The database schema is shown in Figure 3.7. This Entity Relationship Diagram (ERD) shows how the database is logically divided into entities and attributes, together with the relationship between the entities.

Physical Modelling of the Database

Tables 3.5-3.12 present the physical implementation details of the database. All constraints have been implemented once the import was completed (as a speed optimization).

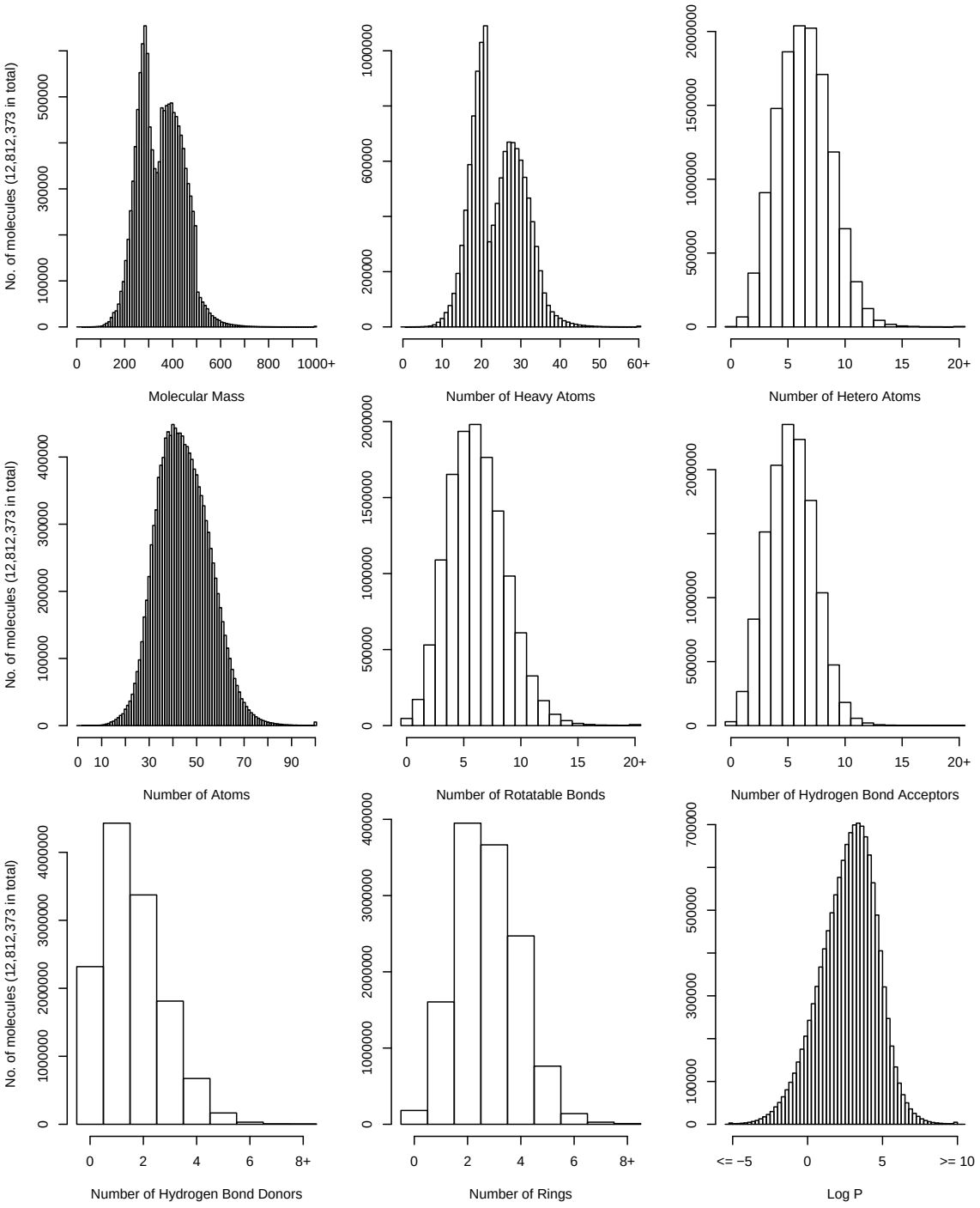


Figure 3.6: Database distributions for molecular mass, number of heavy atoms, number of hetero atoms, number of atoms, number of rotatable bonds, number of hydrogen bond acceptors, number of hydrogen bond donors, number of rings and lipophilicity for all the molecules in the database.

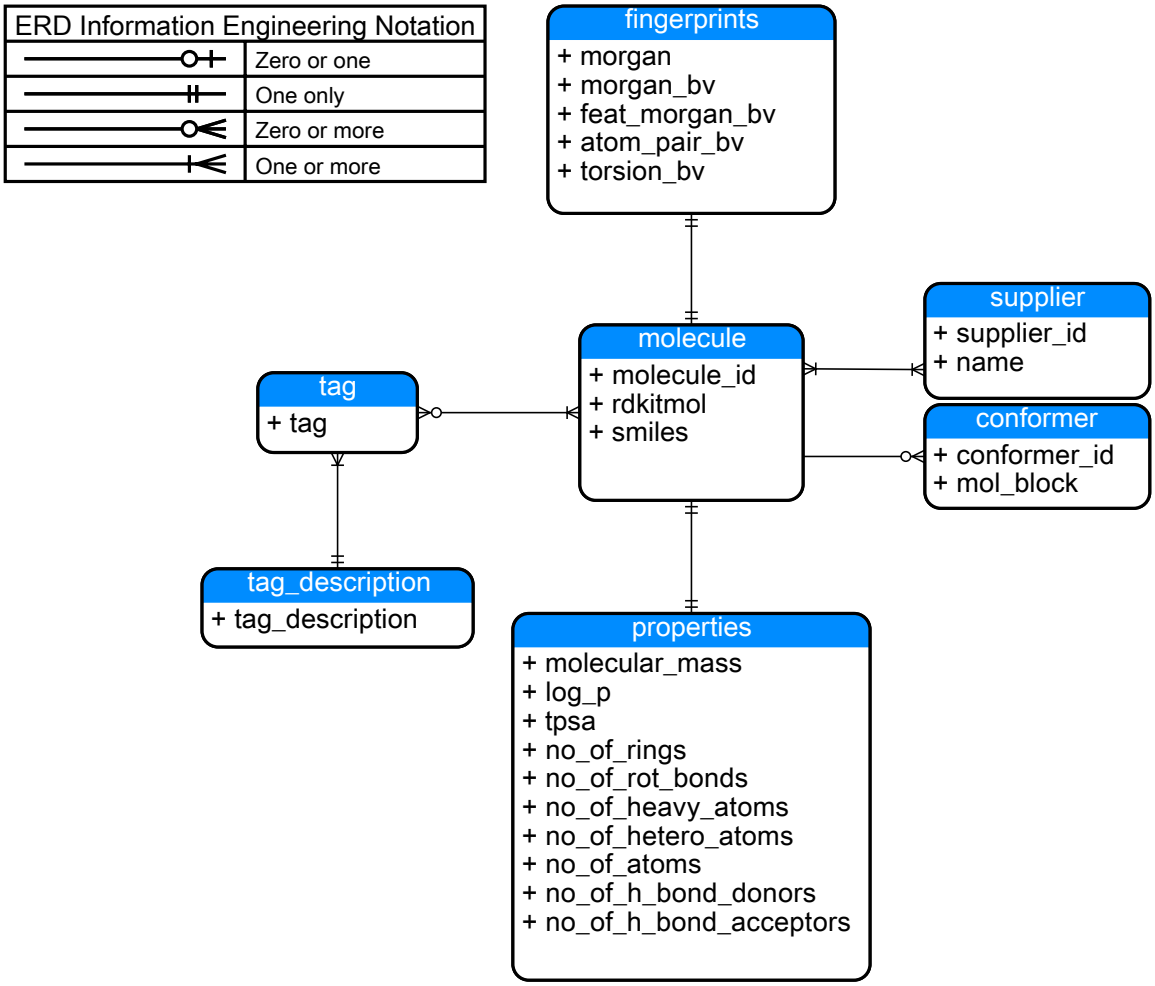


Figure 3.7: Logical Entity Relationship Diagram (ERD) for the Scopius-CSpace database we developed.

Table 3.5: `molecule` database table.

Attribute	Data type	Length	Constraint	Description	Example
`molecule_id`	char	6	Primary Key (Unique and Not Null)	The six letter identifier for the molecule.	AAAAWM
`smiles`	varchar	1000	Unique and Not Null	A textual, canonical representation of the molecule including stereochemistry.	`CC(=O)Nc1cc(F)c(N)cc1F`
`rdkitmol`	mol[†]		Not Null	The molecule instance.	[bytes]

[†] RDKit data type. A RDKit molecule.

Table 3.6: `properties` database table.

Attribute	Data type	Length	Constraint	Description	Example
`molecule_id`	char	6	Primary Key (Unique and Not Null), Foreign Key (references entity `molecule`, on delete cascade)	The six letter identifier for the molecule.	AAAAWM
`mm`	real		Not Null	Molecular Mass.	186.161
`logp`	real		Not Null	Log *P*.	1.505
`hba`	integer		Not Null	Number of Hydrogen Bond Acceptors.	3
`hbd`	integer		Not Null	Number of Hydrogen Bond Donors.	3
`atoms`	integer		Not Null	Number of atoms.	21
`heavy_atoms`	integer		Not Null	Number of heavy atoms.	13
`hetero_atoms`	integer		Not Null	Number of hetero atoms.	5
`rot_bonds`	integer		Not Null	Number of rotational bonds.	2
`rings`	integer		Not Null	Number of rings.	1
`tpsa`	real		Not Null	Topological polar surface area.	55.120

Table 3.7: `fingerprints` database table. Examples are not supplied for this entity because bfp and sfp are stored in hexadecimals in the database.

Attribute	Data type	Length	Constraint	Description
`molecule_id`	char	6	Primary Key (Unique and Not Null), Foreign Key (references entity `molecule`, on delete cascade)	The six letter identifier for the molecule.
`pairbv`	bfp[†]		Not Null, Indexed	Bit vector atom-pair fingerprint.
`torsionbv`	bfp[†]		Not Null, Indexed	Bit vector topological-torsion fingerprint.
`featmorganbv`	bfp[†]		Not Null, Indexed	Bit vector Morgan fingerprint for a molecule using chemical-feature invariants (FCFP-like fingerprint, radius 2).
`morganbv`	bfp[†]		Not Null, Indexed	Bit vector Morgan fingerprint for a molecule using connectivity invariants (ECFP-like fingerprint, radius 2).
`morgan`	sfp[§]		Not Null, Indexed	Count-based Morgan fingerprint for a molecule using connectivity invariants. (ECFP-like fingerprint, radius 2).

[†] RDKit data type. Bit vector fingerprint.
[§] RDKit data type. Sparse count vector fingerprint.

Table 3.8: `tag_descripiton` database table.

Attribute	Data type	Length	Constraint	Description	Example
`tag`	char	3	Primary Key (Unique and Not Null)	Tag identifier.	DRG
`description`	varchar	250	Not Null	Full tag description.	Drug-like compounds

Table 3.9: `tag` database table.

Attribute	Data type	Length	Constraint	Description	Example
`molecule_id`	char	6	Primary Key†; Foreign Key (references entity `molecule`, on delete cascade)	The six letter identifier for the molecule.	AAAAWM
`tag`	char	3	Primary Key†; Foreign Key (references entity `tag_description`, on delete cascade)	Tag for the molecule.	DRG

† Composite key

Table 3.10: `conformer` database table.

Attribute	Data type	Length	Constraint	Description	Example
`molecule_id`	char	6	Primary Key[†], Foreign Key (references entity `molecule`, on delete cascade)	The six letter identifier for the molecule.	AAAAWM
`conformer_id`	char	4	Primary Key[†]	Conformer identifier (sorted by energy minimization).	1
`mol_block`	text		Not Null	Conformer structure formatted as an MDL Molfile.	[MDL Molfile block]

[†] Composite key

Table 3.11: `supplier` database table.

Attribute	Data type	Length	Constraint	Description	Example
`supplier_id`	serial		Primary Key (Unique and Not Null)	Auto-generated integer identifying a supplier.	1
`name`	varchar	1000	Not Null	Supplier's name.	MolPort

Table 3.12: `supplier_id_mapping` database table.

Attribute	Data type	Length	Constraint	Description	Example
`molecule_id`	char	6	Primary Key†, Foreign Key (references entity `molecule`, on delete cascade)	The six letter identifier for the molecule.	AAAAWM
`supplier_mol_id`	varchar	20	Primary Key†, Unique	The identifier used by the supplier for this molecule.	MolPort-000-000-862
`supplier_id`	integer		Foreign Key (references entity `supplier`, on delete cascade), Not Null	The identifier of the supplier of the molecule.	1

† Composite key

3.3.3 Database Access

We have developed a web application to allow easy access to the database back-end for non-technical users. The database may also be accessed directly via the SQL interface by more technical users. The web front-end, called CScape, allows a user to search for molecules using either exact, substructure or similarity searches. Searching may be executed by using the molecule identifier, SMILES or SMARTS expressions or by drawing a chemical structure. Searches using the front-end may be limited to drug-like molecules. For similarity searches the threshold similarity of the molecules returned may be changed by the user. The number of results returned may be limited by the user (this is useful for exploratory work, or, for example, to check that the SMARTS query is correct). An option to download all results as a SMILES file with molecule identifiers is provided. A screenshot of the web front-end is shown in Figure 3.8.

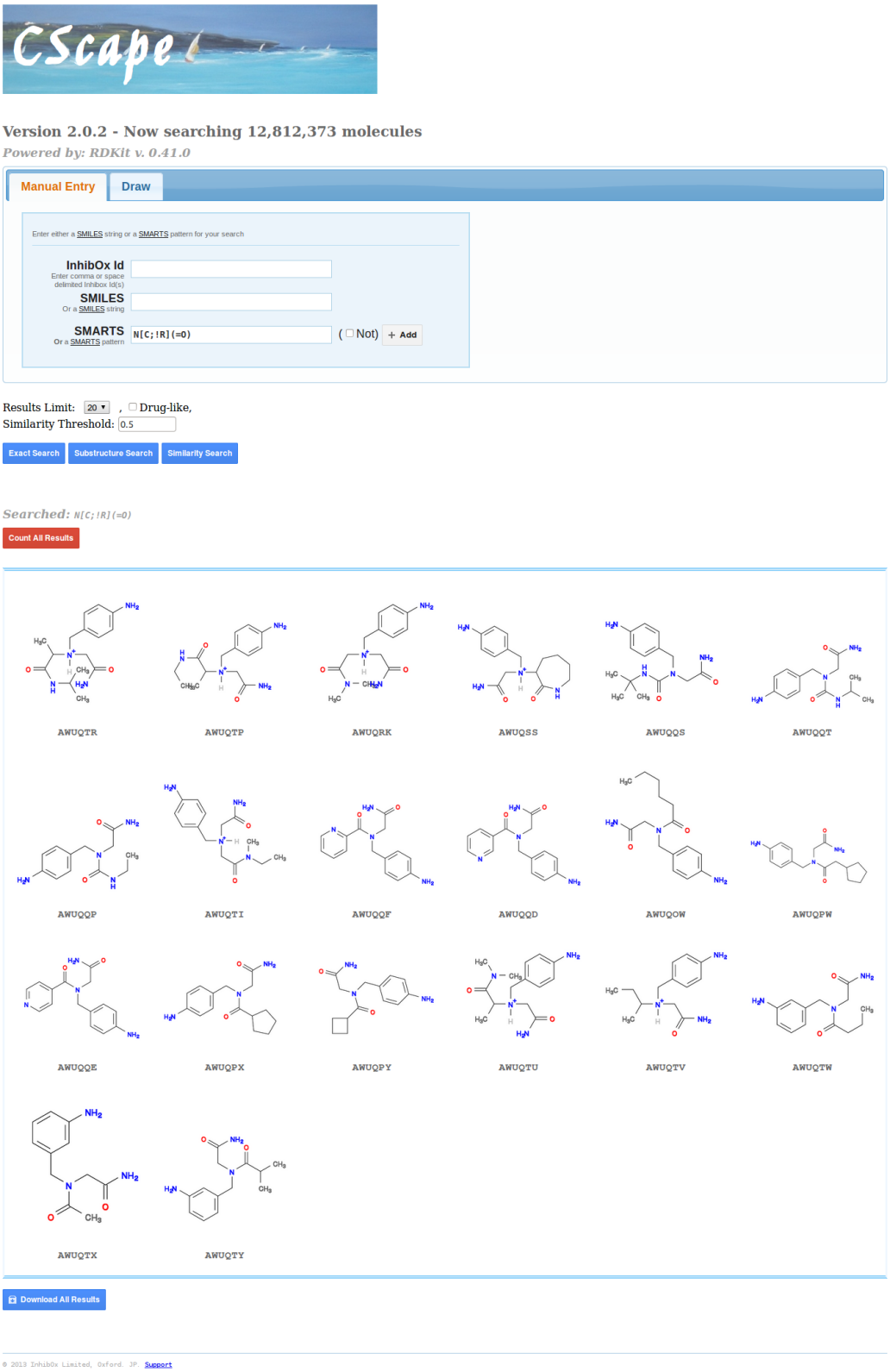


Figure 3.8: CScape web application we developed to access the new version of Scopius-CSpace.

3.3.4 Database Pipeline Software

Table 3.13 lists the main application programs we have developed to construct the database and facilitate its use. Apart from the programs listed in this table, we have developed a number of other utilities which form part of a molecular toolbox. This toolbox, contains utilities to: (i) convert between different formats (*e.g.* `sdf2smi.py` and `smi2sdf.py`), (ii) list molecule names in files, (iii) split molecules files in parts, (iv) find duplicates in a set of molecule files, (v) rename molecules, (vi) filter molecules based on physicochemical properties or SMARTS patterns, (vii) generate various descriptors for molecules (including InChI keys), and (viii) given a list of molecule identifiers, fetch those molecules from the supplier's online database.

3.3.5 On the Cloud

Running the whole pipeline, starting with 13,302,368 molecules, took a few days to complete on a compute cluster with six nodes and 96 processors. This could be easily speeded-up by using cloud computing resources. The database creation tasks may be classified as an 'embarrassingly parallel problem', which means that little or no effort is required to split the job into parallel tasks. This is the case as every molecule in the dataset is independent of all the others. This makes running the database pipeline protocol well suited for cloud computing environments, where individual computing nodes are set to run independently. In a recent article we present a case study of how we built a database of around 30 million compounds using Amazon Web Services (AWS) cloud computing resources (see Section 6.2 from Ebejer et al. [2013]). The software we have developed for this database pipeline is cloud-ready. The most time-consuming tasks of the pipeline, *i.e.* conformer generation, database import and fingerprint generation, can all scale linearly with the number of dedicated computing resources. Indeed, the database may be sharded (tables are partitioned horizontally) and each shard could be processed through the pipeline in parallel.

Table 3.13: Tools developed for database creation and use.

Tool	Description
`sanity.py`	Reads molecules from SMILES, SDF, MDL MOL and Tripos Mol2 files and outputs valid, chemically correct molecules in a new SMILES file. Also writes an error log with invalid molecules.
`desalt.py`	Reads molecules from SMILES files and outputs molecules with only one main chemical component. Removes the salt component from the molecule. If molecule has multiple components that do not match the predefined salt patterns, only the largest component (in terms of the number of heavy atoms) is kept.
`ion.py`	Reads molecules from SMILES files and outputs molecules with a standard ionization (charge) state. The ionization is based on a number of rules defined by SMARTS patterns.
`db_import.py`	Reads molecules from SMILES files and imports them in a PostgreSQL database. As a prerequisite the database schema needs to be created for this program to run. Also assigns identifiers to molecules and deduplicates structures.
`conform.py`	Reads molecules from SMILES, SDF, MDL MOL and Tripos Mol2 files and generates conformer ensembles for each molecule using the protocol described earlier in Chapter 2. Conform uses a multi-step procedure based on RDKit to generate conformational models. The core of the approach is to generate an initial set of 3D coordinates using a distance geometry approach. These coordinates are then optimized using the UFF forcefield. An initial size for the conformational model is based on the number of rotatable bonds. Following energy optimization the conformations are clustered by RMSD to remove similar conformations. An energy cut-off of 10.0 kcal/mol is also employed to prevent high-energy conformations diluting the quality of the conformational model. In addition, knowledge-based heuristics are used to enforce certain geometrical constraints, for example trans-amides.
`conf_import.py`	Reads the conformers generated by Conform and imports them in the database. It assigns an identifier to each conformer based on the molecule identifier and the conformer identifier generated by Conform. Conformer identifiers are sorted based on the UFF energy of the conformer.
`ultra-fetch.py`	Fetches molecules and conformer ensembles from the database. Various options exist that allow the user to retrieve the 2D or 3D molecular data, get only the lowest energy conformer, the full conformer ensemble, use supplier identifiers for searching *etc.*
`fp_gen.py`	Generates the fingerprint descriptors for use in the clustering program, SUBSET.
CScape	Web-based front end written in PHP, which presents a graphical user interface to the database.

3.3.6 Improvements over Previous Version

We have developed version 6 of the Scopius-CSpace small-molecule database. This new version has many advantages over its predecessor, both in terms of quantity and quality of the newly acquired molecular data.

The previous version of Scopius-CSpace (version 5) had 8,447,147 commercially available molecules of which 5,389,545 exhibited drug-like properties. The new version contains 12,812,373 molecules (~52% increase) of which 6,604,127 (~23% increase) exhibited drug-like properties.

The molecules in the previous version were stored in a number of text files in the file system. Retrieving a specific molecule involved parsing all the files until a match is found. This would typically take a few minutes. The new version is stored in a relational database and, using indexing, retrieval of a specific molecule is instantaneous. We precomputed physicochemical properties and fingerprints of molecules and store these in the database, which may be easily queried and searched using SQL's rich syntax. A small subset of physicochemical properties were calculated for the previous version, but these were stored as SDF tags in the molecule files.

The previous database used MOE's `sdwash` to standardize the molecules from the vendor catalogues (*i.e.* the salt removal and ionization steps). We have implemented these steps as independent programs which may be orchestrated together to form a pipeline. Also, in many cases our processes use a more detailed and refined set of rules. For example, MOE uses Lipinski's Ro5 to define drug-like molecules while our drug-like tagging system uses SMART patterns for problematic moieties in addition to the Ro5-like properties. The drug-like portion of the new version of the database is clustered on fingerprint similarity to offer a reduced set of diverse molecules.

No deduplication was carried out in the previous version of Scopius-CSpace, while we remove duplicate molecules in the new version we have developed. We generated the conformers using the protocol described in Chapter 2, which was shown to perform better than MOE (which was used for generating conformers in version 5 of the database).

We developed a web-based front-end, CScape, to easily access the compound database.

CScape allows for exact, substructure and similarity searches which were not immediately possible on the previous version of the database.

3.3.7 The Issue with Tautomers

The main shortcoming of the database is the lack of canonicalizing and enumeration functionality for tautomers. Tautomerism is still a challenging and largely unsolved problem in cheminformatics [Sayle, 2010]. In their recent work, Sitzmann et al. [2010] found that 66% of the 70 million unique chemical structures in the Chemical Structure Database (CSDB), had different tautomeric forms. Also, a total of 680 million tautomers were generated on this dataset (including the original structures). These numbers highlight the scale of the tautomer issue in cheminformatics databases. For more information, the reader is directed to a review by Warr [2010]. In this article, 27 software vendors and database developers were surveyed on how they handle tautomers (if at all).

Different tautomeric forms of a molecule may imply different one dimensional properties such as lipophilicity and pK_a as well as different 3D shape and electrostatic properties. Also, fingerprints might be affected by the molecule's tautomeric state [Martin, 2009]. Different tautomeric states may have a significant influence on the binding of a molecule in a virtual screening exercise. Considering all the possible tautomeric states for every molecule would greatly increase the storage and processing requirements of the database (especially when considering that every tautomeric form should have its own conformers).

Tautomers also produce an undesirable effect in the current database construction process. Duplicates of the same molecules that are not in the same tautomeric form will not be detected by our deduplication program (as the underlying canonical SMILES will be different). Tautomers also affect exact and substructure searches in the database. Searches have to be executed as separate searches by manually entering the different tautomeric forms of the molecule. The tautomer problem is compounded by the fact that tautomers will, in general, not be all equally likely to exist (some tautomers might be more likely than others depending on solvation). Tautomeric states can also be affected by binding site environment. Tautomer handling in the database is scope for future work.

3.3.8 Database Applications

A large database of molecules has multiple applications in cheminformatics. The primary use of this database is for virtual screening studies (we have used it in a prospective study to find novel PfSUB1 inhibitors, see Chapter 5 for more details). The database can be used in both ligand based and structured based virtual screening. Lengthier, more thorough, protocols can make use of the clustered (reduced) version of the drug-like molecules. This diverse set allows for a hierarchical approach, when at a first instance only probes from every cluster are screened and biologically tested and at a second pass the clusters of the most promising hits are analysed.

The 3D structure of the molecules may be used in a pharmacophoric search such as Ligity, a VS method we developed and describe in Chapter 4. The contents of this database may be used to create a library of virtual molecules. Molecules with specific functional groups are extracted from the database and used as reagents to build targeted libraries. This new version of Scopius-CSpace may also be fragmented and used to create a linker database for use in fragment-based drug design.

3.4 Conclusions

We have developed a new version of Scopius-CSpace – a multi-million entry database of commercially available small molecules. Commercial availability is an important aspect as it allows researchers without a synthetic chemistry resource to purchase the compounds for biological testing, the ultimate end-point of any virtual screening study.

The database contains non-redundant structures that have been sanitized, assigned a standard ionization state, stripped of salts, and some of which have been tagged as drug-like. 3D conformer ensembles have also been generated for these molecules. Also, physicochemical descriptors and fingerprints have been calculated for every molecule. ElectroShape descriptors have been calculated for the drug-like set. All the data has been imported in a relational database to enable faster and easier access to the structures. The database functionality includes exact, substructure and similarity structure searching, as

well as searching on physicochemical properties of the molecules. Each molecule has a link to one, or more, suppliers from where it may be purchased.

Apart from the database itself, we have built a set of tools to process compound catalogues which may be assembled into a pipeline to create other databases or targeted molecule libraries. These tools also include a web-based application to access and search the database. This is a valuable asset for non-technical users working with the database.

The underlying relational database technology allows for vast improvements over previous versions of the database. Indexing allows for fast random access to molecular data. The database's size has increased by approximately 50% over the previous version to 12,812,373 molecules (6,604,127 drug-like). The quality of the data has also improved, with stringent sanitization procedures. We also have more complete salt removal and drug-like definition rules.

One of the main applications of this database is virtual screening. Our reduced (clustered), diverse set of molecules (around 10% of the whole dataset) may be used in lengthier and more exhaustive virtual screening protocols.

4 Ligity: a knowledge-based approach to virtual screening

In this chapter we present Ligity, a novel virtual screening method we have developed which makes use of existing *holo* structures in the Protein Data Bank (PDB) to identify bioactive small molecules.

4.1 Background

The background material for this chapter has been presented in the Virtual Screening section in the first chapter of this thesis (for details, please refer to Section 1.3).

4.2 Methods and Materials

In this section we give a detailed description of the Ligity algorithm and of the datasets used for the validation and testing of the method. We also discuss how we measure the performance of the method.

4.2.1 Ligity Algorithm

Ligity is a knowledge-based virtual screening method which uses multiple, existing cognate (or *holo*) protein-ligand complexes as a query to find biologically active molecules

within a large database. A schematic overview of how Ligity works is shown in Figure 4.1. The Ligity suite of programs is written in Python and built upon RDKit (version 2012.12.1), an open-source cheminformatics toolkit [RDKit, online].

The main steps of the algorithm, in a virtual screening context, may be summarized as follows:

1. For the 3D Protein-Ligand complex to use as a **query** to the compound database (note that multiple complexes may be used):
   (a) Standardize the ionization state of the protein and ligand complex
   (b) Find the Pharmacophoric Important Points (PIP) on the cognate ligand side, for each protein-ligand interactions
2. For each molecule in the compound database (typically represented as SMILES):
   (a) Standardize the ionization state of the molecule (using the same rules as for the query protein-ligand complex)
   (b) Generate conformers
   (c) Generate all the PIPs for each conformer
3. For the query and compound database PIPs, generate a descriptor. This descriptor makes use of the distances between 3-PIP or 4-PIP combinations.
4. Generate a similarity score for the query descriptor against each conformer in the compound database. Produce a ranking for each protein-ligand query based on this similarity score.
5. If multiple query complexes are used, fuse the results of each individual ranking results list into one final list.

The details of this algorithm will be explained in the next sections.

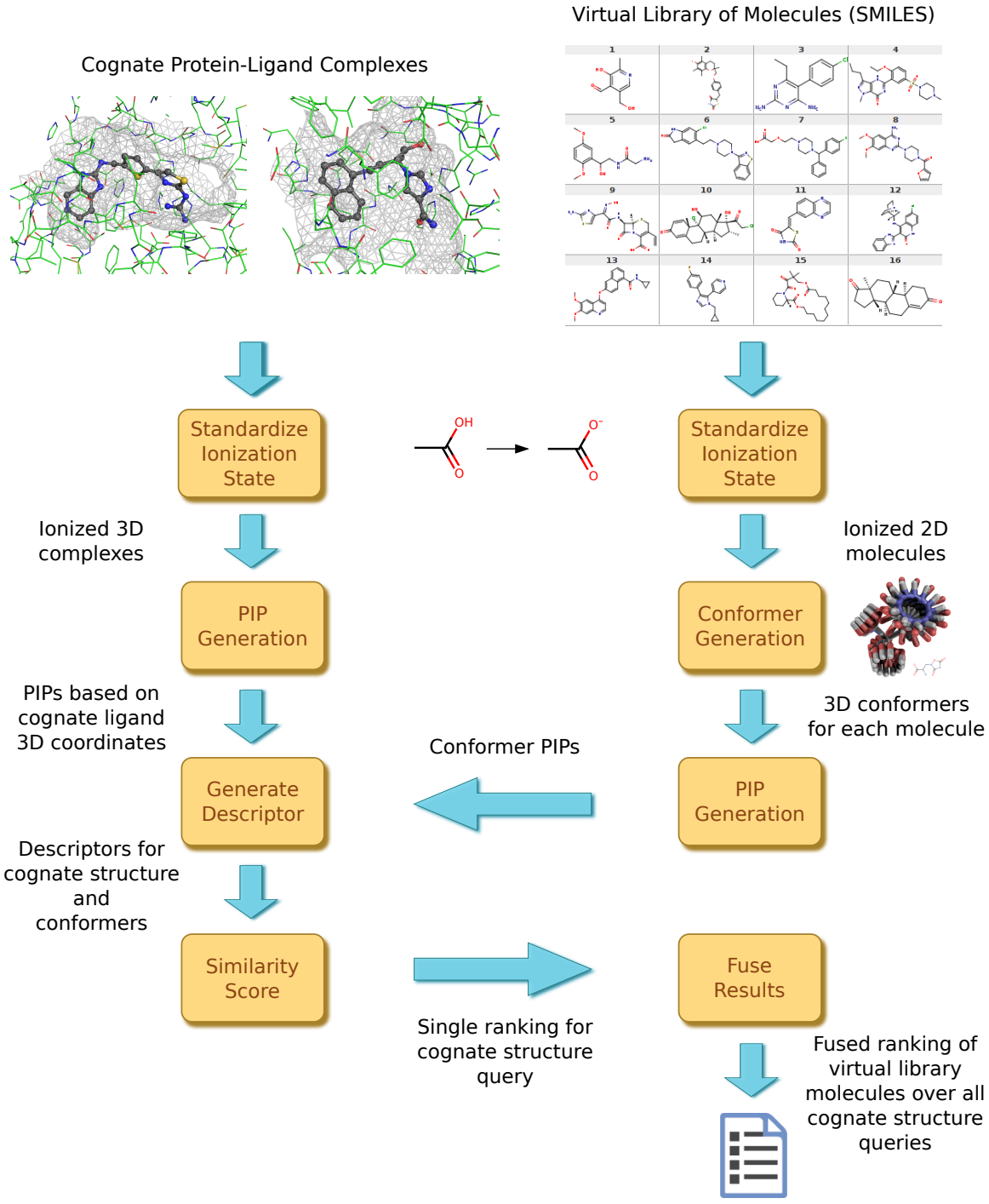


Figure 4.1: Ligity algorithm explained. Note that some processes, *Ionization* and *Generate PIPs*, are repeated because they have different implementations depending on whether their input is a 3D protein-ligand complex or a 2D SMILES molecule.

