A NOVEL KNOWLEDGE BASED CONFORMATION SAMPLING ALGORITHM AND APPLICATIONS IN DRUG DISCOVERY

By

Sandeepkumar Kailas Kothiwale

Dissertation
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
in

Chemistry
August, 2016
Nashville, Tennessee

Approved:
Jens Meiler, PhD
Andes Hess, PhD
Ambra Pozzi, PhD
Terry Lybrand, PhD

ACKNOWLEDGEMENTS

First, I thank my advisor Prof. Jens Meiler, for the continuous support in my Ph.D work, for his patience and insights. I appreciate Jens' time, efforts and funding that has made my PhD experience productive and stimulating. Jens gave me the freedom to work on a diverse set of problems, all the while providing valuable feedback, constructive criticism and advice. Besides my advisor, I would like to thank the rest of my thesis committee: Prof Ambra Pozzi, Prof. Andes Hess, and Dr. Terry Lybrand, for their insightful comments, valuable feedback, support and encouragement.

My sincere thanks also goes to Jefferey Mendenhall for his help and training in software development. I would like to acknowledge my collaborators Prof. Ambra Pozzi and Dr. Corina Borza for a fruitful and successful collaboration on DDR1 kinase. I must also acknowledge Dr. Steven Combs, who has been a great friend, colleague and collaborator. I am especially grateful to him for the opportunity to work in the field of scientific gamification. I would also like to thank my friends both at Vanderbilt University and outside, for their constant support and encouragement.

I owe special thanks to my family, my parents and my brother, for their love and encouragement throughout my life and during my PhD. The last words of acknowledgement are reserved for my loving and caring wife Karuna who has been very encouraging and supportive in the final stages of this PhD. I dedicate this thesis to them.

PREFACE

Computational approaches have become important tools in drug discovery. Drug discovery is a lengthy process that begins with target identification, lead compound discovery, and lead compound optimization, followed by pre-clinical studies. Computational tools have been developed which complement the experimental drug discovery efforts at each of these steps. Target discovery is often achieved by phenotypic screens using micro-array analysis. This is achieved computationally through bioinformatics analysis of gene expression data and protein-protein interaction networks. Often experimental screening for lead compounds is preceded by computational screening of hundreds of thousands of compound. Computational virtual-high throughput compound screening technologies are used to prioritize molecules for experimental testing and are estimated to increase the chance of finding a lead molecule by about ten times. Computational prescreening saves time, resources, and efforts required for experimental screening by reducing the number of compounds to be tested. The goal of lead compound optimization is to improve its potency against the target of interest. This is achieved by medicinal chemistry approaches through synthesis of a number of derivatives. Computational modelling of target-ligand interactions is often used to direct the medicinal chemistry studies. Results from experimental optimization can also be used to create computational models to predict the pharmacophore of the lead compound. Lead optimization can be further aided by computational models that predict drug-likeness and toxicity, saving substantial efforts in the downstream pre-clinical studies.

Development of novel computational technologies for drug discovery has been the primary focus of my PhD thesis. A novel knowledge based conformation sampling algorithm was implemented which derives information from structural databases. Molecular conformation sampling is ubiquitous and critical in computational drug discovery technologies. The new algorithm performs better than other conformation algorithm currently available in the field and has already been incorporated into a major macromolecular modelling software. Another focus of my work has been the application of computational technologies for discovery of novel and selective binders of the kinase domain of Discoidin Domain Receptor (DDR1). At least one novel chemical scaffold was discovered and confirmed as a DDR1 inhibitor through these efforts.

Computer aided-drug discovery is split into domains that focus on either structure-based or ligand-based techniques. Structure-based approaches are feasible when the structure of the biological target protein or its homologues is available. These techniques include ligand docking/design and structure-based pharmacophore maps. In the absence of a structural model, ligand-based approaches provide an alternative way of identifying new active molecules and optimizing their activity. Ligand approaches leverage quantitative structure activity relationship (QSAR) models and pharmacophore maps. A comprehensive review of successful applications of computational technologies in drug discovery processes is provided in the first chapter of this thesis. It is an abridged version of the review article:

Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W., Jr., Computational methods in drug discovery. *Pharmacol Rev* 2014, *66* (1), 334-95. Reprinted with permission of the American Society for Pharmacology and Experimental Therapeutics. All rights reserved.

Molecules are comprised of one or more atoms connected by bonds, many of which are rotatable. This rotation about the bonds allows molecules to adopt distinct orientations, known as conformations, in 3-dimension space. In solution, a given molecule can exist in multiple different low-energy conformations. A small molecule may bind to its protein target in one of the conformations favored in solution, or alternatively one that is induced by the interactions with the target protein of interest. Computational modeling for binding prediction, whether structure-based or ligand-based, thus needs to take into account small molecule flexibility. The success of structure-based drug discovery technologies depends significantly on the availability of high-quality ligand conformations that are necessary to accurately model interactions between the target and the ligand molecules. Ligand conformations are also important for ligand-based methods, for example, to align multiple active molecules for shape matching and developing pharmacophore models. Several conformation sampling methods exist that use either physics-based approaches i.e. molecular mechanics or pre-existing knowledge about small molecular conformations. The primary focus of this doctoral thesis is the development of a high-quality conformation generation algorithm that uses extended fragment conformation information. This algorithm, called BCL::CONF, is described in Chapter 2 of this thesis and is an adaptation of manuscript:

Kothiwale, S.; Mendenhall, J. L.; Meiler, J., BCL::CONF: small molecule conformational sampling using a knowledge based rotamer library. *J Cheminform* 2015, *7*, 47. http://creativecommons.org/licenses/by/4.0/

BCL::CONF uses the knowledge about molecular fragment conformations for conformation sampling. This is analogous to using protein side-chain conformations used in ROSETTA software suite used for modeling protein-related interactions. Previously, ligand conformations needed to be sampled using external software and then imported into ROSETTA. This precluded the implementation of algorithms such as ligand design into ROSETTA that require on-the-fly conformation sampling. The design of the BCL::CONF algorithm has allowed its integration into ROSETTA as an external library.

Foldit is an online scientific game based on ROSETTA where scientific problems are posted as puzzles in the form of a game. Through their collaborative play, players have been able to solve some of the most difficult problems that have been evading scientists for several years. We implemented the drug design game into Foldit i.e. the graphical user interface, interaction maps, toolsets and fragments datasets for drug design. A rapid on-the-fly conformation-sampling algorithm was needed so players can choose best interactions that ligands have with target molecule of interest. BCL::CONF is now an integral part of the Foldit drug design game.

Quantitative Structure Activity Relationship models correlate the structure of molecules to their activities against a particular target of interest. Our hypothesis is that prediction accuracy of the models can be improved if the molecular binding conformation is available. Due to lack of such high quality information, a method was developed to create computational models that predict activities based on conformational space available to a ligand molecule. Chapter 3 describes the research carried out toward using multiple conformations to develop ligand-based quantitative structure activity relationships. Instead of using a single molecular conformation, the neural network was trained on multiple conformations so that it learns conformations that are common only to the active molecules and uses this information to classify them distinctly from inactive molecules.

Another primary focus of this doctoral work was drug discovery efforts for DDR1 kinase protein. This work was done in collaboration with the Pozzi lab at the Nephrology Department at Vanderbilt Medical School. They have substantial evidence that implicates Discoidin Domain Receptor-1 (DDR1) for fibrosis of kidney caused by diabetes. Unregulated expression of DDR1 has also been implicated in several cancers making it a desirable target for anti-cancer therapeutics. Kinase receptors, including DDR1, relay cell-signaling pathways through a kinase domain that phosphorylates substrate proteins. Kinase domains are the most popular targets for inhibiting the activity of uncontrolled receptor activation. In this work, the kinase domain of DDR1 receptor was the target for inhibition using computational drug discovery tools. Homology models and docking studies were performed to study the interaction between the known inhibitors of DDR1 kinase and the kinase domain. Homology models were used to dock DDR1 binders to predict the co-crystal structure, which was found to be in agreement with a crystal structure published later. These studies and models were reported in an article published in Drug Discovery Today. Chapter 4 of this thesis is an adaptation of the article -

Kothiwale, S.; Borza, C. M.; Lowe, E. W., Jr.; Pozzi, A.; Meiler, J., Discoidin domain receptor 1 (DDR1) kinase as target for structure-based drug discovery. *Drug Discov Today* 2015, *20* (2), 255-61. Permission to reprint obtained from copyright clearance center.

In addition to the structure-based studies, chapter 4 describes the ligand-based drug discovery experiments carried out to find DDR1 inhibitors. The present work develops QSAR models that correlate structure of molecules to activity against DDR1, utilizing the several novel DDR1 kinase inhibitors reported in the past three years. Virtual compound libraries were screened using the QSAR models and molecules were prioritized for experimental testing. Molecules were tested using a high-throughput kinase inhibition assay to identify DDR1 binders.

With more than 400 different kinase proteins that are closely related, kinase inhibitors are highly non-specific, that is, they show significant activity across multiple kinase targets. Kinase inhibitors are categorized based on their binding site. Type 1 inhibitors target only the ATP binding pocket and lock the kinase in an active state. Functional kinase domains transfer phosphate from the bound ATP to a substrate protein thereby relaying cellular signals. Type1 inhibitors competitively target the ATP binding pocket thereby rendering the kinase protein inactive. Type II

inhibitors target an allosteric site in addition to the ATP binding site and thus are more selective. Much of the current kinase drug discovery efforts is directed toward idenfying highly-selective inhibitors of specific kinases. Computational models to predict a kinase selectivity profile were developed to aid in the identification of novel DDR1-selective inhibitors, and are reported in Chapter 5.

For each chapter, computational protocols, commands and scripts are included in the appendices under respective headings. The path to scripts, models and code are included in the thesis directory whose structure is detailed in the appendices.

TABLE OF CONTENTS

CHAPTER 1 : COMPUTATIONAL METHODS IN DRUG DISCOVERY

Introduction

Computer aided drug discovery/design methods have played a major role in the development of therapeutically important small molecules for over three decades. These methods are broadly classified as either structure-based or ligand-based methods. Structure-based methods are in principle analogous to high throughput screening in that both target and ligand structure information is imperative. These approaches include ligand-docking, pharmacophore, and ligand-design methods. These approaches include important tools such as target/ligand databases, homology modeling, ligand-fingerprint methods, etc. Ligand-based methods use only ligand information for predicting activity depending on its similarity/dissimilarity to previously known active ligands. Widely used ligand-based methods include ligand-based pharmacophores, molecular descriptors and quantitative structure activity relationships.

On October 5, 1981, Fortune magazine published a cover article entitled the "Next Industrial Revolution: Designing Drugs by Computer at Merck" [6]. Some have credited this as being the start of intense interest in the potential for Computer Aided Drug Design (CADD). While progress was being made in CADD, the potential for high-throughput screening (HTS) had begun to take precedence as a means for finding novel therapeutics. This brute force approach relies on automation to screen high numbers of molecules in search of those that elicit the desired biological response. The method has the advantage of requiring minimal compound design or prior knowledge and technologies required to screen large libraries have become more efficient. However, while traditional HTS often results in multiple hit compounds, some of which are capable of being modified into a lead and later a novel therapeutic, the hit rate for HTS is often extremely low. This low hit rate has limited the usage of HTS to research programs capable of screening large compound libraries. In the past decade, CADD has reemerged as a way of significantly decreasing the number of compounds necessary to screen, while retaining the same level of lead compound discovery. Many compounds predicted to be inactive can be skipped and those predicted to be active can be prioritized. This reduces the cost and workload of a full HTS screen without compromising lead discovery. Additionally, traditional HTS assays often require extensive development and validation before they can be employed. Since CADD requires significantly less preparation time, experimenters can perform CADD studies while the traditional HTS assay is being prepared. The fact that both of these tools can be used in parallel provides an additional benefit for CADD in a drug discovery project.

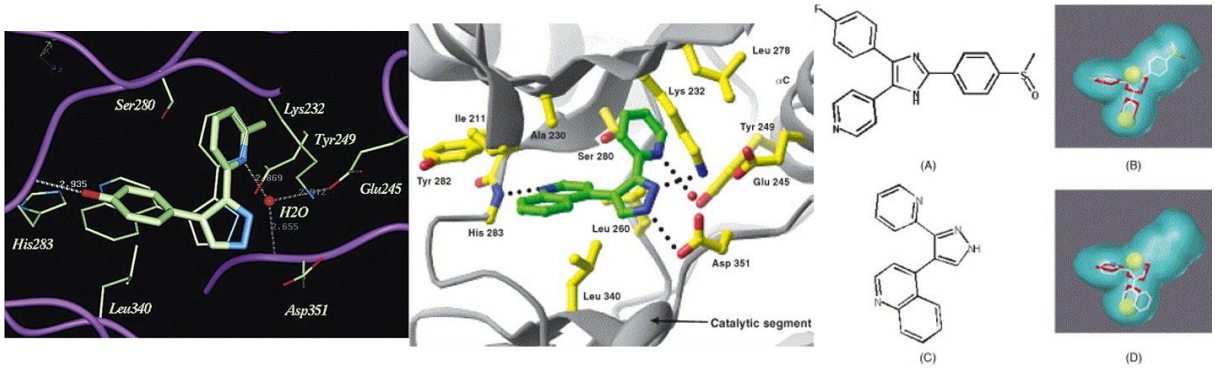


Figure 1-1 Identical lead compounds are discovered in a traditional high-throughput screen and structure-based virtual high-throughput screen. I) X-ray crystal structures of 1 and 18 bound to the ATP-binding site of the TβR-I kinase domain discovered using traditional high-throughput screening. Compound 1, shown as the thinner wire-frame is the original hit from the HTS and is identical to that which was discovered using virtual screening. Compound 18 is a higher affinity compound following lead optimization. Source: (Sawyer, Anderson et al. 2003) II) X-ray crystal structure of compound HTS466284 bound to the TβRI active site. This compound is identical to compound 1 in I but was discovered using structure-based virtual high-throughput screening. Source :Singh et al [4] III) Structure-based virtual screening strategy began with known inhibitor (A), leading to shape-based query (B) based on x-ray crystal structure of inhibitor bound to the TβRI active site. Compound (C) was discovered using this query due to its similarities (D). Source: Singh et al[4]

For example, researchers at Pharmacia (now part of Pfizer) used CADD tools to screen for inhibitors of tyrosine phosphatase-1B, an enzyme implicated in diabetes. Their virtual screen yielded 365 compounds, 127 of which showed effective inhibition, a hit rate of nearly 35%. Simultaneously, this group performed a traditional HTS against the same target. Of the 400,000 compounds tested, 81 showed inhibition, producing a hit rate of only 0.021%. This comparative case effectively displays the power of CADD[7]. CADD has already been used in the discovery of compounds that have passed clinical trials and become novel therapeutics in the treatment of a variety of diseases. Some of the earliest examples of approved drugs that owe their discovery in large part to the tools of CADD include the carbonic anhydrase inhibitor dorzolamide, approved in 1995[8], the angiotensin-converting enzyme (ACE) inhibitor captopril, approved in 1981 as an antihypertensive drug [9], three therapeutics for the treatment of HIV: saquinavir (approved in 1995), ritonavir, and indinavir (both approved in 1996) [6] and tirofiban, a fibrinogen antagonist approved in 1998 [10].

One of the most striking examples of the possibilities presented from CADD occurred in 2003 with the search for novel Transforming Growth Factor-β1 (TGF-β1) receptor kinase inhibitors. One group at Eli Lilly used a traditional HTS to identify a lead compound that was subsequently improved by examination of structure activity relationship (SAR) using *in vitro* assays [11], while a group at Biogen Idec used a CADD approach involving virtual HTS based on the structural interactions between a weak inhibitor and TGF-β1 receptor kinase [4]. Upon the virtual screening of compounds, the group at Biogen Idec identified 87 hits, the best hit being identical in structure to the lead compound

discovered through the traditional HTS approach at Eli Lilly [12]. In this situation CADD, a method involving reduced cost and workload, was capable of producing the same lead as a full-scale HTS.

Position of CADD in the drug discovery pipeline

CADD is capable of increasing the hit rate of novel drug compounds as it employs a much more targeted search than traditional HTS and combinatorial chemistry. It not only aims to explain the molecular basis of therapeutic

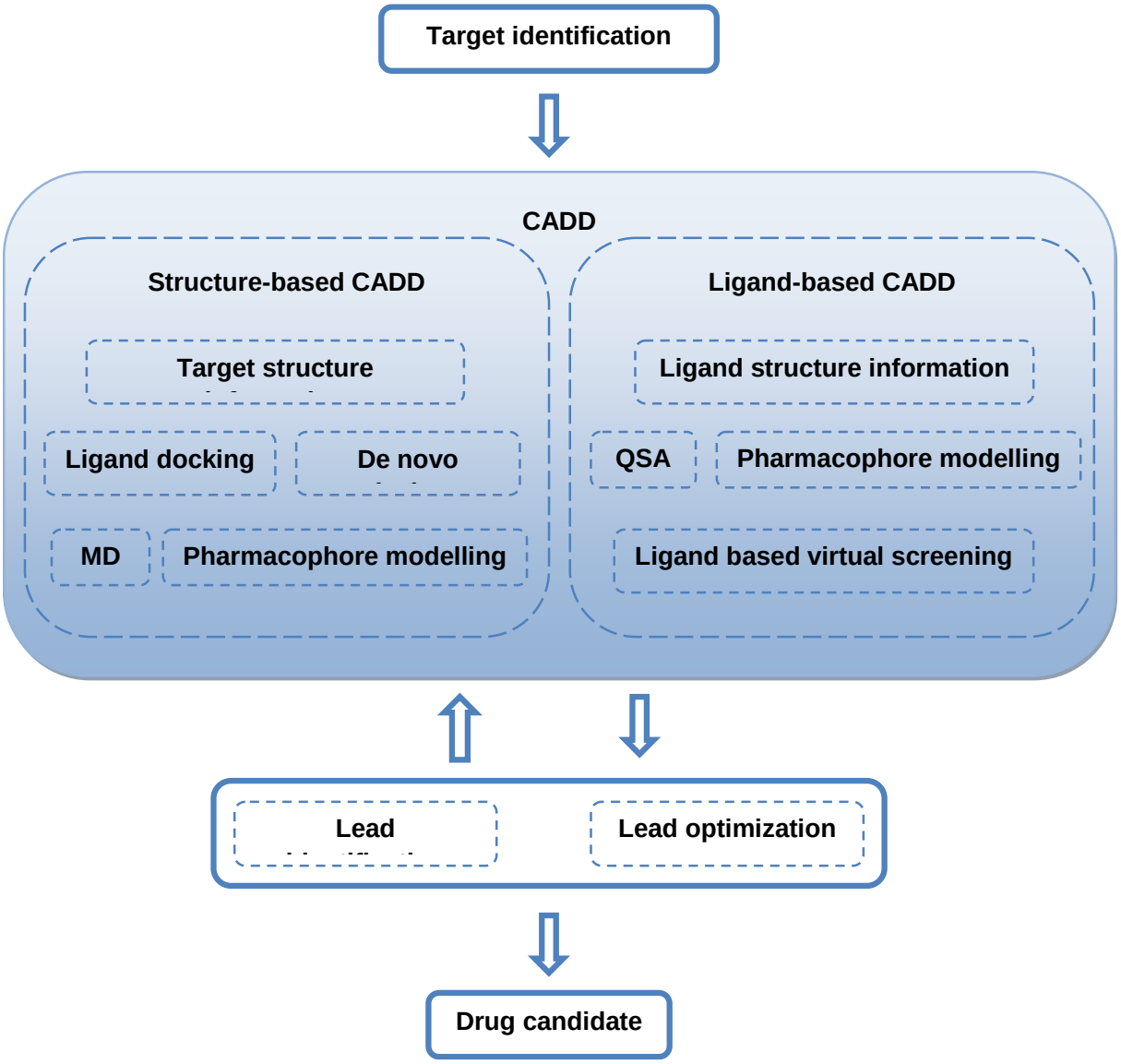


Figure 1-2 CADD in drug discovery/design pipeline – A therapeutic target is identified against which a drug has to be developed. Depending on the availability of structure information, a structure-based approach or a ligand-based approach is employed. A successful CADD campaign will allow identification of multiple lead compounds. Lead identification is often followed by several cycles of lead optimization and subsequent lead identification using CADD. Lead compounds are tested in vivo to identify drug candidates.

activity, but also to predict possible derivatives that could improve activity. In a drug discovery campaign, CADD is usually used for three major purposes: a) filter large compound libraries into smaller sets of predicted active compounds that can be tested experimentally, b) guide the optimization of lead compounds, whether to increase its affinity or optimize Drug Metabolism and Pharmacokinetics (DMPK) properties including Absorption, Distribution, Metabolism, Excretion, and the potential for Toxicity (ADMET), c) design novel compounds, either by 'growing' starting molecules one functional group at a time or by piecing together fragments into novel chemotypes Figure 1-2 illustrates the position of CADD in drug discovery pipeline.

CADD can be classified into two general categories: structure-based and ligand-based. Structure-based CADD relies on the knowledge of the target protein structure to calculate interaction energies for all compounds tested, while ligand-based CADD exploits the knowledge of known active and inactive molecules through chemical similarity searches or construction of predictive, Quantitative Structure-Activity Relation (QSAR) models [13]. Structure-based CADD is generally preferred where high resolution structural data of the target protein is available, i.e. for soluble proteins that can readily be crystallized. Ligand-based CADD is generally preferred when no or little structural information is available, often for membrane protein targets. The central goal of structure-based CADD is to design compounds that bind tightly to the target, i.e. with large reduction in free energy, improved DMPK/ADMET properties, and are target specific, i.e. have reduced off-target effects [14]. A successful application of these methods will result in a compound that has been validated *in vitro* and *in vivo,* and its binding location has been confirmed, ideally through a co-crystal structure.

One of the most common uses in CADD is the screening of virtual compound libraries, also known as virtual high-throughput screening (vHTS). This allows experimentalists to focus resources on testing compounds likely to have any activity of interest. In this way, a researcher can identify an equal number of hits while screening significantly less compounds as compounds predicted to be inactive with high confidence may be skipped. Avoiding a large population of inactive compounds saves money and time as the size of the experimental HTS is significantly reduced without sacrificing a large degree of hits. Ripphausen *et al* note that the first mention of vHTS was in 1997 [15] and chart an increasing rate of publication for the application of vHTS between 1997 and 2010. They also found that the largest fraction of hits has been obtained for G-protein coupled receptors (GPCR's), followed by kinases [16].

vHTS comes in many forms including chemical similarity searches by fingerprints or topology, selecting compounds by predicted biological activity through QSAR models or pharmacophore mapping, and virtual docking of compounds into target of interest, known as structure-based docking[17]. These methods allow the ranking of "hits" from the virtual compound library for acquisition. The ranking can reflect a property of interest such as percent similarity to a query compound or predicted biological activity, or in the case of docking, the lowest energy scoring poses for each ligand bound to the target of interest [18]. Often initial hits are rescored and ranked using higher-level computational techniques that are too time-consuming to be applied to full-scale vHTS. It is important to note that

vHTS does not aim to identify a drug-compound that is ready for clinical testing, but rather to find leads with chemotypes that have not previously been associated with a target. This is not unlike a traditional HTS where a compound is generally considered a hit if its activity is close to 10 μM. Through iterative rounds of chemical synthesis and *in vitro* testing a compound is first developed into a "lead" with higher affinity, some understanding of its structure-activity-relation, and initial tests for DMPK/ADMET properties. Only after further iterative rounds of lead-to-drug optimization and *in vivo* testing does a compound reach a clinically appropriate potency and acceptable DMPK/ADMET properties[19]. For example, the literature survey performed by Ripphausen *et al* revealed that a majority of successful vHTS applications identified a small number of hits that are usually active in the micromolar range, and hits with low nanomolar potency are only rarely identified [16].

The cost benefit of using computational tools in the lead optimization phase of drug development is substantial. Development of new drugs can cost anywhere in the range of 400 million to 2 billion dollars with synthesis and testing of lead analogues being a large contributor to that sum [20]. Therefore, it is beneficial to apply computational tools in hit-to-lead optimization in order to cover a wider chemical space while reducing the number of compounds that must be synthesized and tested *in vitro*. The computational optimization of a hit compound can involve a structure-based analysis of docking poses and energy profiles for hit analogues, ligand-based screening for compounds with similar chemical structure or improved predicted biological activity, or prediction of favorable DMPK/ADMET properties. The comparably low-cost of CADD compared to chemical synthesis and biological characterization of compounds make these methods attractive to focus, reduce, and diversify the chemical space that is explored[17].

De novo drug design is another tool in CADD methods, but rather than screening libraries of previously synthesized compounds, it involves the design of novel compounds. A structure generator is needed to sample the space of chemicals. Given the size of the search space (more than 10^{60} molecules) [21] heuristics are employed to focus these algorithms on molecules that are predicted to be highly active, readily synthesizable, devoid of undesirable properties, often derived from a starting scaffold with demonstrated activity, etc. Additionally, effective sampling strategies are utilized while dealing with large search spaces such as evolutionary algorithms, metropolis search, or simulated annealing [22]. The construction algorithms are generally defined as either linking or growing techniques. Linking algorithms involve docking of small fragments or functional groups such as rings, acetyl groups, esters, etc. to particular binding sites followed by linking fragments from adjacent sites. Growing algorithms, on the other hand, begin from a single fragment placed in the binding site to which fragments are added, removed, and changed to improve activity. Like vHTS, the role of *de novo* drug design is not to design the single compound with nanomolar activity and acceptable DMPK/ADMET properties, but rather to design a lead compound that can be subsequently improved.

Target databases for CADD

The knowledge of the structure of the target protein is required for structure-based CADD. The Protein Data Bank (PDB) [23], established in 1971 at the Brookhaven National Laboratory, and the Cambridge Crystallographic Data Center [24], are among the most commonly used databases for protein structure. PDB currently houses more than 81,000 protein structures, the majority of which have been determined using X-ray crystallography and a smaller set determined using NMR spectroscopy. When an experimentally determined structure of a protein is not available, it is often possible to create a comparative model based on the experimental structure of a related protein. Most frequently, the relation is based in evolution that introduced the term 'homology model'. The Swiss-Model server is one of the most widely used web-based tools for homology modeling[25]. Initially, static protein structures were used for all structure-based design methods. However, proteins are not static structures but rather exist as ensembles of different conformational states. The protein fluctuates through this ensemble depending on the relative free energies of each of these states, spending more time in conformations of lower free energy. Ligands are thought to interact with some conformations but not others, thus stabilizing conformational populations in the ensemble. Therefore, docking compounds into a static protein structure can be misleading, as the chosen conformation may not be representative of the conformation capable of binding the ligand. Recently, it has become state of the art to employ additional computational tools such as molecular dynamics and molecular mechanics to simulate and evaluate a protein's conformational space. Conformational sampling provides a collection of snapshots that can be used in place of a single structure that reflect the breadth of fluctuations the ligand may encounter *in vivo*. This approach was proven to be invaluable in CADD by Schames *et al* in the 2004 identification of novel HIV Integrase inhibitors[26]. Some methods, like ROSETTA-LIGAND [27], are capable of incorporating protein flexibility during the actual docking procedure, omitting the need for snapshot ensembles.

Benchmarking Techniques of CADD

Effective benchmarks are essential for assessment of performance and accuracy of CADD algorithms. Design of the benchmark in terms of number and type of target proteins, size and composition of active and inactive chemicals, and selection of quality measures play a key role when comparing new CADD methods with existing ones. Scientific benchmarks usually involve screening a library of compounds that include a subset of known actives combined with known inactive compounds and then evaluating the number of known actives that were identified by the CADD technique employed [28].

Performance is commonly reported by correlating predicted activities with experimentally observed activities with Receiver Operating Characteristic (ROC) curves. These curves plot the number of true positive predictions on the y-axis versus the false positive predictions on the x-axis. A random predictor would result in a plot of a line with a slope of one, whereas curves with high initial slopes above this line represent increasing performance scores for the method tested[29]. ROC curves are therefore analyzed by determining the area under the curve (AUC), positive

predictive value (PPV) – the ratio of true positives in a subset selected in a vHTS screen, or enrichment – a benchmark that normalizes PPV by the background ratio of positives in the dataset.

For structure-based CADD it is now common to also include decoy molecules that further test a technique's ability to discern actives from inactives at high resolution. Irwin *et al* created the Directory of Useful Decoys (DUD) dataset designed for high resolution benchmarking. It includes experimental data for approximately 3000 ligands covering up to 40 different targets and a set of carefully chosen decoys [30]. These decoys were designed to resemble positive ligands physically but not topologically [31]. These decoys, however, are not experimentally validated and are only postulated to be "inactive" against the targets. Good and Oprea have developed clustered versions of DUD with added data sets from sources such as WOMBAT to avoid challenges in enrichment comparisons between methods due to different parameters and limited diversity [32].

The current chapter covers various established structure-based and ligand-based CADD methods. The applications of various methods are illustrated with recent studies that concluded in compounds that were at least tested *in vivo* and often entered clinical trials.

Structure-Based Computer-Aided Drug Design (SB-CADD)

SB-CADD relies on the ability to determine and analyze 3D structures of biological molecules. The core hypothesis of this approach is that a molecule's ability to interact with a specific protein and exert a desired biological effect depends on its ability to favorably interact with a particular binding site on that protein. Molecules that share those favorable interactions will exert similar biological effects. Therefore, novel compounds can be elucidated through the careful analysis of a protein's binding site. Structural information about the target is a prerequisite for any SB-CADD project. Scientists have been using a target protein's structure to aid in drug discovery since the early 1980s [33]. Since then, SB-CADD has become a commonly used drug discovery technique thanks to advances in genomics and proteomics that have led to the discovery of a large number of candidate drug targets[34]. Extensive use of biophysical techniques such as x-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy has led to the elucidation of a number of 3D structures of human and pathogenic proteins. For example, the PDB has over 81,000 protein structures, while databases like PDBBIND [35] and protein ligand database (PLD) house 5,671 and 129 (as of 2003) ligand-protein co-crystal structures, respectively. Drug discovery campaigns leveraging target structure information have sped up the discovery process and have led to the development of several clinical drugs. A prerequisite for the drug discovery process is the ability to rapidly determine potential binders to the target of biological interest. Computational methods in drug discovery allow rapid screening of a large compound library and determination of potential binders through modeling/simulation and visualization techniques.

Preparation of a Target Structure

A target structure experimentally determined through x-ray crystallography or NMR techniques and deposited in the PDB is the ideal starting point for docking. Structural genomics has accelerated the rate at which target structures are being determined. In the absence of experimentally determined structures, several successful virtual screening campaigns have been reported based on comparative models of target proteins[36]. Efforts have also been made to incorporate information about binding properties of known ligands back into comparative modeling process[37].

Success of virtual screening is dependent upon the amount and quality of structural information known about both the target and the small molecules being docked. The first step is to evaluate the target for the presence of an appropriate binding pocket[38]. This is usually done through the analysis of known target-ligand co-crystal structures or using *in silico* methods to identify novel binding sites [39].

Comparative modeling

Advances in biophysical techniques like X-ray crystallography and NMR techniques have led to increasing availability of protein structures. This has allowed use of structural information to guide drug discovery. In the absence of experimental structures, computational methods are used to predict the 3D structure of target proteins. Comparative modeling is used to predict target structure-based on a template with a similar sequence leveraging that protein structure is better conserved than sequence, i.e. proteins with similar sequences have similar structures. Homology modeling is a specific type of comparative modeling where the template and target proteins share the same evolutionary origin. Comparative modeling involves the following steps: a) Identification of related proteins to serve as template structures, b) sequence alignment of the target and template proteins, c) copying coordinates for confidently aligned regions, d) constructing missing atom coordinates of target structure, and e) model refinement and evaluation. Figure 1-3 illustrates the steps involved in comparative modeling. Several computer programs and web servers exist which automate the comparative modeling process e.g. PSIPRED [40] , MODELLER [41].

Template identification and alignment

In the first step, the target sequence is used as a query for the identification of template structures in the PDB. Templates with high sequence similarity can be determined by a straight-forward PDB-BLAST search [42]. More

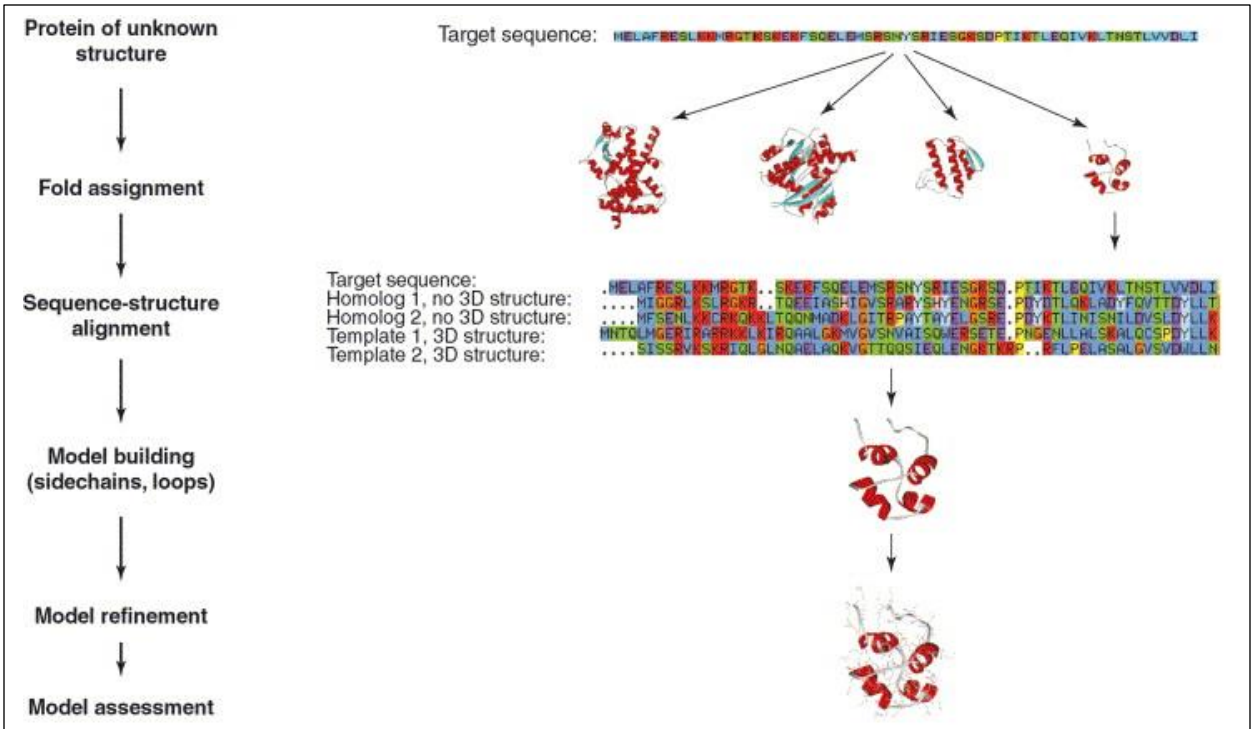


Figure 1-3 Steps in homology model building process. Source: (Hillisch, Pineda et al. 2004).

sophisticated fold recognition methods are available if PDB-BLAST does not yield any hits[43]. Search for template structure is followed by sequence alignment using methods like CLUSTALW [44] which is a multiple sequence alignment tool. For closely related protein structures, structurally conserved regions are identified and used to build the comparative model. Construction and evaluation of multiple comparative models from multiple good-scoring sequence alignments improves the quality of the comparative model[45]. It has been demonstrated that combination of multiple templates can improve comparative models by leveraging well-determined regions that are mutually exclusive[46]. Template selection is key for successful homology modeling. Careful consideration should be given to alignment length, sequence identity, resolution of template structure and consistency of secondary structure between target and templates.

Model building

Gaps or insertions in the original sequence alignment occur most frequently outside secondary structure elements and lead to chain breaks (gaps and insertions) and missing residues (gaps) in the initial target protein model. Modeling these missing regions involves connecting the anchor residues, which are the N- or C- terminal residues of protein segments on either side of the missing region. Two broad classes of loop-modeling methods exist – a) knowledge based methods and b) *de novo* methods. Knowledge based methods use loops from protein structures that have approximately the same anchors as found in target models. Loops from such structures are

applied to the target structure. *De novo* methods generate a large number of loop conformations and use energy functions to judge the quality of predicted loops [47]. Both methods, however, solve the "loop closure" problem – i.e. identifying low energy loop conformations from a large conformational sample space that justify the structural restraint of connecting the two anchor points. Cyclic coordinate descent (CCD) [48] and kinematic closure (KIC) [49] algorithms optimally search for conformations that satisfy constraints for loop closure in a target structure. CCD iteratively changes dihedral angles one at a time such that a distance constraint between anchor residues is satisfied [48]. The KIC algorithm derives from kinematic methods that allow geometric analysis of possible conformations of a system of rigid objects connected by flexible joints. The KIC algorithm generates a Fourier polynomial in N variables for a system of N rotatable bonds by analyzing bond lengths and bond angles constraints [50]. Atom coordinates of the loop are then determined using the polynomial equation.

The loop-modeling step can be affected by two classes of errors: scoring function errors and insufficient sampling. The former arises when nonnative conformations are assigned better scores. Confidence in scoring can be improved by scoring with different functions, assuming that true native conformation will likely be best ranked across multiple scoring methods. Insufficient sampling arises when near native conformations are not sampled. Sufficient sampling can be achieved by running multiple independent simulations to establish convergence.

The next step in comparative modeling is prediction of side chain conformations. A statistical clustering of observed side-chain conformations in PDB, called a rotamer library is used in most side chain construction methods[51]. Methods like dead-end elimination (DEE) [52] implemented in SCRWL [53] and Monte Carlo searches [54] are used for side-chain conformation sampling. DEE imposes conditions to identify rotamers that cannot be members of global minimum energy conformation. For example, the algorithm prunes a rotamer *a*, if a second rotamer *b* exists such that lowest energy conformation containing *a* is greater than highest energy conformations containing *b.* The SCRWL algorithm evaluates steric interactions between side chains by a backbone dependent rotamer library that expresses frequency of rotamers as a function of dihedral angles ϕ and ψ. Monte Carlo algorithms search the side chain conformational space stochastically using the Metropolis criterion to guide the search into energetic minima.

Binding pockets in homology models or even crystal structures are often not amenable for ligand docking due to insufficient accuracy. Ligand information has been used to improve comparative models. Tanrikulu et al used a pseudoreceptor modeling method to improve a homology model of human histamine H_4 receptor. Pseudoreceptor methods map binding pockets around one or more reference ligands by capturing their shape and interactions with the target. Conformation snapshots of the homology model were obtained by MD simulation and pocket-forming coordinates were extracted. Binding pockets of MD frames that matched pseudoreceptor were prioritized for virtual screening. Hits from virtual screening were tested experimentally and two compounds with diverse chemotypes exhibited $pK_i > 4$[55]. Abagyan et al have employed a combined homology modeling and ligand guided backbone ensemble receptor optimization algorithm (LiBERO) for prediction of a protein-ligand complex in CASP experiments.

The approach was identified as the best in that it identified 40% of the 70 contacts that ZMA antagonist makes with adenosine A2a receptor (PDB:3EML). In LiBERO framework multiple models are generated and normal mode analysis (NMA) is used to generate backbone conformation ensembles. Conformers are selected according to docking performance through an iterative process of model building and docking [56]. Ligand information assisted homology modeling is contingent on a) availability of high affinity ligands b) availability of structurally close homologs to ensure good quality initial homology model.

Model refinement and evaluation

Atomic models are refined by introducing ideal bond geometries and by removing unfavorable contacts introduced by the initial modeling process. Refinement involves minimizing models using techniques such as molecular dynamics [57], Monte Carlo Metropolis minimization [58] or genetic algorithms [59]. For example, the ROSETTA refinement protocol fixes bond lengths and angles at ideal values, and removes steric clashes in an initial low-resolution step. ROSETTA then minimizes energy as a function of backbone torsional angles ϕ, ψ and ω using a Monte Carlo minimization strategy [58]. Molecular dynamics-based refinement techniques have been used widely as refinement strategy in drug-design oriented homology model[60].

Model evaluation involves comparison of observed structural features with experimentally determined protein structures. Sali *et al* [61] applied a genetic algorithm that used 21 input model features like sequence alignment scores, measures of protein packing, and geometric descriptors to assess folds of models. Critical Assessment of Techniques for Protein Structure Prediction (CASP) [62] is a worldwide competition in which many groups participate for an objective assessment of methods in the area of protein structure prediction. Models are numerically assessed and ranked by estimating similarity between a model and corresponding experimental structure. Some evaluation methods used in CASP are full model root mean square deviation (RMSD), global distance test-total scores (GDT-TS) and alignment accuracy (AL0 score). GDT-TS is the average maximum number of residues in predicted model that deviate from corresponding residues in the target by no more than a specified distance while AL0 represents the percentage of correctly aligned residues [62].

Binding site detection and characterization

Protein-ligand interaction is a prerequisite for drug activity. Often possible binding sites for small molecules are known from co-crystal structures of the target or a closely related protein with a natural or non-natural ligand. In the absence of a co-crystal structure, mutational studies can pinpoint ligand-binding sites. However, the ability to identify putative high-affinity binding sites on proteins is important if the binding site is unknown or if new binding sites are to be identified, e.g. for allosteric molecules. Computational methods like POCKET, SURFNET, Q-SITEFINDER, etc.[39, 63] are often used for binding site identification. Computational methods for identifying and characterizing binding sites can be divided into three general classes: a) geometric algorithms to find shape concave invaginations

in the target, b) methods based on energetic consideration, c) methods considering dynamics of protein structures and d) by comparison to binding sites in homologous proteins. Detail description about different methods can be found in review article by Kothiwale et al.

Protein-ligand docking

There are three basic methods to represent target and ligand structures *in silico*: atomic, surface, and grid representations[64]. Atomic representation of the surface of the target is usually used when scoring and ranking is based on potential energy functions. An example is DARWIN which uses CHARMM force field to calculate energy [65]. Surface methods represent the topography of molecules using geometric features. The surface is represented as a network of smooth convex, concave, and saddle shape surfaces. These features are generated by mapping part of van der Waals surface of atoms that is accessible to probe a sphere [66]. Docking is then guided by a complementary alignment of ligand and binding site surfaces. Earliest implementation of DOCK [67] used a set of non-overlapping spheres to represent invaginations of target surface and the surface of the ligand (method described earlier in detail for SPHGEN). Geometric matching begins by systematically pairing one ligand sphere a_1 with one receptor sphere b_1. This is followed by pairing a second set of spheres, a_2 and b_2. The move is accepted if the change in atomic distances is less than an empirically determined cut off value. The cut off value specifies the maximum allowed deviation between ligand and receptor internal distance. The pairing step is repeated for a third pair of atoms with the same internal distance checks as above. A minimum of four assignable pairs is essential for determining orientation otherwise the match is rejected. For the grid representation, the target is encoded as physicochemical features of its surface. A grid method described by Katchalski-katzir *et al* [68] digitizes molecules using a 3D discrete function which distinguishes the surface from the interior of the target molecule. Molecules are scanned in relative orientation in three dimensions and the extent of overlap between molecules is determined using a correlation function calculated from a Fourier Transform. Best overlap is determined from a list of overlap functions[68]. Physiochemical properties may be represented on the grid by storing energy potentials on surface grid points.

Sampling Algorithms for Protein-Ligand Docking

Docking methods can be classified as rigid-body docking and flexible docking applications depending on the degree to which they consider ligand and protein flexibility during the docking process[64a 69]. Rigid body docking methods consider only static geometric/physiochemical complementarities between ligand and target, and ignore flexibility and induced-fit [64a] binding models. More advanced algorithms consider several possible conformations of ligand or receptor or both at the same time according to the conformational selection paradigm [70]. Rigid docking simulations are generally preferred when time is critical, i.e. when a large number of compounds are to be docked during an initial vHTS. However, flexible docking methods are still needed for refinement and optimization of poses obtained from an initial rigid docking procedure. With the evolution of computational resources and efficiency,

flexible docking methods are becoming more commonplace. Some of the most popular approaches include systematic enumeration of conformations, molecular dynamic simulations, Monte Carlo search algorithms with Metropolis criterion, and genetic algorithms.

Systematic methods

Systematic algorithms incorporate ligand flexibility through a comprehensive exploration of a molecule's degrees of freedom. In systematic algorithms, the current state of the system determines the next state. Starting from the same exact state and same set of parameters, systematic methods will yield exactly the same final state. Systematic methods can be categorized into a) exhaustive search algorithms and b) fragmentation algorithms.

Exhaustive searches elucidate ligand conformations by systematically rotating all possible rotatable bonds at a given interval. Large conformational space often prohibits an exhaustive systematic search. Algorithms such as GLIDE [71] use heuristics to focus on regions of conformational space that are likely to contain good scoring ligand poses. GLIDE pre-computes a grid representation of target's shape and properties. Next, an initial set of low energy ligand conformations in ligand torsion-angle space is created. Initial favorable ligand poses are identified by approximate positioning and scoring methods (shape and geometric complementarities). This initial screen reduces the conformational space over which the high resolution-docking search is applied. High-resolution search involves the minimization of the ligand using standard molecular mechanics energy function, followed by a Monte Carlo procedure for examining nearby torsional minima.

Fragmentation methods sample ligand conformation by incremental construction of ligand conformations from fragments obtained by dividing the ligand of interest. Ligand conformations are obtained by docking fragments in the binding site one at a time and incrementally growing them, or by docking all fragments into the binding site and linking them covalently. Des Jarlais *et al* modified the DOCK algorithm to allow for ligand flexibility by separately docking fragments into the binding site and subsequently joining them [72]. FLEXX [73] uses the "anchor and grow method" for ligand conformational sampling. A base fragment has to be interactively selected by the user that is followed by automatic determination of placements for the fragment that maximize favorable interactions with the target protein. The base fragment is grown incrementally by adding new fragments in all possible conformations and the extended fragment is selected if no significant steric clashes (overlap volume $\leq$ 4.5 Å^3) are observed between ligand and target atoms. Extended ligands are optimized a) if new interactions are found b) if minor steric interactions exist [73]. Fully automated "anchor and grow" methods have been implemented in several methods like FLOG [74], SURFLEX [75] and SEED [76]. In a benchmark study where performance of eight docking algorithm was compared on 100 protein-ligand complex, GLIDE and SURFLEX were among the methods that showed best accuracy [77]. GLIDE and SURFLEX generated poses close to X-ray conformation for 68 protein-ligand complexes in the Directory of Useful Decoys[78].

Example application in CADD. Human Pim-1 kinase, responsible for cell survival/apoptosis, differentiation and proliferation, is a valuable anticancer target as it is over expressed in a variety of leukemia. Pierce *et al* [79] used GLIDE to dock around 700,000 commercially available compounds and identified four compounds with K_i values less than 5 µM. Chiu *et al* [80] used SURFLEX to identify novel inhibitors of anthrax toxin lethal factor, responsible for anthrax-related cytotoxicity. Docking study of a compound library derived from seven databases including DrugBank [81], ZINC [82], National Cancer Institute (NCI) database [83] etc. identified lead compounds which eventually led to the development of nanomolar inhibitors upon optimization.

Molecular dynamics simulations

Molecular dynamics (MD) simulation calculates the trajectory of a system by the application of Newtonian mechanics. However, standard MD methods depend heavily on the starting conformation and are not readily appropriate for simulation of ligand-target interactions. Due to its nature, MD is not able to cross high-energy barriers within the simulation's lifetime and is not efficient for traversing the rugged hyper surface of protein-ligand interactions. Strategies like simulated annealing have been applied for more efficient use of MD in docking. Di Nola *et al* have described a MD protocol for docking small flexible ligands to flexible targets in water[84]. They separated the center of mass movement of ligand from its internal and rotational motions. The center of mass motion and internal motions were coupled to different temperature baths, allowing independent control to the different motions. Appropriate values of temperature and coupling constants allowed flexible or rigid ligand and/or receptor.

The McCammon group developed a "relaxed-complex" approach which explores binding conformations that may occur only rarely in the unbound target protein. A two ns MD simulation of ligand free target is carried out to extensively sample its conformations. Docking of ligands is then performed in target conformation snapshots taken at different time points of the MD run. This relaxed complex method was used to discover novel modes of inhibition for HIV integrase and led to the discovery of the first clinically approved HIV integrase inhibitor, Raltegravir. This MD method has also been used in several other campaigns to identify inhibitors of target of interest[85].

Metadynamics is a MD-based technique for predicting and scoring ligand binding. The method maps the entire free energy landscape in an accelerated way as it keeps track of history of already sampled regions. During the MD simulation of a protein-ligand complex, a Gaussian repulsive potential are added on explored regions, steering the simulation towards new-free energy regions.[26, 86]

Millisecond timescale MD simulations are now possible with special purpose machines like Anton[87]. Such long simulations have allowed study of drug binding events to their protein target [88]. Anton has been used successfully for full atomic resolution protein folding [89]. Advances in computer hardware capabilities means protein flexibility can be accessed more routinely on longer timescales. This would allow better descriptions of conformational flexibility in future.

Monte Carlo search with metropolis (MCM) criterion

Stochastic algorithms make random changes to either ligand being docked or to its target binding site. These random changes could be translational or rotational in the case of ligand or random conformational sampling of residue side-chains in the target binding site. Whether a step is accepted or rejected in such a stochastic search is decided based on the Metropolis criterion that generally accepts steps that lower the overall energy and occasionally accepts steps that increase energy to enable departure from a local energy minimum. The probability of acceptance of an uphill step decreases with increasing energy gap and depends on the 'temperature' of the MCM simulation[90]. MCM simulations have been adopted for flexible docking applications such as in MCDOCK [91], ICM [92], and ROSETTALIGAND [27, 93]. MCM samples conformational space faster than molecular dynamics in that it requires only energy function evaluation and not the derivative of the energy functions. While traditional MD drives a system towards a local energy minimum, the randomness introduced with Monte Carlo allows hopping over the energy barriers, preventing the system from getting stuck in local energy minima. A disadvantage is that any information about the timescale of the motions is lost.

ROSETTALIGAND [94] uses a knowledge-based scoring procedure with a Monte Carlo-based energy minimization scheme that reduces the number of conformations that must be sampled while providing a more rapid scoring system than offered through molecular mechanics force fields. ROSETTALIGAND incorporates side-chain and ligand flexibility during a high-resolution refinement step through a Monte-Carlo based sampling of torsional angles. All torsion angles of protein and ligand are optimized through gradient-based minimization mimicking an induced fit scenario [93]. MCDOCK uses two stages of docking and a final energy minimization step for generating target-ligand structure. In the first docking stage, the ligand and docking site are held rigid while the ligand is placed randomly into the binding site. Scoring is done completely based on short-contacts. This allows identification of non-clashing binding poses. In the next stage, energy based Metropolis sampling is done to sample the binding pocket[91]. QXP [95] optimizes grid map energy and internal ligand energy for searching ligand-target structure. The algorithm performs a rigid body alignment of ligand-target complex followed by MCM translation and rotation of ligand. This step is followed by another rigid body alignment and scoring using energy grid map. ICM [96] (Internal Coordinate Mechanics) relies on a stochastic algorithm for global optimization of entire flexible ligand in receptor potential grid. The relative positions of ligand and target molecule make up the internal variables of the method. Internal variables are subject to random change followed by local energy minimization and selection by Metropolis criterion. ICM performed satisfactorily in generating protein-ligand complexes for 68 diverse, high-resolution X-ray complexes found in DUD[78].

Example application in CADD. ROSETTALIGAND was used by Kaufmann *et al* [97] to predict the binding mode of serotonin with serotonin transporters. The binding site predicted to be deep within the binding pocket was consistent with mutagenesis studies. QXP has been used to optimize inhibitors of Human β-Secretase (BACE1) [98] which is an important therapeutic target for treating Alzheimer's disease by diminishing β-amyloid deposit

formation. ICM was used successfully to identify inhibitors for a number of targets including Tumor necrosis factor α[99], dysregulation of which is implicated in tumorigenesis and autoinflammatory diseases like rheumatoid arthritis and psoriatic arthritis. Computational screening of 230,000 compounds from the NCI database against neuraminidase using ICM identified 4-(4-((3-(2-amino-4-hydroxy-6-methyl-5-pyrimidinyl) propyl) amino) phenyl)-1-chloro-3-buten-2-one which inhibited influenza virus replication at a level comparable to known neuraminidase inhibitor oseltamivir[100].

Genetic Algorithms

Genetic algorithms (GAs) introduce molecular flexibility through recombination of parent conformations to child conformations. In this simulated evolutionary process, the "fittest" or best scoring conformations are kept for another round of recombination. In this way, the best possible set of solutions evolves by retaining favorable features from one generation to the next. In docking, a set of values that describe the ligand pose in the protein are state variable, i.e. the genotype. State variables may include set of values describing translation, orientation, conformation, number of hydrogen bonds, etc. The state corresponds to the genotype; the resulting structural model of the ligand in the protein corresponds to the phenotype, and binding energy corresponds to the fitness of the individual. Genetic operators may swap large regions of parent's genes, or randomly change (mutate) the value of certain ligand states, to give rise to new individuals.

Genetic Optimization for Ligand Docking (GOLD) [101] explores full ligand flexibility with partial target flexibility using a genetic algorithm. The GOLD algorithm optimizes rotatable dihedrals and ligand-target hydrogen bonds. The fitness of a generation is evaluated based on a maximization of inter-molecular hydrogen bonds. The fitness function is the sum of a hydrogen bonding term, a term for steric energy interaction between the protein and the ligand and a Lennard-Jones potential for internal energy of ligand. AutoDock [102] uses the Lamarckian genetic algorithm (LGA) which allows favorable phenotypic characteristics to become inheritable. GOLD has demonstrated better accuracy than most docking algorithms, except GLIDE, in various benchmark studies.[77, 103]

Example application in CADD. Inhibition of α-glucosidase has shown to retard glucose absorption and decrease postprandial blood glucose level which makes it an attractive target for curing diabetes and obesity. Park *et al* [104] used AUTODOCK to identify four novel inhibitors of α-glucosidase by screening a library of 85000 compounds obtained from INTERBIOSCREEN chemical database (http://www.ibscreen.com) . AUTODOCK was also used to identify inhibitors of RNA Editing Ligase-1 enzyme of *Trypanosoma brucei*, causative agent of Human African trypanosomiasis[105].

Incorporating target flexibility in docking

Conformational variability is seen in unbound form and different apo structures[106]. It is widely believed that the ligand-bound state is selected from an ensemble of protein conformations by the ligand [107]. Accounting for receptor flexibility in the form of protein side-chain and backbone movement is essential for predicting correct binding pose. An ensemble of non-redundant low energy target structures will cover a large conformational space as against a single conformation resulting in more realistic target-ligand bound states. Methods for inducing receptor flexibility include induced-fit docking and ensemble generated from MD simulation snapshots. Induced-fit algorithms allow small overlap between the ligand and the target along with side-chain movements resulting in elasticity. GLIDE uses an induced fit model where all side chain residues are changed to alanine before initial docking. Side-chain sampling is followed by energy minimization of the binding site and ligand. ROSETTALIGAND allows for full protein backbone and side-chain flexibility in the active site. Multiple fix receptor conformations are used in docking protocols, known as ensemble-based screening, to incorporate receptor flexibility [108]. Receptor conformations may be experimentally determined by either crystallography or NMR, or computationally generated from MD simulations, normal mode analysis (NMA) and MC sampling[109]. McCammon et al. used the relaxed complex scheme (RCS) to describe a novel trench in HIV integrase which led to the discovery of the integrase inhibitor raltegravir [110]. In RCS, multiple conformations are determined from MD simulations to perform docking studies. Other sampling methods include umbrella-sampling, metadynamics, accelerated MD etc. [106b].

Scoring Functions for Evaluation Protein-Ligand Complexes

Docking applications need to rapidly and accurately assess protein-ligand complexes, i.e. approximate the energy of the interaction. A ligand docking experiment may generate hundreds of thousands of target-ligand complex conformations, and an efficient scoring function is necessary to rank these complexes and differentiate valid binding mode predictions from invalid predictions. More complex scoring functions attempt to predict target-ligand binding affinities for hit-to-lead and lead-to-drug optimization. Scoring functions can be grouped into four types: a) force-field or molecular mechanics based scoring functions b) empirical scoring functions c) knowledge-based scoring functions d) consensus scoring functions.

Force-field or molecular mechanics based scoring functions

Force-field scoring functions use classical molecular mechanics for energy calculations. These functions use parameters derived from experimental data and *ab initio* quantum mechanical calculations. The parameters for various force terms including pre-factor variables are obtained by fitting to high quality ab initio data on intermolecular interactions [111]. The binding free energy of protein-ligand complexes are estimated by the sum of van der Waals and electrostatic interactions. DOCK uses the AMBER force fields in which van der Waals energy terms are represented by the Lennard-Jones potential function while electrostatic terms are accounted for by coulomb interaction with a distance-dependent dielectric function. Standard force fields are however biased to select highly

charged ligands. This can be corrected by handling ligand solvation during calculations[112] [113]. Terms from empirical scoring functions (discussed below) are often added to force field functions to treat solvation and electronic polarizability. A semi-empirical force field has been implemented in AUTODOCK to evaluate the contribution of water surrounding the receptor-ligand complex in the form of empirical enthalpic and entropic terms, for example[114].

Empirical Scoring Functions

Empirical scoring functions fit parameters to experimental data. An example is binding energy which is expressed as a weighted sum of explicit hydrogen bond interactions, hydrophobic contact terms, desolvation effects, and entropy. Empirical function terms are simple to evaluate and are based on approximations. The weights for different parameters are obtained from regression analysis using experimental data obtained from molecular data. Empirical functions have been used in several commercially available docking suits like LUDI [115] , FLEXX [73] and SURFLEX [75].

Knowledge-Based Scoring Function

Knowledge-Based scoring functions employ the information contained in experimentally determined complex structures. They are formulated under the assumption that inter-atomic distances occurring more often than average distances represent favorable contacts. On the other hand, interactions that are found to occur with lower frequencies are likely to decrease affinity. Several knowledge based potentials have been developed to predict binding affinity like potential of mean force (PMF) [116], DRUGSCORE [117], SMOG [118] and BLEEP [119].

Consensus-Scoring Functions

More recently, consensus-scoring functions have been demonstrated to achieve improved accuracies through a combination of advantages of basic scoring functions. Consensus approaches rescore predicted poses several times using different scoring functions. These results can then be combined in different ways to rank solutions [120]. Some strategies for combining scores include a) weighted combinations of scoring functions b) a voting strategy in which cutoffs established for each scoring method is followed by decision based on number of passes a molecule has c) a rank by number strategy ranks each compound by its average normalized score values d) a rank by rank method sorts compounds based on average rank determined by individual scoring functions. Boyle *et al* [121] evaluated consensus scoring strategies to investigate the parameters for the success of properly combined rescoring strategies. It turns out that combining scoring functions that have complementary strengths leads to better results over those that have consensus in their predictions. For example, scoring functions whose strengths are distinguishing actives from inactive compounds are complemented by scoring functions that can distinguish correct from incorrect binding poses. A disadvantage of consensus scoring methods could be a possible loss of active compound if poorly scored by one of the scoring functions.

Example application in CADD. Okamoto *et al* [122] have used consensus scoring technique for identifying inhibitors of death-associated protein kinases which are targets for ischemic diseases in the brain, kidney, and other organs. The consensus scoring function used in the study was implemented in DOCK4.0 program and included three scoring functions a) empirical scoring function (implemented in FLEXX) b) a knowledge-based PMF scoring function [123] c) a force-field function from DOCK4.0. Around 400,000 compounds from a corporate compound library were docked followed by simultaneous scoring with the three functions. The consensus score was defined as the score that was highest among the three. In another successful application of consensus scoring scheme, Friedman *et al* [124] discovered plasmepsin inhibitors for use as antimalarial agents using a scoring based on median ranking of four field-based scoring functions.

Pharmacophore Model

A pharmacophore model of the target-binding site summarizes steric and electronic features needed for optimal interaction of a ligand with a target. Most common properties that are used to define pharmacophores are hydrogen bond acceptors, hydrogen bond donors, basic groups, acidic groups, partial charge, aliphatic hydrophobic moieties, and aromatic hydrophobic moieties. Pharmacophore features have been used extensively in drug discovery for virtual screening, *de novo* design, and lead optimization [125]. A pharmacophore model of the target-binding site can be used to virtually screen a compound library for putative hits. Apart from querying database for active compounds, pharmacophore models can also be used by de novo design algorithms to guide the design of new compounds.

Structure-based pharmacophore methods are developed based on an analysis of the target-binding site or based on a target-ligand complex structure. LigandScout [126] uses protein-ligand complex data to map interactions between ligand and target. A knowledge based rule-set obtained from the PDB is used to automatically detect and classify interactions into hydrogen bond interactions, charge transfers, and lipophilic regions [126]. The Pocket v.2 [127] algorithm is capable of automatically developing a pharmacophore model from a target-ligand complex. The algorithm creates regularly spaced grids around the ligand and the surrounding residues. Probe atoms which represent a hydrogen bond donor, a hydrogen bond acceptor and a hydrophobic group, are used to scan the grids. An empirical scoring function, SCORE, is used to describe the binding constant between probe atoms and the target. SCORE includes terms to account for van der Waals interactions, metal-ligand bonding, hydrogen bonding and desolvation effects upon binding [128]. A pharmacophore model is developed by rescoring the grids followed by

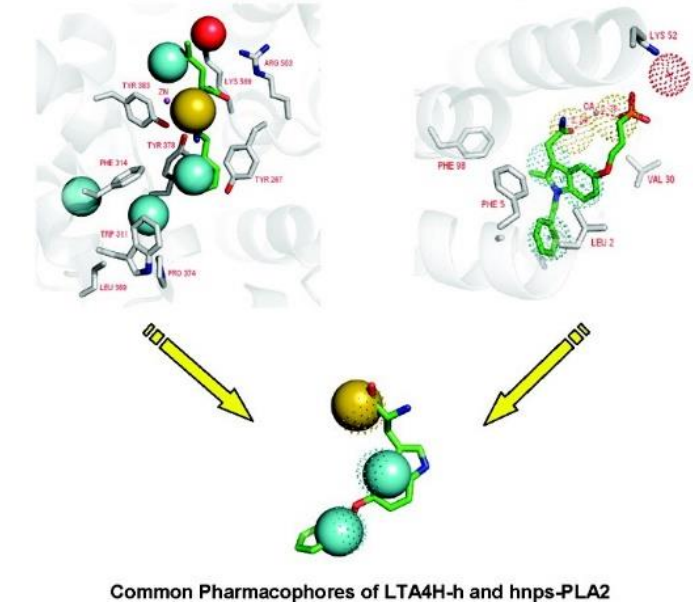


Figure 1-4 Extracting common pharmacophores of LTA4H-h and human-PLA2. Cyan spheres depict hydrophobic centers, red spheres represent H-bond acceptor while yellow spheres stand for feature that coordinates with a metal. Source: Wei et al [1]

clustering and sorting to extract features essential for protein-ligand interaction. During rescoring, hydrogen bond donor/acceptor scores lower than 0.2 and hydrophobic scores lower than 0.47 are reset to zero. Grids with three zero scores are filtered out and the "neighbor number" for each grid is determined by counting the number of grids within 2 Å having non-zero score for a particular type. Grids with less than 50 donor neighbors, 30 acceptor neighbors and 40 hydrophobic neighbors are reset to zero for their donor score, acceptor score and hydrophobic scores respectively. Grids are filtered by eliminating those with three zero scores leaving only those grids that represent key interaction sites. The algorithm then superimposes the ligand on the grid and a given grid is selected as a candidate if it is close to an atom type that can mediate the same interaction. Candidates with non-zero donor, acceptor, or hydrophobic scores are gathered into separate clusters and the grid with highest score is defined as the center of donor, acceptor or hydrophobic property.

Virtual screening using a pharmacophore model

17β-hydroxysteroid dehydrogenase type 1 (17β-HSD1) plays an important role in the synthesis of the most potent estrogen estradiol. Its inhibition could be important for breast cancer prevention and treatment. Schuster *et al* [129] used LigandScout2.0 to generate pharmacophore models of 17β-HSD1 from co-crystallization complexes with inhibitors (PDB code 1EQU and 1I5R). These pharmacophore models represent the binding mode of a steroidal compound and a small hybrid compounds (consisting of a steroidal part and an adenosine) respectively. The 1I5R-

based pharmacophore model was used to screen the NCI and SPECS databases for new inhibitors using CATALYST. Best scoring hit compounds were docked into the binding pocket of 1EQU using GOLD, and final selection for *in vitro* testing was performed according to the best-fit value, visual inspection of predicted docking pose and the ChemScore (GOLD scoring function) value. Four out of 14 compounds tested *in vitro* showed an IC_{50} value of less than 50 μM with the most potent being 5.7 μM. Brvar *et al* [130] applied pharmacophore models to discover novel inhibitors of bacterial DNA gyrase B, a bacterial type II topoisomerase originating from gyrase and a target for antibacterial drugs. A pharmacophore model obtained using LigandScout was used to screen the ZINC database that yielded a novel class of thiazole-based inhibitors with IC_{50} value of 25 μM.

Multi-target inhibitors using common pharmacophore models

Wei *et al* [1] used Pocket v.2 to identify a common pharmacophore for two targets involved in inflammatory signaling, human leukotriene A4 hydrolase (LTA4H-h) and human nonpancreatic secretory phospholipase A2 (PLA2). The co crystal structure (PDB code 1HS6) of LTA4H-h with 2-(3-amino-2-hydroxy- 4-phenylbutyrylamino)-4-methyl-pentanoic acid (bestatin) and the structure (PDB code 1DB4) of PLA2 with [3-(1-benzyl-3-carbamoylmethyl-2-methyl-1H-indol-5-yloxy) propyl] phosphonic acid (indole 8) were used to derive pharmacophores of the two targets. For LTA4H-h, six pharmacophore centers were identified which included four hydrophobic, one hydrogen bond acceptor and one zinc metal coordination pharmacophore. In the binding pocket of PLA2 three hydrophobic centers, one hydrogen bond acceptor and two calcium ion coordination centers were identified. The comparison of two sets of pharmacophore models revealed that two hydrophobic pharmacophores and a pharmacophore that coordinated with metal, shown in Figure 1-5, was common to both proteins. The authors hypothesized that compounds that satisfy the common pharmacophores would inhibit both the proteins. The MDL chemical database was screened virtually with LTA4H-h and PLA2 using Dock4.0 and binding conformation of top 150,000 compounds (60% of database) ranked by Dock score were extracted and checked for conformity to common pharmacophores. This identified 163 compounds whose binding conformations were re-analyzed using Autodock3.5 followed by comparison with common pharmacophores. Finally, nine compounds whose conformations matched the common pharmacophores were tested *in vitro* for binding with PLA2 and LTA4H-h. The best inhibitor, compound 10, shown in Figure 1-4, inhibited LTA4H-h at submicromolar range while PLA2 with an IC_{50} value of 7.3 μM.

Dynamic pharmacophore models that account for protein flexibility

The over expression of murine double minute 2 oncoprotein (MDM2) which inhibits p53 tumor suppressor is responsible for approximately half of all human cancers. Reactivation of p53-MDM2 integration has been shown to be a novel approach for enhancing cancer cell death [131]. Bowman *et al* [131] extracted snapshots at every 100 ps from a 2ns MD simulation of MDM3 bound to p53. The resulting 21 structures for MDM2 were used to generate a 6-site pharmacophore model of the active site that included three aromatic/hydrophobic sites and three hydrogen-bond

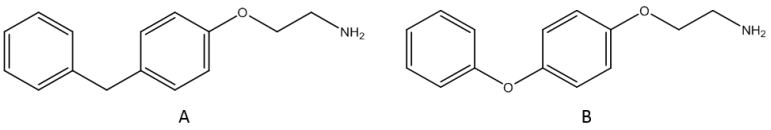


Figure 1-5 A) A reported inhibitor of LTA4H-h B) Compound 11

donor sites. A virtual screening of a library of 35,000 compounds identified 27 hits, 23 of which were tested in a competitive binding assay. Four of the tested compounds were identified as true hits with the best inhibitor having a K_i value of 110 +/- 30 nM. The dynamic pharmacophore model was also used successfully to identify low μM inhibitors of HIV-1 integrase[132].

Automated de novo Design of Ligands

De novo structure-based ligand design can be accomplished by either a ligand growing or ligand linking approach. With the ligand growing approach, a fragment is docked into the binding site and the ligand is extended by adding functional groups added to the fragment. The linking method is similar docks multiple small fragments into adjacent binding pockets of the target. Subsequently, the fragments are linked to each other to form a single compound. This approach is a computational version of the popular SAR by NMR technique introduced by Hadijuk *et al* [133].

Several methods have been developed which implement both ligand growing and ligand linking strategies for designing ligands that can bind to a given target. LigBuilder [134] builds ligands in a step by step fashion using a library of fragments. The design process can be carried out by various operations like ligand growing and linking; and the construction process is controlled by a genetic algorithm. The target-ligand complex binding affinity is evaluated by using an empirical scoring function. The program first reads the target protein and analyzes the binding pocket. Depending on the choice of the user, it can then use either a growing or a linking strategy. In the growing strategy, a seed structure is placed in binding pocket and then the program replaces user defined growing sites with candidate fragments. This gives rise to a new seed structure that can then be used in further rounds of growing. For the linking strategy, several fragments placed at different locations on the target protein serve as seed structure. The growing scheme happens simultaneously on each fragment. In the process, the program seeks to link these fragments. The LUDI [115] algorithm, which precedes LigBuilder, uses primarily a linking strategy for ligand design. It positions seed fragments into binding pockets of the target structure optimizing their interactions individually. This step is followed by linking the fragments into a single molecule. The synthetic accessibility of ligands can be taken into account. For example, LigBuilder 2.0 analyzes designed using a chemical reaction database and a retro-synthesis analyzer [135].

The biggest challenge of *de novo* drug design is inseparable from its greatest advantage. By defining compounds that have never been seen before, one is invariably necessitating synthetic effort for acquisition prior to testing. This forces any *de novo* protocol to incorporate synthesizability metrics into its scoring. This increases the effort required in terms of cost, yield, time, and expertise necessary. Synthesizability thus becomes increasingly important when designing a large number of different compounds and scaffolds. Tools have been designed and employed to approach synthesizability constraints. SYNOPSIS (SYNthesize and OPtimize System in Silico) [136] is a commonly used tool that enforces synthesizability throughout the design process by starting with available compounds and creating novel compounds by virtually employing known chemical reactions. This tool contains a set of 70 reaction types that are selected based on the presence of different functional groups in the evolving molecule. SYNOPSIS also provides additional restraints for desired properties such as solubility. Krier *et al* proposed an approach called the Scaffold-Linker-Functional Group (SLF) approach that has been implemented in *de novo* strategies[137]. This method is designed to create a *de novo* scaffold-focused library that maximizes diversity and minimizes size. A limited number of non-overlapping functional groups were selected that are added or removed from the static scaffold core. The linker plays the role of varying the distances between the scaffold and functional groups. RECAP (Retrosynthetic Combinatorial Analysis Procedure) was the first fragment generation method to incorporate rules that limit the

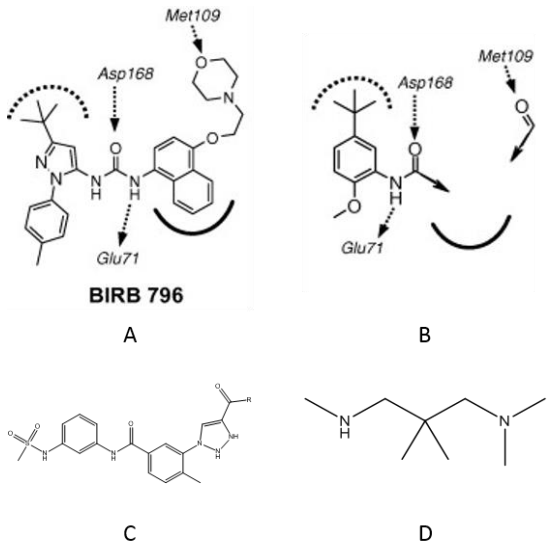


Figure 1-6 Design strategy for inhibitors of p38 MAP kinase. A) Key interactions of BIRB-796 inhibitor with MAP kinase. B) A fragment linking strategy to link two seed structures was applied using LigBuilder. A tert-butyl phenyl fragment was used in the first pocket while a carbonyul fragment was used to access the hydrogen bond with Met109 in the second site. An N-formyl group was attached to the first seed fragment to access hydrogen bonds with Glu 71 and Asp168. C) General structure of optimized structures which showed potent activity. D) R group for compound 28 which showed IC_{50} value of 83 nM. Source: Cogan et al[3]

chemical reactions to ones used in typical combinatorial chemistry techniques, thereby limiting the possible fragments as well as possible recombination patterns[138].

Example application in CADD. *De novo* design by linking fragments has been successfully applied in the design of inhibitors of p38 mitogen-activated protein kinase (MAPK) [3], which is a key regulator in signaling pathways that control the production of cytokines such as TNF-α and IL-1β. Inhibitors of MAPK can potentially be used for the treatment of various autoimmune diseases. The Figure 1-6A shows four classes of interactions of a clinical compound BIRB 796 with MAPK a) interaction with residues in ATP binding site (Met109) b) interaction with the "Phe pocket" (dotted arc) c) hydrophobic interaction with the kinase specificity pocket (solid arc) d) interaction of the urea with backbone NH of Asp168 and carboxylate of Glu71. A design strategy for exploring structurally distinct scaffolds by leveraging the interactions of BIRB 796 was devised as follows: a) A tert-butyl group was used as "Phe pocket" seed structure in place of pyrazole ring of BIRB 796 b) An N-formyl group was appended to tert-butyl fragment to access the hydrogen bonds with Glu71 and Asp168 c) A carbonyl group was used as the second seed fragment to access the hydrogen bond with Met109 as shown in Figure 1-6B. LigandBuilder software was used to link the two seed fragments, the tert-butyl linked to N-formyl group, and the carbonyl group. The program consistently introduced a

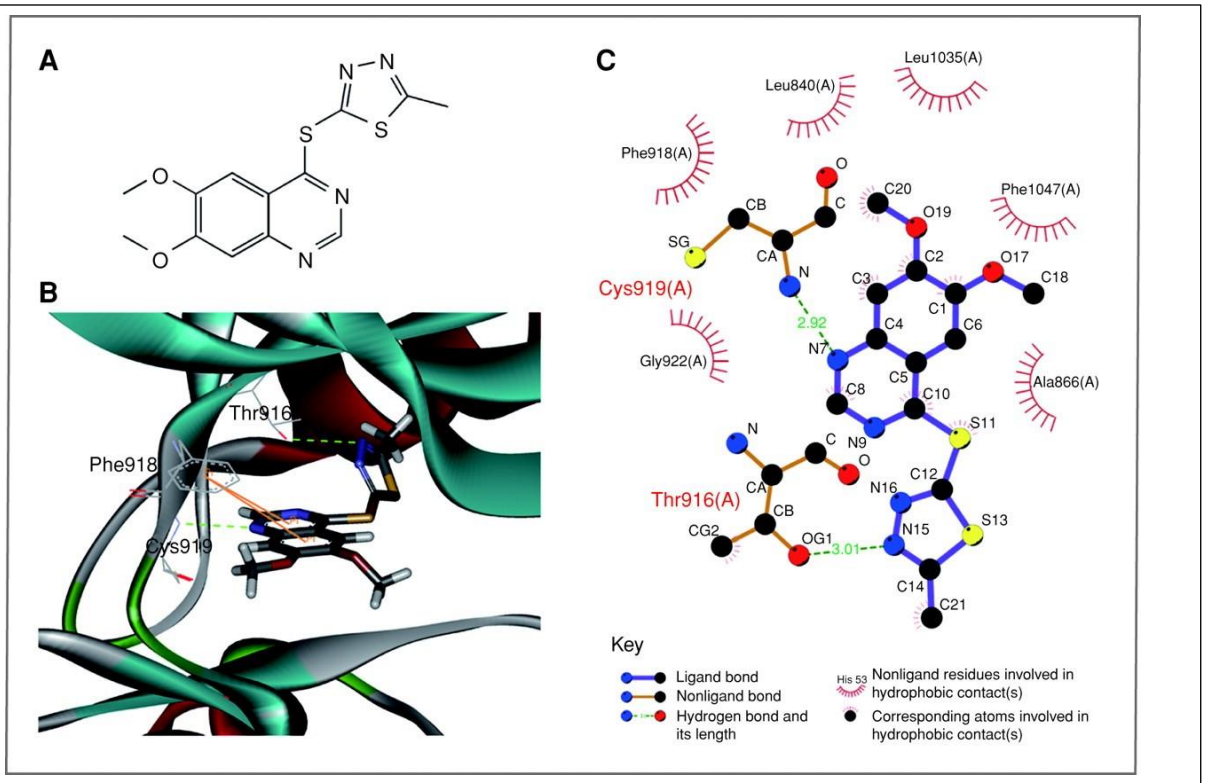


Figure 1-7 A) chemical structure of SKLB1002. B, SKLB1002 is docked into the active site of VEGFR2, showing interactions between SKLB1002 and VEGFR2 by using the in silico model. C, a 2-dimensional interaction map of SKLB1002 and VEGFR2. Source: Zhang et al[2]

4-tolyl group in the kinase specificity pocket. However, LigandBuilder failed to predict favorable rigid linkers for connecting tolyl group to carbonyl group that would be essential for carbonyl display at the proper distance to interact with Met109. Modeling indicated N-linked azoles connected to tolyl group via an N-linkage as a suitable linker. Derivatives of this designed molecule were synthesized leading to the discovery of compound 28 shown in Figure 1-6D, which exhibited IC_{50} value of 83 nM.

Zhang et al have employed the fragment extension approach [2] in the discovery of inhibitors of VEGF Receptor 2 (VEGFR2), a therapeutic target for tumor-induced angiogenesis. The authors used quinazoline as the seed fragment as three of the nine clinically approved kinase inhibitor drugs are 4-anilinoquinazoline derivatives[139]. These inhibitors bind the active site of their respective targets such that the quinazoline ring is located at the front of ATP binding pocket. The ligand building process involved placing the quinazoline fragment in the binding pocket in the same orientation as found for known inhibitors. The design strategy sought to create ligand that would extend to fit a specific hydrophobic pocket at the back of the ATP binding cleft. An NH_2, OH, or SH group was added in the C4 position of the quinazoline ring to allow for a turn owing to orientation of quinazoline and the spatial arrangement of the hydrophobic pocket. A fragment-growth-based de novo method was applied in which various fragments (about 1200 fragments) were allowed to grow on the turn fragment to extend into the hydrophobic pocket. Designed molecules were then re-scored and ranked using GOLD. The design process led to the development of a potent and specific VEGFR2 inhibitor, SKLB1002 shown in Figure 1-7. The inhibitor was successful in inhibiting angiogenic processes in zebra fish embryo and athymic mice with human tumor xenografts.

Ligand-Based Computer-Aided Drug Design (LB-CADD)

The ligand-based computer-aided drug discovery (LB-CADD) approach involves the analysis of ligands known to interact with a target of interest. These methods utilize a set of reference structures collected from compounds known to interact with the target of interest and analyze their 2D or 3D structures. The overall goal is to represent these compounds in such a way that the physicochemical properties most important for their desired interactions are retained while extraneous information not relevant to the interactions is discarded. It is considered an indirect approach to drug discovery in that it does not necessitate knowledge of the structure of the target of interest. The two fundamental approaches of LB-CADD are a) selection of compounds based on chemical similarity to known actives using some similarity measure or b) the construction of a QSAR model that predicts biological activity from chemical structure. The difference between the two approaches is that the latter weighs features of the chemical structure according to their influence on the biological activity of interest, while the former does not. The methods are applied for *in silico* screening for novel compounds possessing the biological activity of interest, hit-to-lead and lead-to drug optimization, and for the optimization of DMPK/ADMET properties. LB-CADD is based on the Similar Property Principle, published by Johnson and Maggiora, which states that molecules that are structurally similar are likely to have similar properties [140]. LB-CADD approaches in contrast to SB-CADD approaches can also be applied

when the structure of the biological target is unknown. Additionally, active compounds identified by Ligand-Based virtual High-Throughput Screening (LB-vHTS) methods are often more potent than those identified in (SB-vHTS) [28].

Molecular Descriptors / Features

LB-CADD techniques utilize different methods for describing features of small molecules using computational algorithms that balance efficiency and information content. The optimal descriptor set depends on the biological function predicted as well as on the LB-CADD technique employed, and therefore many different algorithms for deriving chemical information have been developed and employed. Molecular descriptors can be structural as well as physicochemical, and can be described on multiple levels of increasing complexity. Information described can include properties such as molecular weight, geometry, volume, surface areas, ring content, rotatable bonds, inter-atomic distances, bond distances, atom types, planar and non-planar systems, molecular walk counts, electronegativities, polarizabilities, symmetry, atom distribution, topological charge indices, functional group composition, aromaticity indices, solvation properties, and many others [141]. These descriptors are generated through knowledge-based, graph-theoretical methods, molecular-mechanical, or quantum-mechanical tools [142] and are classified according to the "dimensionality" of the chemical representation from which they are computed [143] – 1D=scalar physicochemical properties such as molecular weight; 2D=molecular constitution-derived descriptors, 2.5D=molecular configuration-derived descriptors; 3D=molecular conformation-derived descriptors. These different levels of complexity, however, are overlapping with the more complex descriptors often incorporating information from the simpler ones. For example, many 2D and 3D descriptors use physicochemical properties to weight their functions and to describe the overall distribution of these properties.

Functional groups

Functional groups are defined by the IUPAC as atoms or groups of atoms that have similar chemical properties across different compounds. These groups are attached to a central backbone of the molecule, also called scaffold or chemotype. The spatial positioning of the functional groups achieved by the backbone defines the physical and chemical properties of compounds. Therefore, the location and nature of functional groups for a given compound contain key information for most ligand-based CADD methods. There are many different kinds of functional groups including those that contain hydrocarbons, halogens, oxygens, nitrogens, sulfur, phosphorous, etc. Functional groups include alcohols, esters, amides, carboxylates, ethers, nitro group, thiols, etc. [144]

Functional groups can either be explicitly described by their atomic composition and bonding or may be implicitly encoded by their general properties. For example, under physiological conditions carboxyl groups are often negatively charged while amine groups are positively charged. This property is accurately reflected in the structure of the functional group, but also in the charge computed from that structure. Since it is the properties conferred by the functional groups that are most important to the biochemical activity of a given compound, many CADD

applications treat functional groups containing different atoms but conferring the same properties as similar or even identical. For example, the capacity for hydrogen bonding can heavily influence a molecule's properties. These interactions frequently occur between a hydrogen atom and an electron donor such as oxygen or nitrogen. Hydrogen bonding interactions influence the electron distribution of neighboring atoms and the site's reactivity, making it an important functional property for therapeutic design. Commonly, hydrogen bonding groups are separated as hydrogen bond donors with strong electron-withdrawing substituents (OH, NH, SH, and CH) and hydrogen bond acceptor groups (PO, SO, CO, N, O, and S) [145]. The applications Phase, Catalyst, DISCO, and GASP as well as Pharmacophore mapping algorithms discussed in greater detail below focus primarily on hydrogen-bond donors, hydrogen-bond acceptors, hydrophobic regions, ionizable groups, and aromatic rings.

Prediction of physio-chemical properties

Descriptors within the same dimensionality can show a range of complexity. The simplest ones such as molecular weight and number of hydrogen bond donors are relatively simple to compute. These can be rapidly and accurately computed. More complex descriptors such as solubility and partial charge are more difficult to compute. However, the higher information content provided by these descriptors makes them extremely useful for model development. [146]. Therefore, prediction of physio-chemical properties is a critical step in developing effective molecular descriptors. The trade-off in computing such descriptors is between the high speed needed to encode thousands of molecules and sufficient accuracy.

Converting properties into descriptors

Molecule properties are converted into numerical vectors of descriptors for analysis. This conversion is needed to ensure that descriptions of molecules have a constant length independent of size. Each position in the vector of descriptors encodes a well-defined property or feature that facilitates comparison by mathematical algorithms.

Binary molecular fingerprints

Fingerprints are bit string representations of molecular structure and/or properties[147]. They encode various molecular descriptors as pre-defined bit settings [148] i.e. representation as 1 or 0, where 1 means descriptor is present or 0 if not. This allows chemical identity to be unambiguously assigned by the presence or absence of features [149]. The features described in a molecular fingerprint can vary in number and complexity (from hundreds of bits for structural fragments to thousands for connectivity fingerprints, and millions for the complex pharmacophore-like fingerprints) [148], depending on the computational resources available and the intended application. Fingerprints which rely solely on interatomic connectivity – i.e. molecular constitution – are known as 2D fingerprints [149]. In the prototypic 2D keyed fingerprint design, each bit position is associated with the presence or absence of a specific

substructure pattern – for example carbonyl group attached to sp^3 carbon, hydroxyl group attached to sp^3 carbon, etc. [150].

2D Description of molecular constitution

2D descriptors can be computed solely from the constitution or topology of a molecule while 3D descriptors are obtained from the 3D structure of the molecule[143]. Many 2D molecular descriptors are based on molecular topology derived from graph-theoretical methods. Topological indices treat all atoms in a molecule as vertices and index specific information for all pairs of vertices. A simple topological index, for example, will contain only constitutional information such as which atoms are directly bound to each other. This is known as an adjacency matrix and an entry of 1 for vertices v_i and v_j if their corresponding atoms are bonded and an entry of 0 for v_i and v_i indicates that the corresponding atoms are not directly bonded [151]. For an adjacency matrix, the sum of all entries is equal to twice the total number of bonds in the molecule.

Complex topological indices are created by performing specific operations to an adjacency matrix that allow for the encoding of more complex constitutional information. These indices are based on local graph invariants which can represent atoms independent of their initial vertex numbering [152]. For example, topological indices may contain entries for the number of bonds linking the vertices. Information gathered from such an index can include the number of bonds linking all pairs of atom and the number of distinct ways a path can be superimposed on the molecular graph. A topological index that includes information such as heteroatoms and multiple bonds through the weighting of vertices and edges was introduced by Bertz [153].

Topological autocorrelation (2D autocorrelation) is designed to represent the structural information of a molecular diagram as a fixed-length vector that can be applied to molecules of any shape or size. It encodes the constitutional information as well as atom property distribution by analyzing the distances between all pairs of atoms. Topological autocorrelations are independent of conformational flexibility because all distances are measured as the shortest path of bonds between the two atoms. The autocorrelation vector is created by summing all products for atom pairs within increasing distance intervals in terms of number of bonds. In other words, it creates a frequency plot for a specific range of atom pair distances. By including atom property coefficients for all atom pairs, autocorrelations are capable of plotting the arrangement of specific atom properties. For example, information such as the frequency at which two negatively charged atoms are three bonds apart versus four bonds apart is stored in an autocorrelation plot that has been weighted by partial atomic charge [154].

3D Description of molecular configuration and conformation

The physicochemical meaning of topological indices and autocorrelations is unclear and incapable of representing some qualities that are inherently three-dimensional (stereochemistry). 3D molecular descriptors were developed to address some of these issues [155].

The 3D Autocorrelation is similar to the 2D autocorrelation but measures distances between atoms as Euclidian distances between their 3D coordinates in space. This allows a continuous measure of distances and encodes the spatial distribution of physicochemical properties. Instead of summing all pairs within discrete shortest path differences, the pairs are summed into interval steps [156].

Radial distribution functions (RDFs) is another very popular 3D descriptor. It maps the probability distribution to find an atom in a spherical volume of radius r. In its simplest form, the RDF maps the interatomic distances within the entire molecule. Often it is combined with characteristic atom properties in order to fit the requirements of the information to be represented [141a]. RDFs Not only provide information regarding interatomic distances between atoms and properties, they reflect other information such as bond distances, ring types, and planar versus non-planar molecules. These functions allow estimation of molecular flexibility by a "fuzziness" coefficient that extends the width of all peaks to allow for small changes in interatomic distances.

Molecular fingerprint and similarity searches

Molecular fingerprint based techniques attempt to represent molecules in such a way as to allow rapid structural comparison in an effort to identify structurally similar molecules or to cluster collections based on structural similarity. These methods are less hypothesis-driven and less computationally expensive than pharmacophore mapping or QSAR models (read below). They rely entirely on chemical structure and omit compound known biological activity, making the approach more qualitative in nature than other LB-CADD approaches [148]. Additionally, fingerprint-based methods consider all parts of the molecule equally and avoid focusing only on parts of a molecule that are thought to be most important for activity. This is less error-prone to over-fitting and requires smaller datasets to begin with. However, model performance suffers from the influence of unnecessary features and the often narrow chemical space evaluated [148]. Despite this drawback, 2D fingerprints continue to be the representation of choice for similarity-based virtual screening [157]. Not only are these methods the computationally least expensive way to compare molecular structures [149], but their effectiveness has been demonstrated in many comparative studies [157].

Similarity searches in LB-CADD

Fingerprint methods may be employed to search databases for compounds similar in structure to a lead query, providing an extended collection of compounds that can be tested for improved activity over the lead. In many situations, 2D similarity searches of databases are performed using chemotype information from first generation hits, leading to modifications that can be evaluated computationally or ordered for *in vitro* testing [9]. Bologa *et al* used 2D fingerprint and 3D shape-similarity searches to identify novel agonists of the estradiol receptor family receptor GPR30. Estrogen is an important hormone responsible for many aspects of development of physiology of tissues[158]. The GPCR GPR30 has recently been shown to bind estrogen with high affinity and its specific role in estrogen-regulated signaling is being studied[159]. This group used virtual screening to identify compounds selective for GPR30 that could be used to study this target. 10,000 molecules provided by Chemical Diversity Labs were enriched with GPCR binding ligands and screened for fingerprint-based similarity to the reference molecule 17β-estradiol. Fingerprints used were Daylight and MDL and similarities were scored using Tanimoto and Tversky scores. The top 100 ranked hits were selected for biological testing and a first-in-class selective agonist with a K_i of 11 nM for GPR30 was discovered. [160].

In addition to the enrichment of lead compound population, fingerprints are also used to increase molecular diversity of test compounds. Fingerprints can be used to cluster large libraries of hits in order to allow the sampling of a wide range of compounds without the need to sample the entire library. In this case, fingerprints are being used to optimize the sampling of diversity space. The Jarvis-Patrick method that calculates a list of nearest neighbors for each molecule has been shown to perform well for chemical clustering. Two structures cluster together if they are in each-others list of nearest neighbors and they have at least K of their J nearest neighbors in common. The MDL keys also provide a way to eliminate compounds which are least likely to satisfy the drug-likeness criterion [161].

Quantitative Structure Activity Relationship (QSAR) models

Quantitative structure-activity relationship (QSAR) models describe the mathematical relation between structural attributes and target response of a set of chemicals [162]. Classical QSAR is known as the Hansch-Fujita approach and involves the correlation of various electronic, hydrophobic, and steric features with biological activity. In the 1960s, Hansch and others began to establish QSAR models using various molecular descriptors to physical, chemical, and biological properties focused on providing computational estimates for the bioactivity of molecules [163]. In 1964, Free-Wilson developed a mathematical model relating the presence of various chemical substituents to biological activity (each type of chemical group was assigned an activity contribution) and the two methods were later combined to create the Hansch/Free-Wilson method[164].

The general workflow of a QSAR-based drug discovery project is to first collect a group of active and inactive ligands and then create a set of mathematical descriptors that describe the physicochemical and structural

properties of those compounds. A model is then generated to identify the relationship between those descriptors and their experimental activity maximizing the predictive power. Finally, the model is applied to predict activity for a library of test compounds that were encoded with the same descriptors. Success of QSAR, therefore, depends not only on the quality of the initial set of active/inactive compounds, but also on the choice of descriptors and the ability to generate the appropriate mathematical relationship. One of the most important considerations regarding this method is the fact that all models generated will be dependent on the sampling space of the initial set of compounds with known activity, the chemical diversity. In other words, divergent scaffolds or functional groups not represented within this "training" set of compounds will not be represented in the final model and any potential hits within the library to be screened that contain these groups will likely be missed. Therefore, it is advantageous to cover a wide chemical space within the training set. For a comprehensive guide on performing a QSAR-based virtual screen, please see the review by Zhang [162].

Multidimensional QSAR: 3D, 4D and 5D QSAR

Multidimensional QSAR (mQSAR) seeks to quantify all energy contributions of ligand binding including removal of solvent molecules, loss of conformational entropy, and binding pocket adaptation.