Ligity's Input and Output

There are two main inputs to the Ligity algorithm: (i) a set of *holo* protein-ligand structures for a particular target of interest, and (ii) a database of molecules to be screened for actives.

Ligity uses three dimensional information from the binding site of the protein-ligand complexes. These may be downloaded from the Protein Data Bank (PDB) [Berman et al., 2000] or from specialised binding site only databases such as the screening PDB (sc-PDB) [Kellenberger et al., 2006]. The sc-PDB has an additional advantage that all binding sites are hierarchically clustered using both the Enzyme Commission (EC) number and their similarity based on the binding sites' 3D superimposition using SiteAlign [Meslamani et al., 2011; Schalon et al., 2008]. This means that any single cluster of similar binding sites can be used as input for Ligity. Note that the *holo* binding sites do not need to be translated or rotated on top of each other as Ligity is a non-superpositional method.

The virtual screening library can be defined as a list of two dimensional simplified molecular-input line-entry system (SMILES) strings.

The output of the algorithm is a ranked list of all the molecules in the virtual library. The list is ordered by the decreasing similarity to the query descriptor. The molecules from the virtual screening library that have the highest scores are the ones that are most similar to the cognate ligand.

Standardization of the Ionization State

We standardize the ionization (charge) state of the protein-bound ligands and the molecules in the virtual screening library. This ionization happens in different dimensional space (in three dimensions for the protein-bound ligands and in two dimensions SMILES for the virtual screening library) but the same rules used in our multi-million small-molecule database (Table 3.3) are applied in both cases.

Conformer Generation

The list of standardized 2D molecules in the virtual screening database are subjected to our conformer generation protocol to generate three dimensional atom coordinates. The number of conformers generated for each molecule is dependent on its flexibility (*i.e.* rotatable bonds), but a maximum of 300 conformers may be generated. This approach is described and validated in Chapter 2.

We also have a 'high-throughput' Ligity mode, where we only store the lowest energy conformer instead of full conformer ensemble. This makes the next steps in the algorithm take a fraction of the time required with the full conformer set.

Pharmacophoric Important Points (PIP) Generation

Pharmacophoric models are automatically generated for the standardized 3D protein-ligand query (or queries) and the conformers for the molecules in the virtual library. These pharmacophoric models are represented by a collection of Pharmacophoric Important Points (PIPs). These highlight the interesting parts of the molecule, or the features necessary for molecular recognition of a ligand by the receptor. We consider the 'classical' six pharmacophore feature-types [Langer and Wolber, 2004]: (i) hydrophobic, (ii) aromatic rings(iii) hydrogen bond donors (HBD), (iv) hydrogen bond acceptors (HBA), (v) anion (-), and (vi) cation (+).

PIP generation for the query protein-ligand complex. In the case of the query protein-ligand complex, we only consider the ligand pharmacophoric features which are close enough to interact with the receptor (*i.e.* the query PIPs are placed on the cognate ligand). Furthermore, the cognate ligand PIP must have a corresponding receptor PIP, *e.g.* a hydrogen bond acceptor PIP on the bound ligand must be matched with a hydrogen bond donor PIP on the protein within 3.9Å. This limits the number of interaction points and ignores the parts of the ligand which are not contributing directly to the binding (*i.e.* the parts of the ligand which are in the solvent). Our definition of 'closeness' is determined by the threshold distances shown in Table 4.1. These geometric rules are partially assembled from literature sources [Bissantz et al., 2010; Marcou and

Table 4.1: Pharmacophoric interaction types and geometric distance constraints for query protein-ligand complexes.

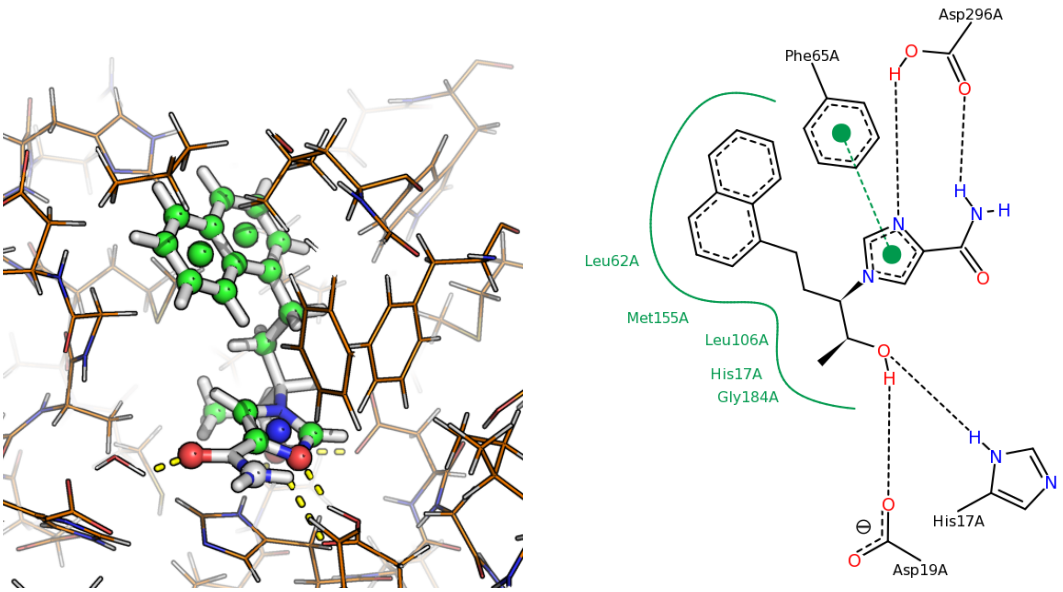
Receptor Pharmacophore	Ligand Pharmacophore	$\leq$ Distance (Å)
Hydrophobe	Hydrophobe	4.5
Acceptor	Donor	3.9
Donor	Acceptor	3.9
Cation	Anion	4.0
Anion	Cation	4.0
Aromatic	Aromatic	4.5
Aromatic	Cation	4.0
Cation	Aromatic	4.0

Rognan, 2007; Schreyer and Blundell, 2009]. Note that we make no distinction between weak, moderate and strong hydrogen bonding, which are typically defined using different distance intervals [Jeffrey, 1997]. Also, these do not include other rare and weaker π interactions such as π donor-acceptor interactions [Meyer et al., 2003]. PIPs may have atom or pseudo-atom 3D coordinates, *e.g.* a benzene ring would have a pseudo-positioned aromatic PIP at the centre of ring.

For hydrogen bond acceptor and donor interactions we also require that the angle between the hydrogen bond donor atom, its attached hydrogen atom and the hydrogen bond acceptor atom form an angle greater than 90°.

The query protein-ligand complex PIPs are therefore defined as those ligand PIPs which interact with the biological receptor (only the ligand spatial coordinates are used for the PIP generation). These are typically a subset of all the possible ligand's PIPs. This is a critical property when selecting a similarity metric.

An example of PIP definition for Adenosine deaminase (ADA) structure with PDB code `2e1w` is shown in Figure 4.2.



(a) PDB Ligand FR6 in ADA structure 2e1w. PIPs shown as spheres, green for hydrophobic, red for acceptor, blue for aromatic and white for donor. Some PIPs may not be shown in this image because their location coincides with others.

(b) Poseview interaction description taken from the PDB. Our PIPs definition match the interactions described here. Note the mediating water interaction is missing here, but shown in our PIP definition on the left.

Figure 4.2: PIP definition example for Adenosine deaminase (ADA), PDB code 2e1w.

PIP generation for virtual library molecules. PIPs are generated for every conformer for each virtual library molecule. We generate the same six PIP types we generate for the query protein-ligand complex, but instead of marking all ligand PIPs within an interaction distance of a complementary receptor PIP we just take all potential PIPs on the virtual library conformer. This makes the PIP lists for the virtual library molecules larger than the ones for the query protein-ligand complexes. Note that each conformer for a single virtual library molecule will have the same number of PIPs which only differ in 3D positioning based on the conformer's atomic coordinates. The virtual library molecule PIPs describe all the potential interaction points on the ligand.

PIP type definitions. The Smiles Arbitrary Target Specification (SMARTS) patterns representing the six PIP types are defined in a feature definition file in RDKit (in `BaseFeatures.fdef`). We modified the existing list of RDKit pharmacophore elucidation definitions and present the resulting full list in Table 4.2. Duplicate PIPs (with identical PIP type and *x*,*y*,*z* coordinates) are filtered from the PIP list describing the molecule.

Table 4.2: SMARTS patterns for each of the six different PIP types. †denotes the feature definitions we added to the RDKit list. §denotes the pharmacophores which have a pseudo-atomic position (typically the centroid of all the atoms involved in the definition)

PIP type	SMARTS Pattern
Donor	`[N&!H0&v3,N&!H0&+1&v4,n&H1&+0]`
	`[$([Nv3&!H0](-C)(-C)-C)]`
	`[$(n[n;H1]),$(nc[n;H1])]`
	`[O,S;H1;+0]`
Acceptor	`[$([Nv3](C)=C)]`†
	`[n;+0;!X3;!$([n;H1](cc)cc)]`
	`[$([N;H0]#[C&v4])]`
	`[N&v3;H0;$(Nc)]`
	`[O;H0;v2;!$(O=N-*)]`
	`[O;-;!$(*-N=O)]`
	`[o;+0]`
	`[O;H1;v2]`
	`[$([O-][CX3])]`

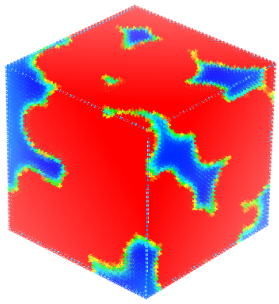
Continued on Next Page...

Table 4.2 – Continued

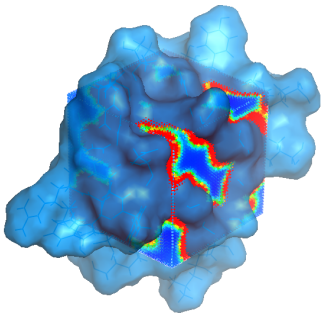
PIP type	SMARTS Pattern
	[$([O][PX4])]†
	[F;$(F-[#6]);!$(FC[F,Cl,Br,I])]
Negative Ionizable	[-]†
	[SX4](=O)(=O)(-[O;H1,HO&-1])†§
	[PX4](=O)(-[O;H1,HO&-1])([!O])([!O])†§
	[PX4](=O)(-[O;H1,HO&-1])(-[O;H1,HO&-1])†§
	[CX3,SX3](=[O,S,P])-[O;H1,HO&-1]§
Positive Ionizable	[+]†
	[$([N;H2&+0][C;!$(C=*)]);!$(N[a])]
	[$([N;H1&+0]([C;!$(C=*)])[C;!$(C=*)]);!$(N[a])]
	[$([N;HO&+0]([C;!$(C=*)])([C;!$(C=*)])[C;!$(C=*)]);!$(N[a])]
	[NX3]=[CX3]([NX3])[!N]§
	NC(=N)N§
	c1ncnc1§
Hydrophobic	[D3,D4;#6;+0;!$([#6] [#7,#8,#9])]
	[CX4](F)(F)(F)
	[R0;D2;#6;+0;!$([#6] [#7,#8,#9])]
	[#6;R]†
	[#17,#35,#53]†
Aromatic	4-membered aromatic rings§
	5-membered aromatic rings§
	6-membered aromatic rings§
	7-membered aromatic rings§
	8-membered aromatic rings§

An alternative grid-based method for PIP detection. Initially, rather than using the cognate ligand to place the 'cavity' PIPs we had a different, grid-based approach to locate these PIPs. We used the negative image of the receptor, by describing the void or space defined by the surface of the receptor. In this case the cavity descriptor is computed by generating an energy grid map of the cavity using different atom probes. We used six different atom types as probes: aromatic (A), aliphatic carbon (C), hydrogen bond donor (HD), nitrogen hydrogen bond acceptor (NA), oxygen hydrogen bond acceptor (OA) and electrostatic (e). This grid map is defined as a cube having 51 points along each

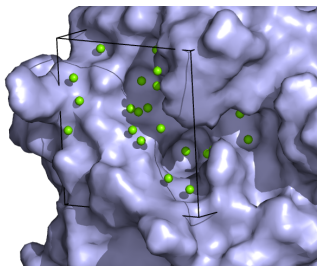
dimension, with a grid spacing of 0.375 Å for a total length of 18.75 Å along each of the three sides. These arbitrary values have been refined and tested. An example of a grid map is shown in Figure 4.3a.



(a) An example of a grid map. An aliphatic carbon grid map coloured by energy values from high (red) to low (blue).



(b) Grid map and protein superimposition.



(c) PIPs from an aliphatic C grid map.

Figure 4.3: Practical example of the grid-based approach to define PIPs.

The centre of a grid box is manually placed in centre of the receptor's known binding site. The six grid maps (one per atom type) were generated using AutoGrid4, a software program used in conjunction with AutoDock [Morris et al., 2009]. Each grid map contains 51^3 energy values, which are then filtered by taking only those that represent a local minimum. A local minimum is defined as any point not on the edge or face of the grid map, which has the smallest value out of its neighbourhood (which has a parametrized size, *e.g.* $3\times3\times3$). These PIPs were then filtered further for each map by using an energy threshold, or by taking the top percentile or number of points from any particular map. PIPs are used to generate a descriptor for the receptor. An example showing aliphatic carbons PIPs including the bounding grid box is shown in Figure 4.3c. These PIPs were then built into a descriptor as previously described. Theoretically, similar cavities should have similar PIPs. And the cavity PIPs should be the hot-spots for ligand interactions, because of their energy favourable positions.

Unfortunately, this approach worked to detect similar cavities but had limited success

when comparing cavity to ligand PIPs. The reasons for this may be that: (i) the generated cavity PIPs were incompatible with the ligand PIPs since they represented an area of interaction rather than a molecule, (ii) there were many cavity PIPs some of which located in clusters but it was difficult to decide which ones to remove, and (iii) there was a large number of continuous parameters, most of which were difficult to estimate even with parameter sweeps.

Generate Descriptor

The list of PIPs, generated either for a protein-ligand complex or a molecule in the virtual library, is used to generate a descriptor. A descriptor is made up of either all the possible 3-PIP or 4-PIP combinations and is assembled in the following ways (Figure 4.4).

3-PIP combinations. In the case of 3-PIP combinations each triplet produces a triangle where the vertices are the PIP types (*e.g.* <HBD,HBA,HBA>) and the edges correspond to the distances between those PIPs. Each triangle is then used to create a 3D cube using the three lengths of the edges as an index in which to store a counter of the PIP triplet combination combination. 3-PIP combinations require a three dimensional cube for the descriptor, as the 3-PIP type counter position in the cube is determined by the length of the three sides of the triangle.

4-PIP combinations. In the case of 4-PIP combinations each quadruplet produces a tetrahedron with four vertices where the vertices are the PIP types (*e.g.* <HBD,HDA,+,HDA>) and the six edges again correspond to the distances between the vertex PIPs (Figure 4.4a). Tetrahedron counts are stored in a six-dimensional data structure, hereafter referred to as a *hypercube* – using the lengths of the edges as an index to a bin which stores the counts of the specific tetrahedron PIP types.

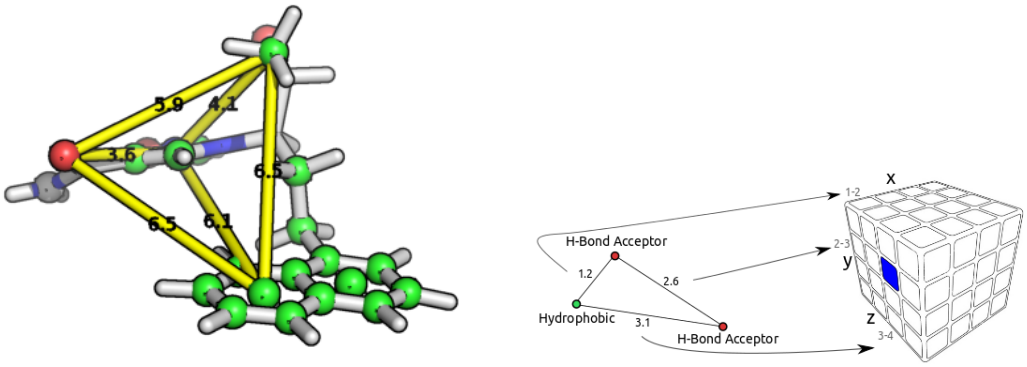
The 3-PIP or 4-PIP descriptor is made up of a number of 3-dimensional or 6-dimensional bins respectively. Each of these bins, identified by using the edges of the triangle (3-PIP) or tetrahedron (4-PIP) geometries as index, stores a counter for each 3 or 4 PIP type combination. In other words, each bin stores multiple counters of the

occurrences of each 3-PIP or 4-PIP type geometries. The 'bin size' determines the resolution of the descriptor. If the maximum distance between two PIPs of a molecule is 15 Å, and a bin size of 1.0 Å is used, we can expect a maximum of 15 bins across any dimension in the descriptor. A very small bin size, *e.g.* 0.1 Å, would give a descriptor with many bins and with most bin counters set to 1. On the other hand, a large bin size, *e.g.* 30.0 Å, would result in only one bin with counters set to all the occurrences of the PIP type combinations.

In both 3-PIP and 4-PIP combinations index determinism is ensured by sorting the the edge labels and edge length tuples by the edge label, this ensures that same 3-PIP triangles <HBA,HBD,+>, <2.6,3.7,3.2> and <HBA,+,HBD>, <2.6,3.2,3.7> populate the same bin counter. Also, when a geometry, *i.e.* a triangle or tetrahedron, has multiple identical PIP types (*e.g.* 3-PIP <HBA,HBD,HBD>) the length of edges for those identical PIP types is sorted. This guarantees that identical geometries, *e.g.* 3-PIP triangles {<HBA,HBD,HBD>, <2.1,3.7,4.2>} and {<HBA,HBD,HBD>, <2.1,4.2,3.7>}, populate the same bin counters. Triangles or a tetrahedrons with any side < 1.5 Å or > 15 Å are filtered out. The lower bounds filtering avoids may repetitive geometries (such as those found in an six-member, hydrophobic ring). We are also able to filter the descriptor from 3-PIP and 4-PIP geometries which appear infrequently (*e.g.* with a bin counter of 1) or the ones which are very common (*e.g.* with a bin counter > 10). Some PIP geometries may be found across all structures, and are therefore not discriminating between actives and decoys.

Bin neighbourhood population. In order to mimic receptor flexibility we do not populate just a single bin in the descriptor but rather have a "neighbourhood population" mode. This option adds partial counts to the neighbouring bins based on either a parametrized ramp or Gaussian function. The value added to each bin counter decreases as the distance of the bin from the edge length increases.

Chirality detection. For 4-PIP descriptors, we also distinguish between different chiralities of a molecule (Figure 4.5). This is done by calculating the volume of the tetrahedrons. Equation 4.1, where a, b, c and d are the vertices of the PIP tetrahedron, will give us



(a) Each possible 4 PIP tetrahedron combination is computed. In this example only one tetrahedron is shown.

(b) The lengths of the edges are used as indexes to the cube data structure. Shown here is the 3-PIP case, with a bin size of 1.0 Å.

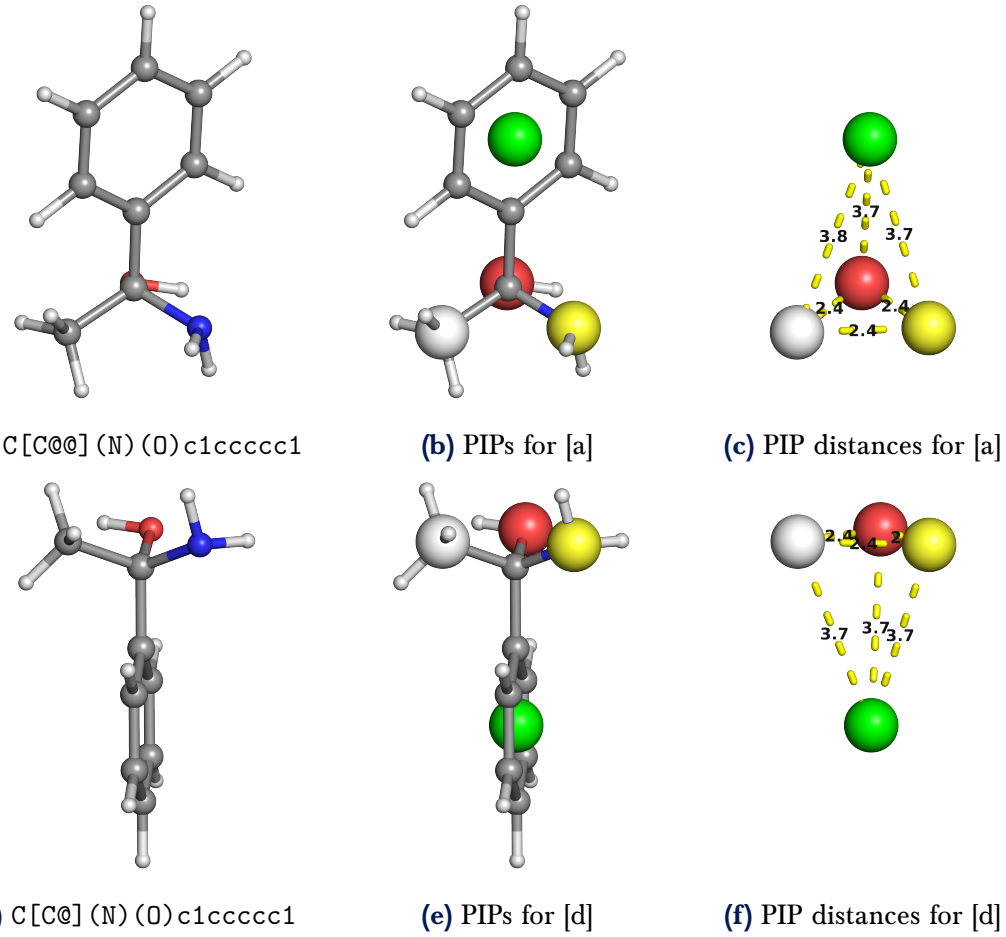
Figure 4.4: Ligity descriptor generation.

positive or negative volumes which distinguish between the different chiral forms of the molecule. In order to do this, the four PIPs are deterministically sorted by their type. Note that distinguishing chirality only applies when we have four different PIP types in the tetrahedron. Also, as a performance optimization, we drop the denominator from the volume calculation as we only need the sign of the result.

$$V = \frac{(a-d)\cdot((b-d)\times(c-d))}{6} \tag{4.1}$$

Similarity Score

Once we have generated descriptors for the protein-ligand complex and for a virtual library molecule we quantify their similarity. This is done using a number of similarity metrics (shown in Equations 1.1 to 1.5). In Equation 1.1, $S_{\alpha\beta}(A,B)$ is the Tversky similarity for the descriptor counts where A is the protein-ligand query descriptor, B is each ligand descriptor in the virtual library, n is the number of bins in the descriptor, A_i and B_i are the count of geometries in the ith bin. In our final implementation we use $\alpha = 1.0$ and $\beta = 0.0$. This asymmetric scoring function works well because, unlike other



(a) `C[C@@](N)(O)c1ccccc1` **(b)** PIPs for [a] **(c)** PIP distances for [a]

(d) `C[C@](N)(O)c1ccccc1` **(e)** PIPs for [d] **(f)** PIP distances for [d]

Figure 4.5: Chirality of PIPs. Molecules (a) and (d) are enantiomers, and differ only in the chirality of the central carbon atom. The PIPs therefore have similar distances from each other but differ in spatial arrangement.

symmetric functions such as Tanimoto, it is able to pick up on substructure nature of the query descriptor, *i.e.* only on those parts of the molecule which are in contact with the protein. We have also tested other similarity metrics such as Tanimoto (Equation 1.2), Dice (Equation 1.3), Cosine (Equation 1.4), and Common Counts (Equation 1.5) and we have found Tversky to be superior. For more details please refer to Section 4.3.4. The similarity score of a molecule in the virtual library is taken to be the highest score in the conformer ensemble.

Fusion of Ranked Results

When multiple starting protein-ligand structure queries are available, we fuse the results from the different queries into a single results list. This is achieved using the `MAX-SIM` method as described by Nasr et al. [2009], and shown in Equation 4.2. The highest scoring instance of a molecule across all ranking lists is used in a final, single ranking. Even if other methods (*e.g.* Exponential Tanimoto Discriminant) are reported by Nasr et al. [2009] to do slightly better, they are much more complex and require one or two parameters to be fit to the data.

In Equation 4.2, S is the similarity score, A is the query vector and B is the vector of the ligand in the multiple lists (where B_i is the vector of the ligand in list i).

$$\textbf{Max. Sim.} \quad S(A, B_1, \ldots, B_{|\bar{B}|}) = \max_i S(\vec{A}, \vec{B}_i) \tag{4.2}$$

4.2.2 Performance Measurement

All ROC curves and AUC calculations were generated using the R software environment for statistical computing and graphics (version 3.0.0) [R Core Team, 2013], and package `ROCR` (version 1.0-4) [Sing et al., 2012]. BEDROC values were generated using R package `enrichvs` (version 0.0.5) [Yabuuchi, 2011]. For BEDROC, we use an α of 20.0 which is the value suggested by the authors of this method [Truchon and Bayly, 2007]. In their work they state that this value for α "means that 80% of the maximum contribution to the BEDROC comes from the first 8% of the list".

4.2.3 Selecting and Preparing Testing and Validation Datasets

To test Ligity, we have used a subset of the Directory of Useful Decoys Enhanced (DUD-E) [Mysinger et al., 2012]. This is an improved and enhanced version of the original Directory of Useful Decoys (DUD) [Huang et al., 2006], which addresses the issues of "analogue bias" noted by Good and Oprea [2008], incorrect partial charges, unbalanced net charges between actives and decoys, putative decoys being actives, and has more varied target classes (*e.g.* membrane proteins). DUD-E contains 102 protein targets and a

total of 22,886 reduced chemotype actives (an average of 224 actives per target). There are 50 decoys for each active having similar physicochemical properties but dissimilar 2D topology. It is worth noting that decoys in DUD-E are still putative inactives (rather than experimentally verified ones). There are still some problems associated with this dataset. For example, the ligand bound to target Adenosine A2a receptor (a GPCR) in DUD-E which is available for download has one O atom less than the ligand ZMA in the PDB structure `3EML` from which it was reportedly taken. Also, for some targets, there are some duplicate instances of decoys with different ionization states (*e.g.* decoy with identifier `C39363328` for receptor `Try1`).

The Maximum Unbiased Validation (MUV) dataset is another high-quality dataset which is available for virtual screening validation studies [Rohrer and Baumann, 2009]. This dataset is constructed from experimental high-throughput primary and low -throughput confirmation bioassay data found in PubChem [Bolton et al., 2008; Wang et al., 2012]. This means that all actives and decoys are experimentally validated. The MUV dataset contains 18 targets, with 30 actives and 15,000 decoys per target. We were unable to use this dataset because: (i) it does not supply a protein structure of its target, (ii) in some cases the binding mechanism is unclear and may not be in the binding site of ligand, and (iii) there is little overlap with the sc-PDB clusters we used as input.

In order to test the performance of Ligity, we created a testing dataset in the following manner. We found all targets in DUD-E that were also present in sc-PDB (2011 release). From these we randomly selected ten targets: Angiotensin-converting enzyme (ACE), Adenosine deaminase (ADA), Cyclin-dependent kinase 2 (CDK2), Coagulation factor X (FA10), Coagulation factor VII (FA7), Glucocorticoid receptor (GCR), Human immunodeficiency virus type 1 integrase (HIVINT), Human immunodeficiency virus type 1 protease (HIVPR), Thrombin (THRB) and Trypsin I (TRY1). For each of the ten targets, we used the sc-PDB cluster of cognate protein-ligand complexes as queries. We removed the few structures that caused errors when read by RDKit (*e.g.* could not kekulize bound ligand) and then standardized the ionization state of the remaining bound ligands. We also standardized the ionization state and generated conformers for the corresponding actives and decoys of each target taken from DUD-E. When present, we removed the

cognate ligand in the sc-PDB structure from the DUD-E actives set. We do this to remove any bias in the method, as a conformer of the cognate ligand scores very highly with the cognate ligand descriptor. This effect may be accentuated when using fusion methods (when taking the highest score for each virtual library molecule across all queries).

The final testing dataset is described in Table 4.3. We list the number of sc-PDB protein-ligand structures to be used as queries, the number of actives and decoy molecules and their conformers. Note that there is a target bias in this set – most of these targets are enzymes with at least four of them being serine proteases.

Table 4.3: Ligity's testing dataset.

Receptor	sc-PDB cluster Id.	sc-PDB structures	# Actives (# Confs.)	# Decoys (# Confs.)
Angiotensin-converting enzyme (ACE)	0132	9	282 (27,346)	16,900 (1,307,531)
Adenosine deaminase (ADA)	0085	20	93 (6,720)	5,450 (371,990)
Cyclin-dependent kinase 2 (CDK2)	1424	109	474 (20,480)	27,850 (1,360,619)
Coagulation factor X (FA10)	0224	81	537 (39,732)	28,325 (1,799,269)
Coagulation factor VII (FA7)	0223	15	114 (9,759)	6,250 (398,145)
Glucocorticoid receptor (GCR)	0367	7	258 (8,682)	14,999 (640,882)
Human immunodeficiency virus type 1 integrase (HIVINT)	1167	3	98 (5,096)	6,650 (327,474)
Human immunodeficiency virus type 1 protease (HIVPR)	0654	166	535 (27,975)	35,750 (2,189,091)
Thrombin (THRB)	0830	113	461 (34,936)	27,004 (2,020,395)
Trypsin I (TRY1)	0850	74	449 (30,311)	25,980 (1,706,265)

Validation Dataset

For validation of the method we create a small subset of sc-PDB and DUD-E targets. We use this validation set to refine the various algorithm parameters and to benchmark the performance of the method. For this purpose, we select sc-PDB three targets (which are also in DUD-E): Adenosine deaminase (ADA), Cyclin-dependent kinase 2 (CDK2), and trypsin (TRY1). We chose these targets because they had a range of performance in our initial testing, ADA did best and CDK2 worst, with TRY1 having median performance. For the CDK2 and TRY1 receptors we randomly select 100 actives and 100 decoys from DUD-E. For ADA this is not possible, as there are only 93 actives in the DUD-E dataset (so we take all of these, and 100 decoys). Note that the number of active and decoy conformers generated is comparable because decoys are matched to actives in DUD-E based on physicochemical properties such as rotatable bonds which is used in our conformer generation protocol.

Table 4.4: A reduced dataset used for validation purposes. Each receptor has approximately a randomly selected 100 actives and 100 decoy molecules taken from DUD-E to be used to evaluate performance.

Receptor	**sc-PDB** cluster id.	**Site 1** PDB code	**Site 2** PDB code	**Site 3** PDB code	**Active** Confs. (#)	**Decoy** Confs. (#)
ADA	0085	`1ndv`	`2e1w`	`3km8`	6,002	6,394
CDK2	1424	`1pxm`	`2bts`	`2c6m`	6,839	6,697
TRY1	1463	`1bjv`	`1o3o`	`3m35`	4,647	4,956

4.3 Results and Discussion

This section presents the results of Ligity and the validation studies used to refine the parameters of the method. Unless otherwise stated, the following computational experiments were carried out using the validation dataset described in Table 4.4. We tested the effects of: (i) using single lowest energy conformer versus multiple conformers for virtual library molecules, (ii) 3-PIP versus 4-PIP combinations in descriptor generation,

(iii) different binning values for the descriptor, (iv) applying different similarity metrics, and (v) single versus fused results rankings. We then compare the performance of Ligity to other VS methods.

4.3.1 Effect of Using Single Lowest Energy Conformer Versus Multiple Conformers for Virtual Library Molecules

We are interested in testing two different ligand conformer representations, one where we have only the lowest energy conformer as a representative of the molecule and the other where we have a full conformer ensemble of up to a maximum of 300 conformers per molecule. This has important repercussions on the speed and storage requirements of the method. Surprisingly, we find that it makes little difference if we use only the lowest energy conformer per molecule rather than the full ensemble on our validation set of three reduced test cases (shown in Figure 4.6). The average difference in AUC between all nine individual structure queries for ADA, CDK2 and TRY1 is of 0.003. For fused results the average difference in AUC between the full conformer ensemble and the lowest energy conformers is 0.011 (with the full conformer model doing just slightly better).