Comparative Field Molecular Analysis (CoMFA) [141h] is a 3D-QSAR technique that aligns molecules and extracts aligned features that can be related to biological activity. This method focuses on the alignment of molecular interaction fields rather than the features of each individual atom. CoMFA was established over 20 years ago as a standard technique for constructing 3D models in the absence of direct structural data of the target. In this method, molecules are aligned based on their 3D structures on a grid and the values of steric (VDW interactions) and electrostatic potential energies (Coulombic interactions) are calculated at each grid point. Comparative Molecular Similarity Indices (CoMSIA) is an important extension to CoMFA. In CoMSIA, the molecular field includes hydrophobic and hydrogen-bonding terms in addition to the steric and coulombic contributions. Similarity indices are calculated instead of interaction energies by comparing each ligand with a common probe and Gaussian-type functions are used to avoid extreme values [165]. One important limitation to these methods, however, is that their applicability is limited to static structures of similar scaffolds while neglecting the dynamical nature of the ligands [142a].

4D-QSAR is an extension of 3D-QSAR that treats each molecule as an ensemble of different conformations, orientations, tautomers, stereoisomers, and protonation states. The fourth dimension in 4D-QSAR refers to the ensemble sampling of spatial features of each molecule. A receptor-independent (RI) 4D-QSAR method was proposed by Hopfinger, et al [166]. This method begins by placing all molecules into a grid and assigning interaction pharmacophore elements (IPE) to each atom in the molecule (polar, nonpolar, hydrogen bond donor, etc.). Molecular dynamic simulations are used to generate a Boltzmann weighted conformational ensemble of each molecule within the grid. Trial alignments are performed within the grid across the different molecules and

descriptors are defined based on occupancy frequencies within each of these alignments. These descriptors are called Grid Cell Occupancy Descriptors (GCODs). A conformational ensemble of each compound is used to generate the GCODs rather than a single conformation.

5D-QSAR has been developed to account for local changes in the binding site that contribute to an induced fit model of ligand binding. In a method developed by Vedani and Dobler [167], induced fit is simulated by mapping a "mean envelope" for all ligands in a training set on to an "inner envelope" for each individual molecule. Their method involves several protocols for evaluating induced-fit models including a linear scale based on the adaptation of topology, adaptations based on property fields, energy minimization, and lipophilicity potential. Using this information, the energetic cost for adaptation of the ligand to the binding site geometry is calculated.

Receptor-Dependent 3D/4D-QSAR

While QSAR methods are especially useful when structural information regarding target-binding site is not available, QSAR methods that specifically include such information have been developed. One method, known as Free Energy Force Field (FEFF) 3D-QSAR trains a ligand-receptor force field QSAR model that describes all thermodynamic contributions for binding[168]. A 4D-QSAR version of FEFF has also been developed to apply this method to the RI-4D-QSAR methods described above[168]. Structurally, the analysis is focused solely on the site of interaction between the ligand and target and all atoms of interest are assigned partial charges. Molecular dynamic simulations are applied to these structures to generate a conformational ensemble following energy minimization. This approach avoids any alignment issues present in the RI-4D-QSAR method since the binding site constrains the three-dimensional orientations of the ligands. The conformation ensembles of receptor-ligand complexes generated are placed in a similar grid-cell lattice as used in RI-4D-QSAR and occupancy profiles are calculated to generate receptor-dependent (RD) 4D-QSAR models. When tested alongside RI-4D-QSAR against a set of glucose analogue inhibitors of glycogen phosphorylase, predictability of RD-4D-QSAR models outperformed those of RI-4D-QSAR[168].

Linear regression and related methods

Linear models used include multivariable linear regression analysis (MLR), principal component analysis (PCA), or partial least square analysis (PLS) [142a]. MLR computes biological activity as a weighted sum of descriptors or features. The method requires typically 4-5 data points for every descriptor used. PCA increases the efficiency of MLR by extracting information from multiple variables into a smaller number of uncorrelated variables. Analysis of results is however not always straightforward[169]. It can be applied with smaller sets of compounds than MLR. PLS combines MLR and PCA and extracts the dependent variable (biological activity) into new components to optimize correlations [170]. PCA or PLS are commonly used for developing models for the molecular interaction field algorithm CoMFA and CoMSIA [142a]. Advantage of these models is that they can be trained rapidly using the tools of linear algebra. The major drawback is that chemical structure often relates with biological activity in a non-linear fashion.

Non-linear models employing machine learning algorithms

Artificial Neural Networks (ANNs) are one of the most popular non-linear regression models applied to QSAR-based drug discovery [171]. These models belong to the class of self-organizing algorithms where the neural network learns the relationship between descriptors and biological activity through iterative prediction and improvement cycles [142a]. A major drawback of neural networks is the fact that they are sensitive to overtraining, resulting in excellent performance within the training set but reduced ability to assess novel compounds. Therefore, care is taken to measure ANN performance always on "independent" datasets not employed for model generation.

SVM is a kernel-based supervised learning method that was introduced by Vapnik and Lerner [172]. It is based on statistical learning theory and the Vapnik-Chervonenkis dimension [173] and seeks to divide sets of patterns (molecules described with descriptors) based on their classification (biological function). Once this separation is performed on a training dataset, novel patterns can be classified based on which side of the boundary they fall. The simplest form of separation can be imagined as a straight line down the center of a graph with the two classes clustered in opposite corners of the graph. Since there are many different lines that can be defined to separate these classes, SVM is described as a maximal margin classifier as it seeks to define the hyperplane with the widest margin between these two classes. The patterns (compounds) which line the closest border of each class define the two hyperplanes separated by that margin. These patterns (molecules) are known as support vectors and represent the maximal margin solution and are used to predict classes for novel unclassified patterns. All patterns that lie further from these boundaries are not support vectors and have no influence on the classification of novel patterns. Hyperplanes defined by the lowest number of support vectors are preferred. The solution is a parallel decision boundary that lies equidistant from the two hyperplanes defined by their respective support vectors[174].

SVM was initially designed for datasets that could be separated linearly. However, especially in CADD application, this is not always possible. Therefore, SVM incorporated a high-dimensional space in which linear classification was once again possible. This involves the preprocessing of input data using feature functions where the input variables are mapped into a Hilbert space of finite or infinite dimension [174a]. While it cannot be predicted which feature functions will allow for linear classification, as the input vector is mapped into higher space, this becomes more possible. This strategy, however, must be offset by the fact that higher dimensional space creates more computational burden and contributes to over-fitting [175].

SVM utilizes kernel functions to ease the computational demand imposed by the existence of higher dimensional data. These special nonlinear functions combine the feature functions in a way that avoids explicit transformation and preprocessing using feature functions [174a]. In other words, the higher dimensional space that allows for linear separation does not need to be dealt with directly.

A kernel is essentially a function in which the solution for two inputs is equal to the dot product of their mapping from input space to Hilbert space. Based on this fact, any novel kernels a researcher seeks to develop must be a dot product in a mapped feature space. This can be tested mathematically applying Mercer's condition [175]. The definition of new kernels, however, is not usually necessary as multiple useful kernels have already been well established for different problem types. Which kernel is necessary for any given problem cannot be predicted, but is generally best selected a priori by researching which kernels have been successfully used in similar applications. It is not recommended to select the best kernel based on performance with the dataset being researched as this can often lead to over-fitting and poor generalizability. Some of the most commonly used kernels include the linear (dot) kernel used mainly as a test of nonlinearity and reference for classification improvement following the application of nonlinear kernels, the polynomial kernel which can be adjusted based on its degree to allow for larger feature space, radial basis function kernel, anova kernel, Fourier series kernel, spline kernel, additive kernel, and tensor product kernel. Addition, multiplication, and composition of these kernels all result in valid kernels [174a]. When implementing a novel kernel function, however, the researcher must ensure that it is the dot product in a feature space for some mapping. This condition can be tested by applying Mercer's condition [175]. It should be considered, however, that over-fitting could be induced with more complex kernel functions.

Decision Tree (DT) learning is a supervised learning algorithm that works by iteratively grouping the training data set into small and more specific groups. The resulting classification resembles a tree where each feature is broken into different values and each of these values is subsequently divided based on values of a different feature. The order in which features are divided is usually based on an information gain (difference between information before and after the branching) parameter with the highest valued features appearing first[176]. Various methods are used to sort the features with the overall goal of the smallest possible decision tree providing the best performance. C4.5 is a widely used DT algorithm that calculates information gain based on information entropy[177]. The information entropy of a given classification that can divide the dataset into two classes is calculated based on the number of compounds in either class. The information entropy of the system when dividing the dataset into two subsets using a specific feature is calculated based on the number of compounds from each class in either of the feature subsets. Finally, the information gain for that specific feature is calculated as the difference between the information entropy of the classification and the information entropy of the system.

Once the decision tree has been optimized for the training set, new compounds can be classified by applying their descriptors to the decision tree and activities can be predicted based on which subset they fall into and the activities of the training compounds that are contained in that subset.

QSAR Application in LB-CADD

Mueller *et al* used ANN QSAR models to identify novel positive and negative allosteric modulators of mGlu5. This receptor has been implicated in neurological disorders including anxiety, Parkinson's disease, and schizophrenia[178]. For the identification of positive allosteric modulators (PAMs), they first performed a traditional high throughput screen of approximately 144,000 compounds. This screen yielded a total of 1,356 hits, a hit rate of 0.94%. The dataset from this HTS was then used to develop a QSAR model that could be used in a virtual screen. To generate the QSAR model, a set of 1,252 different descriptors across 35 categories were calculated using the ADRIANA software package. The descriptors included scalar, 2D, and 3D descriptor categories. The authors iteratively removed the least sensitive descriptors in order to create the optimal set. This final set included 276 different descriptors, including scalar descriptors such as molecular weight up to 3D descriptors including the radial distribution function weighted by lone-pair electronegativity and π electronegativity. A virtual screen was performed against approximately 450,000 commercially available compounds in the ChemBridge database. 824 compounds were tested experimentally for the potentiation of mGlu5 signaling. Of these compounds, 232 were confirmed as potentiators or partial agonists. This hit rate of 28.2% was approximately 30 times greater than that of the original HTS and the virtual screen took approximately one hour to complete once the model had been optimized[5]. In a

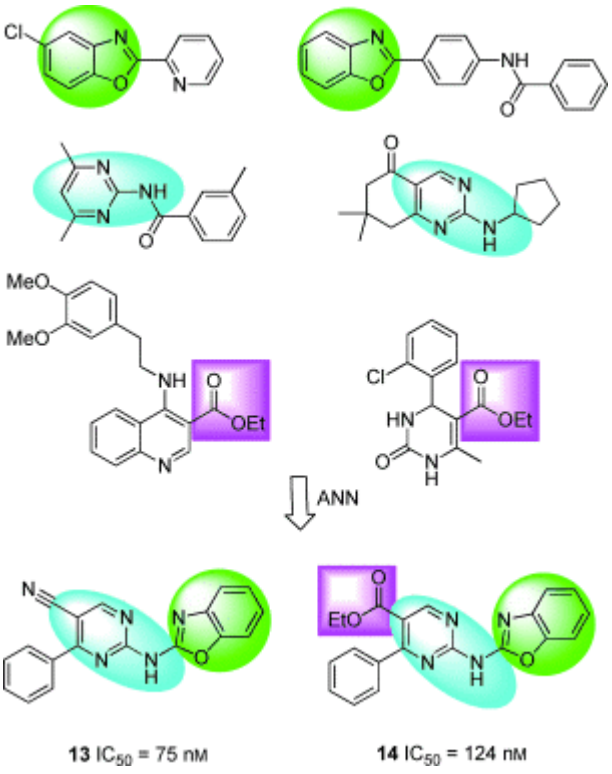


Figure 1-8 QSAR-based virtual screening of mGlu5 negative allosteric modulators yields lead compounds that contain substructure combinations taken across several known actives used for model generation. Mueller et al[5]

separate study, Mueller *et al* [179] used a similar approach to identify negative allosteric modulators for mGlu5. Rodriguez *et al* had previously performed a traditional HTS screen of 160,000 compounds for allosteric modulators of mGlu5 and found 624 antagonists[180]. The QSAR model was used to virtually screen over 700,000 commercially available compounds in the ChemDiv Discovery database. Hits were filtered for drug-like properties, and fingerprint techniques were used to remove hits that were highly similar to known actives in order to identify new chemotypes. 749 compounds were tested *in vitro* and 27 compounds were found to modulate mGlu5 signaling. This hit rate of 3.6% was a significant increase over the 0.2% hit rate of the traditional HTS screen. The most potent of the compounds showed *in vitro* IC_{50}'s of 75 and 124 nM, respectively, and contained a previously unidentified scaffold (Figure 1-8). Following analogue synthesis and stability optimization, the experimenters tested the effect of their best lead *in vivo* against two behaviors known to involve mGlu5: operant sensation seeking behavior [181] and the burying of foreign objects in deep bedding [182]. Both behaviors were found to be inhibited given intra-peritoneal administration of their lead analogue.

CoMFA and CoMSIA 3D-QSAR methods have also been used to predict novel therapeutic compounds for a variety of disease targets. Ke *et al* [183] generated CoMFA and CoMSIA models using 66 previously discovered pyrazole- and furanopyrimidine-based Aurora Kinase inhibitors[184]. Aurora kinase A is a serine/threonine kinase involved in mitosis [185] that has been shown to be involved in various different forms of cancer [186]. Using the model that showed the best predictive performance, the group synthesized a novel compound (compound 67). This compound was tested *in vitro* and displayed an IC50 of 25 nM against Aurora kinase A. Additionally, compound 67 displayed antiproliferative activity with an IC50 of 23 nM against the HCT-116 colon cancer cell line.

Over the past several decades, over 18,000 QSAR models have been reported for a variety of targets with a variety of descriptors. Hansch *et al* have carefully collected these into a comprehensive database of QSAR models called C-QSAR [187]. This collection has provided not only access to models for novel applications, but allows the analysis of QSAR models to identify challenges for the field. Kim *et al* examined the C-QSAR database for outlier patterns – i.e. compounds that showed poor prediction when the average prediction for the model was good. They found that over 47 QSAR models examined, the number of compounds scoring as outliers ranged from 3% to 36%. 26 of the 47 datasets showed 20% or more compound outliers[188]. They presented several theories as to why QSAR models are so sensitive to the generation of outliers. One possibility came from analysis of the RCSB protein databank where they discovered examples where related analogs were shown to bind in very different poses. Another explanation offered was protein flexibility, leading to multiple binding modes and or binding sites on the same protein. These different binding modes/sites may reflect different structure-activity relationships for molecules within a given dataset. In other words, analogous compounds that do not share the same binding mode can present difficulties in the classifications of ligands[188].

Selection of optimal descriptors/features

Hristozov *et al* analyzed the performance of different descriptors across a range of benchmarking datasets and found that the performance of a particular descriptor was often dependent on the activity class. It was found that topological autocorrelation usually offers the best dimensionality/performance ratio. The fusion of the ranked lists obtained with RDF codes and 2D descriptor improved results because RDF codes, while giving similar results, covered different parts of the activity spaces under investigation [29a]. In result, it is not possible to choose a small optimal set of descriptors independent of the problem – a custom-optimized descriptor set is needed for optimal performance of LB-CADD.

Excessive numbers of descriptors or features can add noise to a model reducing its predictive power. Feature selection techniques remove unnecessary features in order to minimize the number of degrees of freedom of the model. Thus, the ratio of data points versus degrees of freedom increases leading to models of increased predictive power. Techniques which have proven successful in QSAR modeling include selecting features by measures such as information gain [189] and F – Score [190], sequential feature forward selection or feature backward elimination [191], genetic algorithm [192], swarm optimization [192a], and input sensitivity analysis [179].

Information gain measures the change of information entropy from the data distribution of two classes (active and inactive compounds) of one feature compared to the entropy of the feature overall. Thus, discriminatory power of the individual feature increases with information gain. An f-score is calculated that considers the mean and standard deviation of each feature across data classes. The higher the f-score value, the greater discriminatory power of that feature. Selecting features by individual benchmarks has the disadvantage, that correlation between features is ignored. For example, let us assume a feature has a high information gain. However, if a second feature highly correlated is already part of the model, no improved model will result from adding the feature. More complex feature selection schemes address this limitation:

Sequential feature forward selection is a deterministic, greedy search algorithm. In each round, the best feature set from the previous round N appends a single feature from the pool of M remaining features and trains the M models using the N+1 features. The best performing feature set from this round then advances to the next round. This continues until all features are used in a final feature set. The best performing model over all iterations is then chosen as the best feature set. This process is time consuming and not guaranteed to yield the optimal feature set – the single best performing feature will always be part of the model. However, there is no guarantee that it is needed. Feature backward elimination inverts the process starting from a model trained from all features eliminating one after the other. While the process is more robust in terms of identifying the optimal model, it also requires substantial computer time. Therefore, alternative approaches have been explored to optimize feature sets:

Genetic algorithms mimic the process of evolution to create an efficient search heuristic. This method uses a population of individuals (distinct feature sets) to encode candidate solutions. The initial individuals can be generated randomly. In each iteration, or generation, the fitness of each individual is evaluated – i.e. the predictive power of the derived LB-CADD model. This fitness function is the performance metric of a model trained using that individual as the feature set. Individuals are then selected based on the fitness and undergo recombination and/or mutation to form the next generation. The algorithm continues until a desired fitness score is achieved or a set number of generations have been completed.

Swarm optimization algorithms, such as ant colony optimization [193], particle swarm optimization, and artificial bee colony optimization [194], are optimization techniques based on the organized behavior of social animals such as birds. The algorithm iteratively searches for a best solution by moving individuals around the search space guided by both the local best solution, as well as the best solutions found so far in the entire population. The best overall solution is constantly updated letting the swarm converge towards the optimal solutions.

Input sensitivity analysis seeks to combine speed of individual benchmark values with accuracy of methods that take correlation into account. First, a model is constructed using all features. Next, the influence of each feature on the model output is determined: Each feature x_i is perturbed and the change in output y is computed. This procedure numerically estimates the partial derivative of the output with respect to each input – a measure that is effective in selecting optimal descriptor sets [179].

Pharmacophore mapping

In 1998, the IUPAC formally defined a pharmacophore as "the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response" [195]. In terms of drug activity, it is the spatial arrangement of functional groups that a compound or drug must contain in order to evoke a desired biological response. Therefore, an effective pharmacophore will contain information about functional groups that interact with the target, as well as information regarding the type of non-covalent interactions and interatomic distances between these functional groups/interactions. This arrangement can be derived either in a structure-based manner by mapping the sites of contact between a ligand and binding site, or using a ligand-based approach. The former can be achieved by analyzing one or several co-crystal structures with lead or drug compounds bound and will not be discussed in more detail here. We focus on the latter, more challenging problem.

To generate a ligand-based pharmacophore, multiple active compounds are overlaid in such a way that a maximum number of chemical features overlap geometrically [196]. This can involve rigid 2D or 3D structural representations or, in more precise applications, incorporate molecular flexibility to determine overlapping sites. This conformational flexibility can be incorporated by pre-computing the conformational space of each ligand and

creating a general-purpose conformational model or conformations can be explored by changing molecule coordinates as needed by the alignment algorithm [196]. For example, one popular pharmacophore-generating software package, Catalyst, uses the "polling" algorithm [197] to generate approximately 250 conformers that it uses in its pharmacophore generation algorithm [142a]. In a study targeting HSP90α, Al-Sha'er *et al* used 83 known reference molecules to generate pharmacophore queries and identified 25 diverse inhibitors including three with IC_{50} values below 10 nM [198].

Superimposing active compounds to create a pharmacophore

Molecules are commonly aligned through either a point-based or a property-based technique. The point-based technique involves superposing pairs of points (atoms or chemical features) by minimizing Euclidean distances. These alignment methods typically use a root-mean-square distance (RMSD) to maximize overlap [199]. Property-based alignment techniques, on the other hand, use molecular field descriptors to generate alignments [196]. These fields define 3D grids around compounds and calculate the interaction energy for a specific probe at each point. The distribution of interaction energies is represented by Gaussian functions and the degree of overlap between Gaussian functions of two aligned compounds is used as the objective scoring function to maximize alignment [199]. One popular field generation method for property-based alignments is GRID [200].

Molecular flexibility is always an important consideration when aligning compounds of interest and several approaches are employed to most efficiently sample conformational space. These approaches include rigid, flexible, and semi-flexible methods. Rigid methods require knowledge of the active conformation of known ligands and align only the active conformations. This is only applicable, however, when the active conformation is known with confidence. Semi-flexible methods begin with pre-generated ensembles of static conformations to overlay and flexible methods, being the most computationally expensive, perform conformational search during the alignment process, often using molecular dynamics or randomly sampling of rotatable bonds. Since the conformational space can increase substantially with an increase in the number of rotatable bonds, strategies are often employed to limit the exploration of conformational space with reference geometry (often an active ligand with low flexibility). This method is known as the Active Analog Approach [201].

Pharmacophore feature extraction

A pharmacophore feature map is carefully constructed to balance generalizability with specificity. A general definition might categorize all functional groups having similar physiochemical properties (i.e. similar hydrogen-bonding behavior, ionizability) into one group, whereas specific feature definitions may include specific atom types at specific locations. More general feature definitions increase the population of compounds that match the pharmacophore. They allow the identification of novel scaffolds but also increase the ratio of false positives. The

level of feature definition generalizability is usually determined by the algorithm used to extract feature maps and through user-specified parameters. The most common features used to define pharmacophore maps are hydrogen bond acceptors and donors, acidic and basic groups, aliphatic hydrophobic moieties, and aromatic hydrophobic moieties [142a]. Features are commonly implemented as spheres with a certain tolerance radius for pharmacophore matching [196].

Pharmacophore Algorithms and Software Packages

The most common software packages employed for ligand-based pharmacophore generation include Phase [202], MOE [203], Catalyst [204], LigandScout [205], DISCO [206], and GASP [207]. These packages utilize different approaches to molecular alignment, flexibility, and feature extraction. Catalyst approaches alignment and feature extraction by identifying common chemical features arranged in certain positions in three-dimensional space. These chemical features focus on those expected to be important for interaction between ligand and protein and include hydrophobic regions, hydrogen-bond donors, hydrogen-bond acceptors, positive ionizable, and negative ionizable regions. Chemical groups that participate in the same type of interaction are treated as identical. Catalyst contains two algorithms that can be used for pharmacophore construction. HipHop is the simpler of the two algorithms and looks for common 3D arrangements of features only for compounds with a threshold activity against the target. It begins with best alignment of only two features (scored by RMS deviations) and continues expanding the model to include more features until no further improvements are possible. This method is only capable of producing a qualitative distinction between active and inactive predictions. HypoGen, on the other hand, employs biological assay data such as IC_{50} values for active compounds as well as a set of inactive compounds. Initial pharmacophore construction in HypoGen is identical to HipHop but includes additional algorithms that incorporate inactive compounds and experimental values. These algorithms compare the best pharmacophore from the 'HipHop' stage with the inactive compounds and features common to the inactive set are removed. Finally, HypoGen performs an optimization routine that attempts to improve the predictive power of the pharmacophore by making adjustments and scoring the accuracy in predicting the specific experimental activities.[204a, 208] This results in models that are capable of quantitative predictions that can predict specific levels of activity. Ten different models are created following a simulated annealing optimization [209]. Both Catalyst methods incorporate molecular flexibility by storing compounds as multiple conformations per molecule. The Poling algorithm published by Smellie *et al* [197] is employed to increase the conformational variation within the set of conformations per molecule. This allows Catalyst to cover the greatest extent of conformational space while keeping the number of conformations at a minimum.

Phase approaches alignment and feature extraction using a tree-based partitioning algorithm and an RMS deviation-based scoring function that considers the volume of heavy atom overlap. It incorporates molecular flexibility through a preparation step where conformational space is sampled using a Monte Carlo or torsional search [199].

DISCO regards compounds as sets of interpoint distances between heavy atoms containing features of interest. Alignments are based on the spatial orientation of common point among all active compounds. DISCO considers multiple conformations that have been pre-specified by the user during the alignments and uses a clique-detection algorithm for scoring alignments [208].

GASP (Genetic Algorithm Superposition Program) uses a genetic algorithm with iterative generations of the best models for pharmacophore construction[207]. Flexibility is handled during the alignment process through random rotations and translations. Conformations are optimized by fitting them to similarity constraints and weighing the conformations that fit these constraints more than conformations that do not [209].

Different software packages can produce different results for the same datasets and their strengths and weaknesses should be considered prior to any application. For example, Catalyst only permits a single bonding feature per heavy atom while LigandScout allows a hydrogen-bond donor or acceptor to be involved in more than one hydrogen-bonding interaction [196]. MOE, on the other hand, allows a more customizable approach to hydrogen-bonding features. Lipophilic areas are generally represented as spheres located on hydrophobic atom chains, branches, or groups in a similar manner across software packages but with slight nuances. While subtle, these differences have important consequences on prediction models. Additionally, software packages that do not attach a hydrophobic feature to an aromatic ring are unable to predict that an aromatic group may be positioned in a lipophilic binding pocket [196]. The level of customizability also differs across pharmacophore software packages and can influence predictions. Catalyst allows the specification of one or more chemical groups that satisfy a particular feature while Phase allows not only matching chemical groups but also a list of exclusions for a given feature. MOE offers a level of customization that allows the user to implement entirely novel pharmacophore schemes as well as modification of existing schemes. However, this requires additional levels of expertise to program[196]. For a comprehensive analysis of the differences between commercial pharmacophore software packages, please see the 2007 review by Wolber *et al* [196] and a 2002 comparison of Catalyst, DISCO, and GASP by Patel *et al* [210].

Ligand-based pharmacophore methods have been used for the discovery of novel compounds across a variety of targets. New compounds can have activity in the micromolar and nanomolar range and reflect proof of concept with *in vivo* disease models. Al-Sha'er *et al* used a diverse set of 83 known Hsp90-α inhibitors and the HypoGen module of Catalyst to generate a pharmacophore model. Hsp90-α is a molecular chaperone that is involved in protein folding, stability, and function [211]. By interacting with many oncogenic proteins, it has been shown to be a valid anticancer drug target[212]. The pharmacophore model was used to screen the NCI list of compounds (238,000) using the "Best Flexible" search option. The top 100 hits were evaluated *in vitro* and their most potent compound had an IC_{50} of 25 nM [198].

Noha *et al* developed 5-point pharmacophore models using the HipHop algorithm of Catalyst based on a training set of compounds with IC_{50} < 100 nM against IKK-β as potential anti-inflammatory and chemosensitizing agents. The

authors used 128 active and 44 inactive compounds to develop a pharmacophore model[213]. Their model was further refined with exclusion volume spheres and shape constraints to improve the scoring of compounds in their virtual high-throughput screen against the National Cancer Institute molecular database. Ten compounds were selected and the most potent compound (NSC719177) showed inhibitory activity against IKK-β in a cell free *in vitro* assay with IC_{50} of 6.95 μM. Additionally, this compound inhibited NF-κB activation induced by TNF-α in HEK293 cells with an IC_{50} of 5.85 μM[213].

Chiang *et al* used the HypoGen module of Catalyst to generate four-feature pharmacophore models based on an indole series of 21 compounds that showed anti-proliferative activity through the inhibition of tubulin polymerization/microtubule depolymerization. Disruption of microtubules during the mitotic phase of the cell cycle can induce cell cycle arrest and apoptosis [214]. Therefore, inhibitors of tubulin polymerization are useful cancer treatments. 130,000 compounds of the ChemDiv database and in-house compound collection were screened and the top 142 hits were tested *in vitro*. Four novel compounds were discovered with anti-proliferative activity. The most potent compound displayed anti-proliferative activity in human cancer KB cells with an IC_{50} of 187 nM. This compound also inhibited the proliferation of other cancer cell types including MCF-7, NCI-H460, and SF-268 and demonstrated anti-cancer effects in a histoculture system. *In vitro* assays revealed that this compound inhibited tubulin polymerization with an IC_{50} of 4.4 μM[215].

Lanier generated pharmacophores containing five feature points using Catalyst and CombiCode software and an exclusion sphere generated in MOE based on a training set of 100 active and 1000 inactive compounds. This model was used to guide and evaluate variations of a core molecule, leading them to a gonadotropin releasing hormone GnRH receptor antagonist with receptor affinity below 10 nM [216]. GnRH is involved in the regulatory pathways of follicle stimulating hormone (FSH) and luteinizing hormone (LH). It is a target for disease therapeutics including endometriosis, uterine fibroids, and prostate cancer[217].

Roche *et al* used known H3 antagonists to generate a pharmacophore model with four features including a distal positive charge, an electron rich position, a central aromatic ring, and either a second basic amine or another aromatic [218]. Histamine is a central modulator in the central and peripheral nervous systems through four receptors (H1-H4) [219]. H3 is a presynaptic autoreceptor that modulates production and release of histamine and other neurotransmitters [220]. H3 antagonists have been studied in Alzheimer's disease, attention deficit disorder, and schizophrenia [221]. Additionally, it has been suggested to be involved in appetite and obesity [222] .This model was used in a *de novo* approach with the Skelgen software [223] to generate novel compounds from fragment libraries that match the pharmacophoric restraints. They found a series of four compounds with high potency and selectivity for H3. Their most potent compound showed inverse agonist activity with an EC_{50} of 200 pM in a GTPγS functional assay and a binding affinity K_i towards H3 of 9.8 nM [218].

Chao *et al* used pharmacophore-based design to take advantage of the therapeutic benefits of Indole-3-carbinol (I3C) in the treatment of cancer. I3C is known to suppress proliferation and induce apoptosis of various cancer cells through the inhibition of Akt activation[224]. I3C, however, has a poor metabolic profile and low potency, likely because its therapeutic behavior comes from only four of its metabolites. By overlaying these low energy conformers of these four metabolites, Chao *et al* was able to identify similar N-N' distances and overlapping indole rings[225]. This led them to design SR13650 that showed an IC_{50} of 80 nM. Tumor xenograft studies using MCF-7 cells revealed antitumor effects at 10 mg/kg for 30 days. Computational analysis was also applied to increase the bioavailability and three compounds showed 45-60% tumor growth inhibition *in vivo* compared to the 26% growth inhibition of SR13650. SR13668 was the most potent compound and also displayed antitumor effects in other xenograft models. *In vitro*, SR13668 was shown to inhibit Akt activation by blocking growth factor stimulated phosphorylation and showed favorable toxicological profiles[225]. This drug is currently in phase 0 trials for the treatment of cancer[226].

Conclusions

The extensive variety of computational tools employed in drug discovery campaigns suggests that there are no fundamentally superior techniques. The performance of methods varies greatly with target protein, available data, and available resources. For example, Kruger and Evers completed a performance benchmark between structure- and ligand-based vHTS tools across four different targets including angiotensin-converting enzyme, cyclooxygenase-2, thrombin and HIV-1 protease [227]. Docking methods including Glide, GOLD, Surflex, and FlexX were used to dock ligands into rigid target crystal structures obtained from PDB. A single ligand was used as a reference for ligand-based similarity search strategies such as 2D (fingerprints and feature-trees) and 3D (Rapid Overlay of Chemical Structures - ROCS), a similarity algorithm that calculates maximum volume overlap of two 3D structures[228]. In general, the authors found that docking methods performed poorly for HIV-1 protease and thrombin due to the flexible nature of the targets and the fact that the known ligands for these proteins have large molecular weight and peptidomimetic character.

Enrichments based on 3D similarity searches were poor for HIV-1 protease and thrombin datasets compared to ACE, which is likely due to the higher level of diversity in the HIV-1 protease and thrombin ligand datasets. Similarity scoring algorithms like ShapeTanimoto, ColorScore, and ComboScore were compared with the performance of ROCS [227]. It was found that even within the scoring algorithm performance varied across targets. For example, ColorScore performed best for ACE and HIV-1 protease while ShapeTanimoto for COX-2 and ComboScore was the method of choice for thrombin. All vHTS tools performed comparatively well for ACE but ligand-based 2D fingerprint approach generally outperformed docking methods. The authors also note an important observation in that, especially for HIV-1 protease, the structure-based and ligand-based approaches yielded complimentary hit lists. Therefore, performance metrics are not the only benchmark to consider when comparing CADD techniques. In some cases, discovery of novel chemotypes is more important than high hit rates or high activity. In the current study, Kruger

and Evers found that ROCS and feature trees were more successful in retrieving compounds with novel scaffolds compared to other fingerprints [227].

Warren *et al* published an in-depth assessment of the capabilities and shortcomings for docking programs and their scoring techniques against eight proteins of seven evolutionarily diverse target types. They found that docking programs were well adept at generating poses that included ones similar to those found in complex crystal structures. In general, while the molecular conformation was less precise across docking programs, they were fairly accurate in terms of the ligand's overall positioning. With regards to scoring, their findings agree with others that docking programs lack reliable scoring algorithms. So while the tools were able to predict a set of poses that included those that were seen in the crystal structure, the preference for the crystal structure pose was not necessarily reflected in the scoring. For five of the seven targets that were evaluated, the success rate, however, was greater than 40%. It was found that the enrichment of hits could be increased by applying previous knowledge regarding the target. However, there was little statistically significant correlation between docking scores and ligand affinity across the targets. The study concluded that a docking program's ability to reproduce accurate binding poses did not necessarily mean that the program could accurately predict binding affinities. This analysis underscores the necessity not only to re-rank the top hits from a docking-based vHTS using computationally expensive tools but also to continue evaluating novel scoring functions that can efficiently and accurately predict binding affinities[229].

Improvements in scoring functions involve the use of consensus scoring methods and free energy scoring with docking techniques. Consensus scoring methods have been shown to improve enrichments and prediction of bound conformations and poses, by balancing out errors of individual scoring functions. In 2008, Enyedy and Egan compared docking scores of ligands with known IC_{50} and found that docking scores were incapable of correctly ranking compounds and were sometimes unable to differentiate active from inactive compounds. They concluded that individual scoring methods can be used successfully to enrich a dataset with increased population of actives but are insufficient to identify actives against inactives[17]. Page *et al* concluded that even though binding energy calculations such as MM-PBSA are one of the more successful methods of estimating free energy of complexes, these techniques are more applicable to providing insights into the nature of interactions rather than prediction or screening [230]. Consensus scoring functions where free energy scores of different algorithms have been combined or averaged have been shown to substantially improve performance[231].

In their literature survey, Ripphausen *et al* reported that structure-based virtual screening was employed much more frequently than ligand-based virtual screening (322 to 107 studies). Despite a preference for structure-based methods, ligand-based methods on average yield hits with higher potency than structure-based methods. Most ligand-based hits had activities better than 1 μM while structure-based hits fall frequently in the range of 1-100 μM [16]. Scoring algorithms in docking functions have been found to be biased towards known protein ligand complexes –

for example more potent hits against protein kinase targets are discovered when compared to other target classes [28].

One CADD approach that has been gaining considerable momentum is the combination of structure-based and ligand-based computation techniques [232]. For example, the GRID-GOLPE method docks a set of ligands at a common binding site using GRID and then calculates descriptors for the binding interactions by probing these docking poses with GOLPE [233]. Multivariate regression is then used to create a statistical model that can explain the biological activity of these ligands. Structure-based interactions between a ligand and target can also be used in similarity based searches to find compounds that are similar only in the regions that participate in binding rather than cross the entire ligand. LigandScout employs such a technique to define a pharmacophore based on hydrogen bonding and charge-transfer interactions between a ligand and its target. Another technique known as the pseudo receptor technique [55a] uses pharmacophore mapping-like overlaying techniques for a collection of ligands that bind to the same binding site in order to establish a virtual representation of the binding site's structure which is then used as a template for docking and other structure-based vHTS. This approach has been utilized by VirtualToxLab [234] for the creation of nuclear receptors and cytochrome P450 binding site models in ADMET prediction tools and by Schneider *et al* in the modeling of the H4 receptor binding site subsequently used to identify novel active scaffolds [55b]. In a recent review by Wilson and Lill [235], these methods are grouped into a major class of combined techniques called interaction based methods. A second major class involves the use of QSAR and similarity methods to enrich a library of virtual compounds prior to a molecular docking project. This can increase the efficiency of the project by reducing the number of compounds to be docked. This is similar to the application of CADD to enrich libraries prior to traditional HTS projects. This chapter also presents comprehensive descriptions of software packages employing a combination of ligand- and structure-based techniques as well as several case studies testing the performance of these tools.

As discussed earlier, these methods are often used in serial where ligand-based methods are first used to enrich libraries that will subsequently be used in structure-based vHTS. The most common application is at the ligand library creation stage through the use of QSAR techniques to filter out compounds with low similarity to a query compound or no predicted activity based on a statistical model. QSAR has also been employed as a means to refine the docking scores of a structure-based virtual screen. 2D and 3D QSAR can also be used to track docking errors. This method has been employed by Novartis where a QSAR model is built from docking scores rather than observed activities and this model is applied to that set to provide additional score weights for each compound [236].

Even though CADD has been applied quite extensively in drug discovery campaigns, certain lucrative therapeutic targets like protein-protein interaction and protein-DNA interactions are still formidable problems mainly due to the relatively massive size of interaction sites (in excess of 1500 $Å^2$) [6]. Lastly, accessibility has also been a problem with CADD, as many tools are not designed with a friendly user interface in mind. In many cases, there can be an

overwhelming number of variables that must be configured on a case-by-case basis and the interfaces are not always straightforward. A great deal of expertise is often required to use these tools to get desired measure of success. Increasingly, efforts are being made to develop user-friendly interfaces especially in commercially available tools. For example, MolSoft is a software package that is designed to be a user-friendly docking tool and replaces the front-end of current docking algorithms with an interface that is manageable to a wider audience [108]. More recently gamification of the ROSETTA folding program, known as FOLDIT [237], has allowed individuals from non-scientific community to help solve the structure of M-PMV retroviral protease [238] and for predicting backbone remodeling of computationally designed biomolecular Diels-Alderase that increased its activity [239]. The successful application of crowd-sourced biomolecule design and prediction suggests further potential of CADD methods in drug discovery.

References

1. Wei, D. G.; Jiang, X. L.; Zhou, L.; Chen, J.; Chen, Z.; He, C.; Yang, K.; Liu, Y.; Pei, J. F.; Lai, L. H., Discovery of Multitarget Inhibitors by Combining Molecular Docking with Common Pharmacophore Matching. *Journal of Medicinal Chemistry* 2008, *51* (24), 7882-7888.
2. Zhang, S.; Cao, Z. X.; Tian, H. W.; Shen, G. B.; Ma, Y. P.; Xie, H. Z.; Liu, Y. L.; Zhao, C. J.; Deng, S. Y.; Yang, Y.; Zheng, R. L.; Li, W. W.; Zhang, N.; Liu, S. Y.; Wang, W.; Dai, L. X.; Shi, S. A.; Cheng, L.; Pan, Y. L.; Feng, S.; Zhao, X.; Deng, H. X.; Yang, S. Y.; Wei, Y. Q., SKLB1002, a Novel Potent Inhibitor of VEGF Receptor 2 Signaling, Inhibits Angiogenesis and Tumor Growth In Vivo. *Clinical Cancer Research* 2011, *17* (13), 4439-4450.
3. Cogan, D. A.; Aungst, R.; Breinlinger, E. C.; Fadra, T.; Goldberg, D. R.; Hao, M. H.; Kroe, R.; Moss, N.; Pargellis, C.; Qian, K. C.; Swinamer, A. D., Structure-based design and subsequent optimization of 2-tolyl-(1,2,3-triazol-1-yl-4-carboxamide) inhibitors of p38 MAP kinase. *Bioorganic & Medicinal Chemistry Letters* 2008, *18* (11), 3251-3255.
4. Singh, J.; Chuaqui, C. E.; Boriack-Sjodin, P. A.; Lee, W. C.; Pontz, T.; Corbley, M. J.; Cheung, H. K.; Arduini, R. M.; Mead, J. N.; Newman, M. N.; Papadatos, J. L.; Bowes, S.; Josiah, S.; Ling, L. E., Successful shape-based virtual screening: the discovery of a potent inhibitor of the type I TGFbeta receptor kinase (TbetaRI). *Bioorganic & medicinal chemistry letters* 2003, *13* (24), 4355-9.
5. Mueller, R.; Dawson, E. S.; Meiler, J.; Rodriguez, A. L.; Chauder, B. A.; Bates, B. S.; Felts, A. S.; Lamb, J. P.; Menon, U. N.; Jadhav, S. B.; Kane, A. S.; Jones, C. K.; Gregory, K. J.; Niswender, C. M.; Conn, P. J.; Olsen, C. M.; Winder, D. G.; Emmitte, K. A.; Lindsley, C. W., Discovery of 2-(2-benzoxazoyl amino)-4-aryl-5-cyanopyrimidine as negative allosteric modulators (NAMs) of metabotropic glutamate receptor 5 (mGlu(5)): from an artificial neural network virtual screen to an in vivo tool compound. *ChemMedChem* 2012, *7* (3), 406-14.
6. Van Drie, J. H., Computer-aided drug design: the next 20 years. *Journal of computer-aided molecular design* 2007, *21* (10-11), 591-601.
7. Doman, T. N.; McGovern, S. L.; Witherbee, B. J.; Kasten, T. P.; Kurumbail, R.; Stallings, W. C.; Connolly, D. T.; Shoichet, B. K., Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *Journal of medicinal chemistry* 2002, *45* (11), 2213-21.
8. Vijayakrishnan, R., Structure-based drug design and modern medicine. *Journal of postgraduate medicine* 2009, *55* (4), 301-4.
9. Talele, T. T.; Khedkar, S. A.; Rigby, A. C., Successful applications of computer aided drug discovery: moving drugs from concept to the clinic. *Current topics in medicinal chemistry* 2010, *10* (1), 127-41.
10. Hartman, G. D.; Egbertson, M. S.; Halczenko, W.; Laswell, W. L.; Duggan, M. E.; Smith, R. L.; Naylor, A. M.; Manno, P. D.; Lynch, R. J.; Zhang, G.; et al., Non-peptide fibrinogen receptor antagonists. 1. Discovery and design of exosite inhibitors. *J Med Chem* 1992, *35* (24), 4640-2.
11. Sawyer, J. S.; Anderson, B. D.; Beight, D. W.; Campbell, R. M.; Jones, M. L.; Herron, D. K.; Lampe, J. W.; McCowan, J. R.; McMillen, W. T.; Mort, N.; Parsons, S.; Smith, E. C.; Vieth, M.; Weir, L. C.; Yan, L.; Zhang, F.; Yingling, J. M., Synthesis and activity of new aryl- and heteroaryl-substituted pyrazole inhibitors of the transforming growth factor-beta type I receptor kinase domain. *Journal of medicinal chemistry* 2003, *46* (19), 3953-6.

12. Shekhar, C., In silico pharmacology: computer-aided methods could transform drug development. *Chemistry & biology* 2008, *15* (5), 413-4.
13. Kalyaanamoorthy, S.; Chen, Y. P., Structure-based drug design to augment hit discovery. *Drug discovery today* 2011, *16* (17-18), 831-9.
14. Jorgensen, W. L., Drug discovery: Pulled from a protein's embrace. *Nature* 2010, *466* (7302), 42-3.
15. Horvath, D., A virtual screening approach applied to the search for trypanothione reductase inhibitors. *Journal of Medicinal Chemistry* 1997, *40* (15), 2412-2423.
16. Ripphausen, P.; Nisius, B.; Peltason, L.; Bajorath, J., Quo vadis, virtual screening? A comprehensive survey of prospective applications. *Journal of medicinal chemistry* 2010, *53* (24), 8461-7.
17. Enyedy, I. J.; Egan, W. J., Can we use docking and scoring for hit-to-lead optimization? *Journal of computer-aided molecular design* 2008, *22* (3-4), 161-8.
18. Joffe, E., Complication during root canal therapy following accidental extrusion of sodium hypochlorite through the apical foramen. *General dentistry* 1991, *39* (6), 460-1.
19. Jorgensen, W. L., The many roles of computation in drug discovery. *Science* 2004, *303* (5665), 1813-8.
20. Basak, S. C., Chemobioinformatics: the advancing frontier of computer-aided drug design in the post-genomic era. *Current computer-aided drug design* 2012, *8* (1), 1-2.
21. Bohacek, R. S.; McMartin, C.; Guida, W. C., The art and practice of structure-based drug design: A molecular modeling perspective. *Medicinal Research Reviews* 1996, *16* (1), 3-50.
22. Schneider, G.; Hartenfeller, M.; Reutlinger, M.; Tanrikulu, Y.; Proschak, E.; Schneider, P., Voyages to the (un)known: adaptive design of bioactive compounds. *Trends Biotechnol* 2009, *27* (1), 18-26.
23. RCSB RCSB Protein Data Bank. http://www.rcsb.org/pdb/home/home.do (accessed July 20).
24. Centre, C. C. D. Cambridge Crystallographic Data Centre. http://www.ccdc.cam.ac.uk/ (accessed July 20).
25. Arnold, K.; Bordoli, L.; Kopp, J.; Schwede, T., The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* 2006, *22* (2), 195-201.
26. Durrant, J. D.; McCammon, J. A., Computer-aided drug-discovery techniques that account for receptor flexibility. *Current opinion in pharmacology* 2010, *10* (6), 770-4.
27. Meiler, J.; Baker, D., ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. *Proteins* 2006, *65* (3), 538-48.
28. Stumpfe, D.; Ripphausen, P.; Bajorath, J., Virtual compound screening in drug discovery. *Future Med Chem* 2012, *4* (5), 593-602.
29. (a) Hristozov, D. P.; Oprea, T. I.; Gasteiger, J., Virtual screening applications: a study of ligand-based methods and different structure representations in four different scenarios. *Journal of computer-aided molecular design* 2007, *21* (10-11), 617-40; (b) Cleves, A. E.; Jain, A. N., Robust ligand-based modeling of the biological targets of known drugs. *Journal of Medicinal Chemistry* 2006, *49* (10), 2921-2938.
30. Huang, N.; Shoichet, B. K.; Irwin, J. J., Benchmarking sets for molecular docking. *Journal of Medicinal Chemistry* 2006, *49* (23), 6789-6801.
31. Irwin, J. J., Community benchmarks for virtual screening. *Journal of computer-aided molecular design* 2008, *22* (3-4), 193-9.
32. Good, A. C.; Oprea, T. I., Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? *Journal of computer-aided molecular design* 2008, *22* (3-4), 169-78.
33. NIH-structure_based http://www.nigms.nih.gov/Education/structure_drugs.htm.
34. (a) Lundstrom, K., Genomics and drug discovery. *Future Medicinal Chemistry* 2011, *3* (15), 1855-1858; (b) Bambini, S.; Rappuoli, R., The use of genomics in microbial vaccine development. *Drug Discovery Today* 2009, *14* (5-6), 252-260.
35. Wang, R. X.; Fang, X. L.; Lu, Y. P.; Wang, S. M., The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *Journal of Medicinal Chemistry* 2004, *47* (12), 2977-2980.
36. (a) Warner, S. L.; Bashyam, S.; Vankayalapati, H.; Bearss, D. J.; Han, H. Y.; Von Hoff, D. D.; Hurley, L. H., Identification of a lead small-molecule inhibitor of the Aurora kinases using a structure-assisted, fragment-based approach. *Molecular Cancer Therapeutics* 2006, *5* (7), 1764-1773; (b) Budzik, B.; Garzya, V.; Shi, D.; Walker, G.; Woolley-Roberts, M.; Pardoe, J.; Lucas, A.; Tehan, B.; Rivero, R. A.; Langmead, C. J.; Watson, J.; Wu, Z.; Forbes, I. T.; Jin, J., Novel N-Substituted Benzimidazolones as Potent, Selective, CNS-Penetrant, and Orally Active M(1) mAChR Agonists. *Acs Medicinal Chemistry Letters* 2010, *1* (6), 244-248; (c) Becker, O. M.; Dhanoa, D. S.; Marantz, Y.; Chen,

D. L.; Shacham, S.; Cheruku, S.; Heifetz, A.; Mohanty, P.; Fichman, M.; Sharadendu, A.; Nudelman, R.; Kauffman, M.; Noiman, S., An integrated in silico 3D model-driven discovery of a novel, potent, and selective amidosulfonamide 5-HT1A agonist (PRX-00023) for the treatment of anxiety and depression. *Journal of Medicinal Chemistry* 2006, *49* (11), 3116-3135.

37. (a) Evers, A.; Gohlke, H.; Klebe, G., Ligand-supported homology modelling of protein binding-sites using knowledge-based potentials. *J Mol Biol* 2003, *334* (2), 327-345; (b) Evers, A.; Klebe, G., Successful virtual screening for a submicromolar antagonist of the neurokinin-1 receptor based on a ligand-supported homology model. *Journal of Medicinal Chemistry* 2004, *47* (22), 5381-5392.

38. (a) Fauman, E. B.; Rai, B. K.; Huang, E. S., Structure-based druggability assessment - identifying suitable targets for small molecule therapeutics. *Current Opinion in Chemical Biology* 2011, *15* (4), 463-468; (b) Hajduk, P. J.; Huth, J. R.; Tse, C., Predicting protein druggability. *Drug Discovery Today* 2005, *10* (23-24), 1675-1682.

39. Laurie, A. T. R.; Jackson, R. M., Methods for the prediction of protein-ligand binding sites for Structure-Based Drug Design and virtual ligand screening. *Current Protein & Peptide Science* 2006, *7* (5), 395-406.

40. Buchan, D. W.; Ward, S. M.; Lobley, A. E.; Nugent, T. C.; Bryson, K.; Jones, D. T., Protein annotation and modelling servers at University College London. *Nucleic Acids Res* 2010, *38* (Web Server issue), W563-8.

41. Marti-Renom, M. A.; Stuart, A. C.; Fiser, A.; Sanchez, R.; Melo, F.; Sali, A., Comparative protein structure modeling of genes and genomes. *Annu Rev Bioph Biom* 2000, *29*, 291-325.

42. Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J., Basic Local Alignment Search Tool. *Journal of Molecular Biology* 1990, *215* (3), 403-410.

43. (a) Soding, J.; Remmert, M., Protein sequence comparison and fold recognition: progress and good-practice benchmarking. *Current Opinion in Structural Biology* 2011, *21* (3), 404-411; (b) Kelley, L. A.; Sternberg, M. J. E., Protein structure prediction on the Web: a case study using the Phyre server. *Nat Protoc* 2009, *4* (3), 363-371.

44. Thompson, J. D.; Higgins, D. G.; Gibson, T. J., Clustal-W - Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. *Nucleic Acids Research* 1994, *22* (22), 4673-4680.

45. (a) Misura, K. M. S.; Chivian, D.; Rohl, C. A.; Kim, D. E.; Baker, D., Physically realistic homology models built with ROSETTA can be more accurate than their templates. *Proceedings of the National Academy of Sciences of the United States of America* 2006, *103* (14), 5361-5366; (b) Chivian, D.; Baker, D., Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection. *Nucleic Acids Research* 2006, *34* (17).

46. Rai, B. K.; Fiser, A., Multiple mapping method: A novel approach to the sequence-to-structure alignment problem in comparative protein structure modeling. *Proteins* 2006, *63* (3), 644-661.

47. Hillisch, A.; Pineda, L. F.; Hilgenfeld, R., Utility of homology models in the drug discovery process. *Drug Discovery Today* 2004, *9* (15), 659-669.

48. Canutescu, A. A.; Dunbrack, R. L., Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Science* 2003, *12* (5), 963-972.

49. Mandell, D. J.; Coutsias, E. A.; Kortemme, T., Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nature Methods* 2009, *6* (8), 551-552.

50. Coutsias, E. A.; Seok, C., Kinematic view of loop closure. *Abstracts of Papers of the American Chemical Society* 2004, *228*, U534-U534.

51. Krivov, G. G.; Shapovalov, M. V.; Dunbrack, R. L., Improved prediction of protein side-chain conformations with SCWRL4. *Proteins* 2009, *77* (4), 778-795.

52. Desmet, J.; Demaeyer, M.; Hazes, B.; Lasters, I., The Dead-End Elimination Theorem and Its Use in Protein Side-Chain Positioning. *Nature* 1992, *356* (6369), 539-542.

53. (a) Dunbrack, R. L.; Karplus, M., Backbone-Dependent Rotamer Library for Proteins - Application to Side-Chain Prediction. *Journal of Molecular Biology* 1993, *230* (2), 543-574; (b) Dunbrack, R. L.; Karplus, M., Conformational-Analysis of the Backbone-Dependent Rotamer Preferences of Protein Side-Chains. *Nature Structural Biology* 1994, *1* (5), 334-340; (c) Bower, M. J.; Cohen, F. E.; Dunbrack, R. L., Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: A new homology modeling tool. *Journal of Molecular Biology* 1997, *267* (5), 1268-1282.

54. Rohl, C. A.; Strauss, C. E. M.; Misura, K. M. S.; Baker, D., Protein structure prediction using ROSETTA. *Numerical Computer Methods, Pt D* 2004, *383*, 66-+.

55. (a) Tanrikulu, Y.; Schneider, G., Pseudoreceptor models in drug design: bridging ligand- and receptor-based virtual screening. *Nat Rev Drug Discov* 2008, *7* (8), 667-77; (b) Tanrikulu, Y.; Proschak, E.; Werner, T.; Geppert, T.; Todoroff, N.; Klenner, A.; Kottke, T.; Sander, K.; Schneider, E.; Seifert, R.; Stark, H.; Clark, T.; Schneider, G., Homology model adjustment and ligand screening with a pseudoreceptor of the human histamine H4 receptor. *Chemmedchem* 2009, *4* (5), 820-7.
56. Katritch, V.; Rueda, M.; Lam, P. C.; Yeager, M.; Abagyan, R., GPCR 3D homology models for ligand screening: lessons learned from blind predictions of adenosine A2a receptor complex. *Proteins-Structure Function and Genetics* 2010, *78* (1), 197-211.
57. Raval, A.; Piana, S.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E., Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. *Proteins-Structure Function and Genetics* 2012.
58. Misura, K. M. S.; Baker, D., Progress and challenges in high-resolution refinement of protein structure models. *Proteins* 2005, *59* (1), 15-29.
59. Xiang, Z., Advances in homology protein structure modeling. *Curr Protein Pept Sci* 2006, *7* (3), 217-27.
60. (a) Serrano, M. L.; Perez, H. A.; Medina, J. D., Structure of C-terminal fragment of merozoite surface protein-1 from Plasmodium vivax determined by homology modeling and molecular dynamics refinement. *Bioorg Med Chem* 2006, *14* (24), 8359-65; (b) Li, W.; Tang, Y.; Liu, H.; Cheng, J.; Zhu, W.; Jiang, H., Probing ligand binding modes of human cytochrome P450 2J2 by homology modeling, molecular dynamics simulation, and flexible molecular docking. *Proteins-Structure Function and Genetics* 2008, *71* (2), 938-49.
61. Melo, F.; Sali, A., Fold assessment for comparative protein structure modeling. *Protein Science* 2007, *16* (11), 2412-2426.
62. Cozzetto, D.; Kryshtafovych, A.; Fidelis, K.; Moult, J.; Rost, B.; Tramontano, A., Evaluation of template-based models in CASP8 with standard measures. *Proteins* 2009, *77*, 18-28.
63. Henrich, S.; Salo-Ahen, O. M. H.; Huang, B.; Rippmann, F.; Cruciani, G.; Wade, R. C., Computational approaches to identifying and characterizing protein binding sites for ligand design. *Journal of Molecular Recognition* 2010, *23* (2), 209-219.
64. (a) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R., Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* 2002, *47* (4), 409-43; (b) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J., Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nature Reviews Drug Discovery* 2004, *3* (11), 935-949.
65. Taylor, J. S.; Burnett, R. M., DARWIN: a program for docking flexible molecules. *Proteins* 2000, *41* (2), 173-91.
66. Connolly, M. L., Analytical Molecular-Surface Calculation. *Journal of Applied Crystallography* 1983, *16* (Oct), 548-558.
67. Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E., A Geometric Approach to Macromolecule-Ligand Interactions. *Journal of Molecular Biology* 1982, *161* (2), 269-288.
68. Katchalskikatzir, E.; Shariv, I.; Eisenstein, M.; Friesem, A. A.; Aflalo, C.; Vakser, I. A., Molecular-Surface Recognition - Determination of Geometric Fit between Proteins and Their Ligands by Correlation Techniques. *Proceedings of the National Academy of Sciences of the United States of America* 1992, *89* (6), 2195-2199.
69. Dias, R.; de Azevedo, W. F., Molecular Docking Algorithms. *Current Drug Targets* 2008, *9* (12), 1040-1047.
70. Changeux, J. P.; Edelstein, S., Conformational selection or induced fit? 50 years of debate resolved. *F1000 Biol Rep* 2011, *3*, 19.
71. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S., Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of Medicinal Chemistry* 2004, *47* (7), 1739-1749.
72. Desjarlais, R. L.; Sheridan, R. P.; Dixon, J. S.; Kuntz, I. D.; Venkataraghavan, R., Docking Flexible Ligands to Macromolecular Receptors by Molecular Shape. *Journal of Medicinal Chemistry* 1986, *29* (11), 2149-2153.
73. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G., A fast flexible docking method using an incremental construction algorithm. *Journal of Molecular Biology* 1996, *261* (3), 470-489.
74. Miller, M. D.; Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P., Flog - a System to Select Quasi-Flexible Ligands Complementary to a Receptor of Known 3-Dimensional Structure. *Journal of Computer-Aided Molecular Design* 1994, *8* (2), 153-174.
75. Jain, A. N., Surflex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. *Journal of Medicinal Chemistry* 2003, *46* (4), 499-511.

76. Majeux, N.; Scarsi, M.; Caflisch, A., Efficient electrostatic solvation model for protein-fragment docking. *Proteins-Structure Function and Genetics* 2001, *42* (2), 256-268.
77. Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D., Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins-Structure Function and Genetics* 2004, *57* (2), 225-42.
78. Cross, J. B.; Thompson, D. C.; Rai, B. K.; Baber, J. C.; Fan, K. Y.; Hu, Y. B.; Humblet, C., Comparison of Several Molecular Docking Programs: Pose Prediction and Virtual Screening Accuracy. *J Chem Inf Model* 2009, *49* (6), 1455-1474.
79. Pierce, A. C.; Jacobs, M.; Stuver-Moody, C., Docking study yields four novel inhibitors of the protooncogene Pim-1 kinase. *Journal of Medicinal Chemistry* 2008, *51* (6), 1972-1975.
80. Chiu, T. L.; Solberg, J.; Patil, S.; Geders, T. W.; Zhang, X.; Rangarajan, S.; Francis, R.; Finzel, B. C.; Walters, M. A.; Hook, D. J.; Amin, E. A., Identification of Novel Non-Hydroxamate Anthrax Toxin Lethal Factor Inhibitors by Topomeric Searching, Docking and Scoring, and in Vitro Screening. *Journal of Chemical Information and Modeling* 2009, *49* (12), 2726-2734.
81. Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J., DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research* 2006, *34*, D668-D672.
82. Irwin, J. J.; Shoichet, B. K., ZINC - A free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling* 2005, *45* (1), 177-182.
83. Milne, G. W. A.; Nicklaus, M. C.; Driscoll, J. S.; Wang, S. M.; Zaharevitz, D., National-Cancer-Institute Drug Information-System 3d Database. *Journal of Chemical Information and Computer Sciences* 1994, *34* (5), 1219-1224.
84. Mangoni, R.; Roccatano, D.; Di Nola, A., Docking of flexible ligands to flexible receptors in solution by molecular dynamics simulation. *Proteins* 1999, *35* (2), 153-162.
85. (a) Durrant, J. D.; Urbaniak, M. D.; Ferguson, M. A. J.; McCammon, J. A., Computer-Aided Identification of Trypanosoma brucei Uridine Diphosphate Galactose 4 '-Epimerase Inhibitors: Toward the Development of Novel Therapies for African Sleeping Sickness. *Journal of Medicinal Chemistry* 2010, *53* (13), 5025-5032; (b) Amaro, R. E.; Schnaufer, A.; Interthal, H.; Hol, W.; Stuart, K. D.; McCammon, J. A., Discovery of drug-like inhibitors of an essential RNA-editing ligase in Trypanosoma brucei. *Proceedings of the National Academy of Sciences of the United States of America* 2008, *105* (45), 17278-17283.
86. (a) Leone, V.; Marinelli, F.; Carloni, P.; Parrinello, M., Targeting biomolecular flexibility with metadynamics. *Curr Opin Struc Biol* 2010, *20* (2), 148-154; (b) Biarnes, X.; Bongarzone, S.; Vargiu, A. V.; Carloni, P.; Ruggerone, P., Molecular motions in drug design: the coming age of the metadynamics method. *J Comput Aid Mol Des* 2011, *25* (5), 395-402.
87. Shaw, D. E.; Deneroff, M. M.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J.; Chao, J. C.; Eastwood, M. P.; Gagliardo, J.; Grossman, J. P.; Ho, C. R.; Ierardi, D. J.; Kolossvary, I.; Klepeis, J. L.; Layman, T.; Mcleavey, C.; Moraes, M. A.; Mueller, R.; Priest, E. C.; Shan, Y. B.; Spengler, J.; Theobald, M.; Towles, B.; Wang, S. C., Anton, a special-purpose machine for molecular dynamics simulation. *Commun Acm* 2008, *51* (7), 91-97.
88. Shan, Y. B.; Kim, E. T.; Eastwood, M. P.; Dror, R. O.; Seeliger, M. A.; Shaw, D. E., How Does a Drug Molecule Find Its Target Binding Site? *J Am Chem Soc* 2011, *133* (24), 9181-9183.
89. Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E., How Fast-Folding Proteins Fold. *Science* 2011, *334* (6055), 517-520.
90. Sousa, S. F.; Fernandes, P. A.; Ramos, M. J., Protein-ligand docking: Current status and future challenges. *Proteins* 2006, *65* (1), 15-26.
91. Liu, M.; Wang, S. M., MCDOCK: A Monte Carlo simulation approach to the molecular docking problem. *Journal of Computer-Aided Molecular Design* 1999, *13* (5), 435-451.
92. Abagyan, R.; Totrov, M.; Kuznetsov, D., Icm - a New Method for Protein Modeling and Design - Applications to Docking and Structure Prediction from the Distorted Native Conformation. *Journal of Computational Chemistry* 1994, *15* (5), 488-506.
93. Davis, I. W.; Baker, D., ROSETTALigand docking with full ligand and receptor flexibility. *J Mol Biol* 2009, *385* (2), 381-92.
94. (a) Kaufmann, K. W.; Lemmon, G. H.; Deluca, S. L.; Sheehan, J. H.; Meiler, J., Practically useful: what the ROSETTA protein modeling suite can do for you. *Biochemistry* 2010, *49* (14), 2987-98; (b) ROSETTACommons ROSETTA - The premier software suite for macromolecular modeling. http://www.ROSETTAcommons.org/ (accessed July 20).

95. McMartin, C.; Bohacek, R. S., QXP: Powerful, rapid computer algorithms for structure-based drug design. *Journal of Computer-Aided Molecular Design* 1997, *11* (4), 333-344.
96. Totrov, M.; Abagyan, R., Flexible protein-ligand docking by global energy optimization in internal coordinates. *Proteins* 1997, *Suppl 1*, 215-20.
97. Kaufmann, K. W.; Dawson, E. S.; Henry, L. K.; Field, J. R.; Blakely, R. D.; Meiler, J., Structural determinants of species-selective substrate recognition in human and Drosophila serotonin transporters revealed through computational docking studies. *Proteins* 2009, *74* (3), 630-42.
98. (a) Malamas, M. S.; Erdei, J.; Gunawan, I.; Barnes, K.; Johnson, M.; Hui, Y.; Turner, J.; Hu, Y.; Wagner, E.; Fan, K.; Olland, A.; Bard, J.; Robichaud, A. J., Aminoimidazoles as Potent and Selective Human beta-Secretase (BACE1) Inhibitors. *Journal of Medicinal Chemistry* 2009, *52* (20), 6314-6323; (b) Nowak, P.; Cole, D. C.; Aulabaugh, A.; Bard, J.; Chopra, R.; Cowling, R.; Fan, K. Y.; Hu, B. H.; Jacobsen, S.; Jani, M.; Jin, G. X.; Lo, M. C.; Malamas, M. S.; Manas, E. S.; Narasimhan, R.; Reinhart, P.; Robichaud, A. J.; Stock, J. R.; Subrath, J.; Svenson, K.; Turner, J.; Wagner, E.; Zhou, P.; Ellingboe, J. W., Discovery and initial optimization of 5,5 '-disubstituted aminohydantoins as potent beta-secretase (BACE1) inhibitors. *Bioorganic & Medicinal Chemistry Letters* 2010, *20* (2), 632-635; (c) Malamas, M. S.; Barnes, K.; Hui, Y.; Johnson, M.; Lovering, F.; Condon, J.; Fobare, W.; Solvibile, W.; Turner, J.; Hu, Y.; Manas, E. S.; Fan, K.; Olland, A.; Chopra, R.; Bard, J.; Pangalos, M. N.; Reinhart, P.; Robichaud, A. J., Novel pyrrolyl 2-aminopyridines as potent and selective human beta-secretase (BACE1) inhibitors. *Bioorganic & Medicinal Chemistry Letters* 2010, *20* (7), 2068-2073.
99. Chan, D. S. H.; Lee, H. M.; Yang, F.; Che, C. M.; Wong, C. C. L.; Abagyan, R.; Leung, C. H.; Ma, D. L., Structure-Based Discovery of Natural-Product-like TNF-alpha Inhibitors. *Angewandte Chemie-International Edition* 2010, *49* (16), 2860-2864.
100. An, J. H.; Lee, D. C. W.; Law, A. H. Y.; Yang, C. L. H.; Poon, L. L. M.; Lau, A. S. Y.; Jones, S. J. M., A Novel Small-Molecule Inhibitor of the Avian Influenza H5N1 Virus Determined through Computational Screening against the Neuraminidase. *Journal of Medicinal Chemistry* 2009, *52* (9), 2667-2672.
101. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R., Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology* 1997, *267* (3), 727-748.
102. Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Olson, A., Automated Docking Using a Lamarckian Genetic Algotithm and an Empirical Binding Free energy Function. *J. Comp. Chem.* 1998, *19* (14)), 1639-1662.
103. (a) Kontoyianni, M.; McClellan, L. M.; Sokol, G. S., Evaluation of docking performance: Comparative data on docking algorithms. *J Med Chem* 2004, *47* (3), 558-565; (b) Li, X.; Li, Y.; Cheng, T. J.; Liu, Z. H.; Wang, R. X., Evaluation of the Performance of Four Molecular Docking Programs on a Diverse Set of Protein-Ligand Complexes. *J Comput Chem* 2010, *31* (11), 2109-2125.
104. Park, H.; Hwang, K. Y.; Kim, Y. H.; Oh, K. H.; Lee, J. Y.; Kim, K., Discovery and biological evaluation of novel alpha-glucosidase inhibitors with in vivo antidiabetic effect. *Bioorganic & Medicinal Chemistry Letters* 2008, *18* (13), 3711-3715.
105. Durrant, J. D.; Hall, L.; Swift, R. V.; Landon, M.; Schnaufer, A.; Amaro, R. E., Novel Naphthalene-Based Inhibitors of Trypanosoma brucei RNA Editing Ligase 1. *Plos Neglected Tropical Diseases* 2010, *4* (8).
106. (a) B-Rao, C.; Subramanian, J.; Sharma, S. D., Managing protein flexibility in docking and its applications. *Drug Discov Today* 2009, *14* (7-8), 394-400; (b) Sinko, W.; Lindert, S.; McCammon, J. A., Accounting for Receptor Flexibility and Enhanced Sampling Methods in Computer-Aided Drug Design. *Chem Biol Drug Des* 2013, *81* (1), 41-49.
107. Carlson, H. A., Protein flexibility and drug design: how to hit a moving target. *Curr Opin Chem Biol* 2002, *6* (4), 447-452.
108. Abagyan, R.; Lee, W. H.; Raush, E.; Budagyan, L.; Totrov, M.; Sundstrom, M.; Marsden, B. D., Disseminating structural genomics data to the public: from a data dump to an animated story. *Trends in Biochemical Sciences* 2006, *31* (2), 76-78.
109. Cozzini, P.; Kellogg, G. E.; Spyrakis, F.; Abraham, D. J.; Costantino, G.; Emerson, A.; Fanelli, F.; Gohlke, H.; Kuhn, L. A.; Morris, G. M.; Orozco, M.; Pertinhez, T. A.; Rizzi, M.; Sotriffer, C. A., Target Flexibility: An Emerging Consideration in Drug Discovery and Design. *J Med Chem* 2008, *51* (20), 6237-6255.
110. Schames, J. R.; Henchman, R. H.; Siegel, J. S.; Sotriffer, C. A.; Ni, H. H.; McCammon, J. A., Discovery of a novel binding trench in HIV integrase. *Journal of Medicinal Chemistry* 2004, *47* (8), 1879-1881.
111. Halgren, T. A., Merck molecular force field .1. Basis, form, scope, parameterization, and performance of MMFF94. *J Comput Chem* 1996, *17* (5-6), 490-519.

112. Shoichet, B. K.; Leach, A. R.; Kuntz, I. D., Ligand solvation in molecular docking. *Proteins-Structure Function and Genetics* 1999, *34* (1), 4-16.
113. Kukic, P.; Nielsen, J. E., Electrostatics in proteins and protein-ligand complexes. *Future Med Chem* 2010, *2* (4), 647-666.
114. Huey, R.; Morris, G. M.; Olson, A. J.; Goodsell, D. S., A semiempirical free energy force field with charge-based desolvation. *J Comput Chem* 2007, *28* (6), 1145-1152.
115. Bohm, H. J., The Computer-Program Ludi - a New Method for the Denovo Design of Enzyme-Inhibitors. *Journal of Computer-Aided Molecular Design* 1992, *6* (1), 61-78.
116. Shimada, J.; Ishchenko, A. V.; Shakhnovich, E. I., Analysis of knowledge-based protein-ligand potentials using a self-consistent method. *Protein Science* 2000, *9* (4), 765-775.
117. Velec, H. F. G.; Gohlke, H.; Klebe, G., DrugScore(CSD)-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *Journal of Medicinal Chemistry* 2005, *48* (20), 6296-6303.
118. DeWitte, R. S.; Shakhnovich, E., SMoG: De novo design method based on simple, fast and accurate free energy estimates. *Abstracts of Papers of the American Chemical Society* 1997, *214*, 6-Comp.
119. Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Forster, M. J.; Thornton, J. M., BLEEP - Potential of mean force describing protein-ligand interactions: II. Calculation of binding energies and comparison with experimental data. *Journal of Computational Chemistry* 1999, *20* (11), 1177-1185.
120. Feher, M., Consensus scoring for protein-ligand interactions. *Drug Discovery Today* 2006, *11* (9-10), 421-428.
121. O'Boyle, N. M.; Liebeschuetz, J. W.; Cole, J. C., Testing Assumptions and Hypotheses for Rescoring Success in Protein-Ligand Docking. *Journal of Chemical Information and Modeling* 2009, *49* (8), 1871-1878.
122. Okamoto, M.; Takayama, K.; Shimizu, T.; Ishida, K.; Takahashi, O.; Furuya, T., Identification of Death-Associated Protein Kinases Inhibitors Using Structure-Based Virtual Screening. *Journal of Medicinal Chemistry* 2009, *52* (22), 7323-7327.
123. Muegge, I.; Martin, Y. C., A general and fast scoring function for protein-ligand interactions: A simplified potential approach. *Journal of Medicinal Chemistry* 1999, *42* (5), 791-804.
124. Friedman, R.; Caflisch, A., Discovery of Plasmepsin Inhibitors by Fragment-Based Docking and Consensus Scoring. *Chemmedchem* 2009, *4* (8), 1317-1326.
125. Yang, S. Y., Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug Discovery Today* 2010, *15* (11-12), 444-450.
126. Wolber, G.; Langer, T., LigandScout: 3-d pharmacophores derived from protein-bound Ligands and their use as virtual screening filters. *Journal of Chemical Information and Modeling* 2005, *45* (1), 160-169.
127. Chen, J.; Lai, L. H., Pocket v.2: Further developments on receptor-based pharmacophore modeling. *Journal of Chemical Information and Modeling* 2006, *46* (6), 2684-2691.
128. Wang, R. X.; Liu, L.; Lai, L. H.; Tang, Y. Q., SCORE: A new empirical method for estimating the binding affinity of a protein-ligand complex. *Journal of Molecular Modeling* 1998, *4* (12), 379-394.
129. Schuster, D.; Nashev, L. G.; Kirchmair, J.; Laggner, C.; Wolber, G.; Langer, T.; Odermatt, A., Discovery of nonsteroidal 17 beta-hydroxysteroid dehydrogenase 1 inhibitors by pharmacophore-based screening of virtual compound libraries. *Journal of Medicinal Chemistry* 2008, *51* (14), 4188-4199.
130. Brvar, M.; Perdih, A.; Oblak, M.; Masic, L. P.; Solmajer, T., In silico discovery of 2-amino-4-(2,4-dihydroxyphenyl)thiazoles as novel inhibitors of DNA gyrase B. *Bioorganic & Medicinal Chemistry Letters* 2010, *20* (3), 958-962.
131. Bowman, A. L.; Nikolovska-Coleska, Z.; Zhong, H. Z.; Wang, S. M.; Carlson, H. A., Small molecule inhibitors of the MDM2-p53 interaction discovered by ensemble-based receptor models. *Journal of the American Chemical Society* 2007, *129* (42), 12809-12814.
132. Deng, J. X.; Lee, K. W.; Sanchez, T.; Cui, M.; Neamati, N.; Briggs, J. M., Dynamic receptor-based pharmacophore model development and its application in designing novel HIT-1 integrase inhibitors. *Journal of Medicinal Chemistry* 2005, *48* (5), 1496-1505.
133. Fesik, S. W.; Shuker, S. B.; Hajduk, P. J.; Meadows, R. P., SAR by NMR: An NMR-based approach for drug discovery. *Protein Engineering* 1997, *10*, 73-73.
134. Wang, R. X.; Gao, Y.; Lai, L. H., LigBuilder: A multi-purpose program for structure-based drug design. *Journal of Molecular Modeling* 2000, *6* (7-8), 498-516.

135. Yuan, Y. X.; Pei, J. F.; Lai, L. H., LigBuilder 2: A Practical de Novo Drug Design Approach. *Journal of Chemical Information and Modeling* 2011, *51* (5), 1083-1091.
136. Vinkers, H. M.; de Jonge, M. R.; Daeyaert, F. F.; Heeres, J.; Koymans, L. M.; van Lenthe, J. H.; Lewi, P. J.; Timmerman, H.; Van Aken, K.; Janssen, P. A., SYNOPSIS: SYNthesize and OPtimize System in Silico. *Journal of medicinal chemistry* 2003, *46* (13), 2765-73.
137. Krier, M.; Araujo-Junior, J. X.; Schmitt, M.; Duranton, J.; Justiano-Basaran, H.; Lugnier, C.; Bourguignon, J. J.; Rognan, D., Design of small-sized libraries by combinatorial assembly of linkers and functional groups to a given scaffold: application to the structure-based optimization of a phosphodiesterase 4 inhibitor. *Journal of medicinal chemistry* 2005, *48* (11), 3816-22.
138. (a) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M., RECAP--retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J Chem Inf Comput Sci* 1998, *38* (3), 511-22; (b) Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; Rarey, M., On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem* 2008, *3* (10), 1503-7.
139. Li, W. W.; Chen, J. J.; Zheng, R. L.; Zhang, W. Q.; Cao, Z. X.; Yang, L. L.; Qing, X. Y.; Zhou, L. X.; Yang, L.; Yu, L. D.; Chen, L. J.; Wei, Y. Q.; Yang, S. Y., Taking Quinazoline as a General Support-Nog to Design Potent and Selective Kinase Inhibitors: Application to FMS-like Tyrosine Kinase 3. *Chemmedchem* 2010, *5* (4), 513-516.
140. Johnson, M. A.; Maggiora, G. M.; American Chemical Society. Meeting, *Concepts and applications of molecular similarity*. Wiley: New York, 1990; p xix, 393 p.
141. (a) Hemmer, M. C.; Steinhauer, V.; Gasteiger, J., Deriving the 3D structure of organic molecules from their infrared spectra. *Vibrational Spectroscopy* 1999, *19* (1), 151-164; (b) Schuur, J. H.; Selzer, P.; Gasteiger, J., The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. *Journal of Chemical Information and Computer Sciences* 1996, *36* (2), 334-344; (c) Pearlman, R. S.; Smith, K. M., Metric validation and the receptor-relevant subspace concept. *Journal of Chemical Information and Computer Sciences* 1999, *39* (1), 28-35; (d) Bravi, G.; Gancia, E.; Mascagni, P.; Pegna, M.; Todeschini, R.; Zaliani, A., MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: A comparative 3D QSAR study in a series of steroids. *J Comput Aided Mol Des* 1997, *11* (1), 79-92; (e) Randic, M., Molecular Profiles - Novel Geometry-Dependent Molecular Descriptors. *New Journal of Chemistry* 1995, *19* (7), 781-791; (f) Roberto Todeschini, V. C., *Molecular Descriptors for Chemoinformatics*
Wiley-VCH Verlag GmbH & Co. KGaA: 2010; p 1-38; (g) Hong, H.; Xie, Q.; Ge, W.; Qian, F.; Fang, H.; Shi, L.; Su, Z.; Perkins, R.; Tong, W., Mold(2), molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. *Journal of Chemical Information and Modeling* 2008, *48* (7), 1337-44; (h) Cramer, R. D.; Patterson, D. E.; Bunce, J. D., Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society* 1988, *110* (18), 5959-67.
142. (a) Acharya, C.; Coop, A.; Polli, J. E.; Mackerell, A. D., Jr., Recent advances in ligand-based drug design: relevance and utility of the conformationally sampled pharmacophore approach. *Current computer-aided drug design* 2011, *7* (1), 10-22; (b) Marrero-Ponce, Y.; Santiago, O. M.; Lopez, Y. M.; Barigye, S. J.; Torrens, F., Derivatives in discrete mathematics: a novel graph-theoretical invariant for generating new 2/3D molecular descriptors. I. Theory and QSPR application. *J Comput Aided Mol Des* 2012, *26* (11), 1229-46.
143. Ekins, S.; Mestres, J.; Testa, B., In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. *British journal of pharmacology* 2007, *152* (1), 9-20.
144. March, J., *Advanced organic chemistry : reactions, mechanisms, and structure*. 2d ed.; McGraw-Hill: New York, 1977; p xv, 1328 p.
145. (a) Pimentel, G. C.; McClellan, A. L., *The hydrogen bond*. W.H. Freeman: 1960; (b) Vinogradov, S. N.; Linnell, R. H., *Hydrogen bonding*. Van Nostrand Reinhold: New York,, 1971; p xi, 319 p.
146. Zhou, T.; Huang, D.; Caflisch, A., Quantum mechanical methods for drug design. *Current topics in medicinal chemistry* 2010, *10* (1), 33-45.
147. (a) Bajorath, J., Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J Chem Inf Comput Sci* 2001, *41* (2), 233-45; (b) Bajorath, J., Integration of virtual and high-throughput screening. *Nat Rev Drug Discov* 2002, *1* (11), 882-94.
148. Auer, J.; Bajorath, J., Molecular similarity concepts and search calculations. *Methods in molecular biology* 2008, *453*, 327-47.

149. Hutter, M. C., Graph-based similarity concepts in virtual screening. *Future medicinal chemistry* 2011, *3* (4), 485-501.
150. Barnard, J. M.; Downs, G. M., Chemical fragment generation and clustering software. *Journal of Chemical Information and Computer Sciences* 1997, *37* (1), 141-142.
151. Trinajstić, N., *Chemical graph theory*. 2nd ed.; CRC Press: Boca Raton, 1992; p 322 p.
152. Devillers, J.; Balaban, A. T., *Topological indices and related descriptors in QSAR and QSPR*. Gordon and Breach: Amsterdam, 1999; p x, 811 p.
153. Bertz, S. H., On the Complexity of Graphs and Molecules. *Bulletin of Mathematical Biology* 1983, *45* (5), 849-855.
154. Moreau, G.; Broto, P., The Auto-Correlation of a Topological-Structure - a New Molecular Descriptor. *Nouveau Journal De Chimie-New Journal of Chemistry* 1980, *4* (6), 359-360.
155. Kubinyi, H.; Folkers, G.; Martin, Y. C., *3D QSAR in drug design*. Kluwer Academic: Dordrecht ; Boston, Mass, 1998; p v. < 2- >.
156. Broto, P.; Moreau, G.; Vandycke, C., Molecular-Structures - Perception, Auto-Correlation Descriptor and Sar Studies - Perception of Molecules - Topological-Structure and 3-Dimensional Structure. *European Journal of Medicinal Chemistry* 1984, *19* (1), 61-65.
157. Willett, P., Similarity-based virtual screening using 2D fingerprints. *Drug discovery today* 2006, *11* (23-24), 1046-53.
158. (a) Osborne, C. K.; Schiff, R., Estrogen-receptor biology: continuing progress and therapeutic implications. *J Clin Oncol* 2005, *23* (8), 1616-22; (b) Hall, J. M.; Couse, J. F.; Korach, K. S., The multifaceted mechanisms of estradiol and estrogen receptor signaling. *J Biol Chem* 2001, *276* (40), 36869-72.
159. Revankar, C. M.; Cimino, D. F.; Sklar, L. A.; Arterburn, J. B.; Prossnitz, E. R., A transmembrane intracellular estrogen receptor mediates rapid cell signaling. *Science* 2005, *307* (5715), 1625-30.
160. Bologa, C. G.; Revankar, C. M.; Young, S. M.; Edwards, B. S.; Arterburn, J. B.; Kiselyov, A. S.; Parker, M. A.; Tkachenko, S. E.; Savchuck, N. P.; Sklar, L. A.; Oprea, T. I.; Prossnitz, E. R., Virtual and biomolecular screening converge on a selective agonist for GPR30. *Nature Chemical Biology* 2006, *2* (4), 207-212.
161. McGregor, M. J.; Pallai, P. V., Clustering of large databases of compounds: Using the MDL ''keys'' as structural descriptors. *Journal of Chemical Information and Computer Sciences* 1997, *37* (3), 443-448.
162. Zhang, S., Computer-aided drug discovery and development. *Methods in molecular biology* 2011, *716*, 23-38.
163. Hansch, C., Citation Classic - Rho-Sigma-Pi-Analysis - a Method for the Correlation of Biological-Activity and Chemical-Structure. *Current Contents/Life Sciences* 1982, (47), 18-18.
164. (a) Free, S. M., Jr.; Wilson, J. W., A Mathematical Contribution to Structure-Activity Studies. *Journal of Medicinal Chemistry* 1964, *7*, 395-9; (b) Tmej, C.; Chiba, P.; Huber, M.; Richter, E.; Hitzler, M.; Schaper, K. J.; Ecker, G., A combined Hansch/Free-Wilson approach as predictive tool in QSAR studies on propafenone-type modulators of multidrug resistance. *Arch Pharm (Weinheim)* 1998, *331* (7-8), 233-40.
165. Klebe, G.; Abraham, U.; Mietzner, T., Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *Journal of Medicinal Chemistry* 1994, *37* (24), 4130-46.
166. Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B. Q.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C., Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *Journal of the American Chemical Society* 1997, *119* (43), 10509-10524.
167. Vedani, A.; Dobler, M., 5D-QSAR: the key for simulating induced fit? *J Med Chem* 2002, *45* (11), 2139-49.
168. Pan, D.; Tseng, Y.; Hopfinger, A. J., Quantitative structure-based design: formalism and application of receptor-dependent RD-4D-QSAR analysis to a set of glucose analogue inhibitors of glycogen phosphorylase. *J Chem Inf Comput Sci* 2003, *43* (5), 1591-607.
169. (a) Wold, S.; Esbensen, K.; Geladi, P., Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems* 1987, *2* (1-3), 37-52; (b) Kubinyi, H., QSAR and 3D QSAR in drug design .1. methodology. *Drug Discovery Today* 1997, *2* (11), 457-467.
170. Zheng, W.; Tropsha, A., Novel variable selection quantitative structure--property relationship approach based on the k-nearest-neighbor principle. *J Chem Inf Comput Sci* 2000, *40* (1), 185-94.
171. Livingstone, D., *Artificial neural networks : methods and applications*. Humana Press: Totowa, NJ, 2008; p ix, 254 p.

172. (a) Vapnik, V.; Lerner, A., Pattern Recognition using Generalized Portrait Method. *Automation and Remote Control* 1963, *24*; (b) Boser, B. E.; Guyon, I. M.; Vapnik, V. N., A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, ACM: Pittsburgh, Pennsylvania, United States, 1992; pp 144-152.
173. (a) Blumer, A.; Ehrenfeucht, A.; Haussler, D.; Warmuth, M. K., Learnability and the Vapnik-Chervonenkis dimension. *J. ACM* 1989, *36* (4), 929-965; (b) Vapnik, V. N., An overview of statistical learning theory. *Neural Networks, IEEE Transactions on* 1999, *10* (5), 988-999.
174. (a) Ivanciuc, O., Applications of Support Vector Machines in Chemistry. In *Reviews in Computational Chemistry*, John Wiley & Sons, Inc.: 2007; pp 291-400; (b) Liang, Y., *Support vector machines and their application in chemistry and biotechnology*. CRC Press: Boca Raton, 2011; p x, 201 p; (c) Boyle, B. H., *Support vector machines : data analysis, machine learning, and applications*. Nova Science Publishers: New York, 2011; p x, 202 p.
175. Cristianini, N.; Shawe-Taylor, J., *An introduction to support vector machines : and other kernel-based learning methods*. Cambridge University Press: Cambridge ; New York, 2000; p xiii, 189 p.
176. (a) Han, J.; Kamber, M., *Data mining : concepts and techniques*. 2nd ed.; Elsevier ;
Morgan Kaufmann: Amsterdam ; Boston
San Francisco, CA, 2006; p xxviii, 770 p; (b) Mitchell, T. M., *Machine Learning*. McGraw-Hill: New York, 1997; p xvii, 414 p.
177. (a) Fukunishi, Y., Structure-Based Drug Screening and Ligand-Based Drug Screening with Machine Learning. *Combinatorial Chemistry & High Throughput Screening* 2009, *12* (4), 397-408; (b) Quinlan, J. R., *C4.5 : programs for machine learning*. Morgan Kaufmann Publishers: San Mateo, Calif., 1993; p x, 302 p.
178. (a) Gasparini, F.; Bilbe, G.; Gomez-Mancilla, B.; Spooren, W., mGluR5 antagonists: discovery, characterization and drug development. *Current opinion in drug discovery & development* 2008, *11* (5), 655-65; (b) Conn, P. J.; Christopoulos, A.; Lindsley, C. W., Allosteric modulators of GPCRs: a novel approach for the treatment of CNS disorders. *Nature reviews. Drug discovery* 2009, *8* (1), 41-54.
179. Mueller, R.; Rodriguez, A. L.; Dawson, E. S.; Butkiewicz, M.; Nguyen, T. T.; Oleszkiewicz, S.; Bleckmann, A.; Weaver, C. D.; Lindsley, C. W.; Conn, P. J.; Meiler, J., Identification of Metabotropic Glutamate Receptor Subtype 5 Potentiators Using Virtual High-Throughput Screening. *ACS Chem Neurosci* 2010, *1* (4), 288-305.
180. Rodriguez, A. L.; Grier, M. D.; Jones, C. K.; Herman, E. J.; Kane, A. S.; Smith, R. L.; Williams, R.; Zhou, Y.; Marlo, J. E.; Days, E. L.; Blatt, T. N.; Jadhav, S.; Menon, U. N.; Vinson, P. N.; Rook, J. M.; Stauffer, S. R.; Niswender, C. M.; Lindsley, C. W.; Weaver, C. D.; Conn, P. J., Discovery of novel allosteric modulators of metabotropic glutamate receptor subtype 5 reveals chemical and functional diversity and in vivo activity in rat behavioral models of anxiolytic and antipsychotic activity. *Molecular pharmacology* 2010, *78* (6), 1105-23.
181. Olsen, C. M.; Childs, D. S.; Stanwood, G. D.; Winder, D. G., Operant sensation seeking requires metabotropic glutamate receptor 5 (mGluR5). *PloS one* 2010, *5* (11), e15085.
182. Deacon, R. M., Digging and marble burying in mice: simple methods for in vivo identification of biological impacts. *Nature protocols* 2006, *1* (1), 122-4.
183. Ke, Y. Y.; Shiao, H. Y.; Hsu, Y. C.; Chu, C. Y.; Wang, W. C.; Lee, Y. C.; Lin, W. H.; Chen, C. H.; Hsu, J. T.; Chang, C. W.; Lin, C. W.; Yeh, T. K.; Chao, Y. S.; Coumar, M. S.; Hsieh, H. P., 3D-QSAR-assisted drug design: identification of a potent quinazoline-based Aurora kinase inhibitor. *ChemMedChem* 2013, *8* (1), 136-48.
184. (a) Coumar, M. S.; Leou, J. S.; Shukla, P.; Wu, J. S.; Dixit, A. K.; Lin, W. H.; Chang, C. Y.; Lien, T. W.; Tan, U. K.; Chen, C. H.; Hsu, J. T.; Chao, Y. S.; Wu, S. Y.; Hsieh, H. P., Structure-based drug design of novel Aurora kinase A inhibitors: structural basis for potency and specificity. *J Med Chem* 2009, *52* (4), 1050-62; (b) Coumar, M. S.; Chu, C. Y.; Lin, C. W.; Shiao, H. Y.; Ho, Y. L.; Reddy, R.; Lin, W. H.; Chen, C. H.; Peng, Y. H.; Leou, J. S.; Lien, T. W.; Huang, C. T.; Fang, M. Y.; Wu, S. H.; Wu, J. S.; Chittimalla, S. K.; Song, J. S.; Hsu, J. T.; Wu, S. Y.; Liao, C. C.; Chao, Y. S.; Hsieh, H. P., Fast-forwarding hit to lead: aurora and epidermal growth factor receptor kinase inhibitor lead identification. *J Med Chem* 2010, *53* (13), 4980-8; (c) Coumar, M. S.; Tsai, M. T.; Chu, C. Y.; Uang, B. J.; Lin, W. H.; Chang, C. Y.; Chang, T. Y.; Leou, J. S.; Teng, C. H.; Wu, J. S.; Fang, M. Y.; Chen, C. H.; Hsu, J. T.; Wu, S. Y.; Chao, Y. S.; Hsieh, H. P., Identification, SAR studies, and X-ray co-crystallographic analysis of a novel furanopyrimidine aurora kinase A inhibitor. *ChemMedChem* 2010, *5* (2), 255-67.
185. Li, M.; Jung, A.; Ganswindt, U.; Marini, P.; Friedl, A.; Daniel, P. T.; Lauber, K.; Jendrossek, V.; Belka, C., Aurora kinase inhibitor ZM447439 induces apoptosis via mitochondrial pathways. *Biochem Pharmacol* 2010, *79* (2), 122-9.