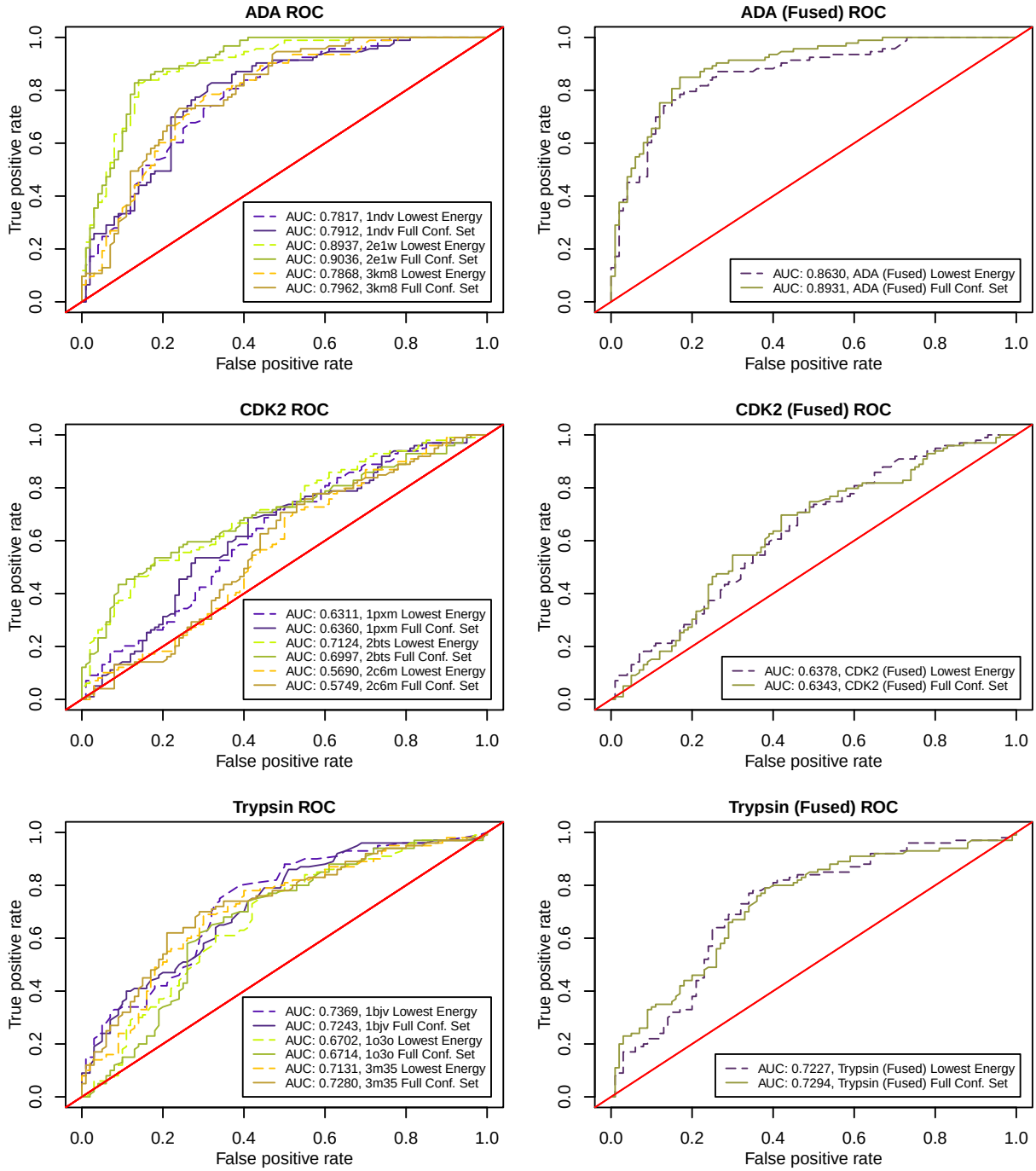


Figure 4.6: Ligity shows little difference when using only lowest energy conformer instead of full conformer ensemble.

Thus, if the method performs just as well when using only the lowest energy conformer is it these low energy conformers that are highest scoring when the full conformer ensemble is given? Put specifically, "Does the method prefer picking lower energy conformers as the highest scoring ones?".

For any molecule, conformer identifiers are always sorted by energy; so the conformer with identifier 1 is the lowest energy, conformer identifier 2 has the second lowest energy, and so on. Therefore, we want to test if Ligity picks lower conformer identifiers more than what would be expected at random. For the actives and decoys sets for ADA, CDK2 and TRY1 we built a theoretical probability density function for the conformer identifiers selection. This model is needed because not all molecules have the same conformer size, and it is therefore more likely to pick up conformer identifier 1, than conformer identifier 300 because every molecule will have conformer 1 but very few will have conformer 300.

The model is described in Equation 4.3 where $P(c_{id})$ is the probability to pick a specific conformer with identifier id for all N molecules, N is the total number of molecules for the receptor (*e.g.* 100 for ADA actives), and C_i is the conformer ensemble set for the ith molecule. The probability of picking a conformer identifier which is larger than the conformer ensemble size for that molecule is zero. In our theoretical model, all conformers in an ensemble are equally likely to be picked. A simplified example is offered in Table 4.5.

$$P(c_{id}) = \frac{\sum_{i=1}^{N} \frac{1}{|C_i|}}{N} \tag{4.3}$$

We generate 1,000 samples (with replacement) using the above described probability density function, all of size N. We then run a Mann-Whitney statistical test for the conformer identifier selection we get in Ligity against each of the 1,000 random samples. In Figure 4.7 we show two histograms, of the 1,000 resulting p-values for the actives and decoys for a specific receptor (ADA). The preference for lower energy conformers is more marked in the actives, where most p-values are below our significance level (α) of 0.05 (Figure 4.7a). This means that for the actives set there is statistical significance between the conformer identifiers picked by Ligity and a random set of generated conformer

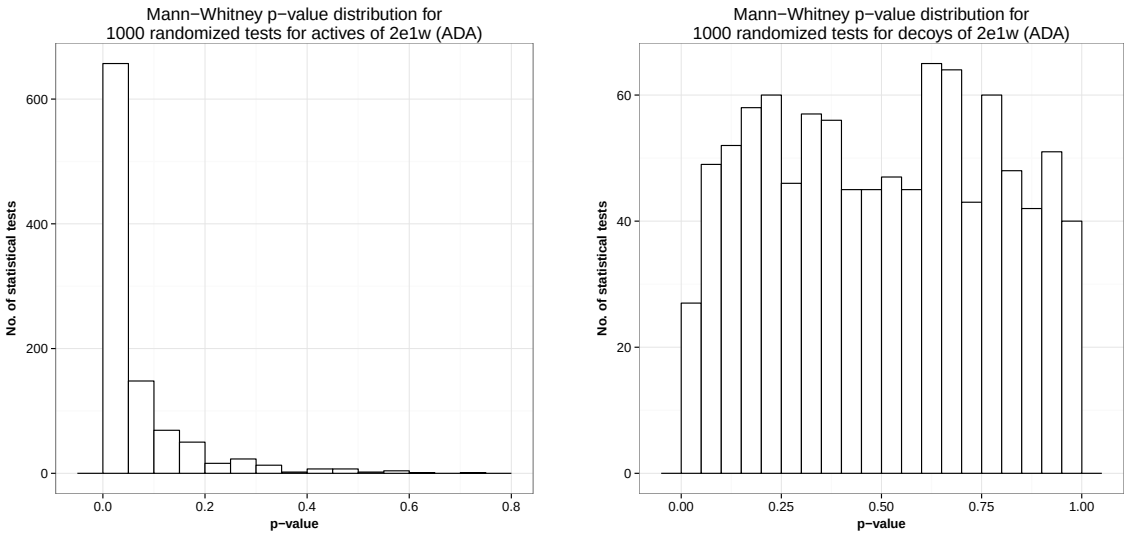
Table 4.5: Practical example of how we build a theoretical model for the probability of picking up a specific conformer identifier over N molecules with different ensemble sizes.

N=3	$\|C_i\|$	$P(1)$	$P(2)$	$P(3)$	$P(4)$	$P(5)$	$P(6)$	$P(7)$	$P(8)$	$P(9)$	$P(10)$
Mol_1	5	$1/5$	$1/5$	$1/5$	$1/5$	$1/5$	0	0	0	0	0
Mol_2	2	$1/2$	$1/2$	0	0	0	0	0	0	0	0
Mol_3	10	$1/10$	$1/10$	$1/10$	$1/10$	$1/10$	$1/10$	$1/10$	$1/10$	$1/10$	$1/10$
$\sum$		$4/5$	$4/5$	$3/10$	$3/10$	$3/10$	$1/10$	$1/10$	$1/10$	$1/10$	$1/10$
$P(c_{id})$		$4/15$	$4/15$	$1/10$	$1/10$	$1/10$	$1/30$	$1/30$	$1/30$	$1/30$	$1/30$

identifiers. We also run a one-sided test, so we specifically test for preference for lower conformer identifier than the random set. For the decoy set (Figure 4.7b) we see no significant preference for lower conformer identifiers as opposed to random selection as there are very few p-values smaller than our chosen α (*i.e.* smaller than 0.05).

Why are low energy conformers 'enough' to distinguish between active and decoys? Even if controversial, perhaps the finding by Butler et al. [2009] that bioactive conformations are close to the energy minimum (two thirds of their 99 molecule dataset lies within 0.5 kcal/mol of an energy minimum) is indeed correct. It is worth noting that others have argued that the energy of the protein-bound ligands is much higher than the global minimum for their unbound form [Perola and Charifson, 2004; Sitzmann et al., 2012].

Other authors have claimed that the effect of the bioactive conformation on virtual screening experiments is small [Renner et al., 2006; Zhang and Muegge, 2006]. Indeed in some of these studies, starting off with a low energy conformation instead of the bioactive conformation yielded little difference in enrichment.



(a) Lower conformer identifiers are selected for actives, rather than one would expect at random.

(b) No difference between conformer identifiers selected for decoys and ones selected at random.

Figure 4.7: Ligity preferentially selects lower energy conformers for actives but not for decoys.

4.3.2 Effect of 3-PIP Versus 4-PIP Combinations in Descriptor Generation

4-PIP combinations give marginally better performance than 3-PIP combinations on our validation dataset (Figure 4.8). The 3-PIP descriptors are on average 90.75% smaller than the 4-PIP descriptors. This can be attributed to the four PIP descriptors having a larger data structure with more entries. With 25 PIPs for a molecule, and without filtering, the descriptor would contain a maximum of 2,300 3-PIP triplets and 12,650 4-PIP quadruplets.

4-PIP combinations do better for individual queries for the ADA (average AUC improvement of 0.018) and TRY1 (average AUC improvement of 0.020) receptors. However there is a decrease in performance for CDK2, where 3-PIP combinations perform better for individual queries (average 3-PIP AUC improvement over 4-PIP of 0.028). The same trait (*i.e.* improvement for ADA and TRY1, but a decline in performance for CDK2) can be seen for the fused results scoring.

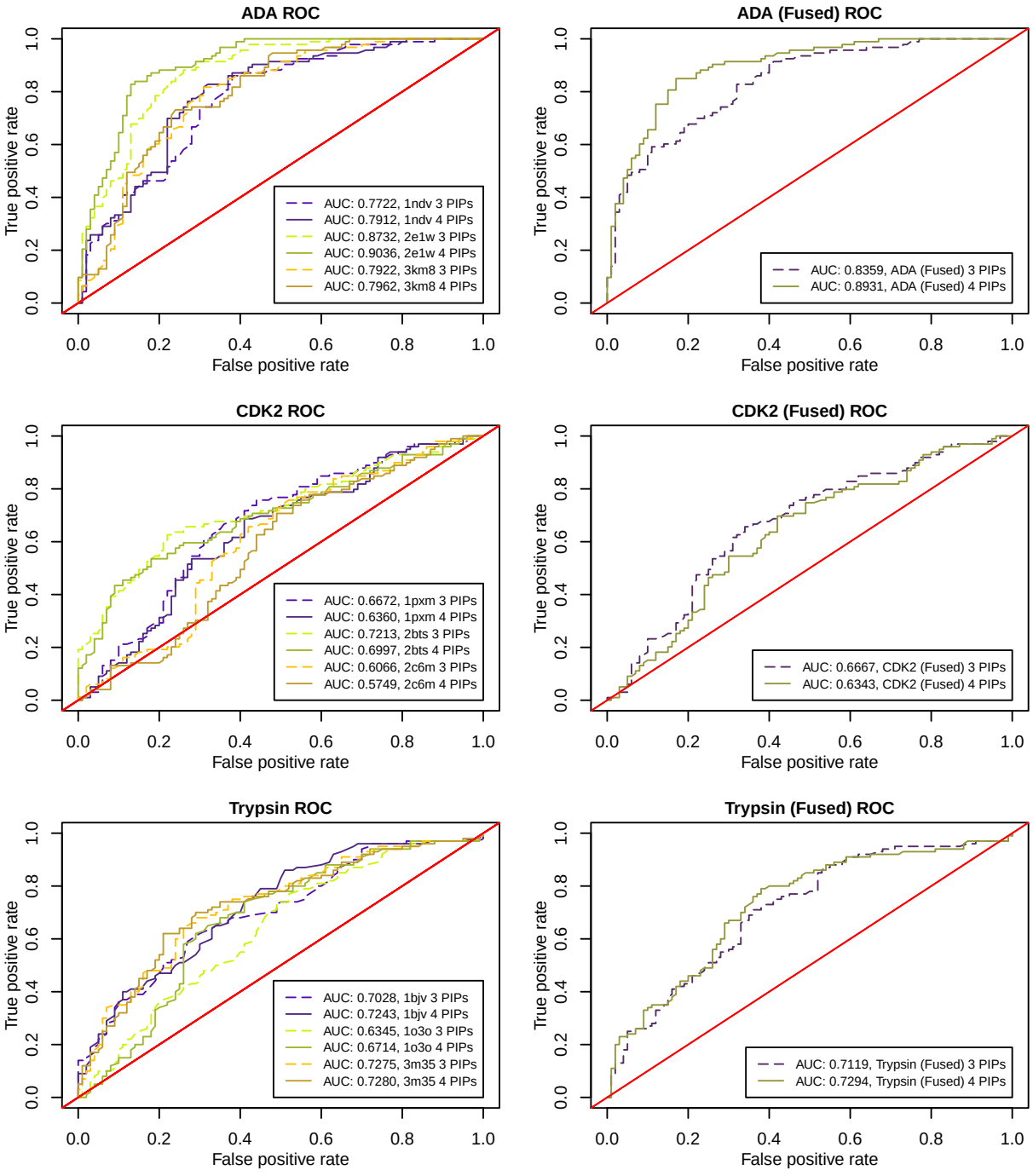


Figure 4.8: ADA, CDK2, and TRY1 receptors tested with 3-PIP and 4-PIP combinations for both separate and fused ranking lists. 4-PIP descriptors perform marginally better for ADA and TRY1 and worst for CDK2.

4.3.3 Effect of Different Binning Values for the Descriptor

We varied the binning of the 4-PIP combination hypercube (the descriptor) using 0.5 Å, 1.0 Å, 1.5 Å and 2.0 Å sizes. The number of query-matching tetrahedrons increases with the bin size. The results are shown in Table 4.6. This table indicates that the bin threshold selection may be dependent on the receptor; Ligity performs slightly better for CDK2 at smaller bins (*i.e.* 0.5 Å) while it performs better at larger bins (*i.e.* 1.5 Å) for ADA and TRY1. A possible explanation for this can be found in the flexibility of the molecules. Flexibility is determined by the number of rotatable bonds in the molecule. CDK2 actives and decoys sets are less flexible than the ADA and TRY1 sets (Figure 4.9). Larger, more flexible molecules have many degrees of freedom and a stricter bin may be too limiting to sample the conformer space of the query molecule adequately. On the other hand smaller, less flexible molecules would match a large number of tetrahedrons with large bins because most of the molecules will populate the same bins. This increases the false positive rate, resulting in a decrease in performance.

Table 4.6: Effect of 4-PIP combination hypercube bin size on Ligity's performance. The best AUC in every row is marked in bold. The mean AUC across all three cognate structure queries for each receptor is shown in the row labelled **mean**. The AUC of the fused results over the three individual queries is shown in the row labelled **fusion**.

Receptor		Query	Bin 0.5 Å AUC	Bin 1.0 Å AUC	Bin 1.5 Å AUC	Bin 2.0 Å AUC
ADA		1ndv	0.763	0.768	**0.791**	0.774
		2e1w	0.886	0.891	**0.904**	0.871
		3km8	0.764	0.814	0.796	**0.831**
	mean		0.804	0.824	**0.830**	0.825
	fusion		0.879	0.884	**0.893**	0.853
CDK2		1pxm	**0.670**	0.626	0.636	0.638
		2bts	**0.717**	0.710	0.700	0.688
		2c6m	**0.597**	0.586	0.575	0.586
	mean		**0.661**	0.641	0.637	0.637
	fusion		**0.698**	0.633	0.634	0.633
TRY1		1bjv	0.710	**0.739**	0.724	0.676
		1o3o	0.636	0.655	**0.671**	0.639
		3m35	0.687	0.714	**0.728**	0.706
	mean		0.678	0.703	**0.708**	0.674
	fusion		0.698	0.726	**0.729**	0.686

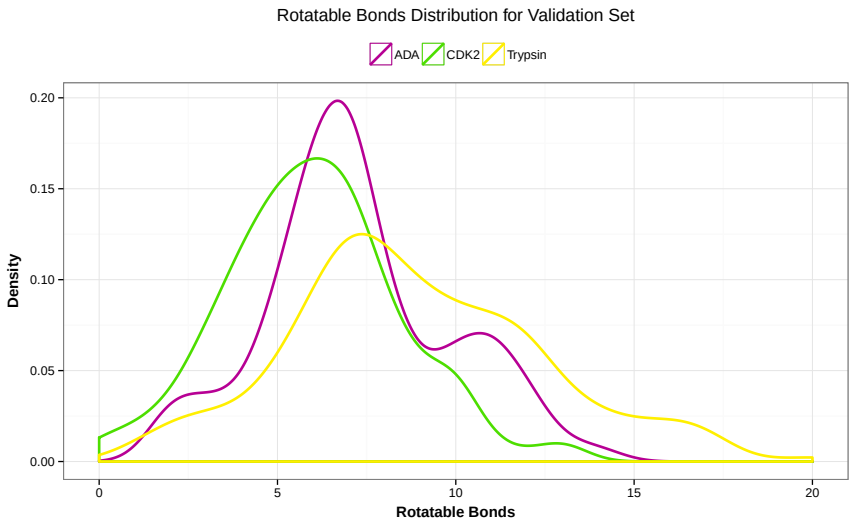


Figure 4.9: Density distribution of rotatable bonds for molecules in the validation set. CDK2 actives and decoys have, on average, fewer rotatable bonds than ADA or TRY1 making them less flexible.

4.3.4 Effect of Applying Different Similarity Metrics

We have analysed the effect of using various similarity metrics to calculate the similarity between the query and virtual library conformers 4-PIP hypercube descriptors, including Tanimoto, Cosine, Dice, simple counts of common bins, and different values for α and β in the Tversky metric. The results are presented in Table 4.7.

The Tanimoto and Dice (not shown) similarity metrics give identical AUCs. These two metrics are very similar in spirit, the only difference is that Tanimoto is the ratio of the number of common features between the two data structures over the total number of features while the Dice metric is the ratio of the total number of common features over the average size of the features in the two data structures. Using Tanimoto and Dice gives the same final rankings.

Tversky (with $\alpha = 1.00$ and $\beta = 0.00$) and Counts also give identical results for each separate receptor query. This is because Tversky with $\beta = 0.0$ is a specialized case of Counts, where the number of common features are divided by the number of features in the cognate protein-ligand query (which is fixed) therefore giving the same ranking lists of actives and decoys. Note that results are different when we consider fused rankings

results. This is because using the Tversky metric the different protein-ligand queries will normalize the common counts between query and database ligand by a different amount (*i.e.* for Tversky with $\alpha = 1.00$ and $\beta = 0.00$ this will be the number of features in the query).

Table 4.7: Effect of different scoring functions on Ligity's performance. The best AUC in every row is marked in bold. Note that Dice (not shown) gave identical results to Tanimoto. The mean AUC across all three cognate structure queries for each receptor is shown in the row labelled **mean**. The AUC of the fused results over the three individual queries is shown in the row labelled **fusion**.

Receptor		Query	Tversky (1) AUC $\alpha = 1.00$ $\beta = 0.00$	Tversky (2) AUC $\alpha = 0.95$ $\beta = 0.05$	Tversky (3) AUC $\alpha = 0.90$ $\beta = 0.10$	Tversky (4) AUC $\alpha = 0.85$ $\beta = 0.15$	Tanimoto AUC	Cosine AUC	Counts AUC
ADA		`1ndv`	0.791	0.814	0.831	**0.839**	0.799	0.810	0.791
		`2e1w`	0.904	0.911	0.919	**0.927**	0.909	0.912	0.904
		`3km8`	**0.796**	0.790	0.784	0.770	0.672	0.714	**0.796**
	mean		0.830	0.838	**0.845**	**0.845**	0.793	0.822	0.830
	fusion		0.893	0.913	0.925	**0.935**	0.912	0.926	0.913
CDK2		`1pxm`	**0.636**	0.581	0.541	0.516	0.508	0.530	**0.636**
		`2bts`	0.700	**0.701**	0.669	0.634	0.539	0.593	0.700
		`2c6m`	**0.575**	0.553	0.539	0.522	0.513	0.481	**0.575**
	mean		**0.637**	0.611	0.583	0.557	0.520	0.535	**0.637**
	fusion		**0.634**	0.575	0.532	0.508	0.488	0.503	0.576
TRY1		`1bjv`	**0.724**	0.511	0.479	0.467	0.412	0.557	**0.724**
		`1o3o`	**0.671**	0.456	0.394	0.363	0.300	0.417	**0.671**
		`3m35`	**0.728**	0.709	0.663	0.615	0.483	0.687	**0.728**
	mean		**0.708**	0.559	0.512	0.482	0.398	0.554	**0.708**
	fusion		**0.729**	0.712	0.662	0.611	0.472	0.687	0.728

Table 4.7 also demonstrates that the performance of the method is more uniform across the different similarity metrics for ADA than for CDK2 and TRY1. For CDK2 and TRY1, similarity metrics which measure global similarity, such as Tanimoto, between the query and virtual library molecule descriptors do poorly in a large number of cases. This can be explained by considering the number of PIPs in the query when compared to the number of target PIPs. Figure 4.10 shows that the number of query PIPs for CDK2 and TRY1 is smaller than the number of ligand PIPs. In order to handle this disparity we need a similarity metric which captures the substructure nature of the query. By setting the α parameter to 1.0 in Tversky we place all the importance on the query PIPs (and effectively ignore ligand's PIPs which do not match).

The number of PIPs in the query and the Tversky score AUCs have a moderate Pearson product-moment correlation coefficient (r) of 0.510. If we remove the `2c6m` query data (as this structure seems to be an outlier), we get a Pearson's $r(6) = 0.741$, $p < 0.05$.

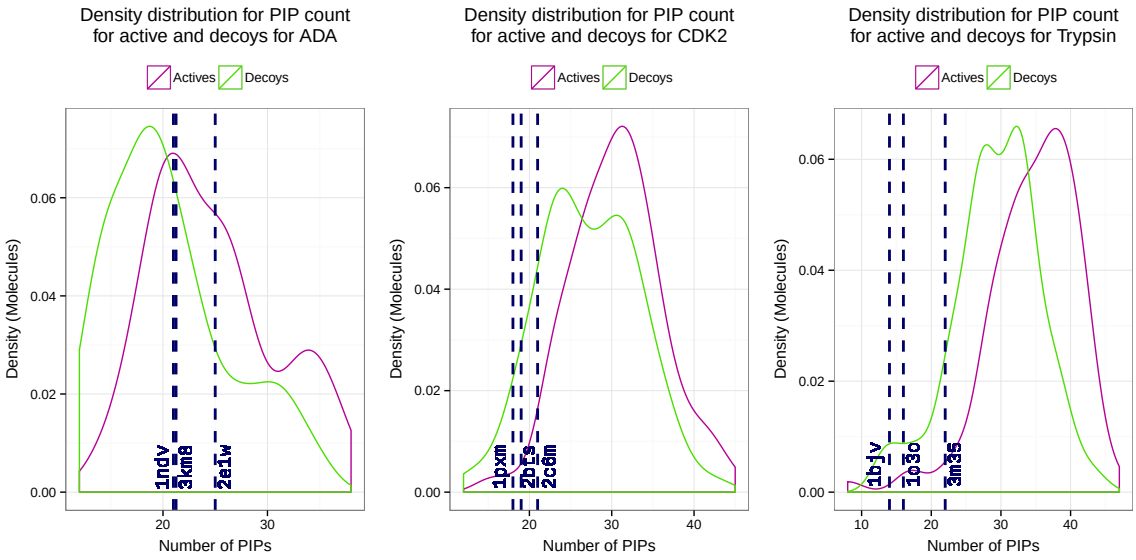


Figure 4.10: Number of PIPs in the validation dataset. For CDK2 and TRY1, we have queries with a small number of PIPs when compared to the rest of the active and decoy molecules. Therefore using one of the traditional symmetric metrics, such as Tanimoto, will inevitably lead to poor results.

4.3.5 Effect of Single Versus Fused Results Rankings

Table 4.7 shows that fusing the results of the multiple queries (or consensus scoring) generally improves performance of the method. This is in accordance with the current view in the field on results fusion.

CDK2 fusion results get worse than the mean AUC when $\beta > 0$ in Tversky scoring. This is because when $\beta > 0$, more weight is given to the features in the ligand descriptor – most of which may be unimportant and may not even interact with the protein. Also, the ligand descriptors are also usually larger in size. Setting $\beta > 0$ for CDK2, effectively introduces noise which in turn worsens the performance of the fusion method.

4.3.6 Validating Ligity Using DUD-E

Ligity has been tested against ten targets (described previously in Table 4.3). We used the protein-ligand complexes in each target's binding site clusters from sc-PDB as input queries. For each of these we extracted the corresponding active and decoy sets from DUD-E. We standardize the ionization state of all the cognate ligands, actives and decoys. First we generate the query descriptor based on the cognate ligand PIPs which interact with the receptor. We then generate conformers for all the active and decoy sets, which we use to find PIPs in three dimensions and generate descriptors for each conformer. All descriptors are 4-PIP combinations hypercubes with a bin of 1.5 Å. Neighbourhood population is disabled. We use the Tversky scoring function, with $\alpha = 1.0$ and $\beta = 0.0$, and fuse results using the maximum score across all separate ranking lists.

In Table 4.8 we offer two AUCs as a testimony of Ligity's performance. We report the **mean** AUC across each individual query run (*i.e.* each single sc-PDB ligand-protein complex is a separate query), together with the standard deviation (σ). We also report the AUC of the **fusion** score based on the ranked results of each individual query. We also report the early enrichment BEDROC score for each individual query and for the fused approach.

The statistical significance of the BEDROC score is shown in Figure 4.11. For each receptor, using the specific number of active and decoy molecules, we generate 10,000

random rankings and calculate the BEDROC score for each random ranking. Using the BEDROC score of the fused results in Table 4.8 as a test statistic we then calculate the p-value using the 10,000 BEDROC scores on the randomly generated ranking lists. Of the ten receptors, nine show a statistically significant early enrichment and only CDK2 shows random early enrichment.

Ligity's performance is moderate to excellent across all receptors tested. The fusion approach improves the AUC by more than 0.05 in 50% of the cases, and is only marginally worse than the mean in just one case (HIVINT).

Table 4.8: Ligity Results

Receptor	Mean AUC ($\pm \sigma$)	Mean BEDROC ($\pm \sigma$) ($\alpha = 20.0$)	Fusion AUC	Fusion BEDROC ($\alpha = 20.0$)
Angiotensin-converting enzyme (ACE)	0.779 ($\pm$ 0.070)	0.424 ($\pm$ 0.181)	0.948	0.776
Adenosine deaminase (ADA)	0.811 ($\pm$ 0.068)	0.302 ($\pm$ 0.102)	0.894	0.557
Cyclin-dependent kinase 2 (CDK2)	0.610 ($\pm$ 0.035)	0.081 ($\pm$ 0.047)	0.643	0.062
Coagulation factor X (FA10)	0.700 ($\pm$ 0.050)	0.195 ($\pm$ 0.079)	0.716	0.208
Coagulation factor VII (FA7)	0.750 ($\pm$ 0.026)	0.277 ($\pm$ 0.540)	0.809	0.270
Glucocorticoid receptor (GCR)	0.790 ($\pm$ 0.094)	0.300 ($\pm$ 0.116)	0.867	0.439
Human immunodeficiency virus type 1 integrase (HIVINT)	0.669 ($\pm$ 0.045)	0.173 ($\pm$ 0.068)	0.637	0.139
Human immunodeficiency virus type 1 protease (HIVPR)	0.874 ($\pm$ 0.018)	0.584 ($\pm$ 0.057)	0.876	0.527
Thrombin (THRB)	0.747 ($\pm$ 0.035)	0.220 ($\pm$ 0.079)	0.752	0.185
Trypsin I (TRY1)	0.725 ($\pm$ 0.060)	0.171 ($\pm$ 0.076)	0.778	0.167

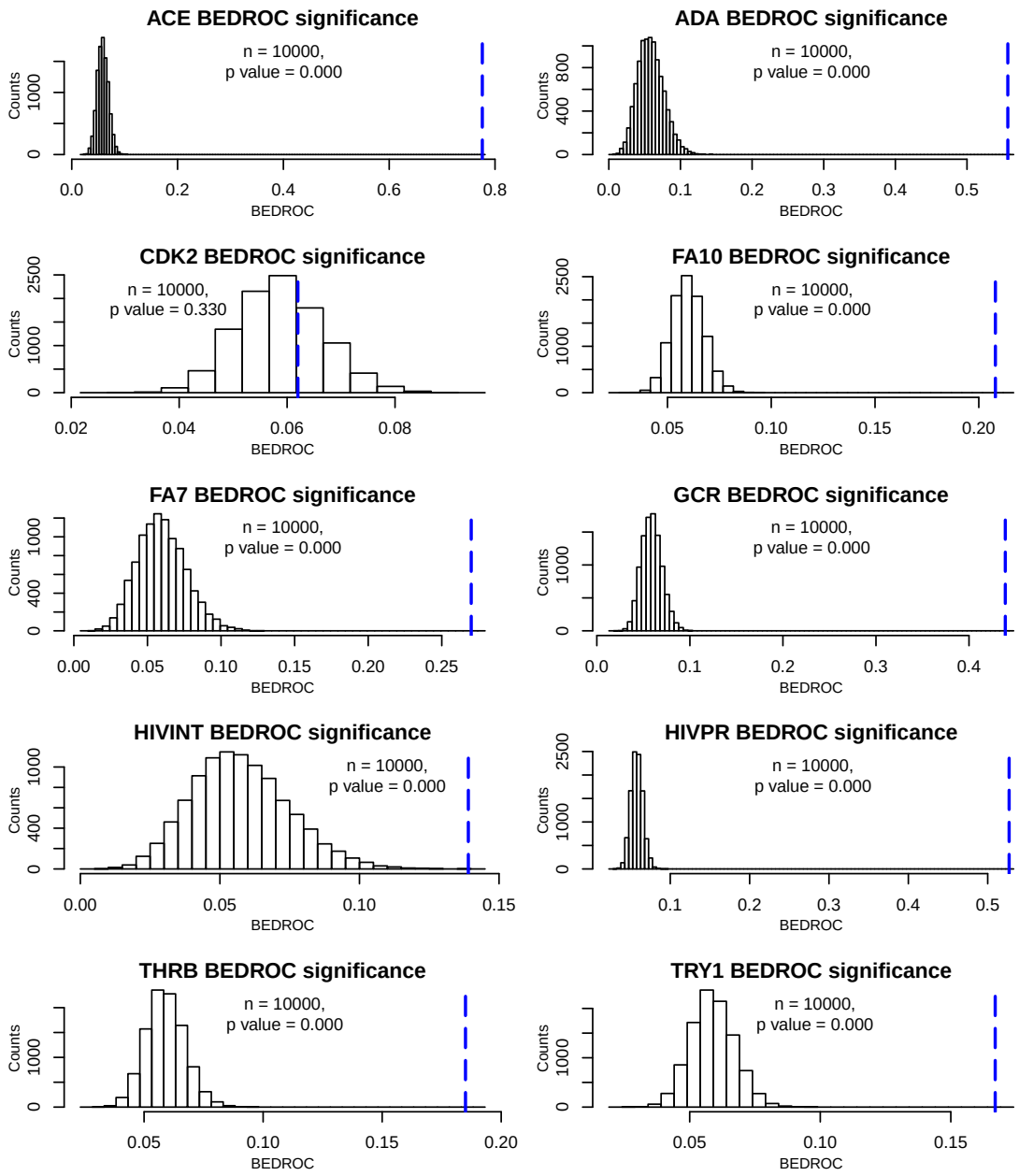


Figure 4.11: Ligity shows statistically significant early enrichment for nine receptors out of ten (all receptors except CDK2). The blue dashed line shows the BEDROC score of the fused ranking. The histograms are made up of 10,000 BEDROC scores calculated on random rankings.

4.3.7 Comparison of Ligity to Existing Methods

In Table 4.9, we compare Ligity's performance to two different ligand fingerprint methods (MACCS and Morgan), a 3D ligand similarity search method (ElectroShape), and a SBVS method (DOCK). For an overview of how these methods work, please refer to Section 1.3.

All comparisons are carried out using a single receptor structure, as defined in DUD-E. For reference, apart from the individual query scores of these PDB structures for Ligity we also give fusion approach score.

The AUC values in the DOCK column in Table 4.9 originate from the literature, specifically from Supplementary Information Table S1 from the DUD-E publication [Mysinger et al., 2012]. In this article, DOCK 3.6 has been tested using the declared PDB structure (second column in the table) and the actives and decoys defined in DUD-E.

We have used InhibOx's proprietary version of ElectroShape (version 2.0.2). The virtual screening study was performed by using the 3D cognate ligand as an input to ElectroShape 4D to generate a query descriptor that is used to rank the active and decoy molecules of each receptor. Similar to Ligity, the ionization state of the query molecule and the active and decoy datasets was standardized at physiological pH. This is particularly important for ElectroShape as it uses partial charges in its similarity computation.

For Morgan and MACCS fingerprints we have implemented an RDKit-based python application which generates these fingerprints from the SMILES representation of the active and decoy molecules in DUD-E. We have used the standardized (in terms of ionization) SMILES of the cognate ligand of the DUD-E receptor as a query, and then executed a Tanimoto similarity with the fingerprints of the standardized molecules for each receptor set. For Morgan fingerprints, we have used a 2048-bit fingerprint with radius 2.

For ADA, MACCS fingerprints do incredibly well. We do not discount the fact that this may be an artefact of how the DUD-E dataset was assembled. To minimise artificial enrichment, the decoys in the DUD-E set are chosen to have similar physicochemical properties (*e.g.* molecular weight, logP, HBD count, HBA count, *etc.*) but different com-

position and connectivity to the actives. The authors of DUD-E used Daylight [Daylight, 2011] fingerprints to remove any similar actives from the decoys set and they warn that this may create an artificial favourable enrichment bias for 2D fingerprinting methods. Also, to avoid that putative decoys are actually active – compounds with known active chemical warheads are removed from the DUD-E decoy sets. Fingerprint methods could easily pick up on this difference in actives and decoys sets. Decoy 2D dissimilarity in DUD-E was enforced using Tanimoto coefficients based on fingerprint methods, so it is hardly surprising that fingerprint methods are able to discriminate between actives and decoys. It is perhaps surprising that MACCS and Morgan fingerprints do not exhibit better performance on our test set.

For fingerprint performance, in 60% of the cases Morgan did better than MACCS. This is in line with the recent suggestion that the choice of the best fingerprinting method may depend on the specific target [Duan et al., 2010].

For CDK2, which is the query receptor for which Ligity does worst, we do not have the `1h00` structure in the CDK2 binding-site cluster in the sc-PDB release we are using. These sc-PDB clusters contain binding sites which are similar to one another (above a certain threshold), and only contain structures which pass a stringent quality filter. This implies that the 2003 deposited structure for `1h00`, might not be similar enough to the sc-PDB binding sites we are using as queries. DOCK might be doing better on this structure because of specific properties (*e.g.* better druggability of `1h00`). For this receptor for Ligity we are not displaying the `1h00` AUC, but instead we present the average across for all the single CDK2 runs (in parentheses).

Perhaps the reason ElectroShape does poorly in this comparison is that the query, active and decoy molecules are similarly sized molecules making it hard to distinguish distributions of interatomic distances from the reference points (*e.g.* the centre of the molecule).

Finally, Ligity does better, on average, than any other 3D method in Table 4.9.

Table 4.9: ROC curve AUC comparison of methods for a single protein-ligand query. Entries in brackets for Ligity are the mean across all the query descriptors, because we did not find the specific structure (second column) in the sc-PDB cluster we use as input. Entries with the best AUC amongst 3D methods (*i.e.* ElectroShape, DOCK and Ligity) are highlighted in bold.

DUD-E Receptor	PDB Code	Fingerprints		ElectroShape	DOCK	Ligity	Ligity (fused)
		MACCS	Morgan				
ACE	3bkl	0.922	0.831	0.452	0.716	**0.749**	0.948
ADA	2e1w	0.978	0.886	0.714	0.764	**0.857**	0.897
CDK2	1h00	0.462	0.638	0.433	**0.791**	(0.610)	0.644
FA10	3kl6	0.808	0.903	0.664	**0.866**	0.716	0.717
FA7	1w7x	0.630	0.681	0.822	**0.879**	0.762	0.809
GCR	3bqd	0.807	0.738	0.521	0.439	**0.807**	0.869
HIVINT	3nf7	0.685	0.679	0.578	0.642	**0.717**	0.637
HIVPR	1xl2	0.640	0.833	0.495	0.596	**0.836**	0.877
THRB	1ype	0.613	0.758	0.646	**0.813**	0.709	0.752
TRY1	2ayw	0.577	0.669	0.320	**0.934**	(0.725)	0.778
mean		0.712	0.762	0.565	0.744	**0.749**	0.793

4.3.8 An Example Result

Using the 2e1w ADA protein-ligand structure as a query, we ran a retrospective virtual screen over the ADA actives and decoys from DUD-E (fusion AUC of 0.948). One of the top hits is shown in Figure 4.12.

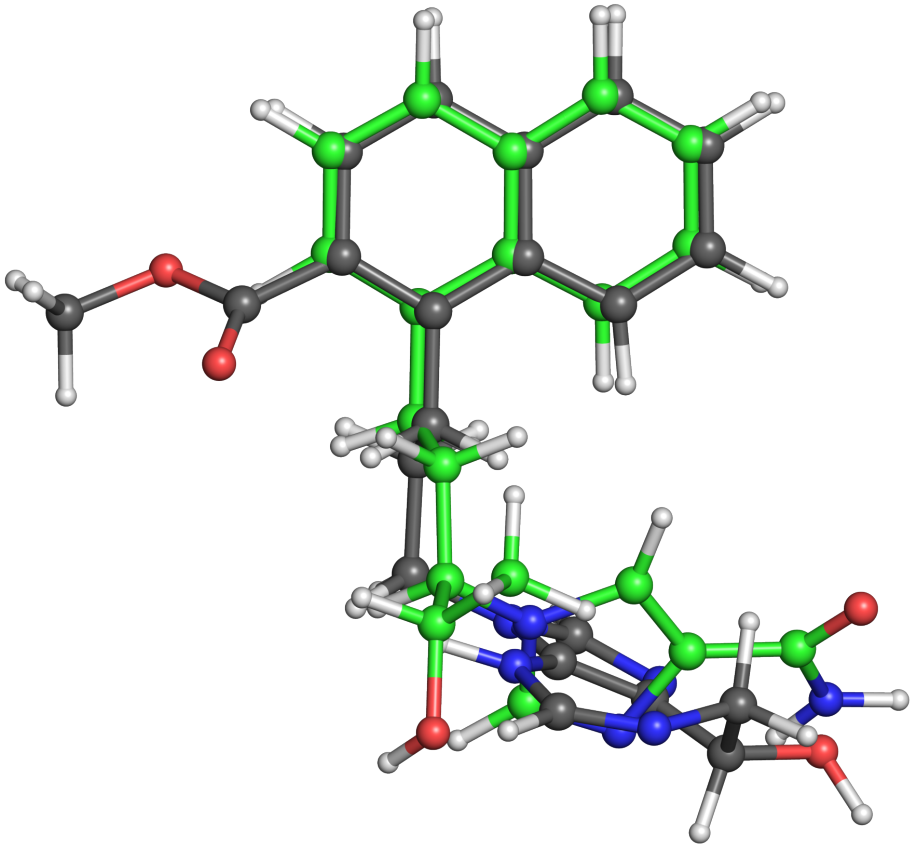


Figure 4.12: Manually superimposed query (green C atoms) and one of the top hits (grey C atoms) in the Ligity virtual screen. The similarity between the two is clear.

4.4 Conclusions

In this chapter we have presented Ligity, a fully automated, pharmacophore-based, high-throughput virtual screening application. Ligity uses information from the cognate ligand-protein structure to build a query descriptor based on the geometric arrangement of ligand pharmacophores which are in contact with the receptor. The novelty of the method is in the descriptor we generate, a six dimensional hypercube for all possible 4-PIP combinations. We compensate for receptor flexibility by populating neighbouring bins in the query descriptor, simulating different PIP arrangements. We also discriminate

the chirality of our 4-PIP tetrahedrons.

We highlight the many different parameter options available; and investigate the parameter space for better enrichment. These tests include the effect of using the single lowest energy conformer versus multiple conformers, using different bin sizes in the descriptor, using 3-PIP versus 4-PIP pharmacophores, the effect of applying different similarity metrics, and using fusion methods to aggregate single results rankings.

Interestingly, we have found that our method preferentially picks lower energy conformers as the highest scoring conformers on our validation dataset. We show that this selection is statistically significant, and is not found in decoys. We postulate that, even if counter intuitive, ligands bind in a relatively low energy conformation.

Based on the retrospective results presented, we think that Ligity has potential as a prospective virtual screening method which is able to screen millions of molecules.

5 Applications: Virtual Screening on a Pan-Malarial Drug Target, PfSUB1

In this chapter we describe our virtual screening (VS) effort to find inhibitors against *Plasmodium falciparum* subtilisin-like protease 1 (PfSUB1), a protease that has attracted attention as a pan-malarial drug target. Both structure-based and ligand-based VS experiments have been carried out. The protocols employed are outlined, and these may be generalized and applied to other biological targets of interest.

Some of this work has been published in **Plasmodium subtilisin-like protease 1 (SUB1): Insights into the active-site structure, specificity and function of a pan-malaria drug target**; Chrislaine Withers-Martinez, Catherine Suarez, Simone Fulle, Samir Kher, Maria Penzo, Jean-Paul Ebejer, Kostas Koussis, Fiona Hackett, Aigars Jirgensons, Paul Finn, Michael J. Blackman; *International Journal for Parasitology*, 42(6):597–612, 2012.

5.1 Biological Testing

The end-point of every virtual screening study should be the biological testing of the top compounds from the ranked results list. This acts as a confirmation that the compounds found computationally are indeed biological actives, and it is especially important because of the high false-positive rates in virtual screening. In this section we describe the setup

of the bioassay used to test the compounds resulting from our virtual screening studies. We also define what constitutes a bioactive hit for us.

5.1.1 Biological Assay for PfSUB1 Inhibitors

We have run a series of virtual screening experiments on PfSUB1. The commercially-available compounds resulting from these studies have been purchased and tested for activity against PfSUB1 by our EU STARS Network collaborators Maria Penzo and Michael J. Blackman at the MRC.

Their bioassay works by using a substrate which is cleaved by PfSUB1 (*i.e.* peptide `CKITAQDDEESC`) with two rhodamine groups attached to its ends. When the peptide is not cleaved, rhodamine groups form a dimer and self-quench resulting in a very low fluorescence. When the peptide is cleaved, the rhodamine groups disassociate, drift apart and fluoresce. This fluorescent signal is detected by a spectrofluorimeter allowing assessment of the cleavage of the substrate in real time. When a small molecule inhibitor binds to the PfSUB1 active site, the substrate is not cleaved and no signal is detected. Recombinant PfSUB1 was incubated in the digestion buffer with the inhibitor diluted in dimethyl sulfoxide (DMSO) in a 96-well microtiter plate. Inhibitors were routinely tested at 50 μM. Compounds that displayed an IC_{50} lower than 30 μM were further tested at a range of serial dilutions and the IC_{50} value was confirmed. The negative control was DMSO alone. The positive control contained *p*-hydroxymercuribenzoate (pHMB) at 1 mM. pHMB is a very weak inhibitor, so its concentration has to be high in order to inhibit PfSUB1. The solutions were tested every 10 minutes for 90 minutes, when the fluorescence intensity reached a steady state [Penzo, private communication; 2013]. For further details on the assay, the reader is directed to Blackman et al. [2002].

5.1.2 Defining Bioactive Hits

Our virtual screening studies produced a ranked list of small-molecules, a selection of which were then purchased and tested for activity in a bioassay (described in Section 5.1.1) by our collaborators at the MRC. If a molecule exhibited an $IC_{50} \leq 50$ μM when

biologically tested, then it was considered bioactive or a 'hit'. Anything above this hit cut-off will be considered a very weak binder, and therefore uninteresting. In a recent study Zhu et al. [2013] surveyed 421 VS publications (all between 2007 and 2011) to suggest possible practical recommendations for hit identification and optimisation. Of these, only 121 studies (30%) reported a clear hit cut-off. From the 121 studies that reported a clear cut-off: 7 used a ≤ 1 μM threshold, 44 used a threshold in the 1-25 μM range, 20 used a threshold in the 25-50 μM range, 25 used a threshold in the 50-100 μM range, 19 used a threshold in the 100-500 μM range and 6 used a threshold > 500 μM. Note that these figures are a close approximation of Figure 1 in the work of Zhu et al. [2013]. The authors argue that the aim of virtual screening is to discover novel chemical scaffolds; sub-micromolar inhibitors while desirable are not necessary as hits may be optimised subsequently. We therefore think that our hit threshold of 50 μM is justified.

5.2 PfSUB1 Structure-Based Virtual Screening

In this section we describe the SBVS experiment we carried out on PfSUB1. First, we describe the setup of the experiment. Second, we present some of the top hits of the virtual screening run. Last, we report the results of the biological testing.

5.2.1 The Ingredients of a Structure-Based Virtual Screening Experiment

There are three components of a SBVS experiment:

1. **A docking protocol**. The docking software and its runtime parameters need to be chosen appropriately. Some docking algorithms may perform better on specific receptor types.

2. **The protein structure**. In the ideal case this is a high-resolution, experimentally-resolved structure. If this is not available, a homology model may be built based on similar proteins.

3. **A database of ligands**. A library of molecules that are to be ranked against the receptor using the selected docking protocol. These should be prepared according to the chosen protocol (*e.g.* ensembles of low energy 3D conformers are required for rigid docking).

In the next sections we describe and justify our choices of these three components.

Docking Protocol

For our PfSUB1 SBVS run, we used the GOLD docking software from the Cambridge Crystallographic Data Centre [Jones et al., 1995, 1997]. A study by Sousa et al. [2006] places this program as the second most popular docking tool after AutoDock. GOLD uses a genetic algorithm to search the ligand conformational space in a binding site. The genetic algorithm operators mutation, crossover and migrate randomly modify a 'chromosome' which encodes the ligand conformation in the binding site. Each chromosome is given a score by a fitness function. There are various fitness functions used to calculate the score of every docking pose, *i.e.* Goldscore (the default), Chemscore, Astex Statistical Potential (ASP) and ChemPLP. Goldscore takes into account hydrogen-bonding energy, van der Waals energy, metal interaction and ligand torsion strain [Jones et al., 1997]. Chemscore is an empirical scoring function based on a regression parametrized on 82 complexes with known binding affinity [Eldridge et al., 1997; Verdonk et al., 2003]. ASP is a knowledge-based atom-atom distance potential derived from a database of protein-ligand complexes [Mooij and Verdonk, 2005]. ChemPLP uses the Chemscore hydrogen-bond and internal energy terms as well as a piecewise linear potential function which captures both the attractive and repulsive forces of neutral contacts and the repulsive part for anti-complementary contacts (*e.g.* donor-donor) [Korb et al., 2009]. ChemPLP is shown to be the fastest (it is a factor of four faster than Goldscore) and performs best when compared to the other three scoring functions [Liebeschuetz et al., 2012]. The fittest (highest-scoring) chromosomes are used as parents to generate fitter children (new chromosomes) in the genetic algorithm.

The 2011 release of the GOLD suite was used for this work (version 5.0.1), which also includes a visualisation program called Hermes and an analysis tool called Goldmine.

GOLD has a 'search efficiency' setting that controls the trade-off of speed versus accuracy, at 100% 30,000 genetic algorithm operations (*i.e.* mutation, crossover and migration) are performed. This search efficiency setting was set at 35% (for a reference, the suggested value for 'virtual screening' mode in the GOLD software manual is of 30% [GOLD User Guide]). ChemPLP was used as the fitness function. Also, the following binding site side-chains were considered flexible: LYS136, PHE162, LYS465, LYS467, PHE491, and PHE493. This is an important consideration because depending on the orientation of these residues, pockets or clefts in the binding site may be enlarged; hence improving the druggability of the target. These residues control access to the S1 and S4 pockets (see next section for more details). Also, side-chain flexibility allows for the 'induced fit' of the protein around the ligand.

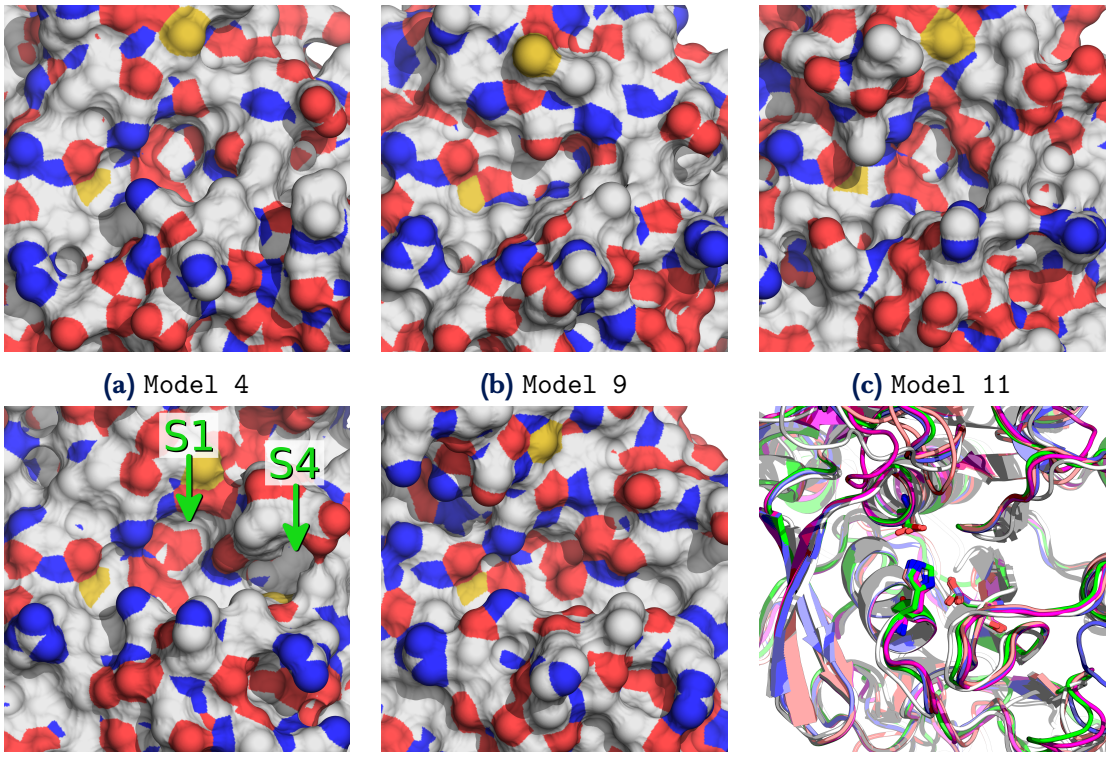
Ensemble docking was set up for our SBVS experiment. In GOLD's ensemble docking a ligand is docked into multiple protein structures, and the best scoring ligand is kept. The multiple structures mimic receptor flexibility. Various studies have shown that using multiple structures improves the overall performance of SBVS, both in terms of actives identification and pose prediction [Huang and Zou, 2007; Rao et al., 2008; Rueda et al., 2009; Totrov and Abagyan, 2008]. Korb et al. [2012] tested the performance of ensemble docking using GOLD and found that using ensembles of proteins almost always improved on the worst result of the single protein structure docking run. They argue that "as the virtual screening performance of a single protein structure is unknown, *a priori*, the use of multiple protein structures in an ensemble docking protocol minimizes the risk of giving the worst virtual screening performance possible". Also in many of their cases the ensemble performance does better than the average of the single protein structure docking runs. In some cases, ensemble docking did better than the best single protein structure docking run. Ensemble docking is shown to work also when homology models are used instead of experimentally resolved structures [Novoa et al., 2010]. GOLD is more efficient at running ensemble docking rather than if each protein structure was run separately. We have also used ligand efficiency when post-processing the GOLD docking results (refer to Section 5.2.2 for more details).

Protein Structure

Ideally, we would use an experimentally resolved structure of PfSUB1 but at the time of writing there is no published structure of this protease. When this study was carried out, only a homology (comparative) model published by Withers-Martinez et al. [2002] existed. We built several homology models, based on more recent, high-quality PDB structures which we used in the GOLD ensemble docking. Lately, one of our MRC collaborators, Chrislaine Withers-Martinez, has successfully crystallized PfSUB1 – a task that has eluded researchers for years. We obtained early access (prior publication) to this experimentally determined structure and found that our models have very similar active sites.

We built thirteen homology models using various alignment methods and modelling software. We used five high-resolution (from 1.4 Å to 2.0 Å) template structures from the PDB (`1DBI` [Smith et al., 1999], `1SCJ` [Jain et al., 1998], `1MEE` [Dauter et al., 1991], `1BH6` [Eschenburg et al., 1998], and `1THM` [Teplyakov et al., 1990]). These had between 28% and 32% sequence identity to the PfSUB1 target. We used pairwise and multiple sequence alignment of the template sequences to the target structure using Cobalt [Papadopoulos and Agarwala, 2007], Modeller [Šali and Blundell, 1993] and MOE [MOE, online]. Finally, we built the homology models using either Modeller or MOE, and we used both single and multiple templates. The thirteen homology models were visually inspected and submitted to SWISS-MODEL model checker [Arnold et al., 2006; Bordoli et al., 2009; Peitsch, 1995]. Out of the thirteen models inspected we kept three models (labelled `Model 4`, `Model 9` and `Model 11`).

In Figure 5.1 we compare the binding sites of the homology models with the recent X-ray crystallographic structure of PfSUB1 (to date, unpublished) we obtained from our collaborators at the MRC [Withers-Martinez, private communication; 2013]. The binding sites are very similar in terms of the protein backbone but their surfaces vary. Loops outside of the binding site vary widely between the models. However this is not problematic for our docking as we are able to confine the search space to the binding site. Withers-Martinez et al. [2012] provide experimental evidence based on substrate composition preferences that a hydrophobic pocket exists at the S4 position (shown in the MRC



(a) `Model 4`

(b) `Model 9`

(c) `Model 11`

(d) MRC Homology model published in Withers-Martinez et al. [2002]. Polar pocket S1 and hydrophobic pocket S4 are highlighted.

(e) Experimentally determined PfSUB1 structure (unpublished)

(f) The backbone of the binding site of the X-ray structure (white) is very similar to our models and the original MRC model (green). Differing side-chain orientations however, produce different surfaces.

Figure 5.1: Comparison of our three PfSUB1 homology models to the MRC model and the recent experimentally resolved structure by Chrislaine Withers-Martinez.

homology model in Figure 5.1d). This hydrophobic pocket is absent in our models (Figures 5.1a, 5.1b and 5.1c), but is present most notably in the MRC model (Figure 5.1d). This pocket is 'lost' in our models because of the orientation of two phenylalanines (`PHE491` and `PHE493`). In our GOLD docking experiment we set these side-chains as flexible to give the ligand access to that pocket. We also set the side-chains of residues `LYS136`, `PHE162`, `LYS465`, and `LYS467` as flexible. This gives better ligand access to the polar S1 pocket (shown in the MRC homology model in Figure 5.1d).

Our three models were complemented by another two models built by Simone Fulle from InhibOx (labelled `SF Model 8` and `SF Model 9`). These five models were superimposed (ensemble docking in GOLD requires superimposed structures) and protonated

using GOLD. The geometries of the added hydrogen atoms will be chemically meaningful (with X-H distances depending on atom types), but in the case of serine, threonine and tyrosine the hydroxyl orientation will be explored during the docking by the GOLD algorithm. The hydrogen atoms on the terminal nitrogen in lysine residues are also optimized during docking. A radius of 12 Å from the delta nitrogen atom of the catalytic histidine was used to define the search space over the binding site. No water molecules were considered in the docking.

Table 5.1 shows the α carbon global and active site RMSD of these models to the PfSUB1 X-ray cyrstallographic structure supplied to us prior to publication by Chrislaine Withers-Martinez (Figure 5.1e). The active site residues were defined as those within 12 Å from the α carbon of the catalytic triad histidine (in the active site). The RMSD values were calculated using the `super` function in PyMOL (version 1.4.1) [Schrödinger, LLC, 2013]. Note that these are the values before refinement.

Other uses of our homology models. Some of the models we built have been used in the Global Online Fight Against Malaria (GOFAM) project [GOFAM website]. This project is a collaboration between Prof. Arthur Olson's laboratory at The Scripps Research Institute and the World Community Grid (a globally distributed computing resource by IBM) and its aim is to discover novel antimalarial compounds. GOFAM uses AutoDock 4.2 and AutoDock Vina to dock approximately 5.6 million commercially available compounds against 204 models of 22 different classes of malarial targets (including PfSUB1) for a total of 1.16 billion docking jobs. This project led to the discovery of `GF13`, a very weak PfSUB1 inhibitor with an IC_{50} of around 200 μM.

Ligand Database

We used two compound databases for our virtual screening, InhibOx Scopius-CSpace and ChEMBL-NTD.

InhibOx Scopius-CSpace (version 3.0) is a predecessor of the database we describe in Chapter 3. It contained approximately 5.2 million commercially available compounds. We randomly selected 10% of this database for our docking study. This selection was

Table 5.1: All α carbon global and active site RMSD of the six homology models to the PfSUB1 structure determined by X-ray crystallography. The active site residues were defined as those within 12 Å from the α carbon of the catalytic triad histidine. The best RMSD is highlighted in bold.

PfSUB1 Homology Model	Global RMSD (Å)	Active Site RMSD (Å)
`Model 4`	2.75	**0.53**
`Model 9`	2.48	0.96
`Model 11`	3.63	2.79
`SF Model 8`	**2.42**	0.79
`SF Model 9`	4.49	1.05
`MRC Model`	4.12	1.86

needed in order to limit the processing time required for docking. If actives from docking are found in a bioassay, we would then proceed with a directed search. That is, we would run similarity searches of the bioactives against the full database and dock the most similar compounds.

We also used compounds from the Glaxo Smith Kline (GSK) Tres Cantos Antimalarial (TCAMS) dataset in the ChEMBL Neglected Tropical Disease (ChEMBL-NTD) archive [ChEMBL-NTD website]. ChEMBL-NTD is a data repository which contains results from primary compound screening and medicinal chemistry regarding neglected tropical diseases in developing regions. The TCAMS dataset (dataset 1 in ChEMBL-NTD), is the result of screening approximately 2 million compounds from the GSK screening library against *Plasmodium falciparum* strain `3D7` in human red blood cells. It contains 13,519 compounds which are confirmed to inhibit the parasite's growth by, at least, 80% at 2 μM concentration [Gamo et al., 2010].

5.2.2 Analysis of Docking Results

The result of docking the virtual library is a ranking of all the molecules based on their ChemPLP score. We also generated a second ranking list based on ligand efficiency. Ligand efficiency is the binding energy normalized by the number of non-hydrogen atoms

in the molecule [Hopkins et al., 2004]. In our case we use the docking score instead of the binding energy (Equation 5.1, where N_{heavy} is the number of heavy atoms in the molecule). This normalization is necessary because the scoring function has an additive nature, so larger compounds may score better simply by virtue of being able to form more hydrogen-bonds and/or van der Waals interactions. Such simplistic additivity is hardly ever observed in nature.

$$\text{Ligand Efficiency (LE)} = \frac{\text{docking score}}{N_{\text{heavy}}} \tag{5.1}$$

We took the top 10,000 results of both docking rankings (*i.e.* results sorted by the ChemPLP scoring function and the ligand efficiency re-ranking) giving us a total of 20,000 compounds and processed these in the following manner:

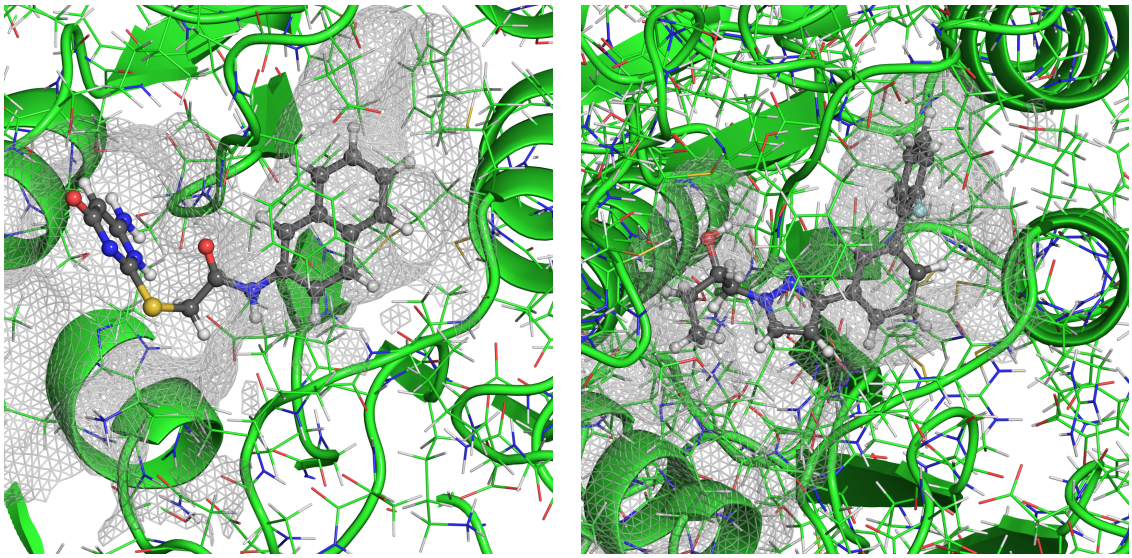
1. Kept those molecules which are common in both lists (and removed duplicate identifiers from the merged list).
2. Kept only molecules with molecular weight ≤ 450 and number of rotatable bonds ≤ 5.
3. Kept only molecules with strain energy ≤ 25 kcal/mol (as calculated using the *Calculate Descriptors* functionality in MOE).
4. Kept only those compounds that were immediately, *off the shelf*, available for purchase (and did not have long delivery times from suppliers).
5. Kept only molecules with ligand efficiency ≥ 3.5 and ChemPLP score ≥ 60.0.
6. Kept only molecules which formed at least two hydrogen bonds with the protein.

After the post-processing we were left with 6,158 compounds, the highest ChemPLP score was 112.690 and the highest ligand efficiency was 6.103.

We visually inspected the top 300 compounds from the docking ranking list and removed anything that may be reactive (*e.g.* molecules with two neighbouring heteroatoms which are not in a ring or which contain a nitrile group). Also, any unlikely and strained geometries (*e.g.* non-planar conjugated system) were also removed.

5.2.3 Example Docking Results

Two of the top results are shown in Figure 5.2. Based on the substrate specificity study we know that S1 has a preference for polar residue side-chains and S4 has a preference for hydrophobic ones (the locations of S1 and S4 in PfSUB1's binding site are shown in Figure 5.1d). It is difficult to find a molecule which spans across both areas because we filtered using rather stringent cut-offs for molecular weight and rotatable bonds. Also, the accessibility of the S4 pocket depends on the rotation of the side-chains (sampled at runtime by GOLD). The binding poses in our top results are consistent with what is known about the binding site.



(a) The polar part of this VS hit interacts with S1 in the active site. The S4 hydrophobic pocket is not fully exploited.

(b) A hydrophobic moiety on the VS hit has a complementary fit to the hydrophobic pocket in S4.

Figure 5.2: Two top hits from the GOLD docking results. Note the different PfSUB1 models.