186. (a) Fu, J.; Bian, M.; Jiang, Q.; Zhang, C., Roles of Aurora kinases in mitosis and tumorigenesis. *Mol Cancer Res* 2007, *5* (1), 1-10; (b) Agnese, V.; Bazan, V.; Fiorentino, F. P.; Fanale, D.; Badalamenti, G.; Colucci, G.; Adamo, V.; Santini, D.; Russo, A., The role of Aurora-A inhibitors in cancer therapy. *Ann Oncol* 2007, *18 Suppl 6*, vi47-52.
187. Kurup, A., C-QSAR: a database of 18,000 QSARs and associated biological and physical data. *J Comput Aided Mol Des* 2003, *17* (2-4), 187-96.
188. Kim, K. H., Outliers in SAR and QSAR: 2. Is a flexible binding site a possible source of outliers? *J Comput Aided Mol Des* 2007, *21* (8), 421-35.
189. KENT, J. T., Information gain and a general measure of correlation. *Biometrika* 1983, *70* (1), 163-173.
190. Chen, Y.-W.; Lin, C.-J., Combining SVMs with Various Feature Selection Strategies
Feature Extraction. Guyon, I.; Nikravesh, M.; Gunn, S.; Zadeh, L., Eds. Springer Berlin / Heidelberg: 2006; Vol. 207, pp 315-324.
191. Mao, K. Z., Orthogonal forward selection and backward elimination algorithms for feature subset selection. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 2004, *34* (1), 629-634.
192. (a) Goodarzi, M.; Freitas, M. P.; Jensen, R., Feature selection and linear/nonlinear regression methods for the accurate prediction of glycogen synthase kinase-3beta inhibitory activities. *J Chem Inf Model* 2009, *49* (4), 824-32; (b) Davis, L., *Handbook of genetic algorithms*. Van Nostrand Reinhold: New York, 1991; p xii, 385 p.
193. Zhou, Y.; Lai, X.; Li, Y.; Dong, W., Ant Colony Optimization With Combining Gaussian Eliminations for Matrix Multiplication. *IEEE Trans Syst Man Cybern B Cybern* 2012.
194. Lv, J.; Wang, Y.; Zhu, L.; Ma, Y., Particle-swarm structure prediction on clusters. *J Chem Phys* 2012, *137* (8), 084104.
195. Wermuth, C. G., Pharmacophores: Historical Perspective and Viewpoint from a Medicinal Chemist. In *Pharmacophores and Pharmacophore Searches*, Wiley-VCH Verlag GmbH & Co. KGaA: 2006; pp 1-13.
196. Wolber, G.; Seidel, T.; Bendix, F.; Langer, T., Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug discovery today* 2008, *13* (1-2), 23-9.
197. Smellie, A.; Teig, S. L.; Towbin, P., Poling - Promoting Conformational Variation. *Journal of Computational Chemistry* 1995, *16* (2), 171-187.
198. Al-Sha'er, M. A.; Taha, M. O., Elaborate ligand-based modeling reveals new nanomolar heat shock protein 90alpha inhibitors. *Journal of Chemical Information and Modeling* 2010, *50* (9), 1706-23.
199. Poptodorov, K.; Luu, T.; Hoffmann, R. D., Pharmacophore Model Generation Software Tools. In *Pharmacophores and Pharmacophore Searches*, Wiley-VCH Verlag GmbH & Co. KGaA: 2006; pp 15-47.
200. Goodford, P. J., A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem* 1985, *28* (7), 849-57.
201. Marshall, G. R.; Barry, C. D.; Bosshard, H. E.; Dammkoehler, R. A.; Dunn, D. A., Conformational Parameter in Drug Design - Active Analog Approach. *Abstracts of Papers of the American Chemical Society* 1979, (Apr), 29-29.
202. Dixon, S. L.; Smondyrev, A. M.; Knoll, E. H.; Rao, S. N.; Shaw, D. E.; Friesner, R. A., PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. *J Comput Aided Mol Des* 2006, *20* (10-11), 647-71.
203. Inc., C. C. G. *Molecular Operating Environment (MOE)*, 2011.10
Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7: 2011.
204. (a) Kurogi, Y.; Guner, O. F., Pharmacophore modeling and three-dimensional database searching for drug design using catalyst. *Curr Med Chem* 2001, *8* (9), 1035-55; (b) Inc., A. *Catalyst*, 2002; Accelrys Inc.; San Diego, CA: 2002.
205. Inte:Ligand, LigandScout - advanced structure-based pharmacophore modeling. 2012.
206. Martin, Y. C.; Bures, M. G.; Danaher, E. A.; DeLazzer, J.; Lico, I.; Pavlik, P. A., A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists. *J Comput Aided Mol Des* 1993, *7* (1), 83-102.
207. Jones, G.; Willett, P.; Glen, R. C., A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *J Comput Aided Mol Des* 1995, *9* (6), 532-49.
208. Güner, O. F., *Pharmacophore perception, development, and use in drug design*. International University Line: LaJolla, CA, 2000; p xiii, 537 p., xx p. of col. plates.
209. Chang, C.; Swaan, P. W., Computational approaches to modeling drug transporters. *Eur J Pharm Sci* 2006, *27* (5), 411-24.

210. Patel, Y.; Gillet, V. J.; Bravi, G.; Leach, A. R., A comparison of the pharmacophore identification programs: Catalyst, DISCO and GASP. *J Comput Aided Mol Des* 2002, *16* (8-9), 653-81.
211. Prodromou, C.; Pearl, L. H., Structure and functional relationships of Hsp90. *Curr Cancer Drug Targets* 2003, *3* (5), 301-23.
212. (a) Solit, D. B.; Rosen, N., Hsp90: a novel target for cancer therapy. *Current Topics in Medicinal Chemistry* 2006, *6* (11), 1205-14; (b) Chiosis, G.; Rodina, A.; Moulick, K., Emerging Hsp90 inhibitors: from discovery to clinic. *Anticancer Agents Med Chem* 2006, *6* (1), 1-8.
213. Noha, S. M.; Atanasov, A. G.; Schuster, D.; Markt, P.; Fakhrudin, N.; Heiss, E. H.; Schrammel, O.; Rollinger, J. M.; Stuppner, H.; Dirsch, V. M.; Wolber, G., Discovery of a novel IKK-beta inhibitor by ligand-based virtual screening techniques. *Bioorganic & medicinal chemistry letters* 2011, *21* (1), 577-83.
214. Valiron, O.; Caudron, N.; Job, D., Microtubule dynamics. *Cell Mol Life Sci* 2001, *58* (14), 2069-84.
215. Chiang, Y. K.; Kuo, C. C.; Wu, Y. S.; Chen, C. T.; Coumar, M. S.; Wu, J. S.; Hsieh, H. P.; Chang, C. Y.; Jseng, H. Y.; Wu, M. H.; Leou, J. S.; Song, J. S.; Chang, J. Y.; Lyu, P. C.; Chao, Y. S.; Wu, S. Y., Generation of ligand-based pharmacophore model and virtual screening for identification of novel tubulin inhibitors with potent anticancer activity. *Journal of medicinal chemistry* 2009, *52* (14), 4221-33.
216. Lanier, M. C.; Feher, M.; Ashweek, N. J.; Loweth, C. J.; Rueter, J. K.; Slee, D. H.; Williams, J. P.; Zhu, Y. F.; Sullivan, S. K.; Brown, M. S., Selection, synthesis, and structure-activity relationship of tetrahydropyrido[4,3-d]pyrimidine-2,4-diones as human GnRH receptor antagonists. *Bioorg Med Chem* 2007, *15* (16), 5590-603.
217. (a) Cheng, K. W.; Leung, P. C., The expression, regulation and signal transduction pathways of the mammalian gonadotropin-releasing hormone receptor. *Can J Physiol Pharmacol* 2000, *78* (12), 1029-52; (b) Huirne, J. A.; Lambalk, C. B., Gonadotropin-releasing-hormone-receptor antagonists. *Lancet* 2001, *358* (9295), 1793-803.
218. Roche, O.; Rodriguez Sarmiento, R. M., A new class of histamine H3 receptor antagonists derived from ligand based design. *Bioorganic & medicinal chemistry letters* 2007, *17* (13), 3670-5.
219. Hough, L. B., Genomics meets histamine receptors: New subtypes, new receptors. *Molecular Pharmacology* 2001, *59* (3), 415-419.
220. Alguacil, L. F.; Perez-Garcia, C., Histamine H3 receptor: a potential drug target for the treatment of central nervous system disorders. *Curr Drug Targets CNS Neurol Disord* 2003, *2* (5), 303-13.
221. Witkin, J. M.; Nelson, D. L., Selective histamine H3 receptor antagonists for treatment of cognitive deficiencies and other disorders of the central nervous system. *Pharmacol Ther* 2004, *103* (1), 1-20.
222. Hancock, A. A.; Brune, M. E., Assessment of pharmacology and potential anti-obesity properties of H3 receptor antagonists/inverse agonists. *Expert Opin Investig Drugs* 2005, *14* (3), 223-41.
223. Stahl, M.; Todorov, N. P.; James, T.; Mauser, H.; Boehm, H. J.; Dean, P. M., A validation study on the practical use of automated de novo design. *J Comput Aided Mol Des* 2002, *16* (7), 459-78.
224. (a) Howells, L. M.; Gallacher-Horley, B.; Houghton, C. E.; Manson, M. M.; Hudson, E. A., Indole-3-carbinol inhibits protein kinase B/Akt and induces apoptosis in the human breast tumor cell line MDA MB468 but not in the nontumorigenic HBL100 line. *Mol Cancer Ther* 2002, *1* (13), 1161-72; (b) Li, Y.; Chinni, S. R.; Sarkar, F. H., Selective growth regulatory and pro-apoptotic effects of DIM is mediated by AKT and NF-kappaB pathways in prostate cancer cells. *Front Biosci* 2005, *10*, 236-43.
225. Chao, W. R.; Yean, D.; Amin, K.; Green, C.; Jong, L., Computer-aided rational drug design: a novel agent (SR13668) designed to mimic the unique anticancer mechanisms of dietary indole-3-carbinol to block Akt signaling. *Journal of medicinal chemistry* 2007, *50* (15), 3412-5.
226. Reid, J. M.; Walden, C. A.; Qin, R.; Ziegler, K. L.; Haslam, J. L.; Rajewski, R. A.; Warndahl, R.; Fitting, C. L.; Boring, D.; Szabo, E.; Crowell, J.; Perloff, M.; Jong, L.; Bauer, B. A.; Mandrekar, S. J.; Ames, M. M.; Limburg, P. J., Phase 0 clinical chemoprevention trial of the Akt inhibitor SR13668. *Cancer prevention research* 2011, *4* (3), 347-53.
227. Kruger, D. M.; Evers, A., Comparison of structure- and ligand-based virtual screening protocols considering hit list complementarity and enrichment factors. *Chemmedchem* 2010, *5* (1), 148-58.
228. (a) Rush, T. S., 3rd; Grant, J. A.; Mosyak, L.; Nicholls, A., A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J Med Chem* 2005, *48* (5), 1489-95; (b) Software, O. S. *ROCS*, OpenEye Scientific Software: Santa Fe, NM.
229. Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S., A critical assessment of docking programs and scoring functions. *Journal of Medicinal Chemistry* 2006, *49* (20), 5912-5931.

230. Page, C. S.; Bates, P. A., Can MM-PBSA calculations predict the specificities of protein kinase inhibitors? *Journal of Computational Chemistry* 2006, *27* (16), 1990-2007.
231. (a) Plewczynski, D.; Lazniewski, M.; von Grotthuss, M.; Rychlewski, L.; Ginalski, K., VoteDock: consensus docking method for prediction of protein-ligand interactions. *Journal of Computational Chemistry* 2011, *32* (4), 568-81; (b) Bar-Haim, S.; Aharon, A.; Ben-Moshe, T.; Marantz, Y.; Senderowitz, H., SeleX-CS: A New Consensus Scoring Algorithm for Hit Discovery and Lead Optimization. *Journal of Chemical Information and Modeling* 2009, *49* (3), 623-633; (c) Fukunishi, H.; Teramoto, R.; Takada, T.; Shimada, J., Bootstrap-based consensus scoring method for protein-ligand docking. *J Chem Inf Model* 2008, *48* (5), 988-96; (d) Teramoto, R.; Fukunishi, H., Consensus scoring with feature selection for structure-based virtual screening. *J Chem Inf Model* 2008, *48* (2), 288-95.
232. Nicolotti, O.; Miscioscia, T. F.; Carotti, A.; Leonetti, F.; Carotti, A., An integrated approach to ligand- and structure-based drug design: development and application to a series of serine protease inhibitors. *Journal of Chemical Information and Modeling* 2008, *48* (6), 1211-26.
233. Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S., Generating Optimal Linear Pls Estimations (Golpe) - an Advanced Chemometric Tool for Handling 3d-Qsar Problems. *Quantitative Structure-Activity Relationships* 1993, *12* (1), 9-20.
234. Vedani, A.; Dobler, M.; Spreafico, M.; Peristera, O.; Smiesko, M., VirtualToxLab - in silico prediction of the toxic potential of drugs and environmental chemicals: evaluation status and internet access protocol. *ALTEX* 2007, *24* (3), 153-61.
235. Wilson, G. L.; Lill, M. A., Integrating structure-based and ligand-based approaches for computational drug design. *Future Med Chem* 2011, *3* (6), 735-50.
236. Klon, A. E.; Glick, M.; Thoma, M.; Acklin, P.; Davies, J. W., Finding more needles in the haystack: A simple and efficient method for improving high-throughput docking results. *Journal of Medicinal Chemistry* 2004, *47* (11), 2743-9.
237. Khatib, F.; Cooper, S.; Tyka, M. D.; Xu, K. F.; Makedon, I.; Popovic, Z.; Baker, D.; Players, F., Algorithm discovery by protein folding game players. *P Natl Acad Sci USA* 2011, *108* (47), 18949-18953.
238. Khatib, F.; DiMaio, F.; Cooper, S.; Kazmierczyk, M.; Gilski, M.; Krzywda, S.; Zabranska, H.; Pichova, I.; Thompson, J.; Popovic, Z.; Jaskolski, M.; Baker, D.; Grp, F. C.; Grp, F. V. C., Crystal structure of a monomeric retroviral protease solved by protein folding game players (vol 18, pg 1175, 2011). *Nat Struct Mol Biol* 2012, *19* (3), 365-365.
239. Eiben, C. B.; Siegel, J. B.; Bale, J. B.; Cooper, S.; Khatib, F.; Shen, B. W.; Players, F.; Stoddard, B. L.; Popovic, Z.; Baker, D., Increased Diels-Alderase activity through backbone remodeling guided by FOLDIT players. *Nat Biotechnol* 2012, *30* (2), 190-192.
240. Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W., Jr., Computational methods in drug discovery. Pharmacol Rev 2014, 66 (1), 334-95. Reprinted with permission of the American Society for Pharmacology and Experimental Therapeutics. All rights reserved

CHAPTER 2 : BCL::CONF SMALL MOLECULE CONFORMATIONAL SAMPLING USING A KNOWLEDGE BASED ROTAMER LIBRARY

Introduction

The interactions between small molecules and proteins are important for receptors, transporters, or enzymes to recognize their substrates as well as for small molecule therapeutics to bind to their target protein. The molecular interaction and, hence, the biological function of a small molecule is related to its three-dimensional structure when interacting with the protein. In solution, small molecules are often flexible and exist as an ensemble of conformations in equilibrium with one another. The biologically active conformation may be a single conformation or a small subset from the conformations sampled in solution or a new conformation, induced by protein binding. A uniform sampling of all energetically accessible small molecule conformations is essential for the success of protein small molecule docking simulations[1] for example in structure-based computer-aided drug discovery/design (CADD)[1-2]. However, also ligand-based CADD applications such as three-dimensional quantitative structure activity relationships (3D-QSAR) predictions[3] or pharmacophore modeling[4] rely on the use of conformational ensembles of molecules that capture the bioactive conformation as one of a diverse set of energetically accessible conformations[5].

Conformational sampling methods

Table 2-1 summarizes some of the existing conformational sampling methods. Conformation sampling methods can be characterized in several ways. First, the allowed search space can be analyzed: Some methods search the entire conformational space, i.e. bond length, angles and torsions can be altered – for example a molecular dynamics simulation in Cartesian space. Other methods restrict the search space to torsion angles only holding bond length and angles fixed. Another approach involves using pre-existing knowledge of small-molecule conformations to restrict the conformational search space even further to likely torsion angles or combinations thereof. Such knowledge-based methods derive torsion angle preferences from molecular mechanics or quantum chemical simulations of small molecules or structural databases like Cambridge Structure Database[6] (CSD) or Protein Data Bank[7] (PDB).

In addition, it is helpful to single out fragment-based approaches: This search strategy splits a molecule of interest and samples conformations of smaller fragments independently. Candidate conformations of the entire molecule are computed by re-combining constituent fragment conformations. In fragment-based methods,

Table 2-1 Commercially available conformation sampling methods.

Method	Search space	Search strategy	Search method	Scoring function
CAESAR[8]	Incremental search of torsion angles combined with distance geometry for ring systems	Fragment based	Systematic	CHARMm force field
CATALYST[9]	Incremental search of torsion angles with subsequent energy minimization	Non-fragment based	Simulation (MD)	CHARMm force field
CONAN[10]	Incremental search of torsion angles	Fragment based	Systematic	-
CONFAB[11]	Incremental search of torsion angles	Non-fragment based	Systematic	MMFF94
CONFGEN[12]	Random walk on energy surface calculated using a truncated version of OPLS_2001	Non-fragment based	Simulation (MC)	MMFFs/OPLs_2001
ENUMERATED TORSIONS (*ET*)[13]	Incremental search of rule-based torsion angles	Non-fragment based	Systematic	-
MIMUMBA[14]	Incremental search of knowledge-based torsion angles from CSD	Non-fragment based	Systematic	Relative frequency of experimentally observed conformations
MOE (LOW MODE MD)[15]	Constant temperature MD	Non-fragment based	Simulation (MD)	MMFF94
MOE (STOCHASTIC SEARCH)[16]	Random perturbations of rotatable bonds in increments biased around 30°	Non-fragment based	Simulation (MC)	MMFF94
MOE (CONFIMPORT)[16]	Pregenerated fragment conformations obtained from stochastic-search	Fragment-based	Simulation	MMFF94
MOE (SYSTEMATIC)[17]	Incremental search of torsion angles	Non-fragment based	Systematic	MMFF94
OMEGA[18]	Knowledge based torsions from analysis of molecules in PDB and conformations generated by MMFF94	Fragment based	Systematic	MMFF94
RDKIT[19]	Distance geometry	Non-fragment based	Simulation (Distance Geometry)	UFF

fragments are reused during conformer generation that improves the time-efficiency of sampling. On the other hand, these methods operate on the assumption that all low energy conformations can be created by combinations of low-energy fragments – an assumption that is not always fulfilled.

An alternative classification approach focuses on whether the search space is sampled systematically in its entirety or a search algorithm follows a trajectory that seeks to restrict the search space to low energy conformations. If the conformational space is sufficiently small, systematic approaches can create all possible conformations iteratively and keep all low-energy conformations. An advantage is complete sampling of the entire search space, one disadvantage is slowness. Trajectory-based methods use random or directed perturbations to alter a starting conformation and the resulting conformation is evaluated energetically. In a feedback loop, this energy and possibly derived forces determine the trajectory of the simulation. Molecular dynamics[8], distance geometry[9], genetic algorithms[4], and Monte Carlo[8b] (MC) are commonly used simulation methods for the conformational sampling of small molecules.

Scoring functions

Most methods score conformations using some form of molecular mechanics energy function. Force field based energy calculations use most frequently the Merck molecular force field (MMFF)[10] or the Chemistry at HARvard Molecular Mechanics (CHARMm) force field[11]. Some methods modify the default versions of these force fields by modifying individual scoring terms or using only a subset of the scoring terms. One alternative approach, as used in MIMUMBA[12], to scoring small molecule conformations can be derived from knowledge-based scoring functions used in protein structure prediction that analyze the frequency of geometric features observed in structural databases such as the PDB or CSD.

Knowledge based conformation sampling

Conformations of small molecules can be restricted in terms of commonly seen conformations of constituent fragments in structure databases like CSD. Brameld[13] et al. have shown that conformations of fragments sampled in the CSD are an accurate representation of conformational space seen in drug-like molecules in complex with protein as observed in the PDB. Fragments occur in these structure databases in different chemical environments, leading to them being observed in different conformations. The central hypothesis of this study is that while not all small molecules have been crystallized in all possible conformations, the conformational space accessible to sufficiently small fragments is adequately sampled.

Existing methods like CONFECT[14] derive torsion profiles for different dihedral bonds types from structure databases. CONFECT treats dihedral bonds as uncorrelated and does not take into account substituent effects. A rule-based proprietary method, developed by Merck research laboratories for internal use, known as *et* for enumerated torsions uses correlated torsion angles to some extent for conformational sampling[15]. The method overlaps multiple fragments containing topologically adjacent rotatable bonds to extend these fragments until they span the entire small molecule. In *et* a proprietary 'atom typer' is used to express molecular fragments as unambiguous patterns[16]. The pattern along with associated data for observed torsion angles and frequency

constitutes a rule. As of 2001, authors reported that 797 rules had been derived over a period of several years. However, these patterns consider only the four atoms involved in a dihedral bond and do not take into account effect of substituents on torsional profile of bonds.

The algorithm BCL::CONF described in the present study goes beyond previous work by using torsional profile of multiple consecutive dihedral bonds and capturing effect of substituents on their torsion profiles. All fragment conformations sampled frequently in the CSD and PDB are considered a knowledge-based 'rule' independent of size or number of rotatable bonds. This fragment conformation approach allows BCL::CONF to capture correlations in torsion states for multiple consecutive dihedral bonds in contrast to other methods that treat likely torsion angle states for consecutive bonds in an uncorrelated way. Conformations observed frequently for one fragment are assumed to represent a local energy minimum and are collected in a database. The use of conformations of fragments has also the advantage that these fragment conformations already reside in locally optimal geometries so that only non-local interactions, i.e. clashes, need to be evaluated when fragments are recombined. Lastly, as explicit fragments are used effects of substituents on torsional profiles of rotatable bonds are taken into account. Brameld et al. have shown the effect of substitution on the torsion distribution of common acyclic organic fragments[13].

We expect that the algorithm is therefore particularly tailored for 'drug-like' small molecules that are overrepresented in the CSD and PDB databases. BCL::CONF mimics the 'rotamer' libraries created to capture amino acid side chain conformations seen in protein structures within the PDB[17] which, ultimately, will ease its integration with protein modeling packages such as ROSETTA[18]. BCL::CONF scoring includes a clash score that avoids atom overlap as well as a knowledge-based scoring function that scores conformations based on probabilities of fragment conformations that it contains.

To benchmark BCL::CONF we use a curated dataset containing drug-like ligands found in complex with proteins in the PDB. The "VERNALIS generic compound set"[19] has been used in several studies to evaluate the performance of conformational sampling methods enabling a direct comparison of BCL::CONF to other methods[20]. The benchmark study tests for recovery of protein-bound conformation of the ligand and the ability of BCL::CONF to produce a diverse set of conformations. To remove any bias during benchmarking, the ligands found in the VERNALIS dataset were removed from the PDB ligand library. Additionally, ligands were removed from the PDB ligand library if bound to proteins or homologues of proteins present in the VERNALIS dataset.

Implementation

BCL::CONF uses fragments generated from decomposing molecules found in CSD and PDB. For this purpose, non-ring bonds of each molecule are broken iteratively to generate all possible fragments. In a second step, all occurrences

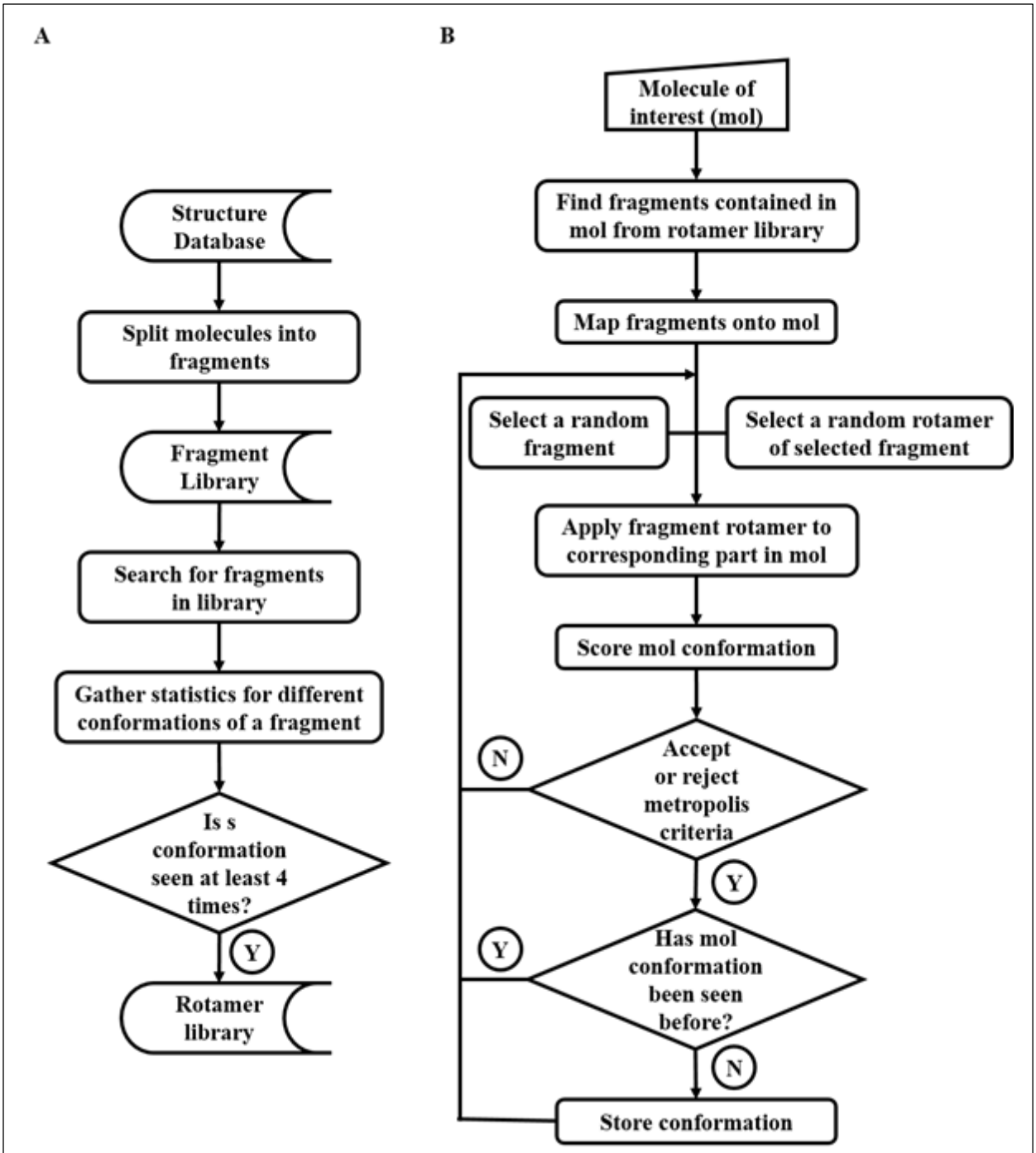


Figure 2-1 General scheme for BCL::CONF conformation generator. A) Scheme for generating the rotamer library. B) Flowchart depicting conformation sampling process. See text for a detailed description. Source: Kothiwale et al[38]

of one fragment within the structure databases are collected and clustered according to discrete dihedral angle bins. A conformer is then defined as a unique conformation represented as a set of integer numbers, one for each dihedral bond, identifying the bin. This procedure is similar to the definition of 'rotamers' that are used to set likely amino acid side chain conformations[17]. A conformer needs to be seen at least four times in the database to be considered as a likely conformation of a fragment. It is then added to the rotamer library for sampling. The flowchart for algorithm implemented in BCL::CONF is shown in Figure 2-1.

Fragment library

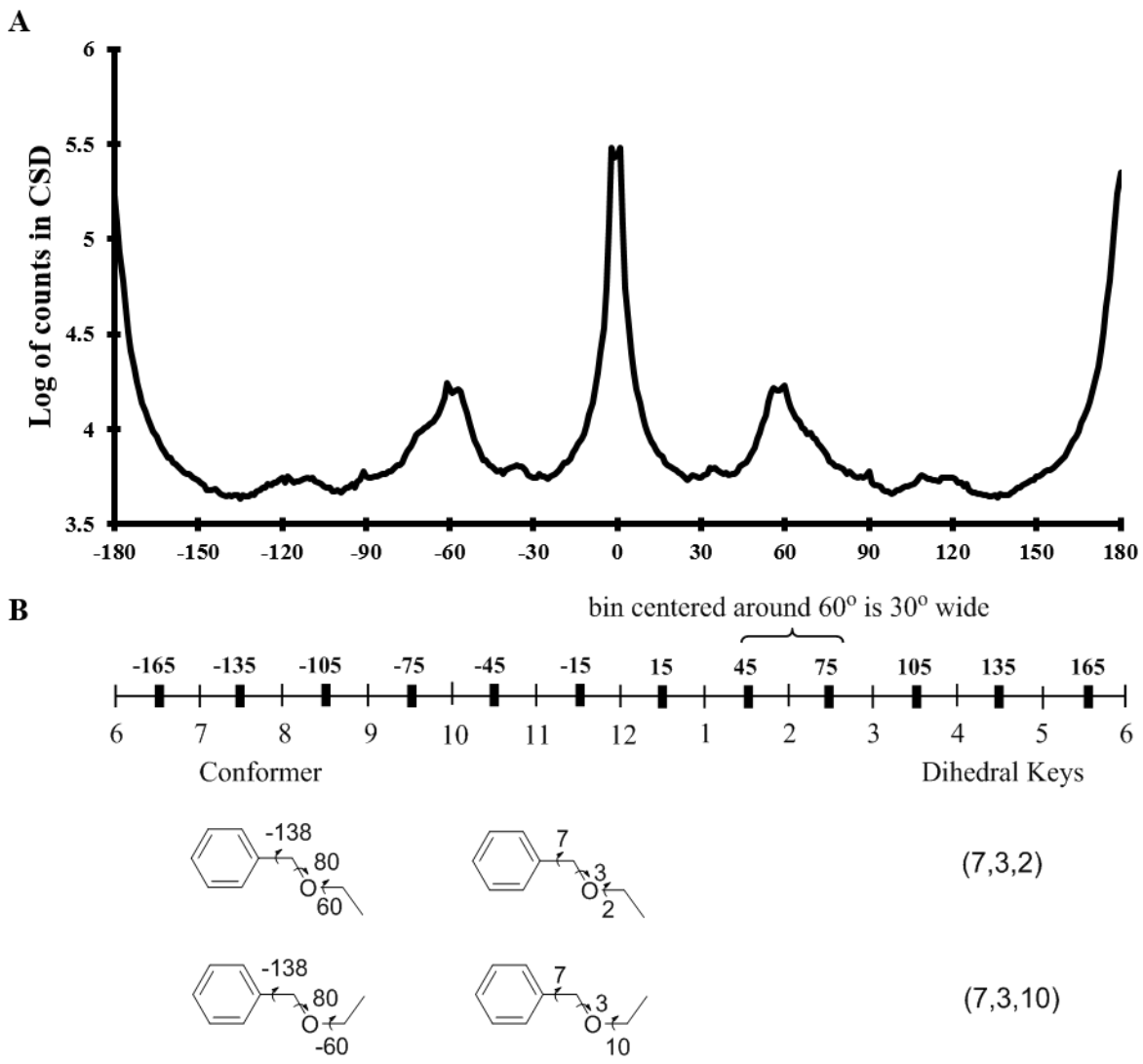


Figure 2-2 Scheme for torsion angle binning. A) The line graph shows the distribution of dihedral angle measurements of all dihedral bonds over all the molecules in the CSD. B) Torsional angles are binned into 12 uniform parts with each bin represented as an integer. For example -135 to -165 belongs to bin number 7. Rotamers can thus be represented by a unique key of integers representing each dihedral angle bin. Source: Kothiwale et al[38]

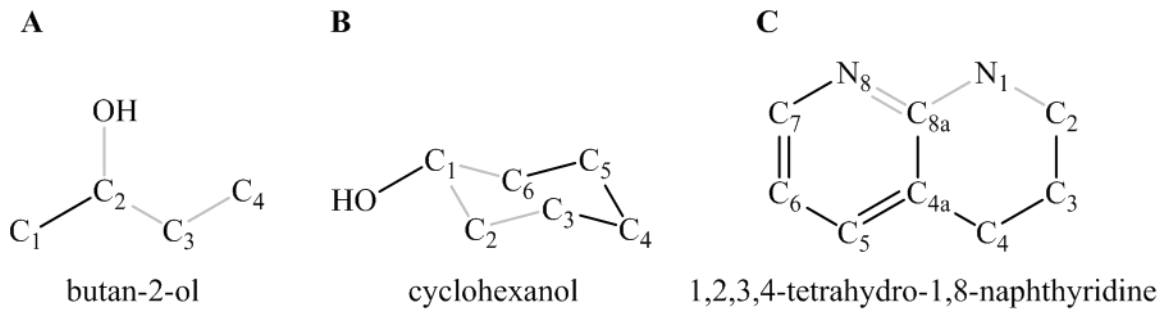


Figure 2-4 Determination of priority dihedral bonds in molecules. Bond priorities are determined using rules analogous to Cahn-Ingold-Prelog (CIP) rules. In the figure priority dihedral bonds are colored in grey. A) The priority dihedral angle of 2-butanol is c4-c3-c2-O. B) Priority dihedral bonds in cyclohexanol are defined such that all atoms that define priority dihedral angles are in the ring. Thus bond C1-C2, C3-C2-C1-C6 is the priority dihedral angle instead of C3-C2-C1-O. C) For multiple ring systems like 1, 2, 3, 4-tetrahydro-1, 8-naphthyridine, priority angles are determined by atom priority using the assumption that all atoms in the multiple ring system are part of one ring. Thus C2-N1-C8a-N8 is the priority dihedral angle instead of C2-N1-C8a-C4a as N8 is counted to be in the same ring system as the N1-C2 bond of interest. Source: Kothiwale et al[38]

Small organic molecules from the CSD and PDB were used for generating fragments. The PDB ligands were obtained from the refined dataset in the PDBBIND database[21]. We removed any molecules for which BCL could not assign correct atom types, molecules with missing 3D coordinates and bad geometries in terms of unrealistic bond-lengths or bond-angles and non-planar aromatic rings or sp^2-sp^2 bonds. This resulted in a database containing

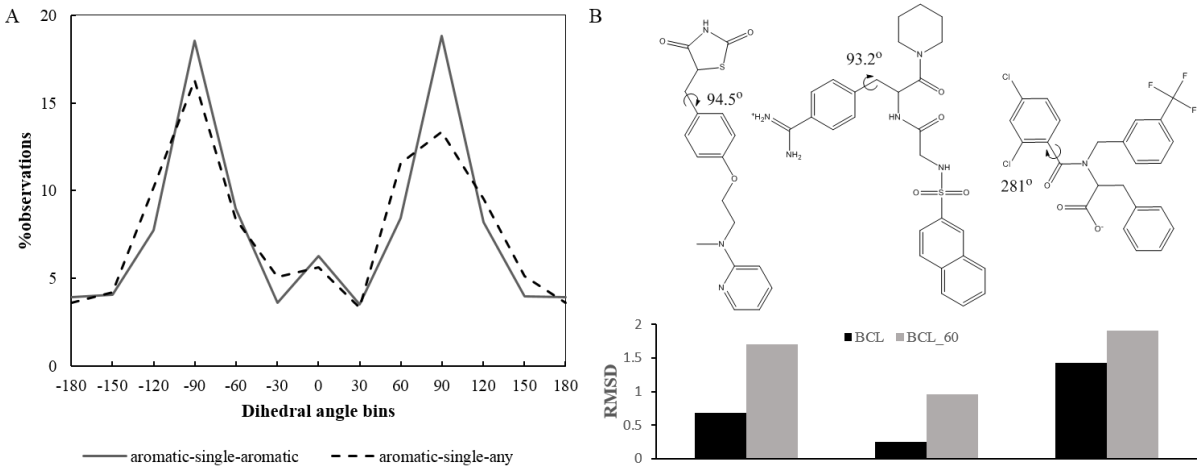
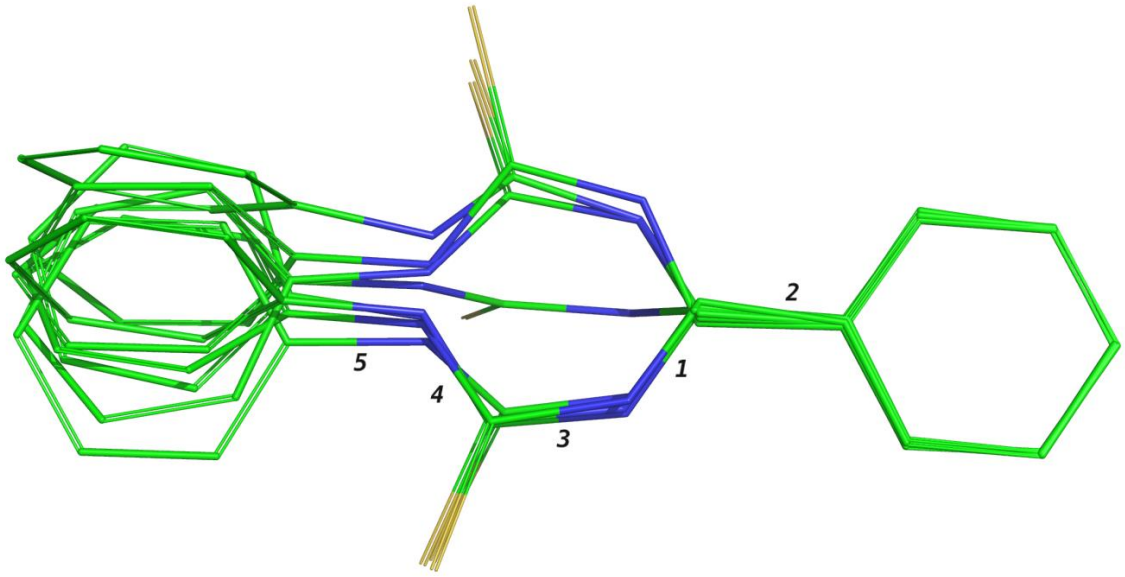


Figure 2-3 Propensity of certain dihedral bonds containing a sequence of aromatic-single-aromatic and aromatic-single-any bonds to measure 90° or 270°. A) The figure shows distribution for seven aromatic-single-aromatic and 129 aromatic-single-any bonds for which there are at-least four 90° or 270° rotamers and are more abundant compared to 30° or 60° rotamers. Average observation of rotamers is plotted for dihedral bonds is plotted. B) Molecules from the VERNALIS dataset containing aromatic-single-single bonds measuring 90° or 270°. The closest to native conformation generated using 30° (BCL) and 60° (BCL_60) binning differ by at least 0.4 Å. Source: Kothiwale et al[38]

113,339 unique molecules. Molecules were broken iteratively at non-ring bonds generating 56,818,272 unique fragments.

Table 2-2 Rotamers of a molecule.



Rotamers	Bond1	Bond2	Bond3	Bond3	Bond4	Bond5
1	6	5	6	12	2	6
2	6	3	6	12	5	6
3	6	1	6	12	2	6
4	6	5	6	12	4	6
5	6	1	6	12	5	6
6	6	5	6	12	5	6
7	6	4	6	12	2	6
8	6	1	6	12	1	6
9	6	5	6	12	1	6

Rotamer library

The rotamer library was generated for fragments that are seen frequently in the same conformations. A unique fragment rotamer/conformation is identified by a set of integers, one for each dihedral bond. The dihedral bonds of a rotamer are represented as a set of integers depending on the angle measure as explained in Figure 2-2B. The frequency of observation of dihedral angle measures seen in CSD, shown in Figure 2-2A, suggests that local minima

for dihedral angles occur at canonical values of 0°, 60°, 120°, and 180° and so on. In addition, bond types such as *aromatic-chain-aromatic* or *aromatic-chain-any* angles of 90° and 270° are likely (Figure 2-3). Hence, while torsion angles of 90° and 270° are not local maxima when summing over all torsions, they are likely conformations for certain types of torsion angles. Therefore, in order to assign as many likely torsion angles as possible unambiguously and close to a bin center, 12 bins each of which is 30° wide are created centered at 0°, 30°, 60°, 90° and so on. Binning strategies using 30° produces closer to native conformations when 60° binning is used (Figure 2-6C). All the bonds including the ones that are inside ring systems are described by an integer so that a rotamer can be described as a string of integers. This string is called the bin-signature of a rotamer.

Determining dihedral angles

Since multiple dihedral angles can be measured at each torsion bond, a scheme is required to prioritize which dihedral angle to use and arrive at unambiguous bin-signatures. Therefore, a priority dihedral angle is defined. This is accomplished using rules analogous to the Cahn-Ingold-Prelog (CIP) system22. For example, as shown in Figure 2-4A, 2-butanol has one torsion bond but two dihedral bonds about the single rotatable bond. According to CIP rules, the O-C-C-C dihedral angle will have a higher priority over the C-C-C-C dihedral angle. If out of three possible dihedral angles, two dihedral angles of equally high priority exist, then the third dihedral angle with lowest priority is used. If ambiguity still exists in assigning unique dihedral bonds, for example in the case where all dihedral angles have the same priority, the one with the smallest angle measure is chosen. Priority dihedral bonds in rings are defined in a special way in that all atoms constituting a priority bond are contained in the ring, as shown in Figure 2-4B for cyclohexanol. This ensures that for the same ring conformation, a substituted ring system have the same dihedral-signature as an un-substituted ring system. If a fused ring system is present, then priority dihedrals are determined using atom priorities and the assumption that all atoms of the ring system are part of one ring (Figure 2-4C). BCL::Conf can identify different ring conformations and use these in conformational sampling. Since dihedral angles are assigned in a unique way for a molecule of interest, a unique rotamer of the molecule has a unique dihedral bin signature. Table 2-2 shows different rotamers for a fragment from the rotamer library and their bin signatures.

Searching rotamers

In building the rotamer library, all instances of every fragment are collected in the molecular database using a graph isomorphism search[23]. For each fragment, all unique rotamers are identified using dihedral bin signatures. Then statistics is gathered for each rotamer including rotamer counts, i.e. the number of times a rotamer is seen in the database, and dihedral angle statistics, i.e. the average angle measure and standard deviation of dihedral bonds within each bin, for each rotamer are extracted. A representative structure for a fragment is obtained by clustering all instances of the most frequently observed rotamer in the structure database based on root mean square

Table 2-3 A) Rotatable bond distribution in the rotamer library. B) Conformation statistics in the rotamer library

Number of rotatable bonds	Number of fragments	Number of rotamers	Number of fragments
0	47,205	1-5	219684
1	38,616	6-10	10840
2	31,225	11-15	1768
3	20,500	16-20	488
4	15,221	21-25	209
5	13,665	26-30	82
6	14,014	31-35	47
7	14,693	36-40	18
8	14,435	41-45	8
9	13,492	46-50	1
>=10	10,064	>50	3

deviation (RMSD) after superposition. In addition, if a fragment contains a ring in different conformations, explicit coordinates are stored for each rotamer. A conformer is added to the rotamer library of a fragment if it is seen at least four times in the structure databases (combined CSD and PDB), i.e. it can be considered a likely conformation for that fragment. A total of 231,049 fragments are observed that have at least one conformer that is seen at least four times in the molecular database and hence these fragments are retained in the rotamer library. Table 2-3 shows the rotatable bond distribution and rotamer distribution of fragments in the rotamer library.

Search fragments from the rotamer library that are contained in the molecule of interest

Conformational sampling begins with searching fragments contained in a molecule of interest. This involves substructure searches to identify all suitable fragments in the rotamer library. A hierarchical search has been implemented to minimize the number of substructure searches. The rotamer library is represented as multiple rooted graphs where each node is a unique constitution. The root nodes are not contained in any other fragments. Child nodes are such that the parent node is an immediate substructure. Figure 2-5 illustrates a rooted graph with benzene as root. Benzene is an immediate substructure of its child nodes i.e. toluene-like fragment that is an immediate substructure of cyclohexylbenzene-like fragment.

The fragment searching begins at the root node of graphs. If the root node is contained in the query molecule, all its immediate child nodes will be searched to determine if they are contained within the molecule of interest. For all child nodes contained, their immediate child notes are considered and so on. In Figure 2-5, fragments that are part of molecule are colored in blue – i.e. a successful substructure search. Fragments colored red indicate that a substructure search was performed but unsuccessful. This terminates further searches in this branch of the tree. Fragments colored in black are not considered for a substructure search, because their parent fragments were not contained within the molecule of interest (colored red). The edges in the graph are directed from parent to child nodes and represent search paths that can be taken to find all constituent fragments in a query molecule. Paths in blue color are actual paths that are taken to identify all the fragments contained in the molecule interest while the paths in red or black are never explored. Search paths in black originate from fragments that are not contained with

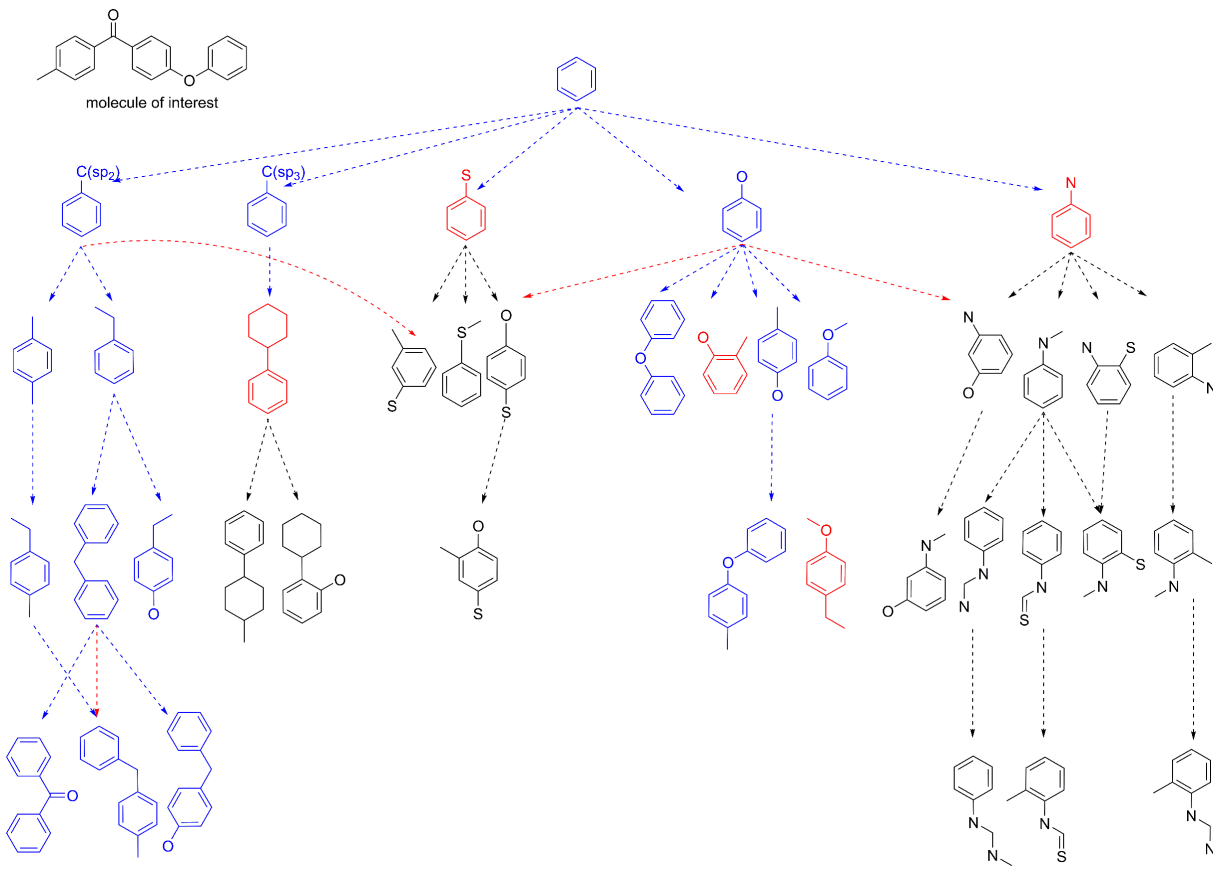


Figure 2-5 Graph database for storing rotamer library for fast searching. The figure illustrates a rooted graph layout of fragments where each node is a unique constitution. The child nodes originating from the root are such that the root (in this case, benzene fragment) is their immediate substructure among all the fragments shown in the graph. Fragments contained in the molecule of interest are colored in blue while those that are not are in red or black. For fragments in black no substructure search is performed because their parent fragments were not found in the molecule of interest. The edges represent all possible search paths for finding fragments contained in the molecule of interest. Paths in blue are the actual searches that were performed for finding fragments for the query molecule. Paths in red and black are never taken during the search. Red colored paths are redundant search paths that have already been covered in a previous search. Source: Kothiwale et al[38]

the molecule. Red paths represent redundant searches in the tree. This hierarchical tree structure of the data enables fast and efficient searching of all the fragments contained within a molecule of interest.