5.2.4 Bioassay Testing Results

We ordered 71 compounds from the top 300 docked molecules we visually inspected. This selection was a compromise between cost of the compounds and the number of molecules purchased. The MRC received 1 mg of each compound with more than 85%

purity which they tested in the PfSUB1 bioassay (described in Section 5.1.1). Unfortunately, none of these 71 compounds showed an $IC_{50} < 50$ μM.

5.2.5 Closing Remarks About the SBVS Experiment

The results of the SBVS study were discouraging, but there are a number of possible explanations for the failure of this experiment.

Firstly, we did not use the experimental structure of PfSUB1 for this study. This was because the X-ray crystallographic structure was determined after we had executed this investigation. Our homology models were built with great attention to make them the highest possible quality, however, a theoretical model may still possess differences to the experimental structure that bias the selected compounds towards inactive molecules.

We used a version of the ligand database (Scopius-CSpace version 3.0) that was not curated and assembled in the rigorous manner described in Chapter 3. Commercially available chemical space is small, and molecules compatible with PfSUB1 inhibition may not be present (in the database).

SBVS has a low percentage success rate in prospective studies, and we have tested only a small number of compounds. In a relatively recent survey, Kolb et al. [2009a], argue that in docking a 5% hit rate is to be considered substantial. Also, our pre-defined hit IC_{50} threshold of 50 μM may be too low. In the aforementioned survey 4 out of 19 studies of docking screens that were followed by a confirmatory experiment identified hits with an $IC_{50} \geq 50$ μM. Molecules with such low levels of activity may be difficult to optimize, and therefore we believe the biological screening threshold is appropriate.

Thus, it is perhaps unsurprising that we have not found any actives in this study.

5.3 PfSUB1 Ligand-Based Virtual Screening

In this section we describe the LBVS experiment we carried out using known inhibitors of the PfSUB1 protease. We first describe the experimental setup and then present the results which have been validated biologically.

5.3.1 A General Overview of the Experiment

The setup of the experiment is shown in Figure 5.3. We started off with four known query molecules, for which we generated Morgan fingerprints and ElectroShape descriptors. This gives eight query descriptors which were used to search the new version of the Scopius-CSpace database (described in Chapter 3) for similar molecules. These searches resulted in a molecule list for each descriptor which is ranked by a similarity score. Only the top compounds (most similar) were considered from each list. The resulting eight reduced lists were each clustered by similarity and a molecule from each cluster partition was selected. The most similar molecule to the original query molecule for that list was selected as a representative from each cluster partition. This whole LBVS study was automated in a pipeline. A more detailed description of each step is given in the following sections.

5.3.2 The Query Molecules

The starting point of any LBVS study is a set of known inhibitors and, therefore, this approach is only possible where known inhibitors exist. Only a handful of PfSUB1 inhibitors have been reported in the literature [Arastu-Kapur et al., 2008; Gemma et al., 2012; Moneriz et al., 2011; Withers-Martinez et al., 2012; Yeoh et al., 2007]. Due to limited chemical synthesis resources, we selected query compounds which were commercially available so that they could be tested by the MRC against the target.

Arastu-Kapur et al. [2008] identified biotinylated chloroisocoumarin (JCP104) as a covalent PfSUB1 inhibitor, with an IC_{50} of 18 μM. This compound was not available for purchase from our chemical suppliers. Also, the covalent nature of the inhibitor makes it less attractive as starting point from a drug discovery perspective. Covalent drugs raise safety profile concerns, because of the possibility of irreversible off-target binding [Singh et al., 2011].

Our work with the MRC has identified a peptidyl α-ketoamide (KS-466) based on the PfSUB1 substrate that was found to inhibit PfSUB1 with an IC_{50} of ~1 μM [Withers-Martinez et al., 2012]. However, KS-466 is not a good starting point for this small-molecule study

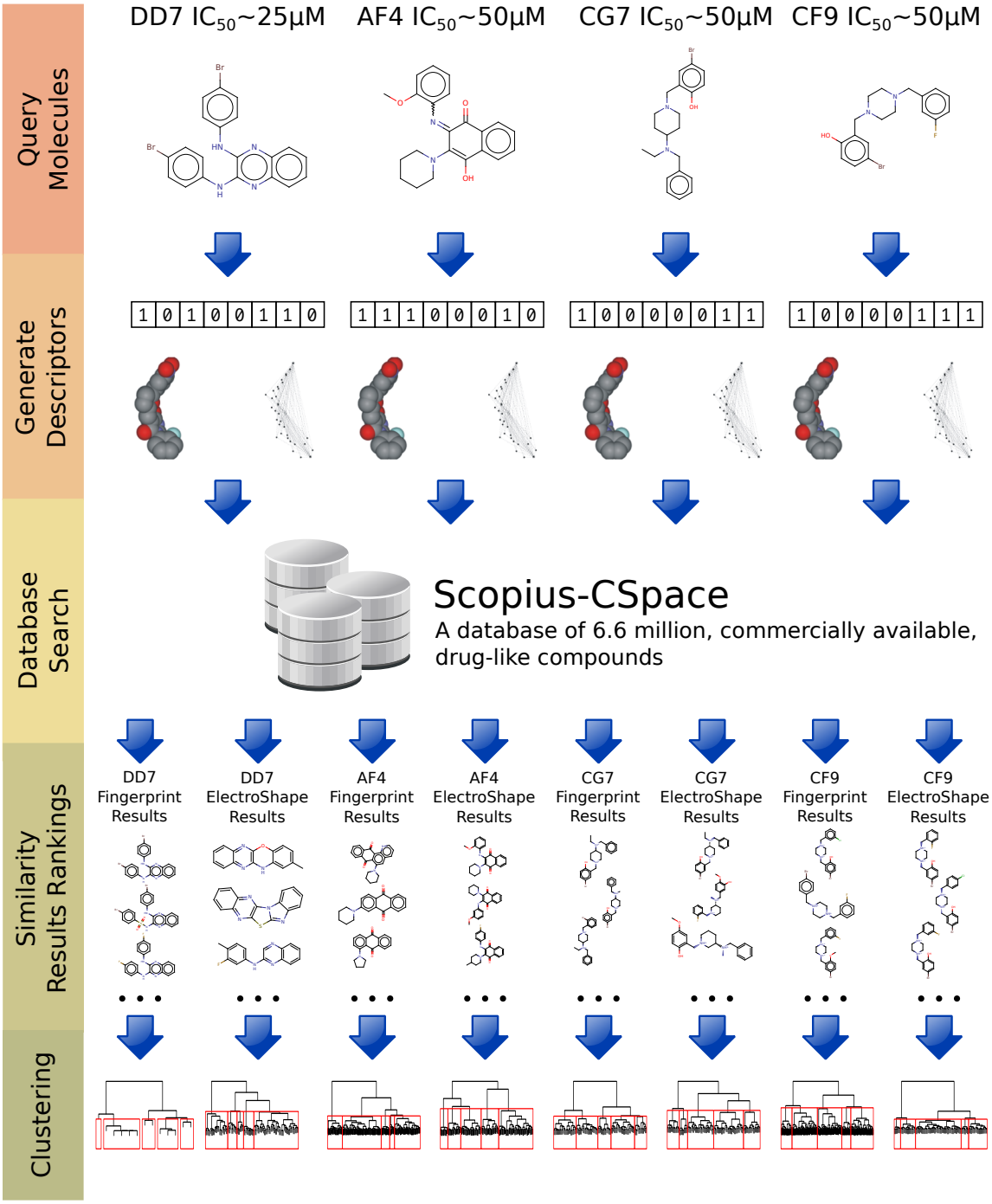


Figure 5.3: The procedure used for the PfSUB1 LBVS experiment.

as it is large (molecular mass of 676.2 Da), too peptide-like and binds covalently to the target. Also, we would be searching a database of drug-like compounds which have a different profile than KS-466.

Yeoh et al. [2007] report a compound, MRT12113, with molecular mass of 595.2 Da and with an IC_{50} of 0.3 μM. This compound was identified by screening 170,000 low molecular weight compounds (both commercial and proprietary libraries). This compound was not commercially available at the time of this study.

Gemma et al. [2012] identified a quinolylhydrazone as a hit compound with an IC_{50} of 20 μM. The authors raise concerns about the safety of the compound because of the presence of a nitro group. This hit was discovered by performing a high throughput screen of a proprietary library, and is therefore not commercially available.

Moneriz et al. [2011] report maslinic acid, a compound derived from natural sources, as a PfSUB1 inhibitor. It is also an inhibitor of subtilisin (with an IC_{50} of ~50 μM). Subtilisin is a serine protease which is closely related to the subtilases 1, 2 and 3 in *Plasmodium Falciparum* (*i.e.* PfSUB1, PfSUB2, and PfSUB3 respectively). We have purchased maslinic acid, and this has been tested by our MRC collaborators in their PfSUB1 bioassay. Unfortunately, maslinic acid showed no inhibition activity against PfSUB1 in the MRC bioassay.

Maria Penzo, a collaborator on this project at the MRC, tested the 400 compounds from the 'Open Access Malaria Box' in the PfSUB1 bioassay [Spangenberg et al., 2013]. This free-of-charge compound box is made up of 200 drug-like compounds and 200 probe-like compounds originally selected from 20,424 compounds in the ChEMBL-NTD set. All these compounds showed activity in phenotype screening (on the parasite) consistent with an EC_{50} of less than 4 μM. The drug-like set have 'Rule-of-5' compliant physicochemical properties, have no known toxicophores and are subjected to Rapid Elimination of Swill (REOS) and Pan Assay Interference Compounds (PAINS) filters. 'Rule-of-5' provides basic guidelines to predict poor absorption or permeation of a molecule [Lipinski et al., 1997]. REOS is a set of filters to remove reactive or otherwise undesirable moieties [Walters et al., 1998]. PAINS is a set of substructure filters to identify promiscuous compounds (also known as 'frequent hitters') which are active in

multiple high throughput screens [Baell and Holloway, 2010]. Any compound which failed these constraints was placed in the probe-like set. To reduce the size of the set and to maximise chemical diversity in the Open Access Malaria Box, the compounds were clustered. Finally, an experienced group of medicinal chemists selected 200 compounds out of each of the drug-like and probe-like sets. It is important to note that, even if these compounds have a confirmed effect on the *Plasmodium Falciparum* parasite, the mechanism of action and the target of these compounds is not known.

The MRC testing of the 400 compounds (from the 'Open Access Malaria Box') in the PfSUB1 bioassay resulted in four moderate actives shown in Table 5.2. We used these four compounds as our query molecules for our LBVS study. The structures of the query molecules are shown in the top part of Figure 5.3.

Table 5.2: The four query molecules from the Open Access Malaria Box we used as starting points for our LBVS study. Note that throughout the text we use the plate identifier to refer to the query molecule.

Plate Identifier	Malaria Box Identifier	Set Origin	IC_{50} (μM)
DD7	MMV007224	probe-like	~25
AF4	MMV085203	probe-like	~50
CG7	MMV019127	drug-like	~50
CF9	MMV000356	drug-like	~50

5.3.3 Descriptor Generation

The four molecules identified through biological testing at the MRC were used as query molecules in similarity searches in our database. In order to achieve this, we needed to represent these four molecules in a homogeneous manner as they are stored in the database.

The ionization state for the four query molecules was standardized, using the rules described in Section 3.2.4. This enabled us to search the database in a consistent manner (as all the database molecules are standardized in the same way). We generated two

types of molecular descriptors for the four query molecules in this LBVS study. The first descriptor used is a 2D Morgan fingerprint with 2048 bit length and radius 2 (equivalent to Extended-Connectivity Fingerprints with diameter 4), generated using RDKit (version 2011.12.1) in Python (version 2.7.2). The second descriptor is an ElectroShape descriptor (version 2.0.2) [ElectroShape, 2010]. ElectroShape is an all-atom 4D ultra-fast shape recognition (USR) method which calculates similarity based on the inter-atomic distances of the 3D atom coordinates and partial charges from five reference points in the molecule [Armstrong et al., 2011, 2010]. Since ElectroShape requires 3D coordinates (and partial charges) and we only had 2D SMILES for our four query molecules we generated a conformer ensemble for each of the queries (using the protocol described in Chapter 2). Partial charges for the molecules were calculated using OpenBabel (version 2.3.1).

5.3.4 Searching the Ligand Database

We searched the ligand database for molecules that are similar to the four query structures. This ligand database, Scopius-CSpace (Chapter 3), consists of ~6.6 million drug-like molecules. It contains both the equivalent fingerprints (for each molecule) and ElectroShape descriptors (for each low-energy conformer stored in the database).

In the case of the AF4 query molecule, both tautomeric forms of the molecule were used to search the database (shown in Figure 5.4). The results lists from both tautomeric forms of this query molecule were then merged together as one list. This was a necessary step because Scopius-CSpace does not inherently support multiple tautomeric states.

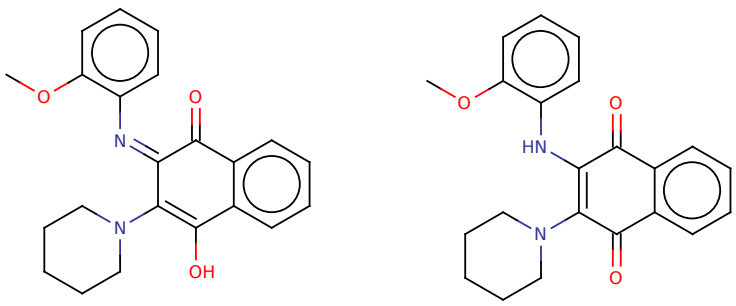


Figure 5.4: The two tautomeric forms of the AF4 query molecule.

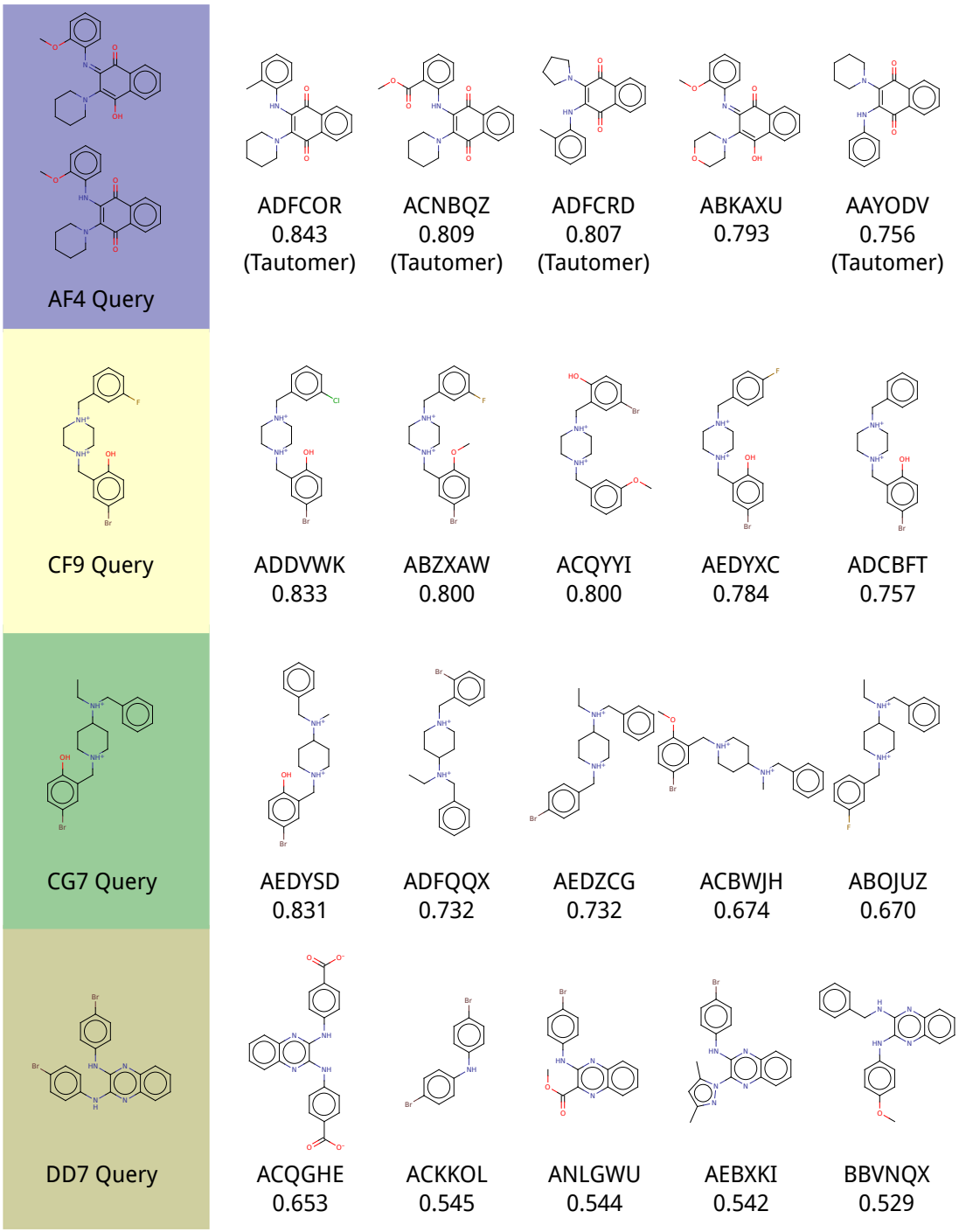
In the case of Morgan fingerprints, similarity of the query fingerprint to the database fingerprint was calculated using the Tanimoto similarity metric.

ElectroShape descriptors were generated for each conformer in the query ensemble and for each of the 189,638,700 conformers present in the database. The similarity score between a query and a database molecule is the maximum ElectroShape score between any of the query conformers and any database conformers for that molecule.

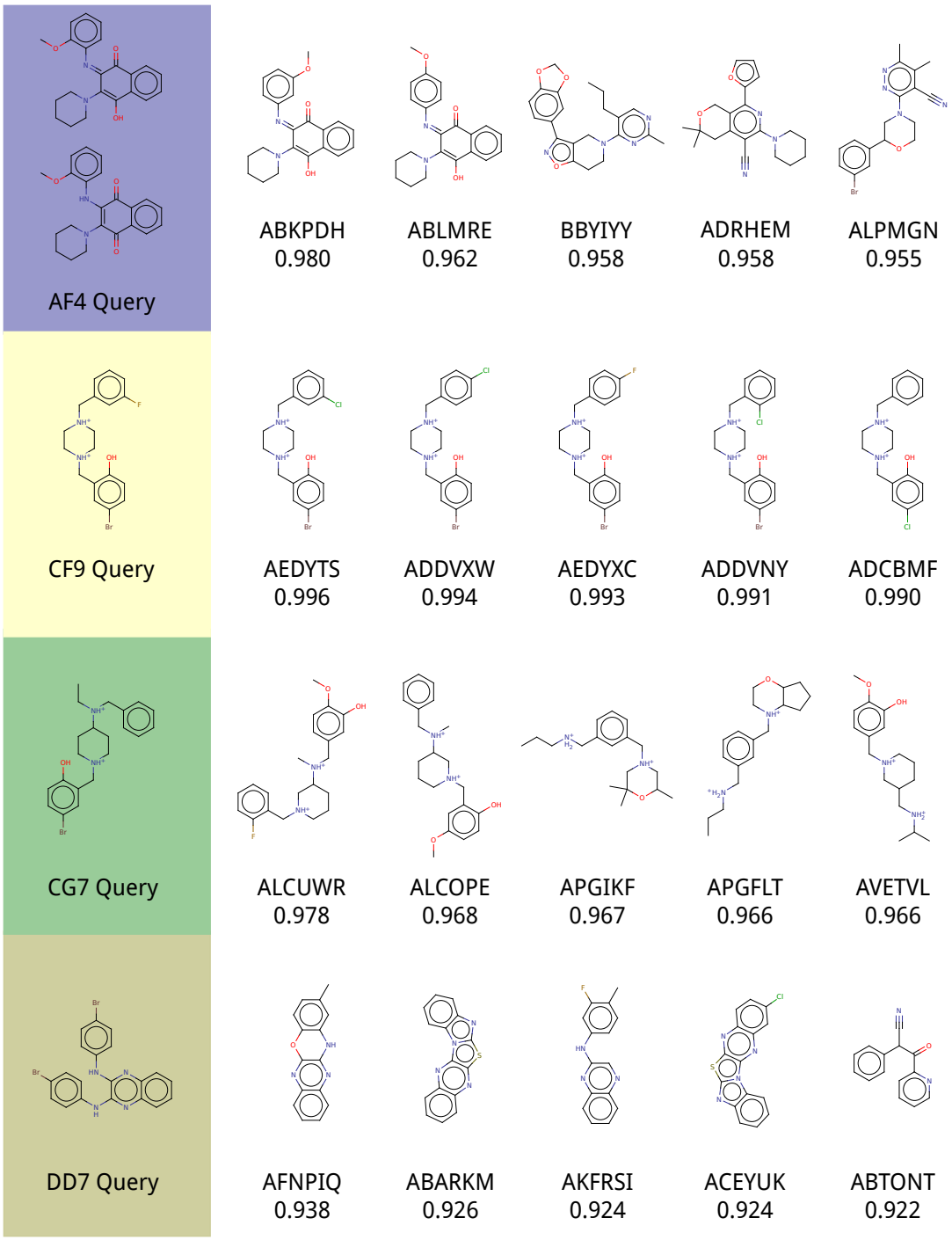
The result of the database search was eight lists ranked by descending similarity; two descriptor types $\times$ four query molecules.

5.3.5 Similarity Results

For the fingerprint results, only molecules with Tanimoto similarity of ≥ 0.5 were kept. This similarity threshold is lower than typically used in LBVS studies, but this value was forced due to the few hits returned (*e.g.* DD7 returned only 13 results at this threshold, with the top scoring molecule in the list having a similarity score of 0.653). Also, the similarity value for the Morgan fingerprints degraded quickly. For the USR method, ElectroShape, the top 100 similar molecules were kept from each ranked result list. The top five results of each query for both fingerprint and USR methods are shown in Figure 5.5. As expected, there was more structural variation in the similarity results for the USR method compared to fingerprints.



(a) Top five results for every query molecule using Morgan fingerprint similarity.



(b) Top five results for every query molecule using ElectroShape similarity.

Figure 5.5: Top five results for Morgan fingerprint and ElectroShape similarity in the PfSUB1 LBVS study.

5.3.6 Compound Clustering

The aim of this LBVS study was to act as an initial exploratory study from which identified hits could be developed into leads. We were more interested in testing different scaffolds than in testing many similar compounds. In an ideal scenario, we would acquire and test all the compounds from the similarity lists, but because we were limited by the purchasing budget and chemical synthesis resources we needed to reduce these compound lists. This was achieved by clustering each of the result lists using Ward's method [Ward, 1963]. Ward's method is an agglomerative hierarchical cluster analysis. In this 'bottom-up' approach each data point starts off in its own cluster and pairs of clusters are then merged together one-by-one based on the optimal value of an objective function. The objective function in Ward's method is the error sum of squares (this is also referred to as Ward's minimum variance method). This function aims to minimise the total within-cluster variance after each cluster merge.

The input of the clustering algorithm is the half-matrix describing all the pairwise similarity scores between the compounds in the results list. We selected a cluster size of five. Each result list was therefore clustered in five partitions. From each partition we selected the compound with the highest similarity score to the original query molecule. An example of the Morgan fingerprints similarity results, using query `DD7`, is shown in Figure 5.6.

All calculations were carried out using R (version 3.0.1) and the `cluster` package (version 1.14.4) [Maechler et al., 2013].

5.3.7 Bioassay Testing Results

After clustering, we selected five molecules (one per partition) from each result list of each query molecule. This gave us 40 molecules (5 molecules $\times$ 4 query molecules $\times$ 2 similarity methods). In addition, we also re-ordered the four query molecules to act as positive controls. For any compound that was not available immediately off-the-shelf from the supplier, we selected the second highest scoring compound and so on. Compounds had to be available in 1 mg amounts and have purity $>$ 85%. Thus, 41 compounds were

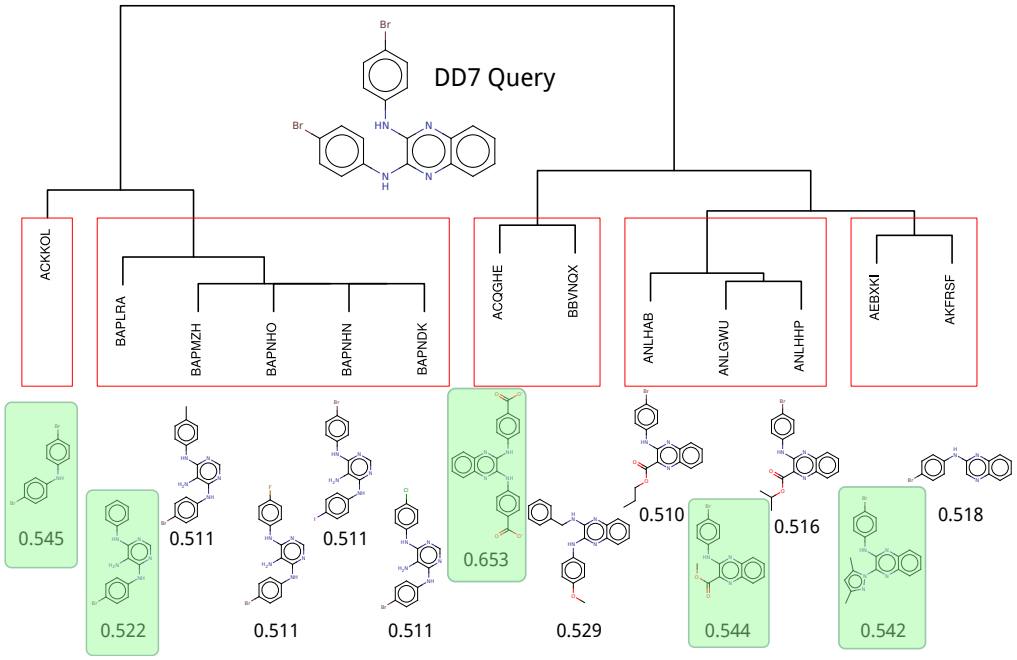


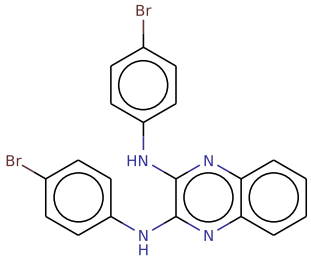
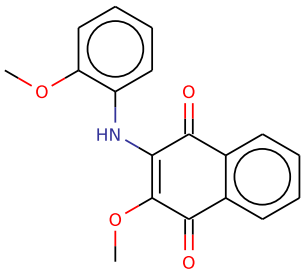
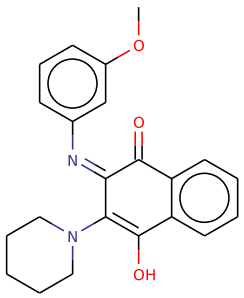
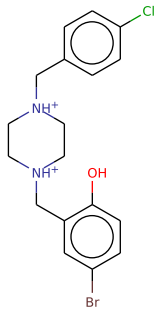
Figure 5.6: Clustering of the result list for the Morgan fingerprint search using query molecule DD7. The 13 resulting molecules are shown clustered in five groups (red rectangles). Each molecule is annotated with its Tanimoto similarity score to the query molecule. The highlighted (green) molecules are the ones which were chosen to represent the partition. These are the molecules which are most similar to the query molecule DD7.

sent to our collaborators at the MRC for bioassay testing with the PfSUB1 protease (some compounds still failed to be delivered by the suppliers).

The results of the biological testing are shown in Table 5.3. 16 out of the 41 molecules tested showed an $IC_{50} < 50$ μM. A few compounds were coloured, and could potentially interfere with (or 'quench') the fluorescent biological assay giving rise to false positives. These coloured compounds were subsequently tested at a lower concentration (6 μM), so that they became transparent or very lightly coloured in the plate. All the query molecules confirmed their activity in the same IC_{50} range of the original testing. The DD7 query molecule still shows the best inhibition with an IC_{50} of 10 μM. The vial of compound ADDVXW also included oxalic acid as a by-product of the synthesis. After testing this molecule on its own, oxalic acid was found to be a very weak inhibitor of PfSUB1 with an IC_{50} of ~140 μM. The bioactive molecules with an $IC_{50} < 50$ μM came from a 50-50 split from the two methods, Morgan fingerprints and ElectroShape. None of the results originate from the most potent query molecule DD7, one originates from CG7, four originate from CF9 and seven originate from the AF4 query. There is literature precedence of fluorescence quenching by bromobenzene [Medinger and Wilkinson, 1965] so the activity of molecules containing this moiety must be treated with caution.

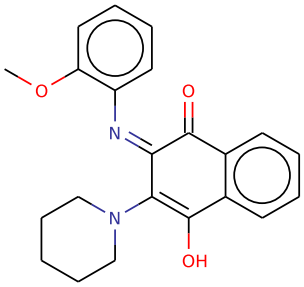
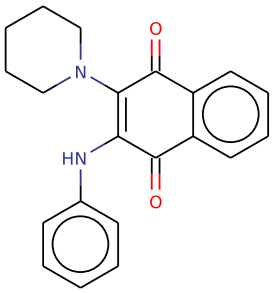
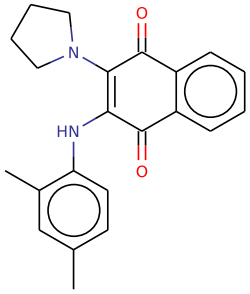
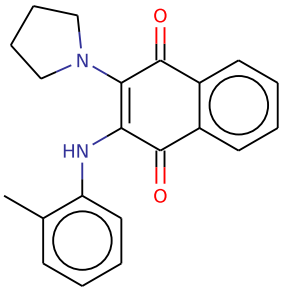
Based on these initial promising results the best hit, DD7, was tested for selectivity at several dilutions against other serine proteases: bovine trypsin, chymotrypsin and subtilisin Carlsberg. Unfortunately, these are all moderately inhibited by the compound with an IC_{50} in the 15-25 μM range. Based on this study our collaborators at LIOS have synthesised a new series. Two of the most potent compounds in this series (KS901 and KS903) have similar IC_{50} to DD7 but do not inhibit trypsin and there is very weak inhibition against chymotrypsin.

Table 5.3: Hits originating from the biological testing of the LBVS top results against the PfSUB1 protease.

Compound	Identifier	Query Molecule	Search Method	IC_{50} (μM)
	DD7	–	–	10
	ABNZUV	AF4	Fingerprints	18
	ABKPDH	AF4	ElectroShape	20
	ADDVXW	CF9	ElectroShape	22

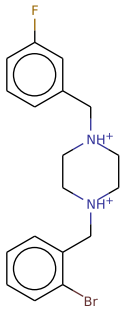
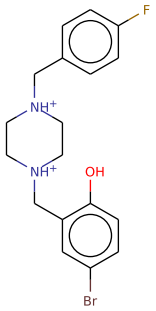
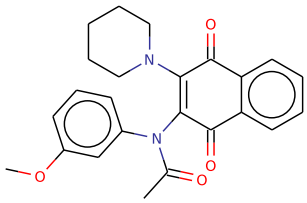
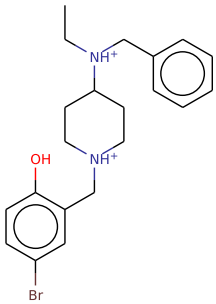
Continued on next page...

Table 5.3 – continued from previous page

Compound	Identifier	Query Molecule	Search Method	IC_{50} (µM)
	AF4	–	–	23
	AAYODV	AF4	Fingerprints	25
	AJOTOT	AF4	ElectroShape	26
	ADFCRD	AF4	ElectroShape	35

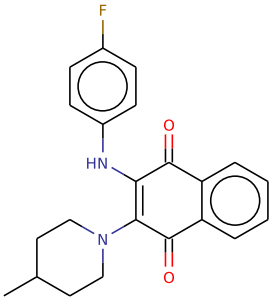
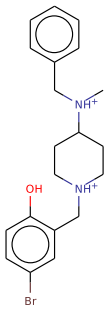
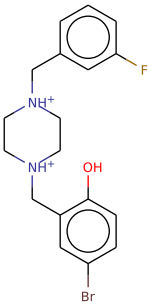
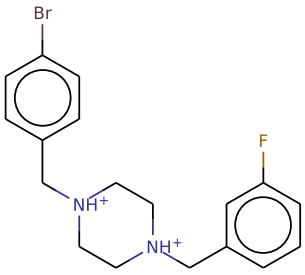
Continued on next page...

Table 5.3 – continued from previous page

Compound	Identifier	Query Molecule	Search Method	IC_{50} (µM)
	ACFGHZ	CF9	Fingerprints	35
	AEDYXC	CF9	ElectroShape	38
	ADFCVT	AF4	Fingerprints	38
	CG7	-	-	40

Continued on next page...

Table 5.3 – continued from previous page

Compound	Identifier	Query Molecule	Search Method	IC_{50} (μM)
	AJOUQQ	AF4	ElectroShape	40
	AEDYSD	CG7	Fingerprints	45
	CF9	-	-	48
	AEFGJA	CF9	Fingerprints	48

5.3.8 Closing Remarks About the LBVS Experiment

We have found an additional 12 compounds that show moderate inhibition against PfSUB1, expanding the structure-activity relationship of our hit compounds. In this LBVS study, we have identified some novel scaffolds that form the starting points for new drug discovery efforts. However, more chemistry work is required to transition some of these molecules from hits to leads.

The original query molecule DD7, which is the most potent compound of the series with an IC_{50} of 10 μM, has a bromobenzene moiety that may have a quenching effect on the bioassay. Potentially, this may mean that the molecule is a false positive and has no real activity against PfSUB1. The second best hit with an IC_{50} of 18 μM does not have this moiety. There are also concerns about the lack of selectivity of the DD7 compound. Our LBVS study provides data to guide further exploration.

Using Scopius-CSpace in a practical LBVS study highlighted possible improvements to the database (*e.g.* we presently need to run separate searches for different tautomeric forms).

5.4 Conclusions

We have carried out two prospective virtual screening studies on PfSUB1, a pan-malarial drug target. These structure-based and ligand-based studies have been partially successful, identifying a number of hit molecules.

For the SBVS study we created homology models of the target, which at a later date turned out to be very similar to the first X-ray crystallographic PfSUB1 structure that was solved after this virtual screening study was conducted. We docked ~520,000 molecules in an ensemble of PfSUB1 homology models. We rigorously filtered and selected the top VS hits to remove problematic (*e.g.* toxic or reactive) molecules. From this exercise we purchased 71 compounds which were tested for bioactivity against PfSUB1 by our collaborators at the MRC. Unfortunately, none of these compounds showed activity at an $IC_{50} < 50$ μM.

For the LBVS study we used four queries from the 'Open Access Malaria Box' exhibiting moderate PfSUB1 inhibition in the 25-50 μM range. We searched 6,604,127 drug-like molecules of the Scopius-CSpace ligand database we built in Chapter 3 using Morgan fingerprints and ElectroShape methods. We filtered the similarity results list and clustered the remaining molecules to achieve more structural diversity. We purchased 41 compounds, out of which 16 (including the original four queries) showed an $IC_{50} < 50$ μM when tested against PfSUB1. As expected, searching for similar molecules resulted in finding molecules with similar properties (including biological activity).

In our studies LBVS worked better than SBVS. This also seems to be the general consensus in the field. The reasons for this might be that fewer assumptions are made in LBVS. The starting point for our LBVS study are confirmed active molecules, whilst in SBVS the structure of the protein directs the docking exercise. We used homology models in our docking because the experimentally-resolved structure was only determined after this study. The virtual screening ligand library used in the LBVS study is much larger (approximately ten-fold) than the one used in the SBVS study (~6.6 versus ~0.5 million molecules respectively). The ligand database in the LBVS study is not only larger, but also of a much higher quality (see Section 3.3.6 for a description of the main improvements).

Finally, although PfSUB1 is a promising drug target it is also a difficult one. So far, hundreds of thousands of molecules have been physically tested against this target with meagre results.

6 Conclusions and Future Directions

This thesis describes the recent journey of a researcher in Virtual Screening (VS). We have developed computational methods, protocols and databases to improve the performance of Computer-Aided Drug Design (CADD). Some of these improvements were successfully tested in a prospective VS study using a pan-malarial drug target called PfSUB1. The methods described herein contributed to the discovery of novel inhibitors against this target.

6.1 Summary

In Chapter 1 we discussed why drug discovery is difficult. We argued that the main reason is the complex multi-objective optimization of a large number of dimensions (*e.g.* potency, safety, absorption *etc.*). We are currently experiencing an exponential growth in many databases relevant to drug discovery such as nucleotide and protein sequence databases, protein structure databases, and small-molecule databases. We assert that as more data becomes available from experimental methods, knowledge-based computational systems will improve in performance.

In Chapter 2 we reviewed a number of small-molecule conformer generation tools. Conformer generation has important implications in cheminformatics, particularly in computational drug discovery where the quality of conformer generation software may affect the outcome of a virtual screening exercise. We examined the performance of four freely available small molecule conformer generation tools (Balloon, Confab, Frog2, and

RDKit) alongside a commercial tool (MOE). The aim of this study was 3-fold: (i) to identify which tools most accurately reproduce experimentally determined structures, (ii) to examine the diversity of the generated conformational set, and (iii) to benchmark the computational time expended. These aspects were tested using a set of 708 drug-like molecules assembled from high quality X-ray crystallographic structures from the OMEGA validation set and the Astex diverse set. These molecules have varying physicochemical properties and at least one known X-ray crystal structure. We found that RDKit and Confab were statistically better than other methods at generating conformers with low RMSD from the known experimentally-resolved structure. RDKit is particularly suited for less flexible molecules while Confab, with its systematic search approach, is able to generate conformers which are geometrically closer to the experimentally determined structure for molecules with a large number of rotatable bonds (≥ 10). In our tests RDKit was also the second fastest method after Frog2. In order to enhance the performance of RDKit, we developed a postprocessing algorithm which builds a diverse and representative set of low energy conformers containing a close conformer to the known structure. Our analysis indicated that, with postprocessing, RDKit is a valid free alternative to commercial, proprietary software. Conformer generation is particularly relevant to this thesis' subsequent chapters. In Chapter 3, we used this work to generate conformer ensembles for every entry in our small-molecule database. In Chapter 4, we used it to build conformer ensembles for our virtual screening libraries. In Chapter 5, we generated conformers for our 3D similarity searches in the ligand-based part of the PfSUB1 virtual screening study. Furthermore, the choice of RDKit for conformer generation was helpful in the subsequent development of internally consistent tools and databases for drug discovery, since it provides a suite of cheminformatics functionality in the form of both a C++ and Python API, and also as a database cartridge for PostgreSQL.

In Chapter 3, we described the development of a compound database of almost 13 million commercially-available molecules (approximately half of these are 'drug-like'). We used stringent quality controls at every step of the database creation. We described the steps taken to build the database: data procurement, sanitization, salt removal, standardization of charges, assigning identifiers, deduplication, descriptor generation,

tagging, clustering, and conformer generation (using the method described in Chapter 2). The molecules were stored in a relational database which allows fast searching using indices. We also built web-based tools to access and query the database. Exact, sub-structure or similarity database searching may be executed using SMILES or SMARTS query strings or by drawing a molecular structure of interest in the user interface. The database does not currently support tautomer enumeration or canonicalization. One of the main applications of this database is for use in virtual screening studies, as described in Chapter 5.

In Chapter 4 we described Ligity, a novel virtual screening method we developed. This pharmacophoric method makes use of the growing number of cognate protein-ligand structures in the PDB. The input to the method is one or more experimentally resolved *holo*-structures for a target protein of interest. The protein and bound ligand are used to determine the interaction hot-spots on the ligand side. An interaction point is referred to as a Pharmacophoric Important Point (PIP). Electrostatic, hydrogen bond donor/acceptor, aromatic, and hydrophobic PIPs are considered. Each combination of four PIPs forms a tetrahedron with six edges. The lengths of these edges are used as indices to update a counter in a six-dimensional data structure known as a hypercube. In order to capture some aspects of receptor flexibility, adjacent bins in the hypercube may be populated using a parametrized ramp or Gaussian function. Chirality of the 4-PIP tetrahedrons is detected by ordering the vertices and calculating the volume of the tetrahedron (a change in sign implies a difference in vertex orientation). Hypercubes are generated for each conformer of every molecule in the virtual screening library we want to search. Similarity of the query hypercube to the virtual screening library hypercubes is calculated using the Tversky similarity metric. This metric captures the asymmetric nature of the query descriptor (which only includes information from the parts of the bound ligand which interact with the protein side) when compared to the hypercube describing the whole molecule in the VS library. We found that Ligity performs better when multiple input structures are used. This is in line with our argument that more data increases performance of knowledge-based systems. When using multiple input structures, the virtual screening ranking results for each input structure are fused into one list by taking

each molecule's maximum score in its conformer ensemble across all input structures. Ligity was validated on a number of protein targets in the DUD-E dataset, and showed better performance (on average) than other 3D virtual screening methods.

In Chapter 5 we described how we conducted structure-based and ligand-based virtual screening experiments to find new small-molecule inhibitors of PfSUB1. PfSUB1 is a serine protease found in *Plasmodium falciparum*, the parasite which causes the deadliest form of malaria. This protein is implicated in the parasite's egress from the red blood cell in the human stages of the parasite's life cycle. It is considered to be an attractive broad-spectrum malarial drug target. Since an experimentally determined structure of PfSUB1 did not exist when this work was begun, we built homology models of PfSUB1 using closely related structures from the PDB (bacterial substilisin with 28%-32% sequence identity). We used these protein models in docking runs with GOLD. The docking protocol included several residues that were defined to have flexible side-chains. We docked approximately half a million compounds in six PfSUB1 models which gave us a ranking based on the scoring function in GOLD. We also re-ranked the docking results based on ligand efficiency (docking score divided by the number of heavy atoms in the ligand). We then post-processed the top 10,000 molecules in both the original and the re-ranked ligand efficiency lists. This post-processing consisted of six steps: (i) we considered molecules found in both lists, (ii) kept only molecules with molecular weight ≤ 450 Da and rotatable bonds ≤ 5, (iii) kept only molecules with strain energy ≤ 25 kcal/mol, (iv) kept only commercially available molecules, (v) kept only molecules with ligand efficiency ≥ 3.5 and GOLD docking score ≥ 60, and (vi) kept only molecules which form more than two hydrogen bonds with the protein. We then visually inspected the resulting top 300 molecules in the post-processed list (ranked by docking score). After this manual selection, we ordered 71 compounds which were sent for biological testing at the MRC. None of these compounds showed activity of $IC_{50} < 50$ μM. This may be due to the use of homology models instead of an experimentally resolved structure of the target.

In the ligand-based virtual screening study, we used four compounds from the Medicines for Malaria Venture (MMV) Open Access Malaria Box as queries. These four compounds, which exhibited activity of $IC_{50} < 50$ μM when tested at the MRC,

were used as queries in several similarity searches using the drug-like subset of the new small-molecule database we developed (described in Chapter 3). The two types of similarity searches were fingerprint-based (ECFP4) and an ultra-fast shape recognition method, called ElectroShape, which uses the atomic 3D coordinates and electrostatics as an added fourth dimension. Approximately 6.6 million molecules were searched using these two methods and the four query molecules. The search results were clustered to remove very similar molecular scaffolds and the 41 top-ranking compounds were biologically tested (including the query molecules as a control). 16 compounds out of the 41 tested showed an activity of $IC_{50} < 50$ μM. This work helped expand the structure-activity relationships of our hit compounds for this important malarial drug target.

This thesis describes some of the main components in virtual screening studies, and their intricate relationships. We have improved the quality of these individual components, thereby improving the holistic process. We reviewed conformer generation tools and developed a post-processing protocol on the best-performing method. This work was used in molecular 3D similarity searches and in the assembly of a large database of small-molecules. This database was used in a successful, prospective virtual screening study on a malarial drug target. The small-molecule database will also be used in future studies using a virtual screening method, Ligity, that we developed and validated in this work.

6.2 Future Directions

In this section we discuss future directions which may enhance and build on the work presented in this thesis.

6.2.1 Conformer Generation

Conformer generation is a key aspect of small-molecule 3D modelling and, by implication, of virtual screening. We have shown that most current methods are able to reproduce experimental structures within a small RMSD tolerance of known X-ray crystallographic structures, at least for smaller, less flexible molecules (with ≤ 7 rotatable

bonds). Many conformer generation protocols may be divided into two distinct steps. First, the 3D atomic coordinates for the molecule are generated. Second, the coordinates are energy minimized using a force field to relax highly-strained structures. 3D coordinates generation approaches, such as distance geometry, are relatively fast compared to the second energy-minimization step. We therefore suggest protocols in the generation step that may help speed up the minimization step. One possible approach is to mine experimentally-resolved structure databases (such as the CSD) and extract fragments (possibly by breaking the molecule's non-terminal rotatable bonds). To generate conformers for a new molecule, the new molecule is fragmented in a similar way and the experimental fragment database is used to find components which match the molecule's fragments. These fragments are then 'stitched' together to determine the new conformer coordinates. For fragments not found in the fragment database, a fallback approach such as distance geometry may be used. The stitching together of the fragments may be based on libraries of experimentally-observed torsion angles. An example of such a library has been recently published by Schärfer et al. [2013]. The final conformers would still need to undergo energy minimization, but this could be limited to the torsion angles, as the fragments coming from the experimental fragment library should already be in a low-energy conformation. Effectively this would reduce the time required for energy minimization, and speed up the overall conformer generation process. The idea described here is similar to the recently published COSMOS method [Andronico et al., 2011; Sadowski and Baldi, 2013] and is in line with our premise that data-driven approaches will improve the performance of current cheminformatics methods.

Another step with scope for improvement is the selection of the number of conformers to generate for a molecule. Many tools require this as an input parameter, implying it is a requirement for conformer generation. In Chapter 2 we offered an empirical value for the number of conformers to generate, but a more rigorous statistical method could be built. Our suggestion generates the same number of conformers for molecules with less than eight rotatable bonds. We then cluster the conformer ensemble based on RMSD, which removes many similar conformers for molecules with few rotatable bonds. This is inefficient, as many redundant conformers that are generated and energy minimized

are then removed by the clustering procedure. The relationship between the molecule and the number of conformers to generate is not linear, as the conformer space increases exponentially with every rotatable bond (at torsional increments of 1° the theoretical conformer space is 360^n, where n is the number of rotatable bonds). The number of conformers to generate should be a built-in calculation in conformer generation methods, and these should not produce conformers which are already represented in the conformer ensemble. The number of generated conformers should be enough to sample the low-energy conformer space adequately.

6.2.2 Building a Small-Molecule Database

Our database of commercially-available small-molecules has been successfully used to find new inhibitors of PfSUB1 in our LBVS study. Still, several enhancements may be implemented to improve the quality and completeness of the database. First, the database has to be updated regularly with only the supplier's catalogue updates being passed through the pipeline (shown in Figure 3.3). This is not only a question of adding molecules to the database, as molecules which have become unavailable need to be updated as well (*i.e.* have their 'available' tag removed). Indeed, keeping the database synchronized with the suppliers' catalogues in real-time is a challenge, and would require real-time access to these catalogues.

Currently, the 3D coordinates of each molecule's conformers are stored in the database as a MDL Molfile block. This is inefficient as it repeats invariant data for every conformer (*e.g.* the connection table does not change between conformers of the same molecule). Also, this chemical file format has no way of handling multiple conformers, tautomeric states and ionization (charge) states at different pH (one would need to store each instance as a separate molecule using this format). There are a plethora of chemical file formats, which would need to be reviewed to determine the best way to efficiently store the molecules. Molecular representation was and continues to be an active area of research in cheminformatics.

One shortcoming in the current database is the lack of tautomer enumeration and canonicalisation. Most publicly available databases do not presently handle tautomers,

because of the large computational processing and storage requirements this would entail. As a starting point, the most stable tautomer for the molecule could be stored. This requires the pK_a to be calculated. Also, multiple stable tautomeric forms may exist. Standardization of a molecule remains an open problem in cheminformatics [O'Boyle et al., 2011b].

The tagging sub-system could be extended from its current binary form (*e.g.* 'drug-like' or not 'drug-like') to store quantitative values. Specifically, adding the QED measure of drug-likeness would be useful (as discussed in Section 3.2.7). Also, to increase the database's coverage of chemical space one could add dynamic queries of 'virtual' space. Using known synthetic chemical reactions and the molecules present in the database, virtual molecules could be built and searched on-the-fly.

Another area where further development is required is the linking of our database with other repositories, such as ChEMBL. This would allow us to annotate individual compounds in our database with biological activity (amongst other things). For example, the user interface could link to ChEMBL, to show available biological activity data of (similar or identical) Scopius-CSpace database molecules. Also, the database should be extended to support the storing and querying of data pertaining to our own biological screens (*e.g.* PfSUB1 LBVS study).

6.2.3 Ligity

Ligity requires one or more cognate ligand-protein structures to determine the location of the pharmacophoric important points (PIP). This necessitates that at least one binder must be known, and an experimentally-resolved structure of this molecule bound to the protein of interest must exist. The first requirement (of the known binder) is a common critique of all LBVS methods, but the second is an additional constraint imposed by the method. This may be circumvented using binding poses of known active compounds generated from docking or molecular dynamics (MD) simulations snapshots – however the method will still have to be validated when using this type of input. Also, the generation of PIPs from multiple MD snapshots which mimic receptor flexibility should be investigated (this is similar to the relaxed complex scheme in docking). Alternatively,

one could determine PIP positions by analysing atomic affinity grids such those employed in docking. We have performed some preliminary experiments in this area, but this approach will require more work to match the performance when using known structures of complexed ligands.

Further work is required to determine if there are specific PIP combinations which are statistically under- or over-represented. This may imply that binding happens because of a small number of PIP combinations, and we could decide to weight these combinations differently. The performance of Ligity may improve if we remove PIP combinations that are present across many different targets (we could consider these as background noise). Removing these common PIP combinations from the descriptors would decrease their size, which in turn would make the method faster by reducing the number of populated bins that would need to be compared.

Currently, query and target descriptors are compared using the Tversky similarity metric which is able to capture the substructure nature of the query (when compared to the target). But many other similarity measures could be useful. Of these, machine learning is of particular interest due to its ability to capture non-linear relationships (such as the ones mentioned in the previous paragraph).

Water in the binding site may be predicted using a method like WaterDock [Ross et al., 2012], which allows to predict the position of water molecules and identifies which water molecules are displaced or conserved. Hydrogen bond acceptor and donor PIPs could be added in the positions of the displaced water molecules. This way Ligity would be searching for molecules that displace and mimic the water, similarly to what happens in the p38α MAP kinase target. Performance analysis and validation of this experiment are material for future work.

Finally, the use of Ligity in a prospective VS study is the foremost priority of this project. Due to the lack of any structural data on PfSUB1, it was not possible to test Ligity on this target. Determining the X-ray crystal structure of PfSUB1 with a natural substrate bound (a decapeptide) is in the pipeline. When this becomes available, we would like to test and extend the method to work using peptides. Handling larger molecules like peptides is a current limitation in many VS methods.

In order to make Ligity a truly high throughput method and achieve an extra speed-up, the hypercube similarity calculation could be ported to the GPU. Each bin in the hypercube is independent of any other in the data structure and processing on each query and target bins can be done in parallel on the many GPU cores.

6.2.4 PfSUB1 Virtual Screening Studies

PfSUB1 is a difficult target for drug discovery. Once the X-ray crystallographic PfSUB1 structure becomes publicly available, the first step will be to repeat the SBVS experiment using this structure rather than the homology models. This should help reduce the degree of uncertainty in the study. Secondly, the new Scopius-CSpace database we have developed should be used in the SBVS study, and cloud computing resources employed to be able to dock more than the original 500,000 compounds. It would be interesting, after including the original virtual screening library in the experiment, to compare where these molecules place in the full final ranking.

The next step in the LBVS study will be to order and test the compounds found in the clusters from which the 12 new hits originated. Each cluster contains the most similar molecules to the original hits in the LBVS study. We have also started developing a method, called Molective, which fragments the molecules and determines the probability of activity for that fragment based on a series' structure-activity relationship (SAR). Unfortunately, we have not been able to apply this new method yet because of the limited (and quite distinct) SAR for this molecular series. The current test set for this method is the modified peptide series (which contains an α-ketoamide) tested on PfSUB1. This series contains 37 α-ketoamide-containing decapeptides of which 18 exhibit an $IC_{50} \leq 50$ μM. The residues at the different sub-site positions may be used to train a Naive-Bayes Classifier to predict activity.

6.3 Final Words

We developed computational methods and resources which are useful to the drug discovery community. For example, the dataset we assembled to test conformer generation

methods is currently being used to validate other methods. Although success rates are still relatively low, we see great promise in the area of Computer-Aided Drug Design. Further contributions are required, which will not only make use of the growing volume of chemical, biological and pharmacological data, but also of the increasing computational resources (*e.g.* made available through cloud-computing). It is time for this area to fulfil its promise, to help bring new drugs to the market and stave off the pathogens' rising resistance against current therapeutics.

References

Abdulla, S., Oberholzer, R., Juma, O., Kubhoja, S., Machera, F., Membi, C., Omari, S., Urassa, A., Mshinda, H., Jumanne, A., Salim, N., Shomari, M., Aebi, T., Schellenberg, D. M., Carter, T., Villafana, T., Demoitié, M.-A., Dubois, M.-C., Leach, A., Lievens, M., Vekemans, J., Cohen, J., Ballou, W. R., and Tanner, M. Safety and Immunogenicity of RTS,S/AS02D Malaria Vaccine in Infants. *N. Engl. J. Med.*, 359(24): 2533–2544, 2008.

Achan, J., Talisuna, A., Erhart, A., Yeka, A., Tibenderana, J., Baliraine, F., Rosenthal, P., and D'Alessandro, U. Quinine, an old anti-malarial drug in a modern world: role in the treatment of malaria. *Malar. J.*, 10(1):144, 2011. ISSN 1475-2875.

Agnandji, S. T., Lell, B., Fernandes, J. F., Abossolo, B. P., Methogo, B. G., Kabwende, A. L., Adegnika, A. A., Mordmuller, B., Issifou, S., Kremsner, P. G., Sacarlal, J., Aide, P., Lanaspa, M., Aponte, J. J., Machevo, S., Acacio, S., Bulo, H., Sigauque, B., Macete, E., Alonso, P., Abdulla, S., Salim, N., Minja, R., Mpina, M., Ahmed, S., Ali, A. M., Mtoro, A. T., Hamad, A. S., Mutani, P., Tanner, M., Tinto, H., D'Alessandro, U., Sorgho, H., Valea, I., Bihoun, B., Guiraud, I., Kabore, B., Sombie, O., Guiguemde, R. T., Ouedraogo, J. B., Hamel, M. J., Kariuki, S., Oneko, M., Odero, C., Otieno, K., Awino, N., McMorrow, M., Muturi-Kioi, V., Laserson, K. F., Slutsker, L., Otieno, W., Otieno, L., Otsyula, N., Gondi, S., Otieno, A., Owira, V., Oguk, E., Odongo, G., Woods, J. B., Ogutu, B., Njuguna, P., Chilengi, R., Akoo, P., Kerubo, C., Maingi, C., Lang, T., Olotu, A., Bejon, P., Marsh, K., Mwambingu, G., Owusu-Agyei, S., Asante, K. P., Osei-Kwakye, K., Boahen, O., Dosoo, D., Asante, I., Adjei, G., Kwara, E., Chandramohan, D., Greenwood, B., Lusingu, J., Gesase, S., Malabeja, A., Abdul, O., Mahende, C., Liheluka, E., Malle, L., Lemnge, M., Theander, T. G., Drakeley, C., Ansong, D., Agbenyega, T., Adjei, S., Boateng, H. O., Rettig, T., Bawa, J., Sylverken, J., Sambian, D., Sarfo, A., Agyekum, A., Martinson, F., Hoffman, I., Mvalo, T., Kamthunzi, P., Nkomo, R., Tembo, T., Tegha, G., Tsidya, M., Kilembe, J., Chawinga, C., Ballou, W. R., Cohen, J., Guerra, Y., Jongert, E., Lapierre, D., Leach, A., Lievens, M., Ofori-Anyinam, O., Olivier, A., Vekemans, J., Carter, T., Kaslow, D., Leboulleux, D., Loucq, C., Radford, A., Savarese, B., Schellenberg, D., Sillman, M., and Vansadia, P. A phase 3 trial of RTS,S/AS01 malaria vaccine in African infants. *N. Engl. J. Med.*, 367 (24):2284–2295, Dec 2012.

Agrafiotis, D. K., Lobanov, V. S., and Salemme, F. R. Combinatorial informatics in the post-genomics ERA. *Nat. Rev. Drug Discov.*, 1(5):337–346, May 2002.

Agrafiotis, D. K., Gibbs, A. C., Zhu, F., Izrailev, S., and Martin, E. Conformational sampling of bioactive molecules: A comparative study. *J. Chem. Inf. Model.*, 47(3):1067–1086, 2007.

Allen, F. H. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr. B*, 58(3 Part 1):380–388, Jun 2002.

Alonso, H., Bliznyuk, A. A., and Gready, J. E. Combining docking and molecular dynamic simulations in drug design. *Med. Res. Rev.*, 26(5):531–568, 2006. ISSN 1098-1128.

Amaro, R. E., Baron, R., and McCammon, J. An improved relaxed complex scheme for receptor flexibility in computer-aided drug design. *J. Comput. Aided Mol. Des.*, 22(9):693–705, 2008. ISSN 0920-654X.

Andronico, A., Randall, A., Benz, R. W., and Baldi, P. Data-Driven High-Throughput Prediction of the 3-D Structure of Small Molecules: Review and Progress. *J. Chem. Inf. Model.*, 51(4):760–776, 2011.

Arastu-Kapur, S., Ponder, E. L., Fonovic, U. P., Yeoh, S., Yuan, F., Fonovic, M., Grainger, M., Phillips, C. I., Powers, J. C., and Bogyo, M. Identification of proteases that regulate erythrocyte rupture by the malaria parasite Plasmodium falciparum. *Nat. Chem. Biol.*, 4(3):203–213, 2008. ISSN 1552-4450.

Armstrong, M. S., Morris, G. M., Finn, P. W., Sharma, R., and Richards, W. G. Molecular similarity including chirality. *Journal of Molecular Graphics and Modelling*, 28(4):368–370, 2009. ISSN 1093-3263.

Armstrong, M. S., Finn, P. W., Morris, G. M., and Richards, W. G. Improving the accuracy of ultrafast ligand-based screening: incorporating lipophilicity into ElectroShape as an extra dimension. *J. Comput. Aided Mol. Des.*, 25(8):785–790, 2011. ISSN 0920-654X.

Armstrong, S. M., Morris, G. M., Finn, P. W., Sharma, R., Moretti, L., Cooper, R. I., and Richards, W. G. ElectroShape: fast molecular similarity calculations incorporating shape, chirality and electrostatics. *J. Comput. Aided Mol. Des.*, 24:789–801, 2010. ISSN 0920-654X.

Arnold, K., Bordoli, L., Kopp, J., and Schwede, T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*, 22(2):195–201, 2006.

Babaoglu, K., Simeonov, A., Irwin, J. J., Nelson, M. E., Feng, B., Thomas, C. J., Cancian, L., Costi, M. P., Maltby, D. A., Jadhav, A., Inglese, J., Austin, C. P., and Shoichet, B. K. Comprehensive Mechanistic Analysis of Hits from High-Throughput and Docking Screens against β-Lactamase. *J. Med. Chem.*, 51(8): 2502–2511, 2008.

Baber, J. C., Shirley, W. A., Gao, Y., and Feher, M. The Use of Consensus Scoring in Ligand-Based Virtual Screening. *J. Chem. Inf. Model.*, 46(1):277–288, 2006.

Baell, J. B. and Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.*, 53(7):2719–2740, 2010.

Balkenhohl, F., von dem Bussche-Hünnefeld, C., Lansky, A., and Zechel, C. Combinatorial synthesis of small organic molecules. *Angewandte Chemie International Edition in English*, 35(20):2288–2337, 1996. ISSN 1521-3773.

Ballester, P. J. and Richards, W. G. Ultrafast shape recognition to search compound databases for similar molecular shapes. *J. Comput. Chem.*, 28(10):1711–1723, 2007. ISSN 1096-987X.

Balloon, online. Balloon website. URL `http://users.abo.fi/mivainio/balloon`. [Online; accessed 7-December-2011].

Ballou, W. R. The development of the RTS,S malaria vaccine candidate: challenges and lessons. *Parasite Immunol.*, 31(9):492–500, 2009. ISSN 1365-3024.

Barnard, J. M. and Downs, G. M. Clustering of chemical structures on the basis of two-dimensional similarity measures. *J. Chem. Inf. Comput. Sci.*, 32(6):644–649, 1992.

Baroni, M., Cruciani, G., Sciabola, S., Perruccio, F., and Mason, J. S. A Common Reference Framework for Analyzing/Comparing Proteins and Ligands. Fingerprints for Ligands And Proteins (FLAP): Theory and Application. *J. Chem. Inf. Model.*, 47(2):279–294, 2007.

Bender, A., Young, D. W., Jenkins, J. L., Serrano, M., Mikhailov, D., Clemons, P. A., and Davies, J. W. Chemogenomic data analysis: prediction of small-molecule targets and the advent of biological fingerprint. *Comb. Chem. High Throughput Screen.*, 10(8):719–731, Sep 2007.

Bender, A. and Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.*, 2:3204–3218, 2004.

Bennani, Y. L. Drug discovery in the next decade: innovation needed {ASAP} . *Drug Discovery Today*, 17, Supplement(0):S31–S44, 2012. ISSN 1359-6446.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.*, 28(1):235–242, 2000.

Biagini, G. A., Bray, P. G., and Ward, S. A. Mechanisms of antimalarial drug resistance. In Mayers, D. L., editor, *Antimicrobial Drug Resistance*, Infectious Disease, pages 561–574. Humana Press, 2009. ISBN 978-1-59745-180-2.

Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S., and Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nat Chem*, 4(2):90–98, Feb 2012. ISSN 1755-4330.

Bienfait, B. SUBSET: computation of a representative subset from a large dataset. http://cactus.nci.nih.gov/subset/, 2001. [Online; accessed 2-August-2013].

Bissantz, C., Kuhn, B., and Stahl, M. A Medicinal Chemist's Guide to Molecular Interactions. *J. Med. Chem.*, 53(14):5061–5084, 2010.

Blackman, M. J. Malarial proteases and host cell egress: an 'emerging' cascade. *Cell. Microbiol.*, 10(10): 1925–1934, 2008. ISSN 1462-5822.

Blackman, M. J., Fujioka, H., Stafford, W. H. L., Sajid, M., Clough, B., Fleck, S. L., Aikawa, M., Grainger, M., and Hackett, F. A subtilisin-like protein in secretory organelles of plasmodium falciparum merozoites. *J. Biol. Chem.*, 273(36):23398–23409, 1998.

Blackman, M. J., Corrie, J. E. T., Croney, J. C., Kelly, G., Eccleston, J. F., and Jameson, D. M. Structural and biochemical characterization of a fluorogenic rhodamine-labeled malarial protease substrate. *Biochemistry (Mosc.)*, 41(40):12244–12252, 2002.

Blaney, J. M. and Dixon, J. S. *Distance Geometry in Molecular Modeling*, volume 5 of *Rev. Comp. Chem.*, chapter 6, pages 299–335. John Wiley & Sons, Inc., Hoboken, New Jersey, USA, 2007.

Blundell, T. L., Jhoti, H., and Abell, C. High-throughput crystallography for lead discovery in drug design. *Nat. Rev. Drug Discov.*, 1(1):45–54, Jan 2002. ISSN 1474-1776.

Böcker, A., Derksen, S., Schmidt, E., Teckentrup, A., and Schneider, G. A hierarchical clustering approach for large compound libraries. *J. Chem. Inf. Model.*, 45(4):807–815, 2005.

Bolton, E. E., Wang, Y., Thiessen, P. A., and Bryant, S. H. Chapter 12 PubChem: Integrated Platform of Small Molecules and Biological Activities . volume 4 of *Annual Reports in Computational Chemistry*, pages 217–241. Elsevier, 2008.

Bordogna, A., Pandini, A., and Bonati, L. Predicting the accuracy of protein-ligand docking on homology models. *J. Comput. Chem.*, 32(1):81–98, 2011. ISSN 1096-987X.

Bordoli, L., Kiefer, F., Arnold, K., Benkert, P., Battey, J., and Schwede, T. Protein structure homology modeling using SWISS-MODEL workspace. *Nat Protoc*, 4(1):1–13, 2009.

Boström, J. Reproducing the conformations of protein-bound ligands: A critical evaluation of several popular conformational searching tools. *J. Comput.-Aided Mol. Des.*, 15:1137–1152, 2001. ISSN 0920-654X.

Brameld, K. A., Kuhn, B., Reuter, D. C., and Stahl, M. Small molecule conformational preferences derived from crystal structure data. a medicinal chemistry focused analysis. *J. Chem. Inf. Model.*, 48(1):1–24, 2008.

Bredel, M. and Jacoby, E. Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet.*, 5(4):262–275, Apr 2004. ISSN 1471-0056.