Generation of initial 3D structure from minimum set of fragments with most likely conformation

An initial 3D conformation is necessary for using the conformer sampler implemented in BCL::CONF. The BCL software suite accepts molecules in the MDL[24] format. A 3D structure generator has been implemented to generate an initial 3D structure if coordinates are not provided. BCL::CONF can generate starting coordinates from connectivity information provided in the MDL format. When coordinates or 3D structure is not available, BCL::CONF first searches for all fragments from the rotamer library that are contained in a molecule of interest. The algorithm identifies the

minimum number of fragments that can be connected to generate molecule of interest. The most likely conformers of fragments are then connected to assemble the molecules of interest and generate an initial 3D structure that may or may not have clashes between atoms. As this conformation only serves as a starting point with the objective to place all torsion angles into a locally reasonable conformation and is not necessarily part of the output ensemble of conformations, atom clashes are not a problem.

Monte-Carlo Metropolis sampling for efficient search of conformational space

Conformational sampling begins by identifying fragments from the rotamer library that are contained in the molecule of interest whose conformations need to be sampled. From the fragments contained in the molecule of interest, a random one is selected and one of its rotamers is applied to change the conformation of the molecule. The rotamer is selected based on probability of its occurrence in the structure database (Figure 2-2B). If the chosen fragment rotamer contains a different ring conformation, then the whole molecule is reassembled by using the chosen conformer as the starting fragment. By default only a subset of rotamers that are observed most frequently are used in sampling. The cutoff value is specified at half of the probability of the most likely rotamer. If more sampling is desired, an option to use the full rotamer set can be specified at the command line.

Starting with the input structure of the molecule of interest, new conformations are created in a continuous MC trajectory. A MC step is accepted or rejected based on the Metropolis criterion. The energy or score used is a combination of atom clashes and propensity of observing constituent fragment rotamers in structure database. The atom clash score is calculated by evaluating non-bonded atom pairs for clashes using equation 1.

Equation 1:

$$Atom\ Clash\ Score\ = \frac{\sum_{i>j}\ 2 * score_{atom_j} \begin{cases} 0, dist\ \geq cov \\ 1, dist\ \leq cov \end{cases}}{Number\ of\ atoms\ in\ the\ molecule}$$

where $dist \stackrel{\text{def}}{=}$ *distance between non-bonded atoms i and j*

$cov \stackrel{\text{def}}{=}$ *sum of covalent radii of atoms i and j*

Rotamer propensity score (Equation 2) leverages the statistics on the rotamer of a particular fragment to estimate the likelihood of a particular conformation. The hypothesis is that there is a correlation between frequency of occurrence and free energy of a fragment conformation. For a given molecular conformation, the observed rotamer of each of the constituent fragments is determined. The observed rotamer propensity for a fragment is calculated by dividing observed rotamer count by average rotamer counts. The overall conformation score is obtained by summing up observed rotamer propensities of all the constituent fragments. If, for a fragment none of the rotamers is seen in a given conformation, then a pseudo rotamer count equal to half of the least common

rotamer count is used instead. The propensity score is normalized by dividing it by absolute value of maximum possible propensity score for the molecule of interest.

Equation 2:

$$Propensity\ Score\ = \sum_{i=0}^{N}\left(-\ln\frac{R_i\ \times\ F_iR_j}{\sum_j F_iR_j}\right)\Big/\sum_{i=0}^{N}\left(\ln\frac{R_i\ \times\ F_iR_{max}}{\sum_j F_iR_j}\right)$$

where $N \stackrel{\text{def}}{=}$ number of fragments that are part of the molecule of interest

$F_i \stackrel{\text{def}}{=}$ ith fragment of molecule

$R_i \stackrel{\text{def}}{=}$ number of rotamers of the ith fragment

$R_{max} \stackrel{\text{def}}{=}$ counts of the most common rotamer

$F_iR_j \stackrel{\text{def}}{=}$ counts of jth rotamer of the ith fragment

Results and Discussion

We assess the performance of BCL::CONF (BCL) with curated generic ligand dataset known as the VERNALIS dataset[19], in comparison with CONFGEN[25], MOE (CONFIMPORT)[26], OMEGA[27] and RDKIT[20B, 28]. The first metric defined as the completeness criteria is the fraction of molecules for which any conformation was generated. The second comparison is the ability of the method to produce ligand conformations within a specified RMSD value to the native conformation of ligands in protein-ligand complexes. This analysis is reported as the percentage of molecules whose conformations are recovered within a given threshold RMSD value. The third criteria for comparison is diversity, that is how similar or different are the generated conformations. Finally, a comparison of the methods on computational speed is provided. We also report results for different flavors of BCL that use different schemes for rotamer library generation – a) using a 60° torsion binning (BCL_60) b) rotamer library derived from only the CSD (BCL_CSD) c) rotamer library containing only single dihedral bond torsion profiles (BCL_D).

Conformational sampling with different methods was performed to yield a symmetry corrected RMSD diversity of 0.25 Å – i.e. no two conformations have a RMSD smaller 0.25 Å – and a maximum of 100 conformers per molecule.

Ligand dataset

VERNALIS dataset is used here to compare BCL::CONF to other existing methods in the field. The VERNALIS Dataset, compound set introduced by Chen and Foloppe[19-20], contains 253 ligands derived from high-resolution protein-ligand complexes found in the PDB and includes the Bostrom[29] ligand set and Perola[30] ligand set. The VERNALIS Dataset has

been used in previous benchmark studies to compare MOE, CATALYST and CONFGEN methods for conformation sampling[19-20].

Conformer generation methods

BCL::CONF (BCL) – Conformation sampling was carried out by providing ligands in the MDL format with all atom coordinates set to zero to remove any initial conformation bias. The rotamer library uses the 30° torsion binning scheme to determine dihedral keys. It is derived from the CSD and the refined set of PDBBIND database minus the VERNALIS dataset ligands to remove any bias. Conformers were generated in 200 iterations of MC fragment sampling at a temperature of 3.0 such that they were at least 0.25 Å away from each other. Table S2 (see supplement) shows parameter optimization for native conformer recovery in terms of RMSD with different temperature and iteration values. The row shaded in gray corresponds to parameters used for comparing to other methods.

BCL_60 – Conformations were sampled using the same settings as described for BCL::CONF except that 60° torsion binning was used instead of 30°. This experiment tests the effect of 60° binning on conformation sampling.

BCL_CSD – Same parameters as used for BCL::CONF with the only difference being that the rotamer library was sourced from only the CSD. This experiment shows the effect of adding PDB fragment conformations.

BCL_D – Conformation sampling was performed by using torsion angle statistics for single dihedral bonds derived from molecules in the CSD and PDBBIND databases. Fragments containing only four atoms and a single dihedral bond from the rotamer library were used for this experiment – i.e. the smallest possible fragments. This experiment tests the impact of the addition of larger fragments that sample the correlation between multiple torsion angles. Initial conformation bias in benchmark dataset molecules was removed by perturbing all dihedral angles to random values. The conformers were generated using the same set of parameters as that for BCL.

CONFGEN – CONFGEN systematically samples rotatable bonds, ring conformations, nitrogen atom inversions and amide bond conformations. Force field OPLS_2001 is used for calculating potential for rotating about each rotatable bond[25]. In the present study, conformer generation was done starting from SMILES string of ligands in the VERNALIS dataset. SMILES string were generated using Maestro from the dataset ligands in MDL format. CONFGEN has been reported to reproduce 93% of molecules within 1.5 Å in the comprehensive mode[25]. 250 Conformers were generated with CONGEN in the comprehensive mode by keeping RMSD cutoff at 0.25 Å, energy cutoff at 104.6 kJ/mol (default value). 100 conformations were saved per ligand for comparison.

MOE-conformation_import (MOE) – Conformational import is a high-throughput conformer generation method in MOE (Molecular Operating Environment). Molecule of interest is divided into overlapping fragments and these are searched in a pregenerated library of fragment conformations. If a fragment is not found, conformations are generated using a stochastic conformation search algorithm available in MOE. For this study, the VERNALIS dataset was provided such that all atom coordinates were set to zero. The default parameters specified with MOE have been determined to perform best in previously reported benchmark studies[19-20]. The MMFF94x force field and Generalized Born solvation model was during ligand conformation generation. Fragment conformation energy cutoff was kept

at a default of 4 kcal/mol. The program was constrained to maintain stereochemistry of the input structures but allowed to sample ring conformations. The stochastic search protocol that conformation import uses for creating conformations of fragments missing in database was modified to generate fragment conformers that were 0.25 Å apart in RMSD. Fragment conformations that were within 15 kcal/mol window of the lowest energy conformer were retained for the stochastic search.

OMEGA – OMEGA is a systematic knowledge based conformer generator developed by OPENEYE Scientific Software. It exhaustively enumerates all rotatable torsions using a knowledge-based list of angles that are then sampled by geometric and energy criteria[27]. The torsion library is derived from analysis of a set of experimental crystal structures from the PDB and from energy scans of torsions against MMFF94. Default parameter values were used except RMSD and MaxConfs which was set to 0.25 and 100 respectively to specify custom conformation diversity level and limit the number of output conformations.

RDKIT – RDKIT uses distance geometry algorithm described by Blaney et al for sampling ligand conformations[31]. A distance bound matrix is calculated for a molecule of interest based on connection table and a set of rules. The matrix is smoothed using a triangle-bounds smoothing algorithm. Random distance matrices that satisfy the bounds matrix are generated followed by embedding in 3D dimension to generate conformations. In a final step, embedded coordinates are cleaned up using a crude force field and the bound matrix[28]. In this study, ligand conformations generated using RDKIT were minimized using the Universal Force Field 'uff' as suggested by Ebejer et al[20b]. 100 conformations were generated followed by minimization and pruning to remove conformations that measure less than 0.25 Å away from each other in RMSD.

BCL::CONF generates conformations for all drug-like small molecules

While BCL, CONFGEN, MOE and RDKIT are able to generate conformations for all the molecules of the VERNALIS dataset, OMEGA could not for 16 molecules due to missing fragments in its library.

Recovery of experimentally observed conformations

The native conformation recovery by BCL, CONFGEN, MOE, OMEGA and RDKIT is plotted in Figure 2-6A. Figure 2-6A shows the percent recovery of native conformation of ligands at different RMSD cutoff values. BCL recovers native conformer for 11 % of ligands within 0.25 Å, 79 % within 1.0 Å and 99% within 2.0 Å. Figure 2-6C shows the effect of rotamer library source (CSD; single dihedral torsion profiles; and CSD+PDB) and binning strategy (30° or 60°) on conformation recovery. Conformation recovery is slightly lower when fragment rotamers observed in only the CSD are used suggesting unique rotamers or significant deviation from canonical values that are observed in ligands bound to proteins. Recovery is not effected significantly when 60° bins are used.

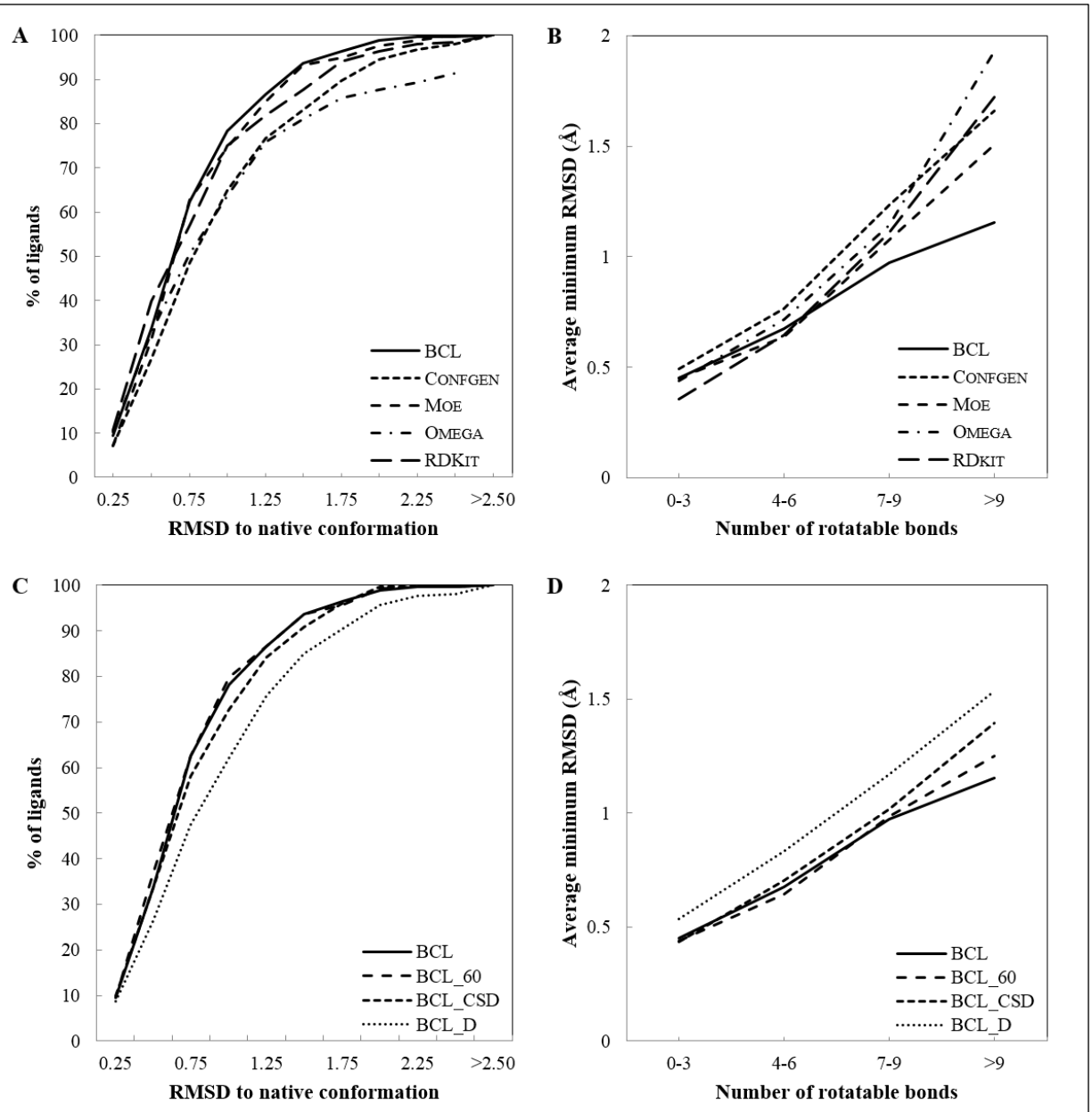


Figure 2-6 Benchmarking results for the VERNALIS dataset. A) The plot represents percentage of ligands (y-axis) for which different methods produce at least one conformer within an RMSD value less or equal to the RMSD value on the x-axis. B) Quality of conformations sampled as the number of rotatable bonds increases. The average RMSD of conformers closest to native structure is plotted on the y-axis as the number of rotatable bonds increases (x-axis). C) Same as A) for different rotamer libraries used. "BCL" refers to recovery using 30° dihedral bins with rotamers derived from both the CSD and PDB. "BCL_CSD" leverages conformations from the CSD only. "BCL_D" refers to experiments in which instead of fragments containing multiple torsion angles, statistics on single dihedral angles were used for sampling conformations. "BCL_60" refers to a rotamer library that uses 60° dihedral bins are. D) Quality of conformations sampled as the number of rotatable bonds increases for BCL, BCL_CSD, BCL_D, and BCL_60. Source: Kothiwale et al[38]

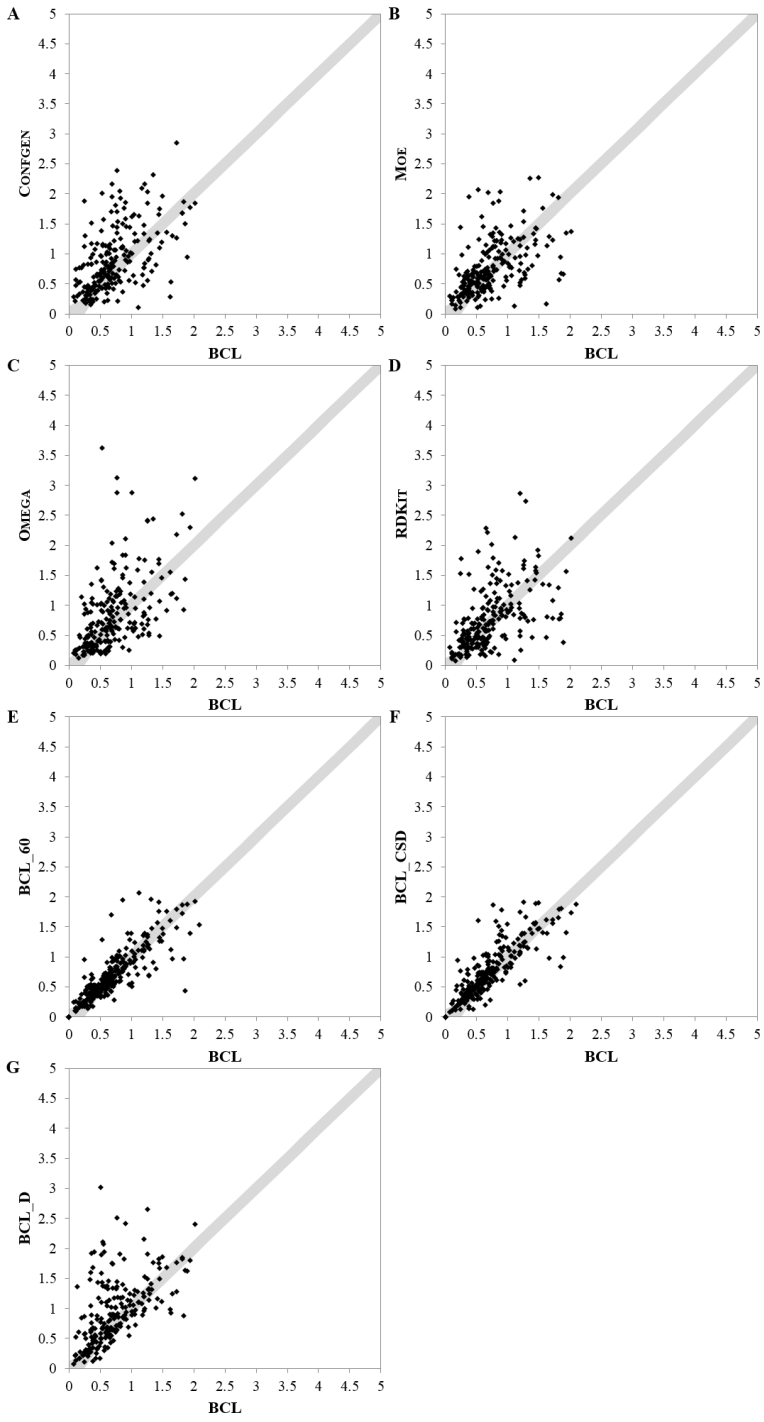


Figure 2-7 Pair-wise comparison of BCL::CONF to other methods. Panels A-G plot the RMSD to native for the BCL on the x-axis, for other methods or flavors of the BCL on the y-axis. BCL::CONF samples closer to native conformations for points that lie above the diagonal. Conformations plotted within the shaded region differ by less than 0.25 Å. Source: Kothiwale et al[38]

Figure 2-7 shows pairwise comparison of CONFGEN, MOE, OMEGA, RDKIT, BCL_60, BCL_CSD and BCL_D to BCL in generating conformer closest to native. Each point corresponds to a molecule in a test set. The coordinates of a point corresponds to the RMSD of closest to native conformer generated by BCL (x-axis) and the method being compared (y-axis). Molecules for which closest to native conformation generated by the pair of methods is within 0.25 Å RMSD of each other are plotted in shaded gray area. For points above the shaded region, BCL recovers lower RMSD conformer compared to the other method referenced. The molecules for which OMEGA could not generate conformations are omitted from the graph and statistical analysis when comparing to BCL. Figure 2-6 and Figure 2-7 suggest that BCL is better than other methods and other flavors of BCL being compared. BCL to those produced by other methods for each molecule in the VERNALIS dataset. Wilcoxon Matched-Pairs Signed-Ranks statistical test was performed to compare conformations generated by different methods. The statistics test was performed using R software package. BCL generated closer to native conformations compared to CONFGEN, MOE, OMEGA and BCL_D at p-value < 0.01 over all the molecules. When compared to BCL_CSD, BCL generates more native like conformations at p-value < 0.05. Statistically there is no significant difference in native recovery between BCL, BCL_60 and RDKIT. However, 30° binning allows recapitulation of frequently observed 90° or 270° rotamers of dihedral bonds containing *aromatic-single-aromatic* or *aromatic-single-any* (Figure 2-3).

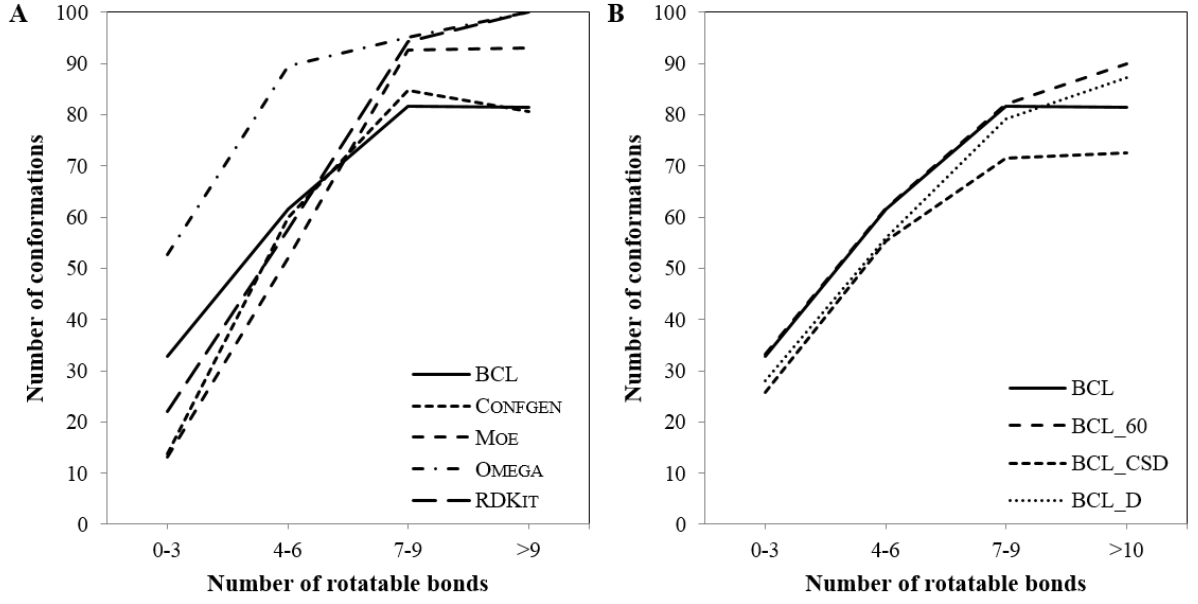


Figure 2-8 Average number of conformations generated by different methods as number of rotatable bonds increase. A) Comparison of commercially available methods to BCL. B) Comparison of different flavors of BCL. Source: Kothiwale et al[38]

Effect of the number of rotatable bonds on native conformation recovery

Figure 2-6B and 6D show the average RMSD of closest to native conformation of molecules plotted against number of rotatable bonds. Figure 2-8 plots the average number of conformations generated by different methods for molecules of different rotatable bonds. BCL is better than other methods at producing closer to native conformers for molecules with greater than six rotatable bonds as suggested by Wilcoxon Paired test at p-value < 0.05. For molecules containing four to six rotatable bonds, BCL performs better than CONFGEN and OMEGA respectively at p-value < 0.01. There is no significant difference between quality of conformations generated between BCL, MOE and RDKIT for molecules with up to six rotatable bonds. For different flavors of BCL, there is no significant difference between BCL and BCL_60 in native conformation recovery based on rotatable bonds. However, statistical analysis clearly shows that using extended fragments improves native conformation recovery compared to using single dihedral bond statistics (BCL_D) for molecules greater than three rotatable bonds at p-value < 0.01. BCL produces closer to native conformations compared to BCL_CSD for molecules with greater than 10 rotatable bonds.

Diversity of conformational space sampled

Diversity of ligand conformations is an important consideration for ligand docking studies. A representative sample that covers ligand's sample space is therefore desired. Figure 2-9A and B show the distribution of RMSDs of

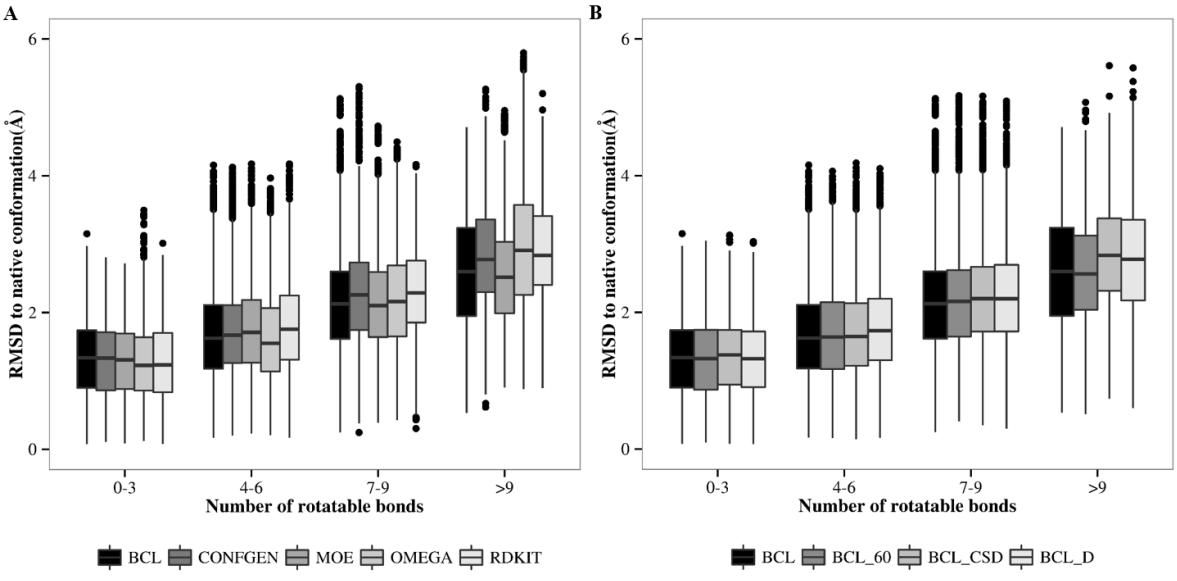


Figure 2-9 The box plots show the diversity of generated molecular conformations depending on the number of rotatable bonds. The upper and lower edges of box correspond to the first and third quartiles. The horizontal dash in the box represents the median value. The whiskers extend from edge to highest/lowest value that is within 1.5 * Inter-Quartile (IQR) of the box, where IQR is the distance between the first and third quartile. A) Conformation diversity produced by different methods B) Conformation diversity obtained by using different flavors of BCL. Source: Kothiwale et al[38]

conformers against the number of rotatable bonds. Box plots show the distribution of conformer RMSD with respect to native structure. The upper and lower edges of box correspond to the first and third quartiles. The whiskers extend from edge to highest/lowest value that is within 1.5 * Inter-Quartile Range (IQR) of the box, where IQR is the distance between the first and third quartile. The data beyond whiskers are plotted as outliers. The horizontal dash in the box represents the median value. Diversity of conformations generated by all the methods is comparable. CONFGEN, MOE and RDKIT sample conformations more efficiently compared to BCL for molecules with up to three rotatable bonds (Figure 2-8). The reason is that smaller fragments have large number of rotamers with similar energy profiles. Larger fragments on the other hand have fewer local minima allowing sampling of relevant conformations in fewer steps.

Comparison of CPU time requirements

The computational run time for the different methods except OMEGA was compared on Intel Xenon model 26 running at 3.2 GHz with 24 GB of RAM. All the methods take less than 2 GB of RAM. BCL generated conformations for a single molecule in 1.6 seconds compared to 1.9 seconds taken by CONFGEN, 5.1 seconds for MOE, 0.5 seconds for OMEGA and 10.2 seconds for RDKIT. Computation time of when using only dihedral torsion profiles i.e. BCL_D is 0.7 s/molecule.

Conclusions

We have developed a conformational search method called the BCL::CONF and validated it against other methods in the field like CONFGEN, MOE, OMEGA and RDKIT. The method utilizes the conformational space seen in the structure databases, CSD and PDB, to sample conformations of small-molecules. BCL::CONF is compared to other methods in three measures that are critical in computational drug discovery process, a) the ability to generate conformation close to experimentally observed structure b) diversity of conformations indication coverage of sample space of molecules c) performance in terms of speed. The benchmark study was performed using a curated dataset of high resolution X-ray crystal structures from the PDB, VERNALIS datasets, containing 253 molecules.

BCL::CONF is capable of reproducing bioactive conformations generating conformers that are structurally close to experimentally determined structures. Analysis of coverage space shows that BCL::CONF generates a diverse set of conformers performing as well as MOE and RDKIT, however in much shorter time. BCL::CONF is better and more efficient in sampling molecules with greater than three rotatable bonds as indicated in Figure 2-6B. Using extended fragments gives BCL::CONF a distinct advantage over other methods in sampling more flexible molecules efficiently. The study shows utility of using explicit fragment conformations to recapitulate protein-bound ligand conformations. A slightly reduced performance is seen when using rotamers derived from only the CSD (Figure 2-6C). The somewhat reduced accuracy could result from biases in the fragment sets between CSB and PDB or biases in dihedral angles between ligands bound to proteins and ligands residing in a crystal. Nonetheless results reported in this paper

suggest that fragment conformations obtained from the CSD seen in structure databases can be used to adequately model small molecule conformations bound to proteins.

BCL::CONF extends the idea of protein side-chain conformer sampling to fragments of small molecules. The method is novel as it takes into account torsion correlations and substituents effects on fragment torsion profiles. It has been designed and developed to be integrated with ROSETTALIGAND that is part of the macromolecular modeling suite ROSETTA.

References

1. Taylor, R. D.; Jewsbury, P. J.; Essex, J. W., A review of protein-small molecule docking methods. *J Comput Aid Mol Des* 2002, *16* (3), 151-166.
2. (a) Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W., Jr., Computational methods in drug discovery. *Pharmacol Rev* 2014, *66* (1), 334-95; (b) Song, C. M.; Lim, S. J.; Tong, J. C., Recent advances in computer-aided drug design. *Brief Bioinform* 2009, *10* (5), 579-591.
3. Shim, J.; Mackerell, A. D., Jr., Computational ligand-based rational design: Role of conformational sampling and force fields in model development. *Medchemcomm* 2011, *2* (5), 356-370.
4. Jones, G.; Willett, P.; Glen, R. C., A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *J Comput Aided Mol Des* 1995, *9* (6), 532-49.
5. (a) Ekins, S.; Mestres, J.; Testa, B., In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. *Br J Pharmacol* 2007, *152* (1), 9-20; (b) Leach, A. R.; Gillet, V. J.; Lewis, R. A.; Taylor, R., Three-dimensional pharmacophore methods in drug discovery. *J Med Chem* 2010, *53* (2), 539-58.
6. Allen, F. H., The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr B* 2002, *58* (Pt 3 Pt 1), 380-8.
7. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Res* 2000, *28* (1), 235-42.
8. (a) Vangunsteren, W. F.; Berendsen, H. J. C., Computer-Simulation of Molecular-Dynamics - Methodology, Applications, and Perspectives in Chemistry. *Angew Chem Int Edit* 1990, *29* (9), 992-1023; (b) Jorgensen, W. L.; TiradoRives, J., Monte Carlo vs molecular dynamics for conformational sampling. *J Phys Chem-Us* 1996, *100* (34), 14508-14513.
9. Lagorce, D.; Pencheva, T.; Villoutreix, B. O.; Miteva, M. A., DG-AMMOS: a new tool to generate 3d conformation of small molecules using distance geometry and automated molecular mechanics optimization for in silico screening. *BMC Chem Biol* 2009, *9*, 6.
10. Halgren, T. A.; Bush, B. L., The Merck molecular force field (MMFF94). Extension and application. *Abstr Pap Am Chem S* 1996, *212*, 2-Comp.
11. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M., Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J Comput Chem* 1983, *4* (2), 187-217.
12. Klebe, G.; Mietzner, T., A Fast and Efficient Method to Generate Biologically Relevant Conformations. *J Comput Aid Mol Des* 1994, *8* (5), 583-606.
13. Brameld, K. A.; Kuhn, B.; Reuter, D. C.; Stahl, M., Small molecule conformational preferences derived from crystal structure data. A medicinal chemistry focused analysis. *J Chem Inf Model* 2008, *48* (1), 1-24.
14. Scharfer, C.; Schulz-Gasch, T.; Hert, J.; Heinzerling, L.; Schulz, B.; Inhester, T.; Stahl, M.; Rarey, M., CONFECT: Conformations from an Expert Collection of Torsion Patterns. *ChemMedChem* 2013, *8* (10), 1690-700.
15. Feuston, B. P.; Miller, M. D.; Culberson, J. C.; Nachbar, R. B.; Kearsley, S. K., Comparison of knowledge-based and distance geometry approaches for generation of molecular conformations. *J Chem Inf Comp Sci* 2001, *41* (3), 754-763.
16. Bush, B. L.; Sheridan, R. P., Patty - a Programmable Atom Typer and Language for Automatic Classification of Atoms in Molecular Databases. *J Chem Inf Comp Sci* 1993, *33* (5), 756-762.
17. Dunbrack, R. L., Rotamer libraries in the 21(st) century. *Curr Opin Struc Biol* 2002, *12* (4), 431-440.

18. Davis, I. W.; Baker, D., ROSETTALigand docking with full ligand and receptor flexibility. *Journal of Molecular Biology* 2009, *385* (2), 381-92.
19. Chen, I. J.; Foloppe, N., Conformational sampling of druglike molecules with MOE and catalyst: implications for pharmacophore modeling and virtual screening. *J Chem Inf Model* 2008, *48* (9), 1773-91.
20. (a) Chen, I. J.; Foloppe, N., Drug-like bioactive structures and conformational coverage with the LigPrep/ConfGen suite: comparison to programs MOE and catalyst. *J Chem Inf Model* 2010, *50* (5), 822-39; (b) Ebejer, J. P.; Morris, G. M.; Deane, C. M., Freely Available Conformer Generation Methods: How Good Are They? *J Chem Inf Model* 2012, *52* (5), 1146-1158.
21. (a) Wang, R.; Fang, X.; Lu, Y.; Wang, S., The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J Med Chem* 2004, *47* (12), 2977-80; (b) Wang, R.; Fang, X.; Lu, Y.; Yang, C. Y.; Wang, S., The PDBbind database: methodologies and updates. *J Med Chem* 2005, *48* (12), 4111-9; (c) Wang, R.; Lu, Y.; Fang, X.; Wang, S., An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein-ligand complexes. *J Chem Inf Comput Sci* 2004, *44* (6), 2114-25.
22. Cahn, R. S.; Ingold, C.; Prelog, V., Specification of Molecular Chirality. *Angewandte Chemie-International Edition* 1966, *5* (4), 385-&.
23. Krissinel, E. B.; Henrick, K., Common subgraph isomorphism detection by backtracking search. *Software Pract Exper* 2004, *34* (6), 591-607.
24. Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J., Description of Several Chemical-Structure File Formats Used by Computer-Programs Developed at Molecular Design Limited. *J Chem Inf Comp Sci* 1992, *32* (3), 244-255.
25. Watts, K. S.; Dalal, P.; Murphy, R. B.; Sherman, W.; Friesner, R. A.; Shelley, J. C., ConfGen: a conformational search method for efficient generation of bioactive conformers. *J Chem Inf Model* 2010, *50* (4), 534-46.
26. MOE (The Molecular Operation Environment). http://www.chemcomp.com.
27. Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; 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* 2010, *50* (4), 572-584.
28. RDKit documentation.
29. (a) Bostrom, J., Reproducing the conformations of protein-bound ligands: a critical evaluation of several popular conformational searching tools. *J Comput Aided Mol Des* 2001, *15* (12), 1137-52; (b) Bostrom, J.; Greenwood, J. R.; Gottfries, J., Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *J Mol Graph Model* 2003, *21* (5), 449-62.
30. Perola, E.; Charifson, P. S., Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *J Med Chem* 2004, *47* (10), 2499-510.
31. Dixon, J. M. B. a. J. S., Distance Geometry in Molecular Modeling. In *Reviews in Computational Chemistry*, 2007; Vol. 5.
32. Li, J.; Ehlers, T.; Sutter, J.; Varma-O'Brien, S.; Kirchmair, J., CAESAR: A new conformer generation algorithm based on recursive buildup and local rotational symmetry consideration. *J Chem Inf Model* 2007, *47* (5), 1923-1932.
33. Kirchmair, J.; Laggner, C.; Wolber, G.; Langer, T., Comparative analysis of protein-bound ligand conformations with respect to catalyst's conformational space subsampling algorithms. *J Chem Inf Model* 2005, *45* (2), 422-30.
34. Smellie, A.; Stanton, R.; Henne, R.; Teig, S., Conformational analysis by intersection: CONAN. *J Comput Chem* 2003, *24* (1), 10-20.
35. O'Boyle, N. M.; Vandermeersch, T.; Flynn, C. J.; Maguire, A. R.; Hutchison, G. R., Confab - Systematic generation of diverse low-energy conformers. *J Cheminformatics* 2011, *3*.
36. J. Gasteiger, C. R., J. Sadowski, Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Computer Methodology* 1990, *3* (6), 537-547.
37. Labute, P., LowModeMD--implicit low-mode velocity filtering applied to conformational search of macrocycles and protein loops. *J Chem Inf Model* 2010, *50* (5), 792-800.
38. Kothiwale, S.; Mendenhall, J. L.; Meiler, J., BCL::CONF: small molecule conformational sampling using a knowledge based rotamer library. J Cheminform 2015, 7, 47. http://creativecommons.org/licenses/by/4.0/

CHAPTER 3 : MULTICONFORMATIONAL 3D – QSAR MODELS

Introduction

Quantitative structure-activity relationship (QSAR) models describe the mathematical relationship between structural attributes of molecules and their target response. A major challenge in the field is incorporating molecular conformation information into the QSAR models. Currently QSAR models are trained on a single low energy conformation that may or may not be the conformation that binds a protein target. This limitation is currently addressed by choosing conformation of active molecules that align structurally and pharmacologically. This study seeks to extend the single conformation models by directly training machine learning algorithms with multiple conformations of molecules to better correlate bioactivities to pharmacologically relevant molecules. Here, we have developed QSAR models using artificial neural networks using numerical descriptors derived from the chemical structure of molecules and those derived from 3-dimensional conformations. Best performance was achieved when model was trained on multiple distinct conformations of active molecules and descriptor averages of inactive conformations.

Classical QSAR is known as the Hansch-Fujita approach and involves the correlation of various electronic, hydrophobic, and steric features with biological activity[1]. In the 1960s, Hansch and others began to establish QSAR models using various molecular descriptors including physical, chemical, and biological properties to computationally estimate bioactivity of molecules [2]. In 1964, Free-Wilson developed a mathematical model relating the presence of various chemical substituents to biological activity (each type of chemical group was assigned an activity contribution) and the two methods were later combined to create the Hansch/Free-Wilson method[3].

Many flavors of QSAR approaches have been developed like the 2D (two-dimensional) and 3D (three-dimensional) QSAR with differences in chemical descriptors and different mathematical approaches that are used to find correlations between the target and the descriptors[4-5]. The general workflow of a QSAR-based drug discovery project is to first collect a group of active and inactive ligands and then create a set of mathematical descriptors that describe the physicochemical and structural properties of those compounds[5]. A model is then generated to identify the relationship between those descriptors and their experimental activity maximizing the predictive power. Finally, the model is applied to predict activity of a library of test compounds which are encoded with the same descriptor sets[5-6]. Success of QSAR, therefore, depends not only on the quality of the initial set of active/inactive compounds, but also on the choice of descriptors and the ability to generate the appropriate mathematical relationship. One of the most important considerations regarding this method is the fact that all models generated will be dependent on the chemical space of the initial set of compounds with known activity, the chemical diversity. In other words, divergent scaffolds or functional groups not represented within this “training” set of compounds will not be

represented in the final model and any potential hits within the library to be screened that contain these groups will likely be missed. Therefore, it is advantageous to cover a wide chemical space within the training set[6-7].

A suite of QSAR algorithms have been implemented in BCL cheminformatics software[8] that is developed in-house. The most successful models have been developed using neural networks trained on dimensional (3D) descriptors derived from a single low-energy 3-D conformations of dataset molecules[8]. However, a molecule may have multiple low energy conformations, one of which could be bioactive. Thus, a single low-energy conformation may not be the one that binds the target of interest and models trained using such inputs may lead to suboptimal performance. In addition, some ligands have been found to bind in multiple conformations within the same pocket of the target. For example, HIV protease inhibitors bind the symmetric binding site of protease dimer in nearly two different but identical binding modes[9]. The work described here extends the use of 3D information by using different approaches to handle conformations of molecules. The hypothesis is to train QSAR models using an ensemble of molecules to identify the correct binding conformation that will improve model performance. Here neural networks were trained using a representation of multiple conformations so that the network will learn the general 3D shape that fits the binding pocket of the target protein. The assumption here is that molecules that bind a particular pocket of a given target protein adopt similar shapes and share common interactions with residues. We hypothesize that QSAR models trained on multiple conformations has the potential to identify 3D pharmacophore that is common between active molecules and not available to inactives, leading to better classification. Different approaches were used in this study to choose most likely conformation by descriptor averaging of multiple conformations, or selection of most positively predicted conformation. Since the number of inactive molecules is large, a descriptor average of all conformations may cover the entire space of inactive pharmacophores.

Quantitative structure activity relationship models correlate biological activity to molecular activity

QSAR models describe mathematical relationships between biological activity and molecular properties. Molecular properties are called descriptors and describe structural and physiochemical properties of ligand molecules such as logP, pK_a, molecular weight, geometry, surface area and volume, polarizability, symmetry, solvation properties etc[10]. These descriptors are generated through knowledge-based, graph-theoretical[11], molecular-mechanics or quantum-mechanics[12] methods and are classified according to the "dimensionality" of chemical representation from which they are computed. QSAR models are accordingly classified according to the dimension of descriptors that are used[4].

1D descriptors derived from molecular formula

One-dimensional descriptors encode numerically generic properties like molecular weight, molar refractivity and octanol/water partition coefficient describing size, shape and lipophilicity of molecules in a low dimension[4].

Even though simple, 1-D descriptors have been used to develop rules for drug-like character of molecules known as the Lipinski's rule [13]. Thus, they are most often found in biological descriptors in any QSAR studies.

2D descriptors derived from topological information

2D-QSAR uses descriptors that are calculated from constitutional representation of molecules[4, 10f]. Many 2D descriptors are based on molecular topology derived from graph-theoretical methods[11]. Topological indices treat all atoms of a molecule as vertices and bonds as edges in a graph, and store atom specific information or pair-wise atomic information. An example of simple topological index is the adjacency matrix that contains only constitutional information such as atoms directly bound to each other. The adjacency matrix is symmetric with each dimension equal to the number of atoms in the molecule. It contains an entry of "one" for atoms that are bonded and zero otherwise making the diagonal zero[14]. A more informative adjacency matrix will code for bond type by an integer as defined by an enumerated list of bond types instead of only ones and zeros[15]. Other topological indices may include entries for number of bonds linking the vertices. A topological index that includes information such as heteroatoms and multiple bonds through the weighting of vertices and edges was introduced by Bertz [16]. Topological correlation (2D) is designed to represent constitutional information as well as atom property distribution by analyzing bond distances between all pairs of atoms. The autocorrelation vector is created by summing all products for atom pairs within increasing distance intervals in terms of number of bonds. In other words, it creates a frequency plot for a specific range of atom pair distances[17]. By including atom property coefficients for all atom pairs, autocorrelations are capable of plotting the arrangement of specific atom properties. For example, information such as the frequency at which two negatively charged atoms are three bonds apart versus four bonds apart is stored in an autocorrelation plot that has been weighted by partial atomic charge [17].

2.5D descriptors incorporate isometry information

2.5D descriptors describe configuration of molecules and therefore encode stereochemistry information. 2D descriptors take into account only the constitution of a molecule that makes it impossible to distinguish between stereoisomers and in particular enantiomers.

3D descriptors represent geometrical properties

3D-QSAR take into account qualities that are three-dimensional[18]. 3D autocorrelation is similar to 2D autocorrelation but the distances are measured as Euclidean distances in 3D space. This allows continuous measure of distances and encodes spatial distribution of physiochemical properties. The atomic pairs are summed into interval steps instead of summing them within discrete shortest path. Radial distribution functions (RDFs) are popular 3D descriptors that map the probability distribution of an atom in a spherical volume of radius r[17, 19]. It is

often combined with atomic properties to provide information regarding interatomic distances between atoms and properties[20]. RDFs allow estimation of molecular flexibility by "fuzziness" coefficient that allows for small changes in interatomic distances. Radial distribution functions have been successfully employed in the development of QSAR models for example in the study of A2A adenosine receptor agonist effect of 29 adenosine analogues[21]. 3D descriptors have proven ability to forecast potency of new scaffolds[18]. A major limitation of 3D QSAR is that it uses lowest energy conformation of the ligand as bioactive conformation, and it is this single conformation of the ligand which exerts the binding effects[22].

4D descriptors are derived in terms of different conformations

4D-QSAR is an extension of 3D-QSAR where each molecule is treated as an ensemble of different conformations, stereoisomers, tautomers and protonation states. The fourth dimension refers to 3D ensemble sampling of features of each molecule accounting for molecular shape analysis[23]. The modelling is likely to involve a number of steps including generation of conformations and molecular alignment or alignment of specific substructure groups. Receptor independent 4D-QSAR is carried out by placing all molecules in a grid and aligning the molecules based on pharmacophore elements (polar, nonpolar, hydrogen bond donor etc.) using conformation sampling using molecular mechanics or Monte Carlo approach[23].