Brenk, R., Vetter, S. W., Boyce, S. E., Goodin, D. B., and Shoichet, B. K. Probing molecular docking in a charged model binding site. *J. Mol. Biol.*, 357(5):1449–1470, 2006. ISSN 0022-2836.

Brown, R. D. and Martin, Y. C. Use of Structure-Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.*, 36(3):572–584, 1996.

Bruno, I. J., Cole, J. C., Kessler, M., Luo, J., Motherwell, W. D. S., Purkis, L. H., Smith, B. R., Taylor, R., Cooper, R. I., Harris, S. E., and Orpen, A. G. Retrieval of crystallographically-derived molecular geometry information. *J. Chem. Inf. Comput. Sci.*, 44(6):2133–2144, 2004.

Bunnage, M. E. Getting pharmaceutical R&D back on target. *Nat. Chem. Biol.*, 7(6):335–339, Jun 2011. ISSN 1552-4450.

Butcher, E. C., Berg, E. L., and Kunkel, E. J. Systems biology in drug discovery. *Nat Biotech*, 22(10): 1253–1259, Oct 2004. ISSN 1087-0156.

Butler, K. T., Luque, F. J., and Barril, X. Toward accurate relative energy predictions of the bioactive conformation of drugs. *J. Comput. Chem.*, 30(4):601–610, 2009. ISSN 1096-987X.

GOLD User Guide. *GOLD User Guide & Tutorial, Version 5.0.1*. Cambridge Crystallographic Data Centre. URL http://www.ccdc.cam.ac.uk/Lists/DocumentationList/gold.pdf. [Online; accessed 3-July-2013].

Cannon, E. O., Nigsch, F., and Mitchell, J. A novel hybrid ultrafast shape descriptor method for use in virtual screening. *Chemistry Central Journal*, 2(1):1–9, 2008.

Carhart, R. E., Smith, D. H., and Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.*, 25(2):64–73, 1985.

Chan, J. N., Nislow, C., and Emili, A. Recent advances and method development for drug target identification. *Trends in Pharmacological Sciences*, 31(2):82–88, 2010. ISSN 0165-6147.

Chang, G., Guida, W. C., and Still, W. C. An internal-coordinate monte carlo method for searching conformational space. *J. Am. Chem. Soc.*, 111(12):4379–4386, 1989.

ChEMBL-NTD website. ChEMBL-NTD Website. URL https://www.ebi.ac.uk/chemblntd. [Online; accessed 5-July-2013].

Chen, I.-J. and Foloppe, N. Conformational sampling of druglike molecules with moe and catalyst: Implications for pharmacophore modeling and virtual screening. *J. Chem. Inf. Model.*, 48(9):1773–1791, 2008.

Chen, J., Swamidass, S. J., Dou, Y., Bruand, J., and Baldi, P. Chemdb: a public database of small molecules and related chemoinformatics resources. *Bioinformatics*, 21(22):4133–4139, 2005.

Chen, J. H., Linstead, E., Swamidass, S. J., Wang, D., and Baldi, P. Chemdb update-full-text search and virtual chemical space. *Bioinformatics*, 23(17):2348–2351, 2007.

Chen, Y. and Shoichet, B. K. Molecular docking and ligand specificity in fragment-based inhibitor discovery. *Nat. Chem. Biol.*, 5(5):358–364, May 2009.

Cheng, T., Li, Q., Zhou, Z., Wang, Y., and Bryant, S. Structure-Based Virtual Screening for Drug Discovery: a Problem-Centric Review. *The AAPS Journal*, 14(1):133–141, 2012.

Cho, C. R., Labow, M., Reinhardt, M., van Oostrum, J., and Peitsch, M. C. The application of systems biology to drug discovery. *Current Opinion in Chemical Biology*, 10(4):294–302, 2006. ISSN 1367-5931.

Clewell III, H. J., Andersen, M. E., and Barton, H. A. A consistent approach for the application of pharmacokinetic modeling in cancer and noncancer risk assessment. *Environ. Health Perspect.*, 110(1): 85–93, Jan 2002.

Codd, E. F. A relational model of data for large shared data banks. *Commun. ACM*, 26(1):64–69, jan 1983. ISSN 0001-0782.

Congreve, M., Murray, C. W., and Blundell, T. L. Keynote review: Structural biology and drug discovery. *Drug Discovery Today*, 10(13):895–907, 2005. ISSN 1359-6446.

Cowman, A. F. and Crabb, B. S. Arresting malaria parasite egress from infected red blood cells. *Nat. Chem. Biol.*, 4(3):161–162, 2008. ISSN 1552-4450.

Cowman, A. F., Berry, D., and Baum, J. The cellular and molecular basis for malaria parasite invasion of the human red blood cell. *The Journal of Cell Biology*, 198(6):961–971, 2012.

Craik, C., Roczniak, S., Largman, C., and Rutter, W. The catalytic role of the active site aspartic acid in serine proteases. *Science*, 237:909–913, August 1987.

Cross, S., Baroni, M., Carosati, E., Benedetti, P., and Clementi, S. FLAP: GRID Molecular Interaction Fields in Virtual Screening. Validation using the DUD Data Set. *J. Chem. Inf. Model.*, 50(8):1442–1450, 2010.

Crowther, G. J., Shanmugam, D., Carmona, S. J., Doyle, M. A., Hertz-Fowler, C., Berriman, M., Nwaka, S., Ralph, S. A., Roos, D. S., Van Voorhis, W. C., and Agüero, F. Identification of attractive drug targets in neglected-disease pathogens using an In Silico approach. *PLoS Negl Trop Dis*, 4(8):e804, 08 2010.

Dauter, Z., Betzel, C., Genov, N., Pipon, N., and Wilson, K. Complex between the subtilisin from a mesophilic bacterium and the leech inhibitor eglin-C. *Acta Crystallographica Section B-Structural Science*, 47:707, 1991. ISSN 0108-7681.

David, E., Tramontin, T., and Zemmel, R. Pharmaceutical R&D: the road to positive returns. *Nat. Rev. Drug Discov.*, 8(8):609–610, Aug 2009. ISSN 1474-1776.

Daylight, 2011. *Daylight Theory Manual, version 4.9*, 2011. URL `http://www.daylight.com/dayhtml/doc/theory/`. [Online; accessed 27-June-2013].

Dearden, J., Cronin, M., and Kaiser, K. How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). *SAR QSAR Environ. Res.*, 20(3-4):241–266, 2009.

Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M., and Ashburner, M. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.*, 36(suppl 1):D344–D350, 2008.

Dimitropoulos, D., Ionides, J., and Henrick, K. *Using MSDchem to Search the PDB Ligand Dictionary*. John Wiley & Sons, Inc., 2006. ISBN 9780471250951.

Dondorp, A. M., Nosten, F., Yi, P., Das, D., Phyo, A. P., Tarning, J., Lwin, K. M., Ariey, F., Hanpithakpong, W., Lee, S. J., Ringwald, P., Silamut, K., Imwong, M., Chotivanich, K., Lim, P., Herdman, T., An, S. S., Yeung, S., Singhasivanon, P., Day, N. P., Lindegardh, N., Socheat, D., and White, N. J. Artemisinin resistance in plasmodium falciparum malaria. *N. Engl. J. Med.*, 361(5):455–467, 2009.

Drew, K. L. M., Baiman, H., Khwaounjoo, P., Yu, B., and Reynisson, J. Size estimation of chemical space: how big is it? *J. Pharm. Pharmacol.*, 64(4):490–495, 2012. ISSN 2042-7158.

Duan, J., Dixon, S. L., Lowrie, J. F., and Sherman, W. Analysis and comparison of 2D fingerprints: Insights into database screening performance using eight fingerprint methods. *J. Mol. Graph. Model.*, 29(2): 157–170, 2010. ISSN 1093-3263.

Durant, J. L., Leland, B. A., Henry, D. R., and Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.*, 42(6):1273–1280, 2002.

Durrant, J. and McCammon, J. A. Molecular dynamics simulations and drug discovery. *BMC Biol.*, 9(1):71, 2011. ISSN 1741-7007.

Ebejer, J.-P., Morris, G. M., and Deane, C. M. Freely available conformer generation methods: How good are they? *J. Chem. Inf. Model.*, 52(5):1146–1158, 2012.

Ebejer, J.-P., Fulle, S., Morris, G. M., and Finn, P. W. The emerging role of cloud computing in molecular modelling. *Journal of Molecular Graphics and Modelling*, 44(0):177–187, 2013. ISSN 1093-3263.

Ekins, S., Mestres, J., and Testa, B. In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. *Br. J. Pharmacol.*, 152(1):9–20, 2007. ISSN 1476-5381.

El-Barghouthi, M. I., Jaime, C., Akielah, R. E., Al-Sakhen, N. A., Masoud, N. A., Issa, A. A., Badwan, A. A., and Zughul, M. B. Free energy perturbation and mm/pbsa studies on inclusion complexes of some structurally related compounds with β-cyclodextrin. *Supramol. Chem.*, 21(7):603–610, Sep 2009. ISSN 1061-0278.

Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V., and Mee, R. P. Empirical scoring functions: I. the development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aided Mol. Des.*, 11(5):425–445, 1997. ISSN 0920-654X.

ElectroShape, 2010. InhibOx ElectroShape. http://www.inhibox.com/ligand-based-screening#ES. [Online; accessed 18-May-2013].

Eschenburg, S., Genov, N., Peters, K., Fittkau, S., Stoeva, S., Wilson, K. S., and Betzel, C. Crystal structure of subtilisin DY, a random mutant of subtilisin Carlsberg. *Eur. J. Biochem.*, 257(2):309–318, 1998. ISSN 1432-1033.

Ewing, T. J., Makino, S., Skillman, A. G., and Kuntz, I. D. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aided Mol. Des.*, 15(5):411–428, May 2001.

Fan, H., Irwin, J. J., Webb, B. M., Klebe, G., Shoichet, B. K., and Sali, A. Molecular docking screens using comparative models of proteins. *J. Chem. Inf. Model.*, 49(11):2512–2527, 2009.

Fawcett, T. ROC Graphs: Notes and Practical Considerations for Researchers. *Machine Learning*, 31:1–38, 2004.

Feher, M. Consensus scoring for protein-ligand interactions. *Drug Discovery Today*, 11(9-10):421–428, 2006. ISSN 1359-6446.

Feher, M. and Schmidt, J. M. Property distributions: Differences between drugs, natural products, and molecules from combinatorial chemistry. *J. Chem. Inf. Model.*, 43(1):218–227, 2003.

Feuston, B. P., Miller, M. D., Culberson, J. C., Nachbar, R. B., and Kearsley, S. K. Comparison of knowledge-based and distance geometry approaches for generation of molecular conformations. *J. Chem. Inf. Model.*, 41(3):754–763, 2001.

Finn, P. W. and Morris, G. M. Shape-based similarity searching in chemical databases. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 3(3):226–241, 2013. ISSN 1759-0884.

Fischer, J. D., Holliday, G. L., and Thornton, J. M. The CoFactor database: organic cofactors in enzyme catalysis. *Bioinformatics*, 26(19):2496–2497, 2010.

Fishman, M. C. and Porter, J. A. Pharmaceuticals: A new grammar for drug discovery. *Nature*, 437(7058): 491–493, Sep 2005. ISSN 0028-0836.

Flemming, A. Antibacterials: Resistance-guided discovery of new antibiotics. *Nat. Rev. Drug Discov.*, 12 (11):826–826, Nov 2013. ISSN 1474-1776.

Fraley, C. and Raftery, A. E. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *The Computer Journal*, 41(8):578–588, 1998.

Frog2, online. Frog2 web interface. URL http://bioserv.rpbs.univ-paris-diderot.fr/cgi-bin/Frog2. [Online; accessed 7-December-2011].

Frolkis, A., Knox, C., Lim, E., Jewison, T., Law, V., Hau, D. D., Liu, P., Gautam, B., Ly, S., Guo, A. C., Xia, J., Liang, Y., Shrivastava, S., and Wishart, D. S. SMPDB: The Small Molecule Pathway Database. *Nucleic Acids Res.*, 38(suppl 1):D480–D487, 2010.

Gamo, F. J., Sanz, L. M., Vidal, J., de Cozar, C., Alvarez, E., Lavandera, J. L., Vanderwall, D. E., Green, D. V., Kumar, V., Hasan, S., Brown, J. R., Peishoff, C. E., Cardon, L. R., and Garcia-Bustos, J. F. Thousands of chemical starting points for antimalarial lead identification. *Nature*, 465(7296):305–310, May 2010.

Gardiner, E. J., Holliday, J. D., OĎowd, C., and Willett, P. Effectiveness of 2D fingerprints for scaffold hopping. *Future Med Chem*, 3(4):405–414, Mar 2011.

Garrett, R. and Grisham, C. *Biochemistry*. Cengage Learning, Fourth edition, 2008. ISBN 9781111798673.

Gasteiger, J., Rudolph, C., and Sadowski, J. Automatic generation of 3d-atomic coordinates for organic molecules. *Tetrahedron Comput. Methodol.*, 3(6, Part 3):537–547, 1990. ISSN 0898-5529.

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., and Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, 40(D1):D1100–D1107, 2012.

Gemma, S., Giovani, S., Brindisi, M., Tripaldi, P., Brogi, S., Savini, L., Fiorini, I., Novellino, E., Butini, S., Campiani, G., Penzo, M., and Blackman, M. J. Quinolylhydrazones as novel inhibitors of Plasmodium falciparum serine protease PfSUB1. *Bioorganic & Medicinal Chemistry Letters*, 22(16):5317–5321, 2012. ISSN 0960-894X.

Ghosh, S., Nie, A., An, J., and Huang, Z. Structure-based virtual screening of chemical libraries for drug discovery. *Curr. Opin. Chem. Biol.*, 10(3):194–202, 2006. ISSN 1367-5931.

GOFAM website. GO Fight Against Malaria Project. URL http://gofightagainstmalaria.scripps.edu. [Online; accessed 5-July-2013].

Golbraikh, A. and Tropsha, A. Beware of q^2! *J. Mol. Graph. Model.*, 20(4):269–276, 2002. ISSN 1093-3263.

Good, A. C., Hodgkin, E. E., and Richards, W. G. Utilization of Gaussian functions for the rapid evaluation of molecular similarity. *J. Chem. Inf. Comput. Sci.*, 32(3):188–191, 1992.

Good, A. C. and Oprea, T. I. Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? *J. Comput. Aided Mol. Des.*, 22(3-4):169–178, 2008. ISSN 0920-654X.

Goto, J., Kataoka, R., and Hirayama, N. Ph4Dock: Pharmacophore-Based Protein-Ligand Docking. *J. Med. Chem.*, 47(27):6804–6811, 2004.

Graves, A. P., Shivakumar, D. M., Boyce, S. E., Jacobson, M. P., Case, D. A., and Shoichet, B. K. Rescoring docking hit lists for model cavity sites: Predictions and experimental testing. *Journal of Molecular Biology*, 377(3):914–934, 2008. ISSN 0022-2836.

Gregori-Puigjané, E. and Keiser, M. J. Chapter 4 chemoinformatic approaches to target identification. In *Designing Multi-Target Drugs*, pages 50–65. The Royal Society of Chemistry, 2012. ISBN 978-1-84973-362-5.

Gu, J. and Bourne, P. E. *Structural Bioinformatics*, chapter 27, page 639. Wiley-Blackwell, Hoboken, New Jersey, USA, second edition, 2009.

Guha, R., Howard, M. T., Hutchison, G. R., Murray-Rust, P., Rzepa, H., Steinbeck, C., Wegner, J. K., and Willighagen, E. L. The Blue Obelisk–Interoperability in Chemical Informatics. *J. Chem. Inf. Model.*, 46: 991–998, 2006.

Gund, P. Three-dimensional pharmacophoric pattern searching. In Hahn, F., Kersten, H., Kersten, W., and Szybalski, W., editors, *Progress in Molecular and Subcellular Biology*, volume 5 of *Progress in Molecular and Subcellular Biology*, pages 117–143. Springer Berlin Heidelberg, 1977. ISBN 978-3-642-66628-5.

Hahn, M. Three-dimensional shape-based searching of conformationally flexible compounds. *J. Chem. Inf. Model.*, 37(1):80–86, 1997.

Halgren, T. A. Merck molecular force field. i. basis, form, scope, parameterization, and performance of mmff94. *J. Comput. Chem.*, 17(5-6):490–519, 1996. ISSN 1096-987X.

Han, J. and Kamber, M. *Data mining: concepts and techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000. ISBN 1-55860-489-8.

Hansson, T. and Åqvist, J. Estimation of binding free energies for hiv proteinase inhibitors by molecular dynamics simulations. *Protein Eng.*, 8(11):1137–1144, 1995.

Hansson, T., Oostenbrink, C., and van Gunsteren, W. Molecular dynamics simulations. *Current Opinion in Structural Biology*, 12(2):190–196, 2002. ISSN 0959-440X.

Hartenfeller, M. and Schneider, G. De novo drug design. In Bajorath, J., editor, *Chemoinformatics and Computational Chemical Biology*, volume 672 of *Methods in Molecular Biology*, pages 299–323. Humana Press, 2011. ISBN 978-1-60761-838-6. doi: 10.1007/978-1-60761-839-3_12.

Hartshorn, M. J., Verdonk, M. L., Chessari, G., Brewerton, S. C., Mooij, W. T. M., Mortenson, P. N., and Murray, C. W. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.*, 50(4):726–741, 2007.

Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., Muthukrishnan, V., Owen, G., Turner, S., Williams, M., and Steinbeck, C. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.*, 41(D1):D456–D463, 2013.

Havel, T., Kuntz, I., and Crippen, G. The theory and practice of distance geometry. *Bull. Math. Biol.*, 45: 665–720, 1983. ISSN 0092-8240.

Hawkins, P. C. D., Skillman, A. G., Warren, G. L., Ellingson, B. A., and Stahl, M. T. Conformer generation with omega: Algorithm and validation using high quality structures from the protein databank and cambridge structural database. *J. Chem. Inf. Model.*, 50(4):572–584, 2010.

Hedstrom, L. Serine protease mechanism and specificity. *Chem. Rev.*, 102(12):4501–4524, 2002.

Hermann, J. C., Marti-Arbona, R., Fedorov, A. A., Fedorov, E., Almo, S. C., Shoichet, B. K., and Raushel, F. M. Structure-based activity prediction for an enzyme of unknown function. *Nature*, 448(7155):775–779, Aug 2007.

Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., and Schuffenhauer, A. New Methods for Ligand-Based Virtual Screening: Use of Data Fusion and Machine Learning to Enhance the Effectiveness of Similarity Searching. *J. Chem. Inf. Model.*, 46(2):462–470, 2006.

Ho, M. and White, N. J. Molecular mechanisms of cytoadherence in malaria. *American Journal of Physiology - Cell Physiology*, 276(6):C1231–C1242, 1999.

Hopkins, A. L., Groom, C. R., and Alex, A. Ligand efficiency: a useful metric for lead selection. *Drug Discovery Today*, 9(10):430–431, 2004. ISSN 1359-6446.

Horvath, D. A virtual screening approach applied to the search for trypanothione reductase inhibitors. *J. Med. Chem.*, 40(15):2412–2423, 1997.

Hou, T. and Wang, J. Structure - ADME relationship: still a long way to go? *Expert Opinion on Drug Metabolism & Toxicology*, 4(6):759–770, 2008.

Hsin, K.-Y., Morgan, H. P., Shave, S. R., Hinton, A. C., Taylor, P., and Walkinshaw, M. D. EDULISS: a small-molecule database with data-mining and pharmacophore searching capabilities. *Nucleic Acids Res.*, 39(suppl 1):D1042–D1048, 2011.

Hu, G., Kuang, G., Xiao, W., Li, W., Liu, G., and Tang, Y. Performance Evaluation of 2D Fingerprint and 3D Shape Similarity Methods in Virtual Screening. *J. Chem. Inf. Model.*, 52(5):1103–1113, 2012.

Huang, N., Shoichet, B. K., and Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.*, 49 (23):6789–6801, 2006.

Huang, S.-Y. and Zou, X. Ensemble docking of multiple protein structures: Considering protein structural variations in molecular docking. *Proteins: Struct., Funct., Bioinf.*, 66(2):399–421, 2007. ISSN 1097-0134.

Hughes, J., Rees, S., Kalindjian, S., and Philpott, K. Principles of early drug discovery. *Br. J. Pharmacol.*, 162(6):1239–1249, 2011. ISSN 1476-5381.

Irwin, J. J. and Shoichet, B. K. ZINC – A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.*, 45(1):177–182, 2005.

Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S., and Coleman, R. G. ZINC: A free tool to discover chemistry for biology. *J. Chem. Inf. Model.*, 52(7):1757–1768, 2012.

Jain, A. K. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010. ISSN 0167-8655.

Jain, S. C., Shinde, U., Li, Y., Inouye, M., and Berman, H. M. The crystal structure of an autoprocessed Ser221Cys-subtilisin E-propeptide complex at 2.0 Å resolution. *Journal of Molecular Biology*, 284(1): 137–144, 1998. ISSN 0022-2836.

Janse, C. J. and Waters, A. P. The exoneme helps malaria parasites to break out of blood cells. *Cell*, 131(6): 1036–1038, 2007. ISSN 0092-8674.

Jean, L., Withers-Martinez, C., Hackett, F., , and Blackman, M. J. Unique insertions within Plasmodium falciparum subtilisin-like protease-1 are crucial for enzyme maturation and activity. *Mol. Biochem. Parasitol.*, 144(2):187–197, 2005. ISSN 0166-6851.

Jeffrey, G. *An Introduction to Hydrogen Bonding*. Topics in Physical Chemistry Series. Oxford University Press, Incorporated, 1997. ISBN 9780195095494.

Johnson, M. and Maggiora, G. *Concepts and applications of molecular similarity*. Wiley-Interscience Publication. Wiley, 1990. ISBN 9780471621751.

Jones, G., Willett, P., and Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.*, 245(1):43–53, 1995.

Jones, G., Willett, P., Glen, R. C., Leach, A. R., and Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.*, 267(3):727–748, 1997. ISSN 0022-2836.

Jorgensen, W. L. The many roles of computation in drug discovery. *Science*, 303(5665):1813–1818, 2004.

Kapetanovic, I. Computer-aided drug discovery and development (CADDD): In silico-chemico-biological approach. *Chemico-Biological Interactions*, 171(2):165–176, 2008. ISSN 0009-2797.

Kellenberger, E., Muller, P., Schalon, C., Bret, G., Foata, N., and Rognan, D. sc-PDB: an annotated database of druggable binding sites from the protein data bank. *J. Chem. Inf. Model.*, 46(2):717–727, 2006.

Khanna, V. and Ranganathan, S. Molecular similarity and diversity approaches in chemoinformatics. *Drug Dev. Res.*, 72(1):74–84, 2011. ISSN 1098-2299.

Kinnings, S. L., Liu, N., Tonge, P. J., Jackson, R. M., Xie, L., and Bourne, P. E. A Machine Learning-Based Method To Improve Docking Scoring Functions and Its Application to Drug Repurposing. *J. Chem. Inf. Model.*, 51(2):408–419, 2011.

Kirchmair, J., Wolber, G., Laggner, C., and Langer, T. Comparative performance assessment of the conformational model generators omega and catalyst: A large-scale survey on the retrieval of protein-bound ligand conformations. *J. Chem. Inf. Model.*, 46(4):1848–1861, 2006.

Klebe, G. and Mietzner, T. A fast and efficient method to generate biologically relevant conformations. *J. Comput.-Aided Mol. Des.*, 8:583–606, 1994. ISSN 0920-654X.

Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., Djoumbou, Y., Eisner, R., Guo, A. C., and Wishart, D. S. DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. *Nucleic Acids Res.*, 39(suppl 1):D1035–D1041, 2011.

Kolb, P., Ferreira, R. S., Irwin, J. J., and Shoichet, B. K. Docking and chemoinformatic screens for new ligands and targets. *Current Opinion in Biotechnology*, 20(4):429–436, 2009a. ISSN 0958-1669.

Kolb, P., Rosenbaum, D. M., Irwin, J. J., Fung, J. J., Kobilka, B. K., and Shoichet, B. K. Structure-based discovery of β2-adrenergic receptor ligands. *Proceedings of the National Academy of Sciences*, 106(16): 6843–6848, 2009b.

Korb, O., Stützle, T., and Exner, T. E. Empirical scoring functions for advanced protein-ligand docking with PLANTS. *J. Chem. Inf. Model.*, 49(1):84–96, 2009.

Korb, O., Olsson, T. S. G., Bowden, S. J., Hall, R. J., Verdonk, M. L., Liebeschuetz, J. W., and Cole, J. C. Potential and limitations of ensemble docking. *J. Chem. Inf. Model.*, 52(5):1262–1274, 2012.

Koutsoukas, A., Simms, B., Kirchmair, J., Bond, P. J., Whitmore, A. V., Zimmer, S., Young, M. P., Jenkins, J. L., Glick, M., Glen, R. C., and Bender, A. From in silico target prediction to multi-target drug design: Current databases, methods and applications. *Journal of Proteomics*, 74(12):2554–2574, 2011. ISSN 1874-3919.

Kraus, C. N. Low hanging fruit in infectious disease drug development. *Current Opinion in Microbiology*, 11 (5):434–438, 2008. ISSN 1369-5274.

Kraut, J. Serine Proteases: Structure and Mechanism of Catalysis. *Annu. Rev. Biochem.*, 46(1):331–358, 1977.

Kristam, R., Gillet, V. J., Lewis, R. A., and Thorner, D. Comparison of conformational analysis techniques to generate pharmacophore hypotheses using catalyst. *J. Chem. Inf. Model.*, 45(2):461–476, 2005.

Krucken, J., Mehnert, L. I., Dkhil, M. A., El-Khadragy, M., Benten, W. P. M., Mossmann, H., and Wunderlich, F. Massive destruction of malaria-parasitized red blood cells despite spleen closure. *Infect. Immun.*, 73(10):6390–6398, 2005.

Kubinyi, H. QSAR and 3D QSAR in drug design Part 2: applications and problems. *Drug Discovery Today*, 2(12):538–546, 1997. ISSN 1359-6446.

Kubinyi, H. Similarity and dissimilarity: A medicinal chemist's view. *Perspectives in Drug Discovery and Design*, 9-11:225–252, 1998. ISSN 0928-2866.

Kubinyi, H. *Success Stories of Computer-Aided Design*, pages 377–424. John Wiley & Sons, Inc., 2006. ISBN 9780470037232.

Kuhn, M., Szklarczyk, D., Franceschini, A., von Mering, C., Jensen, L. J., and Bork, P. Stitch 3: zooming in on protein-chemical interactions. *Nucleic Acids Res.*, 40(D1):D876–D880, 2012.

Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R., and Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.*, 161(2):269–288, 1982. ISSN 0022-2836.

Lang, P. T., Brozell, S. R., Mukherjee, S., Pettersen, E. F., Meng, E. C., Thomas, V., Rizzo, R. C., Case, D. A., James, T. L., and Kuntz, I. D. DOCK 6: Combining techniques to model RNA-small molecule complexes. *RNA*, 15(6):1219–1230, 2009.

Langer, T. and Wolber, G. Pharmacophore definition and 3D searches. *Drug Discovery Today: Technologies*, 1(3):203–207, 2004. ISSN 1740-6749.

Leach, A. R., Shoichet, B. K., and Peishoff, C. E. Docking and scoring. *J. Med. Chem.*, 49:5851–5855, 2006.

Leach, A. R., Gillet, V. J., Lewis, R. A., and Taylor, R. Three-Dimensional Pharmacophore Methods in Drug Discovery. *J. Med. Chem.*, 53(2):539–558, 2010.

Leite, T. B., Gomes, D., Miteva, M., Chomilier, J., Villoutreix, B., and Tufféry, P. Frog: a free online drug 3d conformation generator. *Nucleic Acids Res.*, 35(suppl 2):W568–W572, 2007.

Lell, B., Agnandji, S., von Glasenapp, I., Haertle, S., Oyakhiromen, S., Issifou, S., Vekemans, J., Leach, A., Lievens, M., Dubois, M.-C., Demoitie, M.-A., Carter, T., Villafana, T., Ballou, W. R., Cohen, J., and Kremsner, P. G. A Randomized Trial Assessing the Safety and Immunogenicity of AS01 and AS02 Adjuvanted RTS,S Malaria Vaccine Candidates in Children in Gabon. *PLoS ONE*, 4(10):e7611, 10 2009.

Liebeschuetz, J. W., Cole, J. C., and Korb, O. Pose prediction and virtual screening performance of gold scoring functions in a standardized test. *J. Comput. Aided Mol. Des.*, 26(6):737–748, 2012. ISSN 0920-654X.

Lin, J.-H., Perryman, A. L., Schames, J. R., and McCammon, J. A. Computational drug design accommodating receptor flexibility: The relaxed complex scheme. *J. Am. Chem. Soc.*, 124(20):5632–5633, 2002.

Lin, J.-H., Perryman, A. L., Schames, J. R., and McCammon, J. A. The relaxed complex method: Accommodating receptor flexibility for drug design with an improved scoring scheme. *Biopolymers*, 68 (1):47–62, 2003. ISSN 1097-0282.

Lipinski, C. A., Lombardo, F., Dominy, B. W., and Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 23(1-3):3–25, 1997. ISSN 0169-409X.

Lipkowitz, K. and Boyd, D. Clustering methods and their uses in computational chemistry. In *Reviews in Computational Chemistry*, pages 1–40. Wiley, 2003. ISBN 9780471461425.

Liu, X., Bai, F., Ouyang, S., Wang, X., Li, H., and Jiang, H. Cyndi: a multi-objective evolution algorithm based method for bioactive molecular conformational generation. *BMC Bioinformatics*, 10(1):101, 2009. ISSN 1471-2105.

Lorber, D. M. and Shoichet, B. K. Flexible ligand docking using conformational ensembles. *Protein Sci.*, 7 (4):938–950, 1998. ISSN 1469-896X.

Loving, K., Alberts, I., and Sherman, W. Computational approaches for fragment-based and de novo design. *Curr. Top. Med. Chem.*, 10(1):14–32, 2010.

Ludin, P., Woodcroft, B., Ralph, S. A., and Mäser, P. In silico prediction of antimalarial drug target candidates. *International Journal for Parasitology: Drugs and Drug Resistance*, 2(0):191–199, 2012. ISSN 2211-3207.

Lyne, P. D. Structure-based virtual screening: an overview. *Drug Discovery Today*, 7(20):1047–1055, 2002. ISSN 1359-6446.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. *cluster: Cluster Analysis Basics and Extensions*, 2013. R package version 1.14.4.

Makino, S. and Kuntz, I. D. Automated flexible ligand docking method and its application for database search. *J. Comput. Chem.*, 18(14):1812–1825, 1997. ISSN 1096-987X.

Manion, J. A., Huie, R. E., Levin, R. D., Burgess Jr., D. R., Orkin, V. L., Tsang, W., McGivern, W. S., Hudgens, J. W., Knyazev, V. D., Atkinson, D. B., Chai, E., Tereza, A. M., Lin, C.-Y., Allison, T. C., Mallard, W. G., Westley, F., Herron, J. T., Hampson, R. F., and Frizzell, D. H. Nist chemical kinetics database, nist standard reference database 17, version 7.0 (web version), release 1.4.3, data version 2008.12. http://kinetics.nist.gov/. [Online; accessed 29-August-2013].

Marcou, G. and Rognan, D. Optimizing Fragment and Scaffold Docking by Use of Molecular Interaction Fingerprints. *J. Chem. Inf. Model.*, 47(1):195–207, 2007.

Martin, Y. C., Kofron, J. L., and Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.*, 45(19):4350–4358, 2002.

Martin, Y. C. Let's not forget tautomers. *J. Comput. Aided Mol. Des.*, 23(10):693–704, 2009. ISSN 0920-654X.

Matter, H. Selecting optimally diverse compounds from structure databases: A validation study of two-dimensional and three-dimensional molecular descriptors. *J. Med. Chem.*, 40(8):1219–1229, 1997.

Medinger, T. and Wilkinson, F. Mechanism of fluorescence quenching in solution. part 1. - quenching by bromobenzene. *Trans. Faraday Soc.*, 61:620–630, 1965.

Mekenyan, O., Dimitrov, D., Nikolova, N., and Karabunarliev, S. Conformational coverage by a genetic algorithm. *J. Chem. Inf. Model.*, 39(6):997–1016, 1999.

Melagraki, G. and Afantitis, A. Ligand and structure based virtual screening strategies for hit-finding and optimization of Hepatitis C Virus (HCV) Inhibitors. *Curr. Med. Chem.*, 18(17):2612–2619, 2011.