BCL Quantitative structure activity relationship algorithm

BCL::CHEMINFO is an in house developed cheminformatics library that has modules for developing numerical descriptors for small molecules. The software suite also contains implementations of non-linear modelling algorithms like neural networks, k-nearest neighbors (KNNs), support vector machines (SVMs) etc. Neural networks are popular self-organizing algorithms that can learn non-linear relationships between descriptors and biological activity through iterative prediction and improvement cycles[24]. They are composed of compute nodes known as the neurons (Figure 3-1). A neural network is composed of an input layer, single or multiple hidden layers and an output layer. Each neuron in a given layer is connected to a neuron in the next layer. The neural networks used in this study are called feedforward neural networks because output from one layer is used as an input to the next layer. The input for a neuron is the weighted sum of all the incoming output or activations from neurons in the previous layer. If w_{jk}^{l} denotes weight for connection from k^{th} neuron in $(l-1)^{\text{th}}$ layer to the j^{th} neuron in the l^{th} layer, b_j^l is bias for j^{th} neuron in the l^{th} layer and a_j^l for the activation of j^{th} neuron in the l^{th} layer, then a_j^l is related to activations in $(l-1)^{\text{th}}$ layer by the equation

$$a_j^l = \sigma\left(\sum_k w_{jk}^l a_k^{l-1} + b_j^l\right)$$

where sum is over all the k neurons in the $(l-1)^{th}$ layer. The quantity on the right side of the equation is called *weighted input* to the neurons in layer *l*. The goal of training neural network is to find weights and biases such that output from the network approximates y(x) for all training inputs x. A cost function is defined to quantify this goal

$$C(w,b) = \frac{1}{2n}\sum_x \left||y(x) - a|\right|^2$$

where w denotes the collection of all weights in the network, b all the biases, n is the total number of training inputs, a is the vector of outputs from the network when x is input, and the sum is over all training inputs, x. The training algorithm performs well when $C(w,b) \approx 0$ and by contrast when it is large, it means $y(x)$ is not close to the real output for a large number of units. Thus, the aim of the training algorithm is to minimize the cost $C(w,b)$ as a function of weights and biases. The errors or deltas i.e. the difference between the input and the output values, of all output and hidden neurons are backpropagated to update weights. The output delta and input activation are multiplied to get gradient of the weight and a percent of gradient is subtracted from the weight. The greater the ratio, faster the neuron trains; while lower the ratio, the more accurate the training is. The sign of the gradient indicates where the error is increasing, and so weight is updated in the opposite direction. This process is repeated over and over again until the performance reaches a satisfactory level.

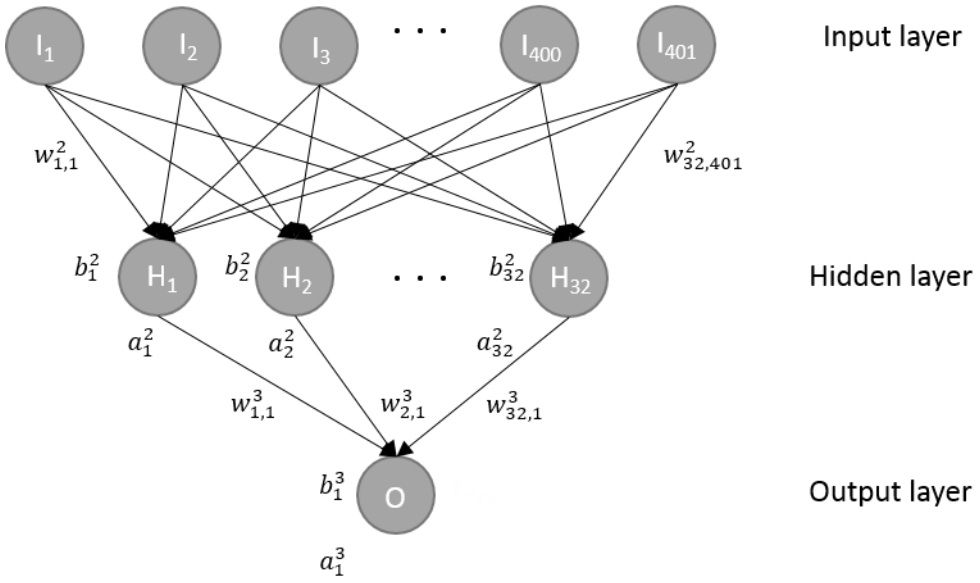


Figure 3-1 Neural network architecture showing input, hidden and output layer with 401, 32 and 1 neuron respectively. Each neuron in the hidden and output layer has an associated bias, weights and activations which are labelled.

Table 3-1 List of descriptors used for describing molecules for QSAR models

1D descriptors	2D autocorrelation descriptors	3D autocorrelation descriptors
Molecular weight	Atom sigma charge	Atom sigma charge
HbondDonor	Atom vcharge	Atom vcharge
HbondAcceptor	Atom in aromatic ring	Atoms in aromatic ring
LogP	Atom in fused aromatic ring	Atoms in fused aromatic ring
Total Charge	Atom signed polarizability	Atom signed polarizability
Number of rotatable bonds	Atom heavy sigma charge	Atom heavy sigma charge
Number of rings	Atom heavy vcharge	Atom heavy vcharge
Topological polar surface area		
Molecular girth		
Maximum ring size		
Bond girth		
Number of atoms in aromatic rings		
Number of atoms in fused aromatic ring		
Number of atoms in fused rings		
Atom Vcharge statistics		
Atom sigma charge statistics		

k-nearest neighbor (KNN) is the simplest pattern recognition method. KNNs utilize Euclidian distance metric to cluster sample data points within close proximity to each other. Output is a class membership and object is classified by a majority vote of its neighbors, so that the object is assigned to the class with most common among its k nearest neighbors where k is a positive integer. Support vector machines (SVMs) is a kernel-based supervised learning method which seeks to divide sets of patterns based on their class[25]. SVM is a maximal range classifier that seeks to define a hyperplane with the widest margin between two classes. The patterns that line the closest border of each class define the two hyperplanes separated by that margin. These patterns are known as support vectors and represent maximal margin solution and used to predict classes for novel unclassified patterns[26]. Decision trees is a supervised learning algorithm that works by iteratively grouping training data into small and more specific groups[27]. The resulting classification resembles a tree where classification is performed based on feature rules. Once a decision tree is optimized for training set, new compounds can be classified by applying decision tree rules on their descriptors[7d].

In the BCL::CHEMINFO, QSAR models with most predictive ability have been developed using artificial neural networks and a set of molecular descriptors described previously[8]. A major drawback of neural networks is that they are sensitive to over-training resulting in excellent predictive ability in the training set but reduced a reduced performance to assess novel compounds. Regularization of BCL::CHEMINFO neural networks through the use of dropouts prevents overfitting and better generalization of the QSAR models[28]. The descriptors are translationally and rotationally invariant geometric functions that describe the distribution of molecular properties. Table 3-1 lists the set of descriptor that are used for developing QSAR models using neural networks. In the current study, QSAR models were developed using multiple molecular conformations generated using BCL::CONF[29]. The neural networks

Table 3-2 Datasets used for benchmarking BCL::CHEMINFO QSAR models.

Protein Target Class	Protein Target	PubChem AID	Number Actives (%)	Number Inactives	Conformations per active molecule	Conformations per inactive molecule
GPC61R	M1 Muscarinic Receptor	1798	188 (0.3)	61,661	43	47
	Orexin1 Receptor	435008	230 (0.1)	218,071	37	47
	M1 Muscarinic Receptor	435034	448 (0.72)	61,407	53	47
Ion Channel	Potassium Ion Channel	1834	172 (0.05)	301,473	55	47
	KCNQ2 potassium channel	2258	213 (0.07)	302,351	38	47
	Cav3 T-type Calcium Channels	463087	703 (0.7)	100,210	56	45
Kinase Inhibitor	Serine/Threonine Kinase	2689	172 (0.05)	319,821	25	47
Transporter	Choline Transporter	488997	252 (0.08)	302,246	48	47

were trained with multiple conformations with the goal of identifying conformation patterns that are important for interactions between the small molecules and the target molecule. The hypothesis is that all the active molecules bind in a similar pose so that their 3D chemical fingerprint will be aligned in terms of the distribution of charge, volume, surface area, polarizability etc. In simplest terms, all the active molecules should be capable of adopting conformation conforming to the active site of target molecule but inactive molecules may not have any conformation that would fit the binding site.

Results and Discussion

Multiple new approaches were implemented to use multiple conformations for training QSAR models. For the sake of completeness all the approaches for developing QSAR models using BCL::CHEMINFO are described below. For each of these approaches same descriptor sets were used for training. The datasets used in this study were compiled by Butkiewicz et al from publically available libraries deposited in PubChem[8]. Datasets were compiled such that the target is one specific protein and contain a minimum of 150 confirmed active compounds. The targets include pharmaceutically relevant small molecule protein targets such as GPCRs, ion channels, transporters and kinase inhibitors. All PubChem confirmatory screens for active molecules are given by PubChem assay ids (AID). An overview of datasets used in this study is provided in Table 3-2(adapted from Butkiewicz et al.)[8]. The table provides the number of active/inactive and percentage of active molecules in the dataset. It also lists the average number of conformations generated by BCL::CONF for each dataset.

Models

CORINA – Single low energy conformation is generated for each of the molecules in the dataset. ANNs were trained using dropout parameter of 0.25 for hidden layer and 0.05 for visible layer.

BCL_ONE – Single best scoring conformation was selected from among ~50 conformations (Table 1) generated using BCL::CONF. ANN was trained using features of single conformations using dropout parameters (H: 0.25, V: 0.05).

BCL_CONFORMATION_AVG – In this experiment, molecular descriptors calculated for multiple ligand conformations were averaged for training the neural network. At most top 100 best scoring and diverse BCL::CONF conformations were used which differ from each other by an rmsd of at least 0.25 Å conformations.

BCL_CONFORMATION_AVGStdDev – In addition to the average descriptor value, standard deviations were used for training neural network. Essentially the descriptor size is double compared to one used in BCL_CONFORMATION_AVG.

BCL_FIVE_PREVROUND – Five diverse and lowest scoring conformations were used for training. A new back-propagation scheme was implemented for this experiment. In each iteration, predictions are computed for conformations of a molecule using the current state of the network. Errors from only one conformation is back-propagated when activity prediction for a conformation satisfy condition that it is better than it is better than the one seen for molecule in the last round and is better than any other conformation prediction in the current round. This experiment was performed with dropout parameters.

BCL_FIVE_NODP – The backpropagation scheme is an improved implementation compared to the BCL_FIVE_PREVROUND experiment. For a molecule, errors from only the best predicted conformation is backpropagated after predictions on all the five conformations have been made. Dropout scheme was not implemented in this experiment.

BCL_ONE_NODP – This experiment was performed with single conformation generated using BCL::CONF but with no neuron dropout during training. Results from this experiment is control for BCL_FIVE_NODP.

BCL_FIVE – The backpropagation scheme to update weights based on errors associated with the best predicted conformation in the current round was implemented with dropout functionality.

BCL_TWOSTAGE – In this experiment, BCL_CONFORMATION_AVG and BCL_ONE are used sequentially for developing the QSAR model. As shown in figure 3-2, QSAR models developed using BCL conformations perform best when conformation descriptors are averaged. This model was used to predict activity of every conformation for all the molecules in a dataset. Conformation predicted to be most active is then used to train BCL_ONE model.

BCL_AvgInactive_MultiActive – Inactive molecules were represented by averaged descriptor values calculated from multiple conformations. Active molecule conformations were used explicitly for training. The hypothesis here is that a single representation of all inactive molecules will suffice for covering the inactive space. For cross-validation

predictions, multiple conformations of all molecules were scored and the best predicted conformation was used for classifying actives from inactive molecules.

Model evaluation

ANN models are typically analyzed by receiver operator characteristics (ROC) curves to assess their predictive power. In this study, QSAR models were evaluated by means of a false positive rate – true positive rate (FPR-TPR) curve. Figure 3-2 shows ROC curves in green for experiments BCL_AvgInactive_MultiActive and BCL_five for dataset AID1834. If a ROC curve is a diagonal, it represents performance of a random predictor and has an area under the curve (AUC) value of 0.5. Higher integral value means a better model. However, since in a typical virtual screening experiment predictions are made for a large compound library and a small fraction of compounds are used for experimental testing (1% or 10^3), it is important that the initial 1000 compounds predicted as most active are actually active. Thus a global AUC value is of much less value compared to initial integral of TNR-TPR curve. For this reason, the initial section of the ROC curve is the most important and AUC value for this region is reported for comparison. To achieve this the area under the curve of a logarithmic x-axis ROC curve is used to quantify high confidence

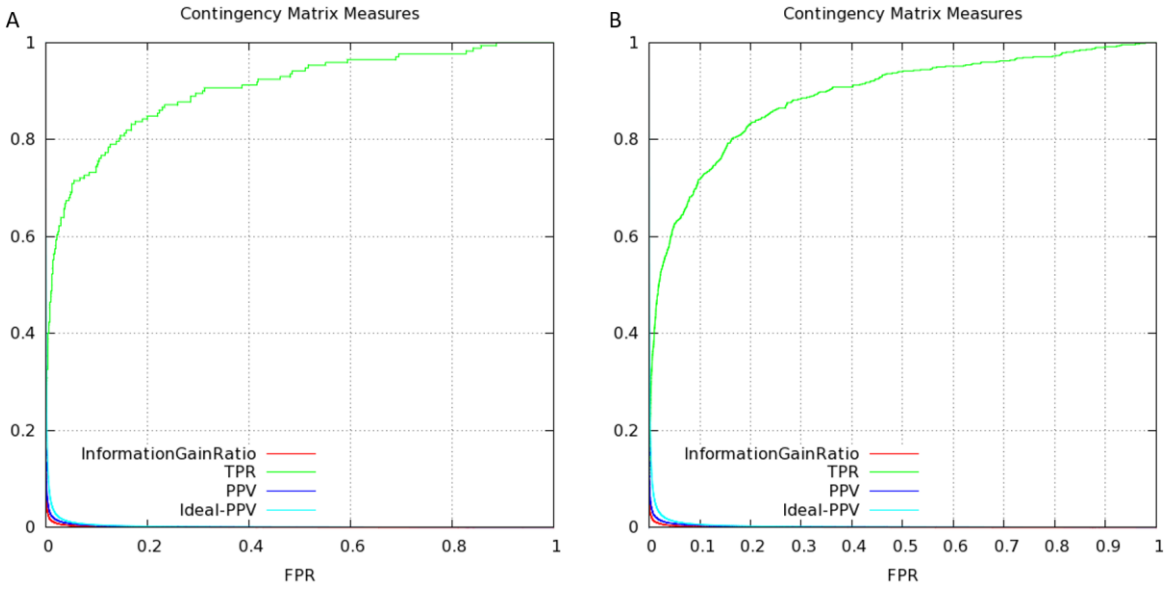


Figure 3-2 Area under curve for predictive performance of model A) BCL_AvgInactive_MultiActive B) BCL_five for dataset 1834. The area under curve is the plot of true positive rate (TPR) vs false positive rate (FPR). A value of one corresponds to perfect prediction and a value of 0.5 corresponds to a model that produces random results.

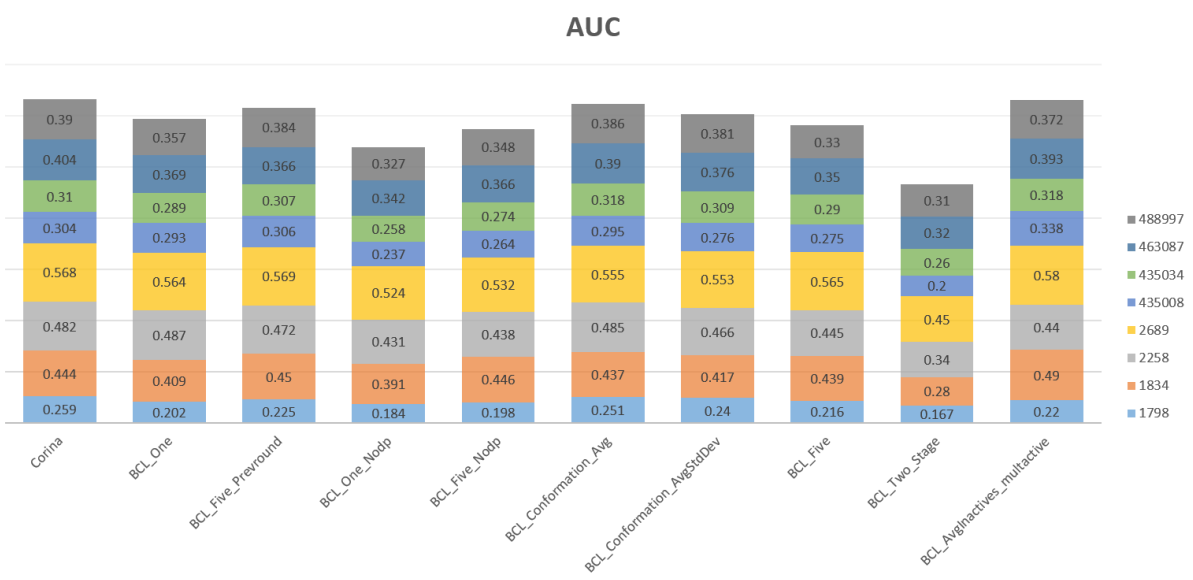


Figure 3-3 Predictive ability of neural networks trained using different routines for generating multi-conformation QSAR models. Each model was tested for prediction on each of the datasets which are color coded as indicated in the legend. The metric for describing predictive power is the log (AUC) value which is reported in the stacked column for each dataset. For example, for experiment BCL_Five, model generated for dataset 1834 had a log (AUC) value of 0.439.

predictions. Figure 3-3 compares model performance using log AUC values across eight datasets when using the different neural network training approaches of training with multiple conformations described in this paper.

The current QSAR models in BCL::CHEMINFO are trained using single low-energy conformation obtained from Corina. The routine BCL_ONE can be directly compared to BCL_CORINA as a single BCL::CONF generaed conformation is used for training the neural network. BCL_ONE performs slightly worse (~10%) compared to BCL_CORINA for datasets AID1834, AID435034, AID463087 and AID488997. Two multi-conformation QSAR models that perform at par with BCL_CORINA were BCL_CONFORMATION_AVG and BCL_AvgInactive_MultiActive. BCL_CONFORMATION_AVG used average descriptor calculated from an average of ~50 conformations (Table 3-2) of molecules in the dataset while training of BCL_AvgInactive_MultiActive used average descriptor of inactive conformations and explicit active conformations. According to Wilcoxon paired test, there is no significant difference between logAUC values obtained for different datasets by models BCL_CORINA, BCL_AvgInactive_MultiActive and BCL_CONFORMATION_AVG. However, we do see some improvement using multiple conformations over the use of single conformations generated using BCL::CONF. According to Student's t-test, there is statistically significant improvement in predictions of BCL_Five_Prev_Round over BCL_ONE for only two datasets AID1834 and AID488997. BCL_CONFORMATION_AVG and BCL_AvgInactive_MultiActive have significantly better predictive ability compare to BCL_ONE for datasets AID1834, AID1798, AID488997 and AID 463087. However, a surprising result was that BCL_Twostage performed much worse

compared to BCL_CONFORMATION_AVG and BCL_ONE even though it is composed of these two experiments performed sequentially.

One of the reasons why we could not get a great lift in predictive performance could be that the descriptor set used or the general parameters of neural network training were not optimal. The modified neural networks described in this work were trained using parameters found to be optimal when training with CORINA generated conformations. The descriptor set was also a reduced set optimized for training standard BCL::CHEMINFO QSAR models. Future experiments could be performed to benchmark neural network training parameters to optimize training with conformations generated using BCL::CONF. These experiments can be done by training the neural networks with the actual binding conformations of active molecules. No significant improvement will indicate that parameter tuning is required. Another approach to improve model performance is to train using common 3D shapes that the active molecules can assume. This can be achieved by aligning conformations of different molecules with each other and identifying the ones that are common to most active molecules. Reduced noise in the input data may lead to better models.

Conclusions

The goal of this study was to develop QSAR models using multiple molecular conformations. Current BCL::CHEMINFO QSAR models use a single low energy conformer generated by CORINA to develop QSAR model. The hypothesis for this work is that QSAR model performance could be improved if binding conformation of active molecules is known. The single low energy Corina generated conformation is most likely not the binding conformation. Here we have trained neural networks using multiple conformations such that the network can identify conformation closest to binding conformation. The goal of this research is to improve the performance of neural networks by allowing the network to learn about the pharmacophore from conformational space of active molecules. Several different approaches were implemented including modification of the backpropagation algorithm, use of average descriptors calculated over ensemble of conformations and iterative model building process.

Using average descriptors calculated over ensemble of BCL::CONF generated conformations performed better than modified backpropagation approaches. We then used a hybrid approach which produced the best result. Here an average descriptor was calculated from on average of ~50 conformations of each inactive molecule. We hypothesize that this approach covers most of the inactive pharmacophore space given the large number of inactive molecules. For the active molecules, explicit ~50 conformations were used. The neural network was trained using a modified backpropagation algorithm where only the errors of the best predicted conformation are backpropagated. Further studies in parameter tuning and descriptor set benchmarking may be required to improve QSAR models using conformations generated using BCL::CONF.

References

1. Hansch, C.; Maloney, P. P.; Fujita, T.; Muir, R. M., Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature* 1962, *194* (4824), 178-180.
2. Hansch, C., Citation Classic - Rho-Sigma-Pi-Analysis - a Method for the Correlation of Biological-Activity and Chemical-Structure. *Current Contents/Life Sciences* 1982, (47), 18-18.
3. (a) Free, S. M., Jr.; Wilson, J. W., A Mathematical Contribution to Structure-Activity Studies. *Journal of Medicinal Chemistry* 1964, *7*, 395-9; (b) Tmej, C.; Chiba, P.; Huber, M.; Richter, E.; Hitzler, M.; Schaper, K. J.; Ecker, G., A combined Hansch/Free-Wilson approach as predictive tool in QSAR studies on propafenone-type modulators of multidrug resistance. *Arch Pharm (Weinheim)* 1998, *331* (7-8), 233-40.
4. Ekins, S.; Mestres, J.; Testa, B., In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. *British journal of pharmacology* 2007, *152* (1), 9-20.
5. Zhang, S., Computer-aided drug discovery and development. *Methods in molecular biology* 2011, *716*, 23-38.
6. Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W., Jr., Computational methods in drug discovery. *Pharmacol Rev* 2014, *66* (1), 334-95.
7. (a) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, II; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A., QSAR modeling: where have you been? Where are you going to? *J Med Chem* 2014, *57* (12), 4977-5010; (b) Bajorath, J., Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J Chem Inf Comput Sci* 2001, *41* (2), 233-45; (c) Bajorath, J., Integration of virtual and high-throughput screening. *Nat Rev Drug Discov* 2002, *1* (11), 882-94; (d) Bajorath, J.; Barreca, M. L.; Bender, A.; Bryce, R.; Hutter, M.; Laggner, C.; Laughton, C.; Martin, Y.; Mitchell, J.; Padova, A.; Renner, S.; Selzer, P. M.; Sherman, W.; Sippl, W.; Taft, C.; Tuccinardi, T.; Vistoli, G.; Willett, P., Ask the experts: focus on computational chemistry. *Future medicinal chemistry* 2011, *3* (8), 909-21.
8. Butkiewicz, M.; Lowe, E. W., Jr.; Mueller, R.; Mendenhall, J. L.; Teixeira, P. L.; Weaver, C. D.; Meiler, J., Benchmarking ligand-based virtual High-Throughput Screening with the PubChem database. *Molecules* 2013, *18* (1), 735-56.
9. (a) Mobley, D. L.; Dill, K. A., Binding of small-molecule ligands to proteins: "what you see" is not always "what you get". *Structure* 2009, *17* (4), 489-98; (b) Lewi, P. J.; de Jonge, M.; Daeyaert, F.; Koymans, L.; Vinkers, M.; Heeres, J.; Janssen, P. A.; Arnold, E.; Das, K.; Clark, A. D., Jr.; Hughes, S. H.; Boyer, P. L.; de Bethune, M. P.; Pauwels, R.; Andries, K.; Kukla, M.; Ludovici, D.; De Corte, B.; Kavash, R.; Ho, C., On the detection of multiple-binding modes of ligands to proteins, from biological, structural, and modeling data. *J Comput Aided Mol Des* 2003, *17* (2-4), 129-34.
10. (a) Cramer, R. D.; Patterson, D. E.; Bunce, J. D., Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society* 1988, *110* (18), 5959-67; (b) Randic, M., Molecular Profiles - Novel Geometry-Dependent Molecular Descriptors. *New Journal of Chemistry* 1995, *19* (7), 781-791; (c) Schuur, J. H.; Selzer, P.; Gasteiger, J., The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. *Journal of Chemical Information and Computer Sciences* 1996, *36* (2), 334-344; (d) Bravi, G.; Gancia, E.; Mascagni, P.; Pegna, M.; Todeschini, R.; Zaliani, A., MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: A comparative 3D QSAR study in a series of steroids. *J Comput Aided Mol Des* 1997, *11* (1), 79-92; (e) Hong, H.; Xie, Q.; Ge, W.; Qian, F.; Fang, H.; Shi, L.; Su, Z.; Perkins, R.; Tong, W., Mold(2), molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. *Journal of Chemical*

Information and Modeling 2008, *48* (7), 1337-44; (f) Roberto Todeschini, V. C., *Molecular Descriptors for Chemoinformatics*
Wiley-VCH Verlag GmbH & Co. KGaA: 2010; p 1-38.
11. Marrero-Ponce, Y.; Santiago, O. M.; Lopez, Y. M.; Barigye, S. J.; Torrens, F., Derivatives in discrete mathematics: a novel graph-theoretical invariant for generating new 2/3D molecular descriptors. I. Theory and QSPR application. *J Comput Aided Mol Des* 2012, *26* (11), 1229-46.
12. Zhou, T.; Huang, D.; Caflisch, A., Quantum mechanical methods for drug design. *Current topics in medicinal chemistry* 2010, *10* (1), 33-45.
13. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J., Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews* 1997, *23* (1-3), 3-25.
14. Trinajstić, N., *Chemical graph theory*. 2nd ed.; CRC Press: Boca Raton, 1992; p 322 p.
15. Devillers, J.; Balaban, A. T., *Topological indices and related descriptors in QSAR and QSPR*. Gordon and Breach: Amsterdam, 1999; p x, 811 p.
16. Bertz, S. H., On the Complexity of Graphs and Molecules. *Bulletin of Mathematical Biology* 1983, *45* (5), 849-855.
17. Moreau, G.; Broto, P., The Auto-Correlation of a Topological-Structure - a New Molecular Descriptor. *Nouveau Journal De Chimie-New Journal of Chemistry* 1980, *4* (6), 359-360.
18. Kubinyi, H.; Folkers, G.; Martin, Y. C., *3D QSAR in drug design*. Kluwer Academic: Dordrecht ; Boston, Mass, 1998; p v. < 2- >.
19. Broto, P.; Moreau, G.; Vandycke, C., Molecular-Structures - Perception, Auto-Correlation Descriptor and Sar Studies - Perception of Molecules - Topological-Structure and 3-Dimensional Structure. *European Journal of Medicinal Chemistry* 1984, *19* (1), 61-65.
20. Hemmer, M. C.; Steinhauer, V.; Gasteiger, J., Deriving the 3D structure of organic molecules from their infrared spectra. *Vibrational Spectroscopy* 1999, *19* (1), 151-164.
21. Gonzalez, M. P.; Teran, C.; Teijeira, M.; Helguera, A. M., Radial distribution function descriptors: an alternative for predicting A2 A adenosine receptors agonists. *Eur J Med Chem* 2006, *41* (1), 56-62.
22. Verma, J.; Khedkar, V. M.; Coutinho, E. C., 3D-QSAR in Drug Design - A Review. *Curr Top Med Chem* 2010, *10* (1), 95-115.
23. Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B. Q.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C., Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *Journal of the American Chemical Society* 1997, *119* (43), 10509-10524.
24. Livingstone, D., *Artificial neural networks : methods and applications*. Humana Press: Totowa, NJ, 2008; p ix, 254 p.
25. Vapnik, V.; Lerner, A., Pattern Recognition using Generalized Portrait Method. *Automation and Remote Control* 1963, *24*.
26. Boser, B. E.; Guyon, I. M.; Vapnik, V. N., A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, ACM: Pittsburgh, Pennsylvania, United States, 1992; pp 144-152.
27. Han, J.; Kamber, M., *Data mining : concepts and techniques*. 2nd ed.; Elsevier ;

Morgan Kaufmann: Amsterdam ; Boston

San Francisco, CA, 2006; p xxviii, 770 p.
28. Mendenhall, J.; Meiler, J., Improving quantitative structure-activity relationship models using Artificial Neural Networks trained with dropout. *J Comput Aided Mol Des* 2016, *30* (2), 177-89.
29. Kothiwale, S.; Mendenhall, J. L.; Meiler, J., BCL::CONF: small molecule conformational sampling using a knowledge based rotamer library. *J Cheminform* 2015, *7*, 47.

CHAPTER 4 : DISCOIDIN DOMAIN RECEPTOR1 (DDR1) STRUCTURE-BASED AND LIGAND-BASED DRUG DISCOVERY

Introduction

DDR1 and DDR2 are tyrosine kinase receptors composed of an extracellular Discoidin (DS) homology domain which encompasses the collagen binding site, a DS-like domain which contributes to collagen-induced receptor activation, an extracellular juxtamembrane region which contains N- and O-glycosylation sites and matrix metalloproteinase cleavage sites. In addition, DDRs have a single transmembrane helix, a cytoplasmic tyrosine kinase domain and additional carboxy-terminal and juxtamembrane regulatory regions (Figure 4-1A)[1]. The DDR family consists of two distinct members, DDR1 and DDR2. DDR1 has five isoforms while DDR2 has a single one[1]. DDR1 receptor is important for cell survival, migration, and differentiation in development and pathological conditions[8]. Current research in Pozzi lab at Vanderbilt University focusses on DDR1 as therapeutic strategy for renal fibrosis. This chapter describes the structure-based studies of DDR1 that improved our understanding of the structure of the DDR1 kinase domain and its interaction with potential inhibitors. These studies involved homology modeling and docking studies with the goal of performing high-throughput docking studies to discover new binders for probing or inhibition of DDR1 kinase. Recently, a large number of new novel inhibitors of DDR1 were reported in scientific literature. We developed QSAR models using DDR1 inhibition data submitted in PUBCHEM and CHEMBL, and used the models to screen virtual compound libraries to identify new scaffolds that inhibit DDR1. Predicted molecules were tested experimentally to verify and identify real DDR1 kinase binders.

Upon activation by binding of fibrillar collagens I-III & V, DDR1 undergoes phosphorylation and initiates various downstream signaling pathways. Multiple tyrosine residues within the intracellular juxtamembrane region and tyrosine kinase domain of DDR1 can be phosphorylated and recruit proteins such as ShcA, SHP-2 and the p85 subunit of PI3K[2]. DDR1 stimulates several signaling pathways in a context and cell type-dependent manner. For example, DDR1 activates ERK signaling in vascular smooth muscle cells[3], but inhibits ERK in mesangial cells[4], and has no effect on ERK activation in T47D breast cancer cells[2d]. In addition, DDR1 modulates signaling pathways initiated by other matrix receptors (e.g., integrins) [5], cytokines (e.g., TGF-β) [6], and transmembrane receptors (e.g., insulin receptors and Notch1) [7]. Interaction of DDR1 with various receptors is important for the regulation of cell survival, migration, and differentiation in development and pathological conditions [8].

Our understanding of the role of DDR1 in development, tissue homeostasis and disease has been significantly enhanced by availability of DDR1-deficient mice. These mice have defects in mammary gland morphogenesis and inability of blastocysts to implant properly in the uterine wall[9]. In contrast to these findings, DDR1 ablation has been

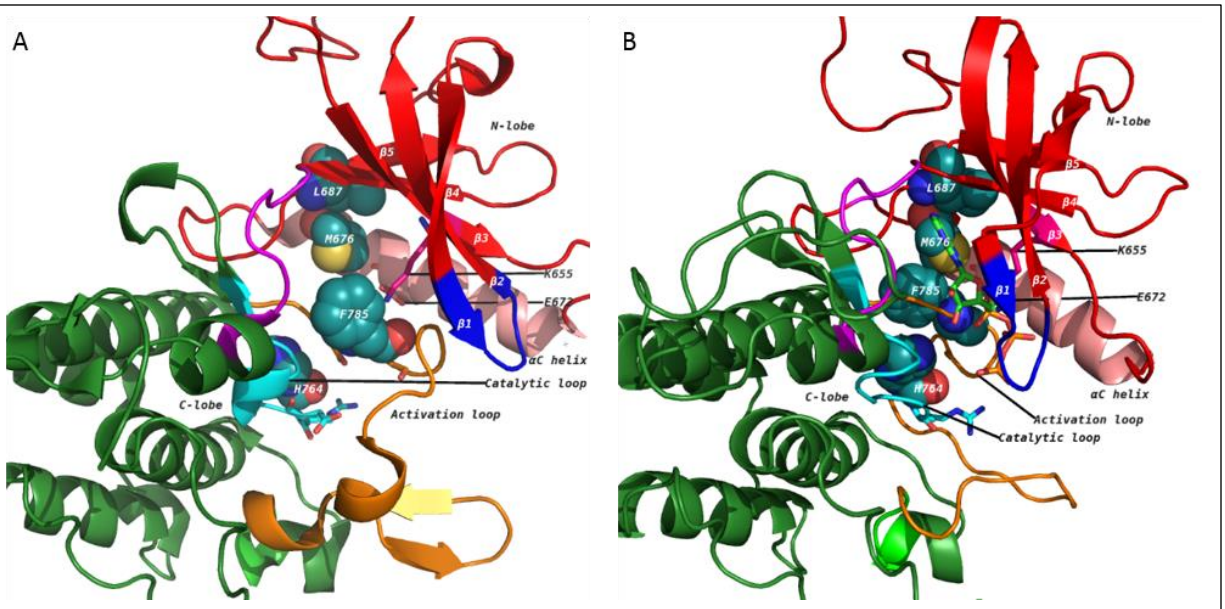


Figure 4-1 A) The DDR1 kinase domain (3ZOS) has the characteristic bilobal architecture. The image shows the DFG-out (F785 belongs to the DFG motif in the catalytic loop) or the "inactive" state. The N-terminal lobe contains five beta strands (1 through 5, in red) and a universally conserved helix called αC (in red). The C-terminal lobe is primarily helical (in green). The catalytic loop is colored in cyan while the activation loop is in orange. The disrupted hydrophobic spine characteristic of inactive kinases is shown in spheres. B) Active DDR1 conformation homology model (DFG-in) showing intact hydrophobic spines in spheres. ATP occupies the cleft between N-lobe and C-lobe. The ATP binding pocket is bound by a glycine rich loop and C-terminal hinge region. Source : Kothiwale et al[33]

shown to have a beneficial role in various mouse models of fibrotic diseases including atherosclerosis [10], pulmonary fibrosis [11], and renal fibrosis [8]. Thus, inhibiting DDR1 may be a promising therapeutic strategy for fibrotic diseases.

The DDR1 kinase domain

DDR1 intracellular kinase domain shares the typical structure of other kinase domains (Figure 4-1). However how DDR1 kinase is activated upon collagen binding is poorly understood. It is discussed that the process is fundamentally different from the accepted paradigm of ligand-induced RTK dimerization. Unlike the typical RTKs, DDR1 exists as a preformed dimer and following collagen binding undergoes receptor oligomerization, internalization and is phosphorylated unusually slowly. A recent study showed that collagen binding to DDR1 fails to induce a major conformational change that could explain kinase activation, and instead proposed that collagen-induced receptor oligomerization may be responsible for the kinase activation [12]. In support of this hypothesis, events that presumably reduce receptor oligomerization such as antibodies that bind to DS-like domain or enforced covalent receptor dimerization at residues within the DS-like domain reduce DDR1 phosphorylation and activation. However, mutation of Asn211, a conserved glycosylation site within the DS-like domain, results in ligand-

independent activation of DDR1, enhanced receptor dimerization, and internalization, suggesting that, in addition to receptor clustering, ligand-induced internalization may also contribute to receptor activation [13].

Collagen binding to DDR1 induces a slow receptor tyrosine auto-phosphorylation of multiple tyrosine residues including Tyr792, Tyr796, and Tyr797 in the activation loop which likely causes the kinase domain to switch from the inactive to the active state[14]. The active state satisfies chemical restraints that allow for the transfer of γ-phosphate of ATP to hydroxyl group of tyrosine on the loop that allows recruitment of substrate proteins. In this chapter the residues of DDR1 kinase domain (PDB: 3ZOS) will be referenced to corresponding residues in protein kinase A (PKA, PDB: 1ATP) and the kinase domain of Abl tyrosine kinase. Positions indicated in italics and normal in square brackets are PKA and Abl kinase residue numbers, respectively, equivalent to those in DDR1.

The tyrosine kinase domain consists of an N-terminal (N-lobe) and a C-terminal (C-lobe) lobe [15]. Figure 4-1A shows the kinase domain of DDR1 (PDB: 3ZOS). The N-lobe consists of a five-stranded beta-sheet and a prominent α-helix, called αC helix while C-lobe is mostly helical [15]. The ATP binding pocket lies in the cleft between the two lobes and sits beneath a highly conserved glycine-rich loop which is between β1 and β2 strands [15] (Figure 4-1A). In

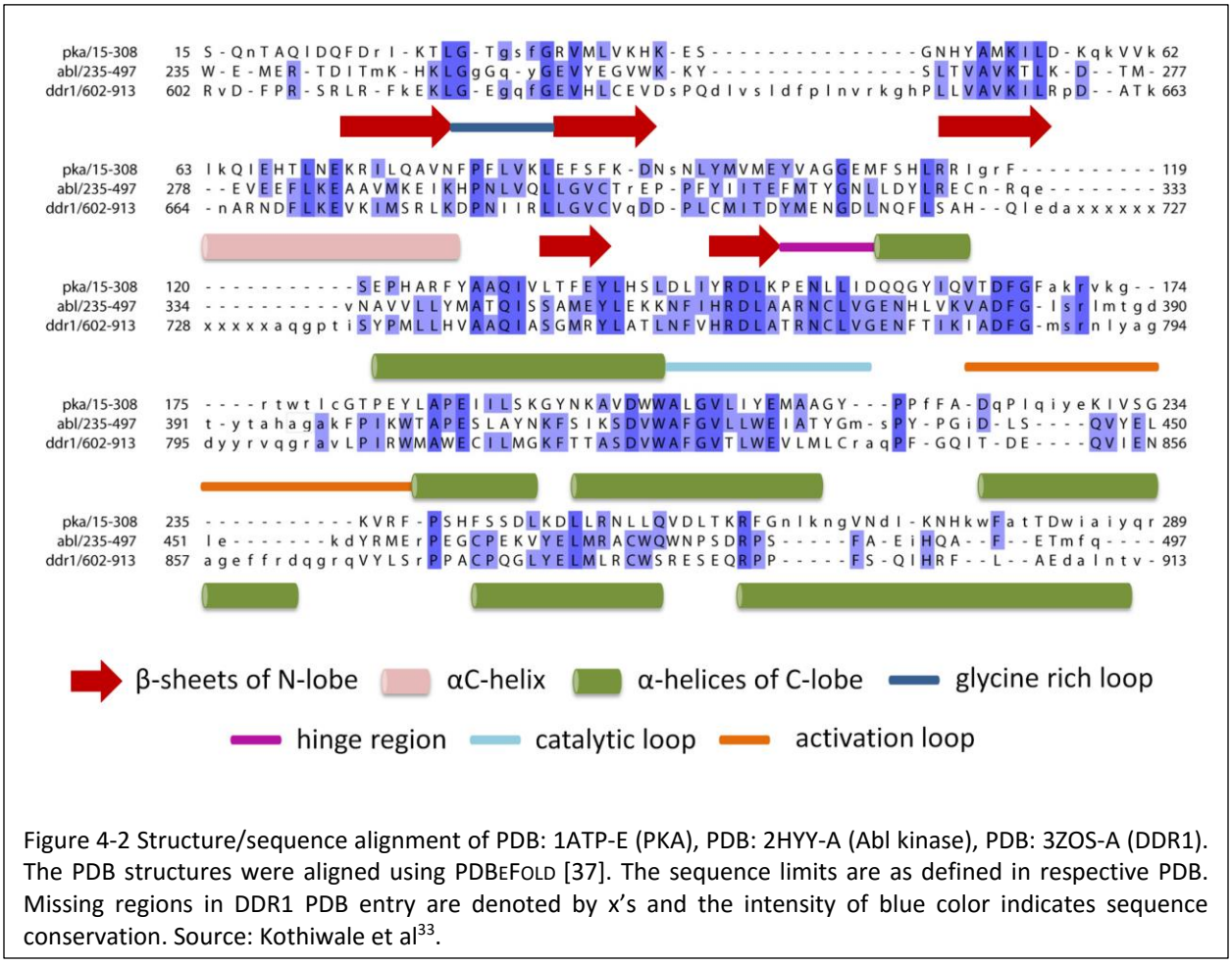


Figure 4-2 Structure/sequence alignment of PDB: 1ATP-E (PKA), PDB: 2HYY-A (Abl kinase), PDB: 3ZOS-A (DDR1). The PDB structures were aligned using PDBeFold [37]. The sequence limits are as defined in respective PDB. Missing regions in DDR1 PDB entry are denoted by x's and the intensity of blue color indicates sequence conservation. Source: Kothiwale et al[33].

the active conformation (Figure 4-1B), this loop positions the γ-phosphate of ATP for catalysis and a conserved valine, Val624 [*Val57*, Val254] makes a hydrophobic contact to the base of ATP. As with other domains, the DDR1 kinase domain contains the highly conserved DFG and HRD (YRD in case of PKA) motifs on activation and catalytic loops respectively. Asp784 [*Asp184*, Asp381] of the Asp-Phe-Gly (DFG) motif forms polar contacts with all three ATP phosphates. The phenylalanine of Asp784-Phe785-Gly786 [*Asp184-Phe185-Gly186*, Asp381-Phe382-Gly383] makes hydrophobic contacts with the Met676 [*Leu95*, Met290] of αC helix and the histidine of conserved His764-Arg765-Asp766 [*Tyr164-Arg165-Asp166*, His361-Arg362-Asp363] motif. The His-Arg-Asp (HRD) motif of DDR1 and Abl (Tyr-Arg-Asp (YRD) in PKA) is part of the "activation loop" which provides a platform for the peptide substrate binding. Phosphorylation of tyrosine residues within the activation loop is required to support a configuration that allows binding and phosphorylation of substrate protein. A conserved glutamate residue Glu672 [*Glu91*, Glu286] located on αC helix forms an ion pair with Lys655 [*Lys72*, Lys271] side chain that coordinates the α- and β-phosphates of ATP. In a number of active kinases, αC makes direct contact with N-terminal region of activation loop, with its conformation often linked to the DFG motif. While there is no experimental structure of DDR1 in its active conformation, it is expected that this feature is preserved. The C-lobe consists of mostly α-helices that surround a central β-sheet and serves as a docking site for substrate proteins. The residues in the interface of C-lobe with N-lobe are involved in catalytic machinery associated with transfer of phosphate from ATP. Typically disruption of interactions within N-lobe and between the two lobes immobilizes kinase activity. Figure 4-2 shows the structure/sequence alignment of PKA, Abl, DDR1 and DDR2 kinase domain where conserved residues are capitalized.

In addition to conformational changes to the activation loop, αC-helix position and orientation of catalytic residues, conserved spatial arrangement patterns of residues have been identified in the active kinase conformation. Four residues in the ATP-binding site link together N and C lobes of the kinase domain[16]. Residues His764, Phe785, Met676, and Leu687 [*Tyr164, Phe185, Leu95, and Leu106*; His361, Phe382, Met290, and Leu301] form the so-called hydrophobic spine. An intact conformation of the hydrophobic spine is essential for maintaining the active state conformation of kinase domain, while a disruption of the arrangement leads to inactive conformation. The hydrophobic spine supports the relative orientation of the two lobes as a hinge for inter-conversion of the open and closed conformations required for binding ATP and releasing ADP [16b, 17]. Figure 4-1A and B illustrate the hydrophobic spines in inactive conformation of DDR1 (PDB: 3ZOS) and active DDR1 conformation (homology model).

Targeting DDR1 for inhibition

As DDR1 plays a key role in pathological conditions, including atherosclerosis, cancer, inflammation and fibrosis, blocking DDR1-mediated downstream signaling by inhibiting the transfer of γ-phosphate of ATP to the hydroxyl group of the tyrosine kinase on protein substrates is an appealing strategy to prevent DDR1 activation.