Merski, M. and Shoichet, B. K. The impact of introducing a histidine into an apolar cavity site on docking and ligand recognition. *J. Med. Chem.*, 56(7):2874–2884, 2013.

Meslamani, J., Rognan, D., and Kellenberger, E. sc-PDB: a database for identifying variations and multiplicity of 'druggable' binding sites in proteins. *Bioinformatics*, 27(9):1324–1326, 2011.

Meyer, E. A., Castellano, R. K., and Diederich, F. Interactions with Aromatic Rings in Chemical and Biological Recognition. *Angewandte Chemie International Edition*, 42(11):1210–1250, 2003. ISSN 1521-3773.

Miteva, M. A., Guyon, F., and Tufféry, P. Frog2: Efficient 3d conformation ensemble generator for small compounds. *Nucleic Acids Res.*, 38(suppl 2):W622–W627, 2010.

MOE, online. MOE (the molecular operating environment) version 2010.10. URL http://www.chemcomp.com. [Online; accessed 11-December-2011].

Molecular Networks, online. Molecular networks. URL http://www.molecular-networks.com. [Online; accessed 6-December-2011].

Moneriz, C., Mestres, J., Bautista, J. M., Diez, A., and Puyet, A. Multi-targeted activity of maslinic acid as an antimalarial natural compound. *FEBS J.*, 278(16):2951–2961, 2011. ISSN 1742-4658.

Mooij, W. T. M. and Verdonk, M. L. General and targeted statistical potentials for protein-ligand interactions. *Proteins: Struct., Funct., Bioinf.*, 61(2):272–287, 2005. ISSN 1097-0134.

Morgan, H. L. The generation of a unique machine description for chemical structures - A technique developed at chemical abstracts service. *J. Chem. Doc.*, 5(2):107–113, 1965.

Moro, S., Bacilieri, M., and Deflorian, F. Combining ligand-based and structure-based drug design in the virtual screening arena. *Expert Opinion on Drug Discovery*, 2(1):37–49, 2007.

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., and Olson, A. J. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.*, 30(16):2785–2791, 2009. ISSN 1096-987X.

Moustakas, D. T., Lang, P. T., Pegg, S., Pettersen, E., Kuntz, I. D., Brooijmans, N., and Rizzo, R. C. Development and validation of a modular, extensible docking program: DOCK 5. *J. Comput. Aided Mol. Des.*, 20(10-11):601–619, 2006. ISSN 0920-654X.

Murray, C. J., Rosenfeld, L. C., Lim, S. S., Andrews, K. G., Foreman, K. J., Haring, D., Fullman, N., Naghavi, M., Lozano, R., and Lopez, A. D. Global malaria mortality between 1980 and 2010: a systematic analysis. *The Lancet*, 379(9814):413–431, 2012. ISSN 0140-6736.

Murray, C. W. and Rees, D. C. The rise of fragment-based drug discovery. *Nat Chem*, 1(3):187–192, Jun 2009. ISSN 1755-4330.

Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.*, 55(14):6582–6594, 2012.

Nasr, R., Swamidass, S. J., and Baldi, P. Large scale study of multiple-molecule queries. *Journal of Cheminformatics*, 1(1):7, 2009. ISSN 1758-2946.

National Institute of Allergy and Infectious Diseases. NIAID Division of AIDS Anti-HIV/OI/TB Therapeutics Database. http://chemdb.niaid.nih.gov/. [Online; accessed 29-August-2013].

Nicholls, A. What do we know and when do we know it? *J. Comput. Aided Mol. Des.*, 22(3-4):239–255, 2008. ISSN 0920-654X.

Nichols, S. E., Baron, R., and McCammon, J. On the use of molecular dynamics receptor conformations for virtual screening. In Baron, R., editor, *Computational Drug Discovery and Design*, volume 819 of *Methods in Molecular Biology*, pages 93–103. Springer New York, 2012. ISBN 978-1-61779-464-3.

Nicklaus, M. C., Wang, S., Driscoll, J. S., and Milne, G. W. A. Conformational changes of small molecules binding to proteins. *Bioorg. Med. Chem.*, 3(4):411–428, 1995. ISSN 0968-0896.

Nilakantan, R., Bauman, N., Dixon, J. S., and Venkataraghavan, R. Topological torsion: a new molecular descriptor for SAR applications. comparison with other descriptors. *J. Chem. Inf. Comput. Sci.*, 27(2): 82–85, 1987.

Novoa, E. M., Pouplana, L. R. d., Barril, X., and Orozco, M. Ensemble docking from homology models. *J. Chem. Theory Comput.*, 6(8):2547–2557, 2010.

O'Boyle, N., Morley, C., and Hutchison, G. Pybel: a python wrapper for the OpenBabel cheminformatics toolkit. *Chem. Cent. J.*, 2(1):5, 2008. ISSN 1752-153X.

O'Boyle, N., Banck, M., James, C., Morley, C., Vandermeersch, T., and Hutchison, G. Open Babel: An open chemical toolbox. *J. Cheminf.*, 3(1):33, 2011a. ISSN 1758-2946.

O'Boyle, N., Guha, R., Willighagen, E., Adams, S., Alvarsson, J., Bradley, J.-C., Filippov, I., Hanson, R., Hanwell, M., Hutchison, G., James, C., Jeliazkova, N., Lang, A., Langner, K., Lonie, D., Lowe, D., Pansanel, J., Pavlov, D., Spjuth, O., Steinbeck, C., Tenderholt, A., Theisen, K., and Murray-Rust, P. Open Data, Open Source and Open Standards in chemistry: The Blue Obelisk five years on. *Journal of Cheminformatics*, 3(1):37, 2011b. ISSN 1758-2946.

O'Boyle, N., Vandermeersch, T., Flynn, C., Maguire, A., and Hutchison, G. Confab - systematic generation of diverse low-energy conformers. *J. Cheminf.*, 3(1):8, 2011c. ISSN 1758-2946.

Okimoto, N., Futatsugi, N., Fuji, H., Suenaga, A., Morimoto, G., Yanai, R., Ohno, Y., Narumi, T., and Taiji, M. High-performance drug discovery: Computational screening by combining docking and molecular dynamics simulations. *PLoS Comput. Biol.*, 5(10):e1000528, 10 2009.

Olah, M. M., Bologa, C. G., and Oprea, T. I. Strategies for compound selection. *Curr. Drug Discov. Technol.*, 1(3):211–220, Oct 2004.

OpenBabel, online. The open babel package, version 2.3.1. URL http://openbabel.sourceforge.net. [Online; accessed 7-December-2011].

OpenEye Scientific, online. Openeye scientific software. URL http://www.eyesopen.com. [Online; accessed 6-December-2011].

Oprea, T. I. Property distribution of drug-related chemical databases. *J. Comput.-Aided Mol. Des.*, 14: 251–264, 2000. ISSN 0920–654X.

Oprea, T. I. and Matter, H. Integrating virtual screening in lead discovery. *Curr. Opin. Chem. Biol.*, 8(4): 349–358, 2004. ISSN 1367-5931.

Osmond, R. I., Crouch, M. F., and Dupriez, V. J. An emerging role for kinase screening in GPCR drug discovery. *Curr. Opin. Mol. Ther.*, 12(3):305–315, Jun 2010.

Ou-Yang, S.-s., Lu, J.-y., Kong, X.-q., Liang, Z.-j., Luo, C., and Jiang, H. Computational drug discovery. *Acta Pharmacol. Sin.*, 33(9):1131–1140, Sep 2012. ISSN 1671-4083.

Pammolli, F. and Magazzini, M. Lauraand Riccaboni. The productivity crisis in pharmaceutical R&D. *Nat. Rev. Drug Discov.*, 10(6):428–438, Jun 2011. ISSN 1474-1776.

Papadopoulos, J. S. and Agarwala, R. COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics*, 23(9):1073–1079, 2007.

Payne, D. J., Gwynn, M. N., Holmes, D. J., and Pompliano, D. L. Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nat. Rev. Drug Discov.*, 6(1):29–40, Jan 2007. ISSN 1474-1776.

Pearlman, D. A. Evaluating the molecular mechanics poisson-boltzmann surface area free energy method using a congeneric series of ligands to p38 map kinase. *J. Med. Chem.*, 48(24):7796–7807, 2005.

Peitsch, M. C. Protein modeling by e-mail. *Biotechnology (N. Y.)*, 13:658–660, 1995.

Pencheva, T., Lagorce, D., Pajeva, I., Villoutreix, B., and Miteva, M. AMMOS: Automated molecular mechanics optimization tool for in silico screening. *BMC Bioinformatics*, 9(1):438, 2008. ISSN 1471-2105.

Perola, E. and Charifson, P. S. Conformational analysis of drug-like molecules bound to proteins: An extensive study of ligand reorganization upon binding. *J. Med. Chem.*, 47(10):2499–2510, 2004.

Polhemus, M. E., Magill, A. J., Cummings, J. F., Kester, K. E., Ockenhouse, C. F., Lanar, D. E., Dutta, S., Barbosa, A., Soisson, L., Diggs, C. L., Robinson, S. A., Haynes, J. D., Stewart, V. A., Ware, L. A., Brando, C., Krzych, U., Bowden, R. A., Cohen, J. D., Dubois, M.-C., Ofori-Anyinam, O., De-Kock, E., Ballou, R. W., and Heppner, G. D. Jr. Phase I dose escalation safety and immunogenicity trial of Plasmodium falciparum apical membrane protein (AMA-1) FMP2.1, adjuvanted with AS02A, in malaria-naive adults at the Walter Reed Army Institute of Research. *Vaccine*, 25(21):4203–4212, 2007. ISSN 0264-410X.

Powers, R. A. and Shoichet, B. K. Structure-based approach for binding site identification on ampc β-lactamase. *J. Med. Chem.*, 45(15):3222–3234, 2002.

Price, R. N., Douglas, N. M., and Anstey, N. M. New developments in Plasmodium vivax malaria: severe disease and the rise of chloroquine resistance. *Curr. Opin. Infect. Dis.*, 22(5):430–435, Oct 2009.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL `http://www.R-project.org/`. R version 3.0.0.

Rao, S., Sanschagrin, P. C., Greenwood, J. R., Repasky, M. P., Sherman, W., and Farid, R. Improving database enrichment through ensemble docking. *J. Comput. Aided Mol. Des.*, 22(9):621–627, 2008. ISSN 0920-654X.

Rappe, A. K., Casewit, C. J., Colwell, K. S., Goddard, W. A., and Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.*, 114(25): 10024–10035, 1992.

Rastelli, G., Degliesposti, G., Del Rio, A., and Sgobba, M. Binding estimation after refinement, a new automated procedure for the refinement and rescoring of docked ligands in virtual screening. *Chemical Biology & Drug Design*, 73(3):283–286, 2009. ISSN 1747-0285.

RDKit manual, online. Getting started with the RDKit in Python, Version Q1 2011. URL http://www.rdkit.org/GettingStartedInPython.pdf. [Online; accessed 12-December-2011].

RDKit, online. RDKit: Open-source cheminformatics. http://www.rdkit.org. [Online; accessed 11-April-2013].

Reddy, M. B., Harvey J. Clewell III, T. L., and Andersen, M. E. *Physiologically Based Pharmacokinetic Modeling: A Tool for Understanding ADMET Properties and Extrapolating to Human.* InTech, Jan 2013. ISBN 978-953-51-0946-4.

Renner, S., Schwab, C. H., Gasteiger, J., and Schneider, G. Impact of Conformational Flexibility on Three-Dimensional Similarity Searching Using Correlation Vectors. *J. Chem. Inf. Model.*, 46(6):2324–2332, 2006.

Reymond, J.-L. and Awale, M. Exploring Chemical Space for Drug Discovery Using the Chemical Universe Database. *ACS Chemical Neuroscience*, 3(9):649–657, 2012.

Reynolds, C. H., Druker, R., and Pfahler, L. B. Lead Discovery Using Stochastic Cluster Analysis (SCA): A New Method for Clustering Structurally Similar Compounds. *J. Chem. Inf. Comput. Sci.*, 38(2):305–312, 1998.

Reynolds, C. R., Amini, A. C., Muggleton, S. H., and Sternberg, M. J. E. Assessment of a Rule-Based Virtual Screening Technology (INDDEx) on a Benchmark Data Set. *The Journal of Physical Chemistry B*, 116(23):6732–6739, 2012.

Riniker, S. and Landrum, G. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of Cheminformatics*, 5(1):26, 2013. ISSN 1758-2946.

Rogers, D. and Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.*, 50(5):742–754, 2010.

Rohrer, S. G. and Baumann, K. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. *J. Chem. Inf. Model.*, 49(2):169–184, 2009.

Roiko, M. S. and Carruthers, V. B. New roles for perforins and proteases in apicomplexan egress. *Cell. Microbiol.*, 11(10):1444–1452, 2009. ISSN 1462-5822.

Rokach, L. and Maimon, O. Clustering methods. In Maimon, O. and Rokach, L., editors, *Data Mining and Knowledge Discovery Handbook*, pages 321–352. Springer US, 2005. ISBN 978-0-387-24435-8.

Roques, B. P., Noble, F., Daugé, V., Fournié-Zaluski, M. C., and Beaumont, A. Neutral endopeptidase 24.11: structure, inhibition, and experimental and clinical pharmacology. *Pharmacol. Rev.*, 45(1):87–146, 1993.

Ross, G. A., Morris, G. M., and Biggin, P. C. Rapid and Accurate Prediction and Scoring of Water Molecules in Protein Binding Sites. *PLoS ONE*, 7(3):e32036, 03 2012.

Ruddigkeit, L., Blum, L. C., and Reymond, J.-L. Visualization and Virtual Screening of the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.*, 53(1):56–65, 2013.

Rueda, M., Bottegoni, G., and Abagyan, R. Consistent Improvement of Cross-Docking Results Using Binding Site Ensembles Generated with Elastic Network Normal Modes. *J. Chem. Inf. Model.*, 49(3): 716–725, 2009.

Sachs, J. and Malaney, P. The economic and social burden of malaria. *Nature*, 415(6872):680–685, Feb 2002. ISSN 0028-0836.

Sadowski, J. and Boström, J. MIMUMBA Revisited: Torsion Angle Rules for Conformer Generation Derived from X-ray Structures. *J. Chem. Inf. Model.*, 46(6):2305–2309, 2006.

Sadowski, J. and Gasteiger, J. From atoms and bonds to three-dimensional atomic coordinates: automatic model builders. *Chem. Rev.*, 93(7):2567–2581, 1993.

Sadowski, J., Gasteiger, J., and Klebe, G. Comparison of automatic three-dimensional model builders using 639 x-ray structures. *J. Chem. Inf. Model.*, 34(4):1000–1008, 1994.

Sadowski, P. and Baldi, P. Small-molecule 3D Structure Prediction Using Open Crystallography Data. *J. Chem. Inf. Model.*, 2013. Early Access.

Sayle, R. A. So you think you understand tautomerism? *J. Comput. Aided Mol. Des.*, 24(6-7):485–496, 2010. ISSN 0920-654X.

Schalon, C., Surgand, J.-S., Kellenberger, E., and Rognan, D. A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins: Struct., Funct., Bioinf.*, 71(4):1755–1778, 2008. ISSN 1097-0134.

Schärfer, C., Schulz-Gasch, T., Ehrlich, H.-C., Guba, W., Rarey, M., and Stahl, M. Torsion Angle Preferences in Druglike Chemical Space: A Comprehensive Guide. *J. Med. Chem.*, 56(5):2016–2028, 2013.

Schlitzer, M. Antimalarial drugs - what is in use and what is in the pipeline. *Arch. Pharm. (Weinheim)*, 341 (3):149–163, 2008. ISSN 1521-4184.

Schomburg, I., Chang, A., Placzek, S., Söhngen, C., Rother, M., Lang, M., Munaretto, C., Ulas, S., Stelzer, M., Grote, A., Scheer, M., and Schomburg, D. BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic Acids Res.*, 41(D1):D764–D772, 2013.

Schreyer, A. and Blundell, T. CREDO: A Protein-Ligand Interaction Database for Drug Discovery. *Chemical Biology & Drug Design*, 73(2):157–167, 2009. ISSN 1747-0285.

Schreyer, A. and Blundell, T. Usrcat: real-time ultrafast shape recognition with pharmacophoric constraints. *Journal of Cheminformatics*, 4(1):27, 2012. ISSN 1758-2946.

Schrödinger, LLC. The PyMOL molecular graphics system, version 1.4.1. August 2013.

Schulz-Gasch, T. and Stahl, M. Scoring functions for protein-ligand interactions: a critical perspective. *Drug Discovery Today: Technologies*, 1(3):231–239, 2004. ISSN 1740-6749.

Schwab, C. H. Conformations and 3d pharmacophore searching. *Drug Discovery Today: Technologies*, 7(4): e245–e253, 2010. ISSN 1740-6749.

Scior, T., Bender, A., Tresadern, G., Medina-Franco, J. L., Martínez-Mayorga, K., Langer, T., Cuanalo-Contreras, K., and Agrafiotis, D. K. Recognizing Pitfalls in Virtual Screening: A Critical Review. *J. Chem. Inf. Model.*, 52(4):867–881, 2012.

Seidel, T., Ibis, G., Bendix, F., and Wolber, G. Strategies for 3D pharmacophore-based virtual screening. *Drug Discovery Today: Technologies*, 7(4):e221–e228, 2010. ISSN 1740-6749.

Seifert, M. H., Wolf, K., and Vitt, D. Virtual high-throughput in silico screening. *BIOSILICO*, 1(4):143–149, 2003. ISSN 1478-5382.

Seiler, K. P., George, G. A., Happ, M. P., Bodycombe, N. E., Carrinski, H. A., Norton, S., Brudz, S., Sullivan, J. P., Muhlich, J., Serrano, M., Ferraiolo, P., Tolliday, N. J., Schreiber, S. L., and Clemons, P. A. ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res.*, 36 (suppl 1):D351–D359, 2008.

Shoichet, B. K. Virtual screening of chemical libraries. *Nature*, 432(7019):862–5, 2004.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. *ROCR: Visualizing the performance of scoring classifiers*, 2012. URL http://CRAN.R-project.org/package=ROCR. R package version 1.0-4.

Singh, J., Petter, R. C., Baillie, T. A., and Whitty, A. The resurgence of covalent drugs. *Nat. Rev. Drug Discov.*, 10(4):307–317, Apr 2011.

Sitzmann, M., Ihlenfeldt, W.-D., and Nicklaus, M. C. Tautomerism in large databases. *J. Comput. Aided Mol. Des.*, 24(6-7):521–551, 2010. ISSN 0920-654X.

Sitzmann, M., Weidlich, I. E., Filippov, I. V., Liao, C., Peach, M. L., Ihlenfeldt, W.-D., Karki, R. G., Borodina, Y. V., Cachau, R. E., and Nicklaus, M. C. PDB Ligand Conformational Energies Calculated Quantum-Mechanically. *J. Chem. Inf. Model.*, 52(3):739–756, 2012.

Smith, C. A., Toogood, H. S., Baker, H. M., Daniel, R. M., and Baker, E. N. Calcium-mediated thermostability in the subtilisin superfamily: the crystal structure of Bacillus Ak.1 protease at 1.8 Å resolution. *Journal of Molecular Biology*, 294(4):1027–1040, 1999. ISSN 0022-2836.

Sotriffer, C., Mannhold, R., Kubinyi, H., and Folkers, G. *Virtual Screening: Volume 48 - Principles, Challenges, and Practical Guidelines.* Methods and Principles in Medicinal Chemistry. Wiley, 2011. ISBN 9783527633340.

Sousa, S. F., Fernandes, P. A., and Ramos, M. J. Protein-ligand docking: Current status and future challenges. *Proteins: Struct., Funct., Bioinf.*, 65(1):15–26, 2006. ISSN 1097-0134.

Spangenberg, T., Burrows, J. N., Kowalczyk, P., McDonald, S., Wells, T. N. C., and Willis, P. The Open Access Malaria Box: A Drug Discovery Catalyst for Neglected Diseases. *PLoS ONE*, 8(6):e62906, 06 2013.

Spek, A. L. Structure validation in chemical crystallography. *Acta Crystallogr. D Biol. Crystallogr.*, 65(Pt 2): 148–155, Feb 2009.

Spellmeyer, D. C., Wong, A. K., Bower, M. J., and Blaney, J. M. Conformational analysis using distance geometry methods. *J. Mol. Graphics Modell.*, 15(1):18–36, 1997. ISSN 1093-3263.

Stahura, F. L. and Bajorath, J. New methodologies for ligand-based virtual screening. *Curr. Pharm. Des.*, 11:1189–1202(14), 2005.

Stockwell, B. R. Exploring biology with small organic molecules. *Nature*, 432(7019):846–854, Dec 2004.

Stolze, S. C., Meltzer, M., Ehrmann, M., and Kaiser, M. Ahp Cyclodepsipeptides: The Impact of the Ahp Residue on the "Canonical Inhibition" of S1 Serine Proteases. *Chembiochem*, 14(11):1301–1308, 2013. ISSN 1439-7633.

Swamidass, S. J. and Baldi, P. Bounds and Algorithms for Fast Exact Searches of Chemical Fingerprints in Linear and Sublinear Time. *J. Chem. Inf. Model.*, 47(2):302–317, 2007.

Tang, Y., Zhu, W., Chen, K., and Jiang, H. New technologies in computer-aided drug design: Toward target identification and new chemical entity discovery. *Drug Discovery Today: Technologies*, 3(3):307–313, 2006. ISSN 1740-6749.

Teotico, D. G., Babaoglu, K., Rocklin, G. J., Ferreira, R. S., Giannetti, A. M., and Shoichet, B. K. Docking for fragment inhibitors of AmpC β-lactamase. *Proceedings of the National Academy of Sciences*, 106(18): 7455–7460, 2009.

Teplyakov, A. V., Kuranova, I. P., Harutyunyan, E. H., Vainshtein, B. K., Frömmel, C., Höhne, W. E., and Wilson, K. S. Crystal structure of thermitase at 1.4 Å resolution. *J. Mol. Biol.*, 214(1):261–279, July 1990.

Terstappen, G. C. and Reggiani, A. In silico research in drug discovery. *Trends in Pharmacological Sciences*, 22(1):23–26, 2001. ISSN 0165-6147.

Tetko, I. V., Bruneau, P., Mewes, H.-W., Rohrer, D. C., and Poda, G. I. Can we estimate the accuracy of ADME-Tox predictions? *Drug Discovery Today*, 11(15-16):700–707, 2006. ISSN 1359-6446.

Thera, M. A., Doumbo, O. K., Coulibaly, D., Diallo, D. A., Kone, A. K., Guindo, A. B., Traore, K., Dicko, A., Sagara, I., Sissoko, M. S., Baby, M., Sissoko, M., Diarra, I., Niangaly, A., Dolo, A., Daou, M., Diawara, S. I., Heppner, D. G., Stewart, V. A., Angov, E., Bergmann-Leitner, E. S., Lanar, D. E., Dutta, S., Soisson, L., Diggs, C. L., Leach, A., Owusu, A., Dubois, M.-C., Cohen, J., Nixon, J. N., Gregson, A., Takala, S. L., Lyke, K. E., and Plowe, C. V. Safety and immunogenicity of an ama-1 malaria vaccine in malian adults: Results of a phase 1 randomized controlled trial. *PLoS ONE*, 3(1):e1465, 01 2008.

Totrov, M. and Abagyan, R. Flexible ligand docking to multiple receptor conformations: a practical alternative. *Current Opinion in Structural Biology*, 18(2):178–184, 2008. ISSN 0959-440X.

Truchon, J.-F. and Bayly, C. I. Evaluating virtual screening methods: Good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model.*, 47(2):488–508, 2007.

Ursu, O., Rayan, A., Goldblum, A., and Oprea, T. I. Understanding drug-likeness. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(5):760–781, 2011. ISSN 1759-0884.

Vainio, M. J. and Johnson, M. S. Generating conformer ensembles using a multiobjective genetic algorithm. *J. Chem. Inf. Model.*, 47(6):2462–2474, 2007.

Valler, M. J. and Green, D. Diversity screening versus focussed screening in drug discovery. *Drug Discovery Today*, 5(7):286–293, 2000. ISSN 1359-6446.

Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., and Taylor, R. D. Improved protein-ligand docking using GOLD. *Proteins: Struct., Funct., Bioinf.*, 52(4):609–623, 2003. ISSN 1097-0134.

Verma, J., Khedkar, V. M., and Coutinho, E. C. 3D-QSAR in Drug Design - A Review. *Curr. Top Med. Chem.*, 10:95–115(21), 2010.

Villoutreix, B. O., Eudes, R., and Miteva, M. A. Structure-based virtual ligand screening: Recent success stories. *Combinatorial Chemistry & High Throughput Screening*, 12(10):1000–1016, 2009.

Šali, A. and Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *Journal of Molecular Biology*, 234(3):779–815, 1993. ISSN 0022-2836.

Walters, P. W., Stahl, M. T., and Murcko, M. A. Virtual screening-an overview. *Drug Discovery Today*, 3(4): 160–178, 1998. ISSN 1359-6446.

Wang, L., Ma, C., Wipf, P., Liu, H., Su, W., and Xie, X.-Q. TargetHunter: An In Silico Target Identification Tool for Predicting Therapeutic Potential of Small Organic Molecules Based on Chemogenomic Database. *The AAPS Journal*, 15(2):395–406, 2013. doi: 10.1208/s12248-012-9449-z.

Wang, R. and Wang, S. How Does Consensus Scoring Work for Virtual Library Screening? An Idealized Computer Experiment. *J. Chem. Inf. Comput. Sci.*, 41(5):1422–1426, 2001.

Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., Zhou, Z., Han, L., Karapetyan, K., Dracheva, S., Shoemaker, B. A., Bolton, E., Gindulyte, A., and Bryant, S. H. PubChem's BioAssay Database. *Nucleic Acids Res.*, 40(D1):D400–D412, 2012.

Ward, J. H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.*, 58(301):236–244, 1963.

Warr, W. A. Tautomerism in chemical information management systems. *J. Comput. Aided Mol. Des.*, 24 (6-7):497–520, 2010. ISSN 0920-654X.

Wei, B. Q., Baase, W. A., Weaver, L. H., Matthews, B. W., and Shoichet, B. K. A model binding site for testing scoring functions in molecular docking. *J. Mol. Biol.*, 322(2):339–355, 2002. ISSN 0022-2836.

Wei, B. Q., Weaver, L. H., Ferrari, A. M., Matthews, B. W., and Shoichet, B. K. Testing a flexible-receptor docking algorithm in a model binding site. *J. Mol. Biol.*, 337(5):1161–1182, 2004. ISSN 0022-2836.

Weininger, D. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Model.*, 28(1):31–36, 1988.

Wermuth, C. G., Ganellin, C. R., Lindberg, P., and Mitscher, L. A. Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998). *Pure Appl. Chem.*, 70(5):1129–1143, 1998.

White, N. J. Delaying antimalarial drug resistance with combination chemotherapy. *Parassitologia*, 41(1-3): 301–308, Sep 1999.

White, N. J. Antimalarial drug resistance. *The Journal of Clinical Investigation*, 113(8):1084–1092, 4 2004.

WHO World Malaria Report, 2010. Technical report, World Health Organisation (WHO), 2010. URL http://www.who.int/malaria/world_malaria_report_2010/worldmalariareport2010.pdf. [Online; accessed 1-July-2013].

Willett, P. Enhancing the effectiveness of ligand-based virtual screening using data fusion. *QSAR & Combinatorial Science*, 25(12):1143–1152, 2006. ISSN 1611-0218.

Willett, P. Similarity Searching Using 2D Structural Fingerprints. In Bajorath, J., editor, *Chemoinformatics and Computational Chemical Biology*, volume 672 of *Methods in Molecular Biology*, pages 133–158. Humana Press, 2011. ISBN 978-1-60761-838-6.

Willett, P., Barnard, J. M., and Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.*, 38 (6):983–996, 1998.

William, T., Menon, J., Rajahram, G., Chan, L., Ma, G., Donaldson, S., Khoo, S., Frederick, C., Jelip, J., Anstey, N. M., and Yeo, T. W. Severe Plasmodium knowlesi malaria in a tertiary care hospital, Sabah, Malaysia. *Emerging Infect. Dis.*, 17(7):1248–1255, Jul 2011.

Williams, M. Productivity Shortfalls in Drug Discovery: Contributions from the Preclinical Sciences? *J. Pharmacol. Exp. Ther.*, 336(1):3–8, 2011.

Wilson, S. R., Cui, W., Moskowitz, J. W., and Schmidt, K. E. Applications of simulated annealing to the conformational analysis of flexible molecules. *J. Comput. Chem.*, 12(3):342–349, 1991. ISSN 1096-987X.

Wirth, D. F. The parasite genome: Biological revelations. *Nature*, 419(6906):495–496, Oct 2002. ISSN 0028-0836.

Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.*, 34(suppl 1):D668–D672, 2006.

Withers-Martinez, C., Saldanha, J. W., Ely, B., Hackett, F., O'Connor, T., and Blackman, M. J. Expression of Recombinant Plasmodium falciparum Subtilisin-like Protease-1 in Insect Cells. *J. Biol. Chem.*, 277(33): 29698–29709, 2002.

Withers-Martinez, C., Jean, L., and Blackman, M. J. Subtilisin-like proteases of the malaria parasite. *Mol. Microbiol.*, 53(1):55–63, 2004. ISSN 1365-2958.

Withers-Martinez, C., Suarez, C., Fulle, S., Kher, S., Penzo, M., Ebejer, J.-P., Koussis, K., Hackett, F., Jirgensons, A., Finn, P., and Blackman, M. J. Plasmodium subtilisin-like protease 1 (SUB1): Insights into the active-site structure, specificity and function of a pan-malaria drug target. *International Journal for Parasitology*, 42(6):597–612, 2012. ISSN 0020-7519.

Wolber, G. and Langer, T. LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters. *J. Chem. Inf. Model.*, 45(1):160–169, 2005.

Wolber, G., Seidel, T., Bendix, F., and Langer, T. Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug Discovery Today*, 13(1-2):23–29, 2008. ISSN 1359-6446.

Wongsrichanalai, C. and Meshnick, S. R. Declining artesunate-mefloquine efficacy against falciparum malaria on the Cambodia-Thailand border. *Emerging Infect. Dis.*, 14(5):716–719, May 2008.

Yabuuchi, H. *enrichvs: Enrichment assessment of virtual screening approaches*, 2011. URL http://CRAN.R-project.org/package=enrichvs. R package version 0.0.5.

Yang, S.-Y. Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug Discovery Today*, 15(11-12):444–450, 2010. ISSN 1359-6446.

Yang, T., Wu, J. C., Yan, C., Wang, Y., Luo, R., Gonzales, M. B., Dalby, K. N., and Ren, P. Virtual screening using molecular simulations. *Proteins: Struct., Funct., Bioinf.*, 79(6):1940–1951, 2011. ISSN 1097-0134.

Yang, Y., Adelstein, S. J., and Kassis, A. I. Target discovery from data mining approaches. *Drug Discovery Today*, 14(3-4):147–154, 2009. ISSN 1359-6446.

Yeoh, S., O'Donnell, R. A., Koussis, K., Dluzewski, A. R., Ansell, K. H., Osborne, S. A., Hackett, F., Withers-Martinez, C., Mitchell, G. H., Bannister, L. H., Bryans, J. S., Kettleborough, C. A., and Blackman, M. J. Subcellular discharge of a serine protease mediates release of invasive malaria parasites from host erythrocytes. *Cell*, 131(6):1072–1083, 2007. ISSN 0092-8674.

Young, D., Martin, T., Venkatapathy, R., and Harten, P. Are the Chemical Structures in Your QSAR Correct? *QSAR & Combinatorial Science*, 27(11-12):1337–1345, 2008. ISSN 1611-0218.

Yousef, G. M., Kopolovic, A. D., Elliott, M. B., and Diamandis, E. P. Genomic overview of serine proteases. *Biochem. Biophys. Res. Commun.*, 305(1):28–36, 2003. ISSN 0006-291X.

Zhang, Q. and Muegge, I. Scaffold Hopping through Virtual Screening Using 2D and 3D Similarity Descriptors: Ranking, Voting, and Consensus Scoring. *J. Med. Chem.*, 49(5):1536–1548, 2006.

Zhu, T., Cao, S., Su, P.-C., Patel, R., Shah, D., Chokshi, H. B., Szukala, R., Johnson, M. E., and Hevener, K. E. Hit Identification and Optimization in Virtual Screening: Practical Recommendations Based on a Critical Literature Analysis. *J. Med. Chem.*, 56(17):6560–6572, 2013.