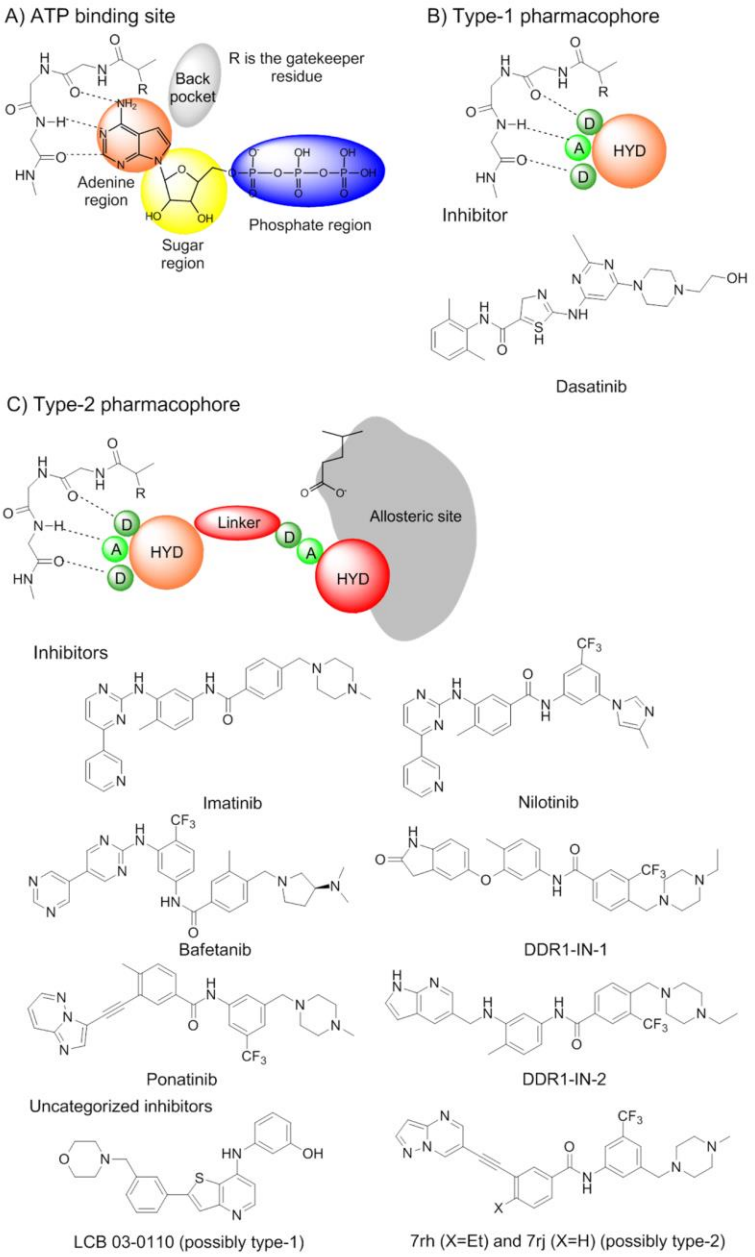


Figure 4-3 A) ATP binding site region and ATP interactions with the hinge residues of a kinase domain. Hydrogen bonds are represented by dashed lines. B) Type-1 kinase inhibitors mimic the binding of adenine moiety of ATP and lock the kinase domain in its active-state conformation. The type-1 inhibitor pharmacophore is shown representing the potential hydrogen bonds with the hinge region. C) Type-2 kinase inhibitors lock the kinase domain in the inactive-conformation by leveraging the ATP binding site as well as the allosteric site that is accessible in the inactive state. The pharmacophore is shown representing the interactions with the hinge region and the allosteric site present in the "DFG-out" conformation. Hydrogen bond donors are represented by circles labeled D, hydrogen bond acceptors by circles labeled A. The larger circles labeled HYD indicate hydrophobic moieties. The moiety that occupies the adenine ring region is colored in orange. The allosteric site is represented in gray. Adapted from: Fabio Zuccotto; Elena Ardini; Elena Casale; Mauro Angiolini; J. Med. Chem. 2010, 53, 2681-2694. Source: Kothiwale et al[33]

DDR1 inhibitors reported so far are ATP competitive inhibitors that bind to either the active (type-1 inhibitors) or inactive (type-2 inhibitors) conformations, preventing transfer of terminal phosphate group of ATP to protein substrate. Screening for inhibitory activity against a panel of kinases identified imatinib [18], nilotinib [19], dasatinib [19] and bafetanib [20] as DDR1 inhibitors (Figure 4-3). Day et al have reported inhibition of DDR1 by imatinib, nilotinib and dasatinib with IC_{50} values of 43 $\pm$ 2.4 nM, 3.7 $\pm$ 1.2 nM and 1.35 $\pm$ 0.2 nM respectively[21]. However, these inhibitors are not selective, as they were originally designed to target Abl kinase. Sun et al identified (3-(2-(3-(morpholinomethyl) phenyl) thieno [3, 2-b] pyridin-7-ylamino) phenol (LCB 03-0110 in Figure 4-3) as a potent inhibitor of both DDR1 and DDR2 along with several other tyrosine kinases[22]. Recently Ding et al have identified a series of 3-(2-(pyrazolo [1, 5-*a*] pyrimidin-6-yl) ethynyl) benzamides as potent DDR1 inhibitors, most potent of which (7rh and 7rj in Figure 4-3) have IC_{50} values of 6.8 and 7.0 nM respectively[23]. Kim et al have reported two inhibitors DDR1-IN-1 and DDR1-IN-2 (Figure 4-3) which exhibit an IC_{50} of 105 nM and 47 nM respectively [24].

DDR1-inhibitor complexes

Dasatinib is a type-1 inhibitor which targets kinase domains in the active form which is characterized by an open conformation of the activation loop (see below for details). Type-1 inhibitors bind the ATP site by mimicking the adenine ring's interaction with the "hinge" residues of protein. Even though there is no co-crystal structure of the DDR1-dasatinib complex, it is expected that dasatinib will bind in the so-called open conformation of DDR1 kinase domain which is characterized by "DFG-in" configuration of the conserved triad DFG at the beginning of activation loop (see also Figure 4-1B for details). Imatinib and nilotinib, on the other hand, are type-2 inhibitors which bind to

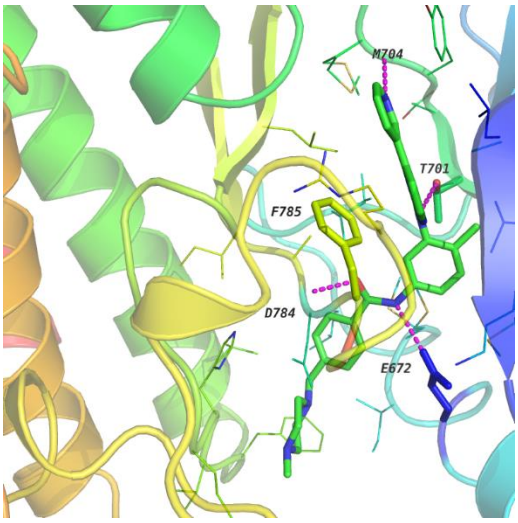


Figure 4-4 Co-crystal complex of DDR1-imatinib (PDB: 4BKJ). All type 2 inhibitors form conserved hydrogen-bond pairs with – a) side chain of a conserved glutamic acid in the αC-helix b) backbone amide of asparatic acid in the DFG motif. Figure shows hydrogen bonds between imatinib and DDR1. The glutamic acid and DFG motif are labeled. Source : Kothiwale et al[33]

and stabilize an inactive kinase form that is characterized by "DFG-out" conformation (see below for details). The "DFG-out" motif opens an additional cavity, a hydrophobic allosteric site which, in addition to the ATP binding pocket, is targeted by type-2 inhibitors (see also Figure 4-1B for details).

Type-1 inhibitors

Generally, type-1 inhibitors bind to the ATP site (Figure 4-3A) by mimicking the interactions of the adenine moiety. Figure 4-3B shows the type-1 kinase pharmacophore which is made up of hydrogen bond acceptor, two hydrogen bond donors, and a hydrophobic moiety. Type-1 inhibitors typically form one to three hydrogen bonds with kinase hinge residues and some hydrophobic interactions with residues which occupy region around the adenine ring of ATP [25]. Below we discuss a homology model of DDR1 in complex with a type-1 inhibitor that illustrates these important interactions.

Type-2 inhibitors

Recently the inactive conformation of DDR1 bound to imatinib has been reported in the PDB (PDB: 4BKJ)[26]. Co-crystal structure of DDR1-IN-1 with DDR1 kinase (PDB: 4CKR) suggests a comparable binding mode as imatinib, suggesting it is a type-2 inhibitor which locks the kinase in the inactive "DFG-out" conformation [24]. Type-2 inhibitors leverage the ATP binding pocket as well as an allosteric site created by a conformational change of the activation loop. The conformational change moves the phenylalanine residue (Phe785 [*Phe185*, Phe382]) more than 10 Å from its position in kinase active conformation creating hydrophobic site adjacent to ATP binding pocket. The co-crystal structure of DDR1 (PDB: 4BKJ) and imatinib displays the frequently observed hydrogen-bond interactions with the residues in the allosteric site (Figure 4-4).

Type-2 inhibitors target additional allosteric sites

Type-1 inhibitors have high cross-reactivity within the kinase family due to high degree of sequence and structural similarity in ATP binding site. In general, type-1 inhibitors tend to be promiscuous, because they tend to target well-conserved active kinase binding sites. However, type-1 inhibitors have the advantage of inhibiting kinases that have acquired mutations resistant to type-2 inhibitors. Type-2 inhibitors tend to be more selective because the inactive "DFG-out" kinase conformation allows additional interactions between the inhibitor and specific, not-well-conserved exposed hydrophobic sites within the kinase domain (Figure 4-3A). A third class of inhibitors have been identified that target either the catalytically active ("DFG-in" and αC-helix-in) or inactive ("DFG-out" and αC-helix-out) kinases by leveraging a hydrophobic back cavity [27]. The back cavity is accessible in kinases which have a small gatekeeper residue which is the first residue of the hinge connecting the C-lobe and N-lobe. The small gatekeeper

residue, Thr701 [*Met120*, Thr315] in DDR1 (PDB:3ZOS) will allow the design of selective and potent binders that engage with the back pocket which becomes available due to small side chain.

An intact conformation of the regulatory spine is essential for kinase activity. Efforts have been made to develop non-ATP competitive, i.e. truly allosteric inhibitors that target the regulatory spine. In this context ARQ197 [28], a non-ATP competitive inhibitor of met proto-oncogene (c-Met) is currently in Phase III clinical trials for non-small-cell lung cancer [17]. The co-crystal structure of c-Met and ARQ197 reveals disrupted interactions between Met1131, Leu1142, His1202 and Phe1223 in the regulatory spine [17]. The inhibitor has shown exceptional exclusivity against a panel of 230 human kinases of which only four are inhibited to any significant degree[28]. Eathiraj et al. created a generalized computational model of this inactive kinase conformation which was successfully applied to identify a series of fibroblast growth factor receptor (FGFR) TRK inhibitors[29]. Being tyrosine kinases, the novel mode of kinase inhibition is pertinent to DDR kinases and can be explored for identifying selective inhibitors.

DDR1 models for structure-based drug discovery

The sequence similarity in kinase domains has allowed homology modeling of DDR1. Day et al [21] used homology models based on multiple templates to describe inhibition of DDR1 by imatinib, nilotinib and dasatinib. Fu et al [1] built two homology models of DDR1 based on the "DFG-in" (active) and "DFG-out" (inactive) conformations of the DFG motif. For the current work we have developed homology models for DDR1 using the ROSETTA macromolecular modeling suite [30].

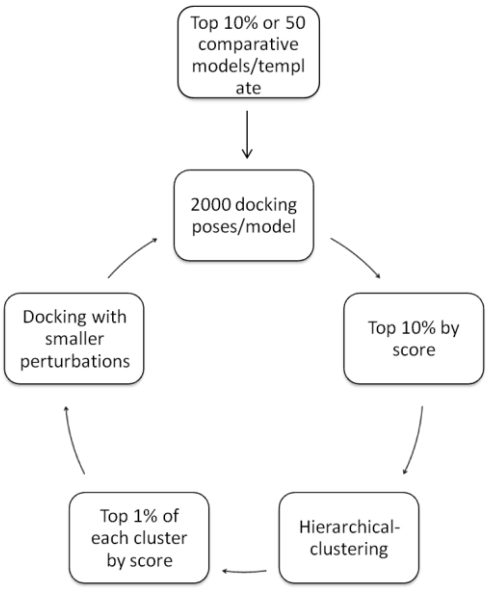


Figure 4-5 Hierarchical docking protocol used for homology modeling. Source : Kothiwale et al[33]

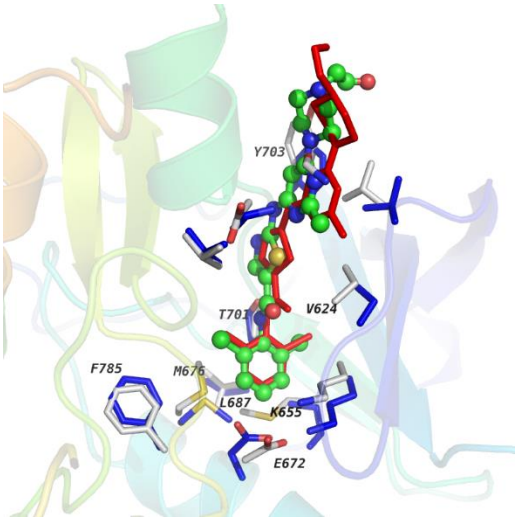


Figure 4-6 DDR1 homology model complexed with dasatanib is shown along with the co-crystal structure of Abl kinase and dasatinib (PDB: 2GQG). Dasatinib docked into DDR1-active state homology model is shown in the multicolored ball and stick model while the native binding pose of dasatinib in 2GQG is shown in red stick model. Residues in the binding pocket for DDR1 and Abl kinase are shown in grey and blue sticks respectively. Residues that interact with ligands are labeled along with those that form the hydrophobic spine. Source: Kothiwale et al[33]

The homology model for the "DFG-out" conformation was built using three templates (PDB: 3BEA, PDB: 4AT5 and PDB: 4HVS). The top 50 models from each template were used for docking the type-2 inhibitor imatinib. A hierarchical docking protocol (Figure 4-5) was used to identify favorable homology models based on the ability to recover native poses for the ligand of interest. Models were clustered on the basis of RMSD of docked imatinib to native imatinib pose. Top 1% models were taken from each cluster which recovered the native imatinib binding pose within 0.2 Å and recovered side chain conformations (details in Appendix). The resulting four homology were used for further docking studies. Meanwhile, the structure was also determined experimentally (PDB: 4BKJ) which allowed us to compare our model in a blind experiment (Figure 4-7). The best-scoring homology models achieved a RMSD of 2.5 Å to PDB: 4BKJ. Ligand docking into DDR1-inactive homology models yielded docking poses which had a RMSD of 0.2 Å to the experimental imatinib pose in PDB: 4BKJ (Figure 4-7A) and was successful in recovery of side chain conformations in the binding pocket (Figure 4-7B).

These results give us confidence in creating a model for DDR1 in the "DFG-in" conformation. Four "DFG-in" conformation template structures (PDB: 2PVF, PDB: 2X2L, PDB: 3C4F and PDB: 3RHX) were used to create homology models using the hierarchical docking protocol represented in Figure 4-5. Five models obtained after clustering and model selection as explained above for "DDR-out" conformation. Docking dasatinib into five different homology models recovered the binding pose of dasatinib reported in PDB: 2GQG within RMSD of 0.18 Å (Figure 4-6 shows the dasatanib docked pose in multi-colored ball and stick model and in red is its native pose co-crystallized with Abl kinase domain (PDB: 2GQG). The docked model recovers the position of the benzene group of dasatinib deep in the

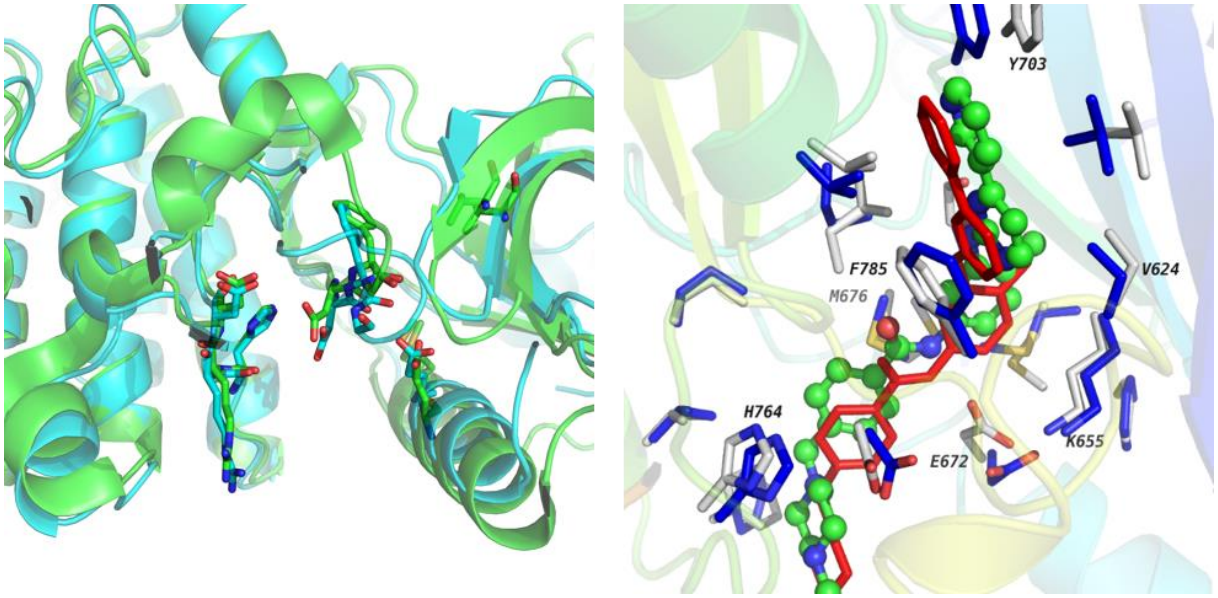


Figure 4-7 A) DDR1 homology model created using ROSETTA is shown in green and is aligned with PDB: 4BJK. 4BJK is the DDR1 kinase domain in the in-active state bound with imatinib. B) Docking of imatinib in DDR1–inactive homology model shows good pose recovery along with recovery of rotamers of residues in the binding site. The multi-colored ball and stick model is the docked imatinib pose while the pose in red is the native pose of imatinib in PDB: 4BKJ. Source : Kothiwale et al[33]

ATP-binding cleft. The hydrogen bond between the nitrogen of carboxamide group and Thr315 (PDB: 2GQG) is observed in the docked model with Thr701 of DDR1. In the binding pocket, side chain orientations of residues Glu672 [Glu286], Lys655 [Lys271], Val624 [Val256] and Tyr703 [Phe317] were captured as in the experimental structure.

Availability of experimental structures as well as high-quality homology models makes DDR1 amenable for application of structure-based drug discovery methods. Flexible docking with ROSETTALIGAND into crystal structures would improve docking accuracy. Also as the number of known DDR1 inhibitors grows[18-20, 22-24, 31] ligand-based drug-discovery methods such as quantitative structure-activity relation (SAR) models can be developed.

Ligand based probe development

The number of known DDR1 inhibitors has grown[18-20, 22-24, 31-32] allowing development of ligand-based drug-discovery methods such as quantitative structure-activity relation (SAR) models. In this study, we developed QSAR models using BCL::CHEMINFO to identify DDR1 inhibitors by computational screening of virtual small molecule libraries. QSAR models correlate structure of molecules to their experimental activity. BCL::CHEMINFO developed in the Meiler lab is a software suite containing cheminiformatics methods including algorithms to develop QSAR models. QSAR models are developed using artificial neural networks (ANNs) implemented in BCL. BCL::CHEMINFO QSAR algorithms are described in detail in Chapter 3.

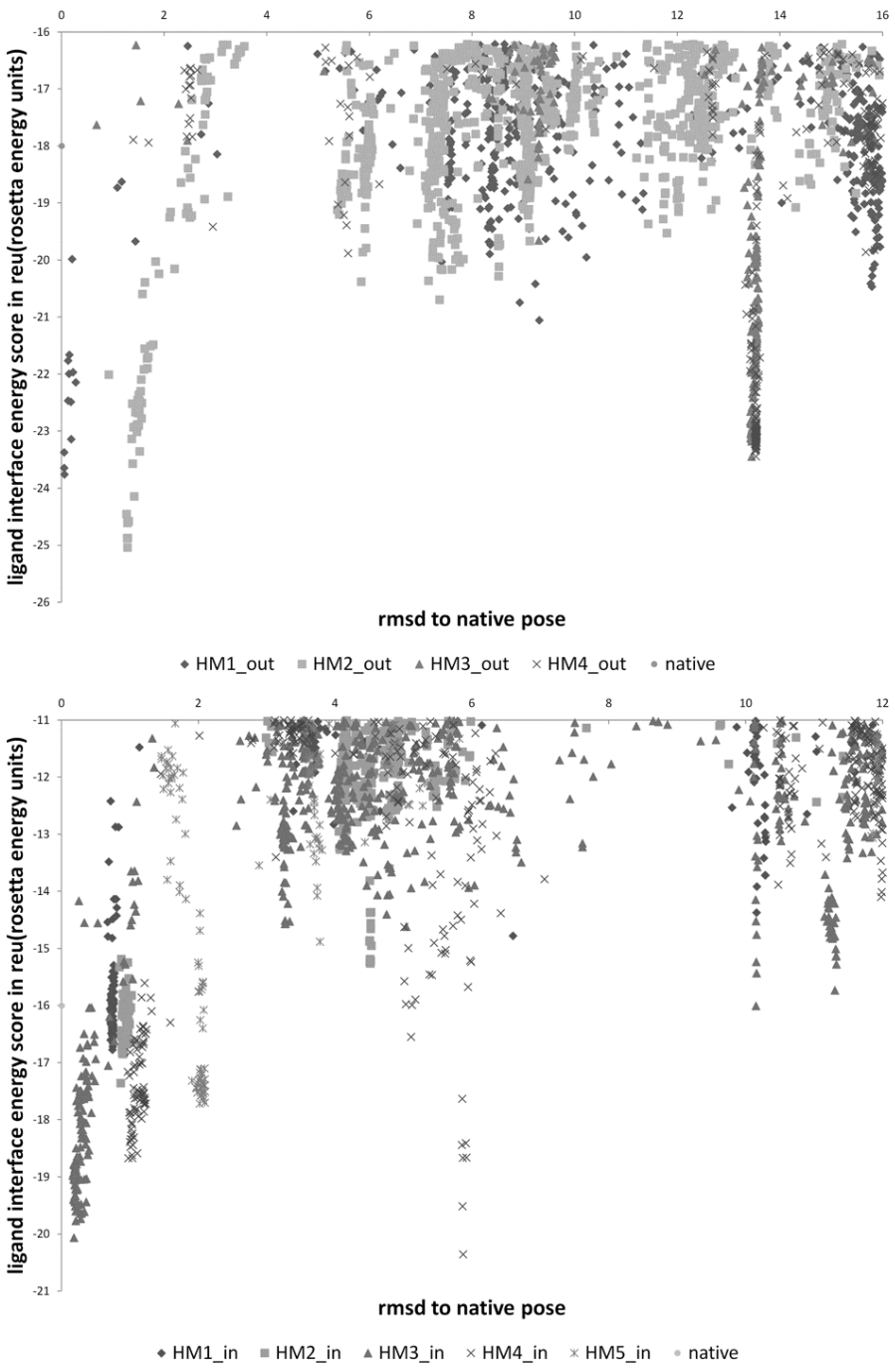


Figure 4-8 The figure shows the RMSD vs score plot for docking studies done with homology models of DDR1. The x-axis is root mean square deviation from the native binding pose in Anstrom for the ligand of interest. On the y-axis is the ROSETTALIGAND score for docking poses. A) Imatinib docked into DDR1–inactive "DFG-out" conformation homology models. Four homology models were used for docking studies and the docking poses derived from each are shown separately as HM1_out, HM2_out, HM3_out and HM4_out. B) Dasatanib docked into DDR1-active ("DFG-in") conformation homology model. Five homology models were used and the docking poses are plotted for each separately. Source : Kothiwale et al[33]

A pre-requisite for developing QSAR models is access to structures of small molecules that are active or inactive against a target of interest. The structure of these small-molecules is converted to numerical description known as features. Mathematical/statistical models are developed that are able to distinguish features of active molecules from features of inactive molecules. These mathematical/statistical models can then be used to predict whether a molecule with an unknown activity against the target of interest is active or inactive. The advantage of these models is that these can be used to screen large compound libraries containing millions of compounds in couple of hours. Molecules predicted to be active can be prioritized and experimentally tested. Virtual screening of huge libraries to prioritize molecules before experimental testing saves resources, labor and time. Traditional high throughput experimental screens have a hit rate of 0.2%. Using computational prediction to prioritize molecules for experimental screening has been shown to significantly increase the hit rate to about 3%.

Using QSAR methods implemented in BCL::CHEMINFO, we have been able to achieve a hit rates of 28.2% and 3.6% for predicting mGlu5 partial agonists and allosteric modulators for mGlu5 respectively. The QSAR model for positive allosteric modulators (PAMs) was developed using experimental screen performed at Vanderbilt University in which approximately 144,000 compounds screened yielded 1,356 hits, a hit rate of 0.94%. A virtual screen against approximately 450K compounds was performed and 824 compounds were prioritized for experimental testing. Of these compounds, 232 were confirmed as mGlu5 partial agonists accounting for 28.2% hit-rate, approximately 30 times greater than original HTS performed at Vanderbilt University. In another study, Rodriguez et al screened 160,000 compounds for allosteric modulators of mGlu5 found 624 at a hit rate of 0.2%. QSAR model developed using this experimental data was used to virtually screen 700,000 commercially available compounds and prioritize 749 compounds out of which 27 compounds were found to modulate mGlu5 signaling indicating a hit rate of 3.6%.

In the current study, an iterative computational virtual screening followed by feedback from experimental results was used to identify DDR1 active molecules. Figure 4-9 shows the scheme of the study. Since no high throughput screening assay has yet been performed for DDR1, there is limited activity data available for DDR1. The first QSAR model was developed using molecules which have been reported as DDR1 active in the literature. Inactive molecules were used from AID – 2689, a serine-threonine kinase-33 screening data stored in the PUBCHEM database. Since most kinase-inhibitors bind the conserved ATP binding site of the kinase domain, there is high likelihood that inactive molecules do not have activity against any other kinase. The QSAR model was used to screen the Vanderbilt high throughput-screening library, called the VICB library, to prioritize 25 molecules for experimental testing. All of the 25 molecules were found to be inactive against DDR1 kinase.

For the second round of QSAR model development the datasets were updated with the 25 inactive molecules found in the first round of experimental screening. The model was used to screen the VICB library again to select 25 molecules to be tested experimentally. Experimental testing identified four molecules to be active with an inhibitory activity of 60%. These molecules were also active against DDR2. With the goal of identifying more potent and

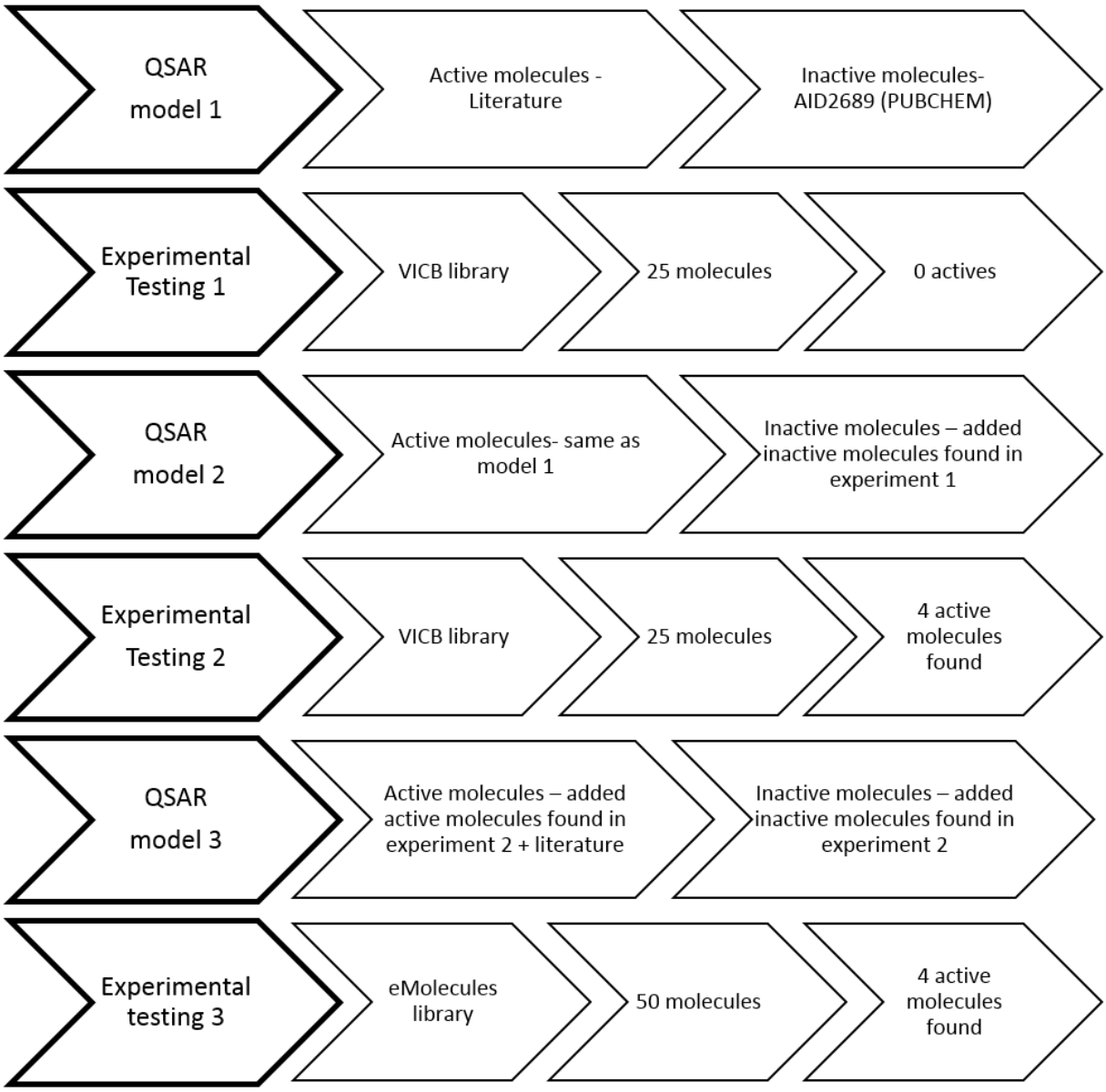


Figure 4-9 Iterative computational and experimental screening used in this study to identify DDR1 binders. QSAR model 1 was developed using molecular activities reported in the literature or reported in PUBCHEM. VICB library was virtually screened to prioritize 25 molecules which were tested experimentally among which none were found to be active. QSAR model was updated using feedback from experiment 1 result and used to screen the VICB library again to prioritize 25 molecules for experimental testing. Active molecules identified in the experimental screening was used to update QSAR models. Computational screening of eMolecules was carried out and 50 molecules prioritized for screening. Four molecules were identified which had some inhibitory activity against DDR1.

selective inhibitors of DDR1, an external library of five million compounds, called eMoleclues library, was screened. Another QSAR model was developed with the feedback incorporated from the second round of screening.

A new set of potent molecules reported by Elkamhawy et al[32] was also included in the third round of models. The external library called the eMolecules was screened and 50 molecules were prioritized for experimental testing. Two molecules showed an inhibitory activity of greater than 80% in experimental testing while two others had an activity of less than 60% inhibition. The two potent inhibitors have no reported activity against DDR1 but have been reported as kinase inhibitors. The two other molecules with activity of less than 60% have no reported activity against any protein targets in the literature. Further confirmatory testing is being carried out at the Pozzi laboratory at Vanderbilt University.

References

1. Fu, H. L.; Valiathan, R. R.; Arkwright, R.; Sohail, A.; Mihai, C.; Kumarasiri, M.; Mahasenan, K. V.; Mobashery, S.; Huang, P.; Agarwal, G.; Fridman, R., Discoidin domain receptors: unique receptor tyrosine kinases in collagen-mediated signaling. *J Biol Chem* 2013, *288* (11), 7430-7.
2. (a) Vogel, W.; Gish, G. D.; Alves, F.; Pawson, T., The discoidin domain receptor tyrosine kinases are activated by collagen. *Mol Cell* 1997, *1* (1), 13-23; (b) Koo, D. H.; McFadden, C.; Huang, Y.; Abdulhussein, R.; Friese-Hamim, M.; Vogel, W. F., Pinpointing phosphotyrosine-dependent interactions downstream of the collagen receptor DDR1. *FEBS Lett* 2006, *580* (1), 15-22; (c) Wang, C. Z.; Su, H. W.; Hsu, Y. C.; Shen, M. R.; Tang, M. J., A discoidin domain receptor 1/SHP-2 signaling complex inhibits alpha2beta1-integrin-mediated signal transducers and activators of transcription 1/3 activation and cell migration. *Mol Biol Cell* 2006, *17* (6), 2839-52; (d) L'Hote C, G.; Thomas, P. H.; Ganesan, T. S., Functional analysis of discoidin domain receptor 1: effect of adhesion on DDR1 phosphorylation. *FASEB J* 2002, *16* (2), 234-6.
3. Lu, K. K.; Trcka, D.; Bendeck, M. P., Collagen stimulates discoidin domain receptor 1-mediated migration of smooth muscle cells through Src. *Cardiovasc Pathol* 2011, *20* (2), 71-76.
4. Curat, C. A.; Vogel, W. F., Discoidin domain receptor 1 controls growth and adhesion of mesangial cells. *J Am Soc Nephrol* 2002, *13* (11), 2648-2656.
5. Xu, H.; Bihan, D.; Chang, F.; Huang, P. H.; Farndale, R. W.; Leitinger, B., Discoidin domain receptors promote alpha1beta1- and alpha2beta1-integrin mediated cell adhesion to collagen by enhancing integrin activation. *PLoS One* 2012, *7* (12), e52209.
6. Roarty, K.; Serra, R., Wnt5a is required for proper mammary gland development and TGF-beta-mediated inhibition of ductal growth. *Development* 2007, *134* (21), 3929-39.
7. (a) Iwai, L. K.; Chang, F.; Huang, P. H., Phosphoproteomic analysis identifies insulin enhancement of discoidin domain receptor 2 phosphorylation. *Cell Adh Migr* 2013, *7* (2), 161-4; (b) Kim, H. G.; Hwang, S. Y.; Aaronson, S. A.; Mandinova, A.; Lee, S. W., DDR1 receptor tyrosine kinase promotes prosurvival pathway through Notch1 activation. *J Biol Chem* 2011, *286* (20), 17672-81.
8. Borza, C. M.; Pozzi, A., Discoidin domain receptors in disease. *Matrix Biology* 2013.
9. Vogel, W. F.; Aszodi, A.; Alves, F.; Pawson, T., Discoidin domain receptor 1 tyrosine kinase has an essential role in mammary gland development. *Mol Cell Biol* 2001, *21* (8), 2906-17.
10. Franco, C.; Hou, G.; Ahmad, P. J.; Fu, E. Y.; Koh, L.; Vogel, W. F.; Bendeck, M. P., Discoidin domain receptor 1 (ddr1) deletion decreases atherosclerosis by accelerating matrix accumulation and reducing inflammation in low-density lipoprotein receptor-deficient mice. *Circ Res* 2008, *102* (10), 1202-11.
11. Avivi-Green, C.; Singal, M.; Vogel, W. F., Discoidin domain receptor 1-deficient mice are resistant to bleomycin-induced lung fibrosis. *Am J Respir Crit Care Med* 2006, *174* (4), 420-7.
12. Xu, H.; Abe, T.; Liu, J. K.; Zalivina, I.; Hohenester, E.; Leitinger, B., Normal activation of discoidin domain receptor 1 mutants with disulphide cross-links, insertions or deletions in the extracellular juxtamembrane region: mechanistic implications. *J Biol Chem* 2014.
13. Fu, H. L.; Valiathan, R. R.; Payne, L.; Kumarasiri, M.; Mahasenan, K. V.; Mobashery, S.; Huang, P.; Fridman, R., Glycosylation at Asn211 regulates the activation state of the discoidin domain receptor 1 (DDR1). *J Biol Chem* 2014, *289* (13), 9275-87.

14. Rabiller, M.; Getlik, M.; Kluter, S.; Richters, A.; Tuckmantel, S.; Simard, J. R.; Rauh, D., Proteus in the world of proteins: conformational changes in protein kinases. *Arch Pharm (Weinheim)* 2010, *343* (4), 193-206.
15. Johnson, L. N.; Noble, M. E.; Owen, D. J., Active and inactive protein kinases: structural basis for regulation. *Cell* 1996, *85* (2), 149-58.
16. (a) Taylor, S. S.; Kornev, A. P., Protein kinases: evolution of dynamic regulatory proteins. *Trends Biochem Sci* 2011, *36* (2), 65-77; (b) Kornev, A. P.; Taylor, S. S., Defining the conserved internal architecture of a protein kinase. *Biochim Biophys Acta* 2010, *1804* (3), 440-4.
17. Norman, R. A.; Toader, D.; Ferguson, A. D., Structural approaches to obtain kinase selectivity. *Trends Pharmacol Sci* 2012, *33* (5), 273-8.
18. Bantscheff, M.; Eberhard, D.; Abraham, Y.; Bastuck, S.; Boesche, M.; Hobson, S.; Mathieson, T.; Perrin, J.; Raida, M.; Rau, C.; Reader, V.; Sweetman, G.; Bauer, A.; Bouwmeester, T.; Hopf, C.; Kruse, U.; Neubauer, G.; Ramsden, N.; Rick, J.; Kuster, B.; Drewes, G., Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors. *Nat Biotechnol* 2007, *25* (9), 1035-44.
19. Rix, U.; Hantschel, O.; Durnberger, G.; Remsing Rix, L. L.; Planyavsky, M.; Fernbach, N. V.; Kaupe, I.; Bennett, K. L.; Valent, P.; Colinge, J.; Kocher, T.; Superti-Furga, G., Chemical proteomic profiles of the BCR-ABL inhibitors imatinib, nilotinib, and dasatinib reveal novel kinase and nonkinase targets. *Blood* 2007, *110* (12), 4055-63.
20. Rix, U.; Remsing Rix, L. L.; Terker, A. S.; Fernbach, N. V.; Hantschel, O.; Planyavsky, M.; Breitwieser, F. P.; Herrmann, H.; Colinge, J.; Bennett, K. L.; Augustin, M.; Till, J. H.; Heinrich, M. C.; Valent, P.; Superti-Furga, G., A comprehensive target selectivity survey of the BCR-ABL kinase inhibitor INNO-406 by kinase profiling and chemical proteomics in chronic myeloid leukemia cells. *Leukemia* 2010, *24* (1), 44-50.
21. Day, E.; Waters, B.; Spiegel, K.; Alnadaf, T.; Manley, P. W.; Buchdunger, E.; Walker, C.; Jarai, G., Inhibition of collagen-induced discoidin domain receptor 1 and 2 activation by imatinib, nilotinib and dasatinib. *Eur J Pharmacol* 2008, *599* (1-3), 44-53.
22. Sun, X.; Phan, T. N.; Jung, S. H.; Kim, S. Y.; Cho, J. U.; Lee, H.; Woo, S. H.; Park, T. K.; Yang, B. S., LCB 03-0110, a novel pan-discoidin domain receptor/c-Src family tyrosine kinase inhibitor, suppresses scar formation by inhibiting fibroblast and macrophage activation. *J Pharmacol Exp Ther* 2012, *340* (3), 510-9.
23. Gao, M.; Duan, L.; Luo, J.; Zhang, L.; Lu, X.; Zhang, Y.; Zhang, Z.; Tu, Z.; Xu, Y.; Ren, X.; Ding, K., Discovery and optimization of 3-(2-(Pyrazolo[1,5-a]pyrimidin-6-yl)ethynyl)benzamides as novel selective and orally bioavailable discoidin domain receptor 1 (DDR1) inhibitors. *J Med Chem* 2013, *56* (8), 3281-95.
24. Kim, H. G.; Tan, L.; Weisberg, E. L.; Liu, F.; Canning, P.; Choi, H. G.; Ezell, S. A.; Wu, H.; Zhao, Z.; Wang, J.; Mandinova, A.; Griffin, J. D.; Bullock, A. N.; Liu, Q.; Lee, S. W.; Gray, N. S., Discovery of a Potent and Selective DDR1 Receptor Tyrosine Kinase Inhibitor. *ACS Chem Biol* 2013.
25. Liu, Y.; Gray, N. S., Rational design of inhibitors that bind to inactive kinase conformations. *Nat Chem Biol* 2006, *2* (7), 358-64.
26. Canning, P.; Tan, L.; Chu, K.; Lee, S. W.; Gray, N. S.; Bullock, A. N., Structural mechanisms determining inhibition of the collagen receptor DDR1 by selective and multi-targeted type II kinase inhibitors. *J Mol Biol* 2014.
27. Zuccotto, F.; Ardini, E.; Casale, E.; Angiolini, M., Through the "gatekeeper door": exploiting the active kinase conformation. *J Med Chem* 2010, *53* (7), 2681-94.
28. Munshi, N.; Jeay, S.; Li, Y.; Chen, C. R.; France, D. S.; Ashwell, M. A.; Hill, J.; Moussa, M. M.; Leggett, D. S.; Li, C. J., ARQ 197, a novel and selective inhibitor of the human c-Met receptor tyrosine kinase with antitumor activity. *Mol Cancer Ther* 2010, *9* (6), 1544-53.
29. Eathiraj, S.; Palma, R.; Volckova, E.; Hirschi, M.; France, D. S.; Ashwell, M. A.; Chan, T. C., Discovery of a novel mode of protein kinase inhibition characterized by the mechanism of inhibition of human mesenchymal-epithelial transition factor (c-Met) protein autophosphorylation by ARQ 197. *J Biol Chem* 2011, *286* (23), 20666-76.
30. Kaufmann, K. W.; Lemmon, G. H.; Deluca, S. L.; Sheehan, J. H.; Meiler, J., Practically useful: what the Rosetta protein modeling suite can do for you. *Biochemistry* 2010, *49* (14), 2987-98.
31. Davis, M. I.; Hunt, J. P.; Herrgard, S.; Ciceri, P.; Wodicka, L. M.; Pallares, G.; Hocker, M.; Treiber, D. K.; Zarrinkar, P. P., Comprehensive analysis of kinase inhibitor selectivity. *Nat Biotechnol* 2011, *29* (11), 1046-51.
32. Elkamhawy, A.; Park, J. E.; Cho, N. C.; Sim, T.; Pae, A. N.; Roh, E. J., Discovery of a broad spectrum antiproliferative agent with selectivity for DDR1 kinase: cell line-based assay, kinase panel, molecular docking, and toxicity studies. *J Enzyme Inhib Med Chem* 2016, *31* (1), 158-66.
33. Kothiwale, S.; Borza, C. M.; Lowe, E. W., Jr.; Pozzi, A.; Meiler, J., Discoidin domain receptor 1 (DDR1) kinase as target for structure-based drug discovery. Drug Discov Today 2015, 20 (2), 255-61.

CHAPTER 5 : KINASE SELECTIVITY MODEL

Introduction

Kinases are a large and diverse multigene family involved in the regulation of multicellular aspects of organisms [1]. For example, tyrosine phosphorylation is a ubiquitous mechanism utilized by intra- and inter- cellular communication pathways in metazoans, and a family of kinases known as the tyrosine kinases catalyze the transfer of phosphate group from ATP to select tyrosine residues in target proteins which leads to signal transduction [1b, 2]. Similarly, kinases belonging to the serine/threonine family phosphorylate the hydroxyl group on the side chain of a serine or threonine amino acid residue in a protein substrate [1a]. Kinase enzymes have two distinct lobes – an amino-terminal lobe comprising a five-stranded β sheet and one α helix, and a carboxy-terminal lobe that is mainly α-helical Figure 5-1). The ATP-binding cleft is located at the interface of the two lobes which is lined with several highly conserved residues [1b, 3]. The heterocyclic ring of ATP interacts with the hinge region through hydrogen bonds (Figure 5-1). Kinases undergo conformational changes due to ATP binding leading to the phosphorylation of substrate proteins [3a, 4]. Most notably catalytic and activation loop attain conformations that align important residues involved in transfer of phosphate from the ATP molecule to a target substrate protein [1b, 4-5]. Figure 5-1 shows ATP bound to a kinase in the active state that allows transfer of phosphate group from ATP to phosphorylation site of activation loop (orange) [3]. The aspartate of conserved motif HRD in the catalytic loop (cyan) accepts proton from substrate hydroxyl group during phosphotransfer mechanism [3b, 4].

There are more than 500 kinases in the human genome out of which 92 belong to the tyrosine kinase family while the rest belong to the serine/threonine family. There are 12 genes in the human genome encoding receptors that have intrinsic serine/threonine kinase domains. These receptors respond to the transforming growth factor β (TGF β) family [1a, 1c]. Serine/threonine kinase receptors (RSTK) are activated by ligand-induced assembly into heterotetrameric receptor complexes. RSTKs often activate growth inhibitory and apoptotic signals for example by activating members of Smad transcription factor family [1c]. Non-RSTKs are involved in cellular regulation through posttranslational modification of proteins by phosphorylation including metabolism, growth, differentiation, motility, membrane transport, learning, and memory [1a, 5]. Serine/threonine kinases interact with diverse substrates including other kinases, enzymes, transcription factors, receptors and other regulatory proteins. For example, cAMP-dependent kinase or protein kinase A is activated by downstream signaling transmitted by G protein-coupled receptor (GPCR). Protein kinase A in turn regulates glycogen, sugar and lipid metabolism [1a, 5].

Of about 92 tyrosine kinases that have been identified, 58 are transmembrane receptor type and 34 are cytoplasmic non-receptor type (Non-RTK)[2a]. Receptor tyrosine kinases (RTKs) are membrane-spanning cell surface proteins that play critical roles in transducing extracellular signals to the cytoplasm [1b, 2a]. The RTKs have an extracellular ligand-binding domain, a single pass transmembrane hydrophobic helix and the cytoplasmic portion

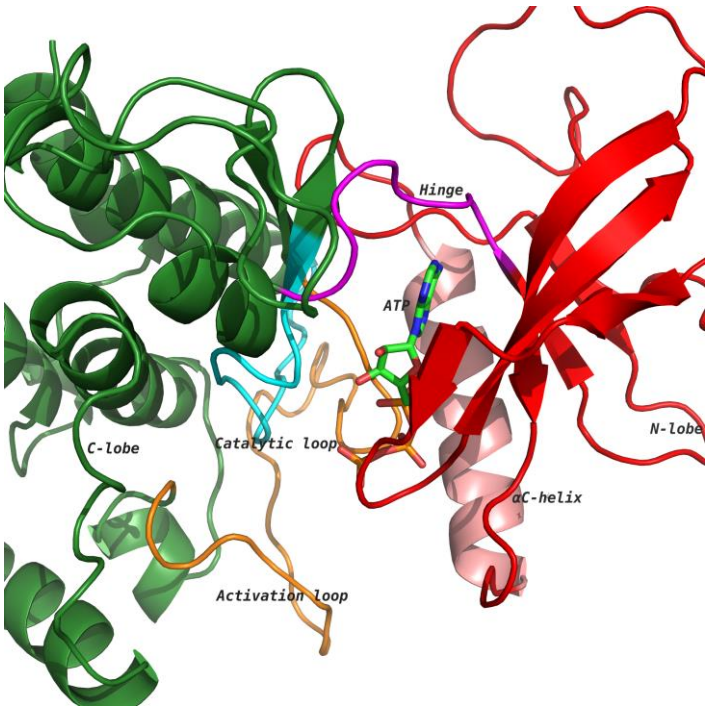


Figure 5-1 General structure of a kinase domain consists of two lobes, helix-rich C-terminal lobe (green) and sheet-rich N-terminal lobe (red). The ATP binding pocked is present at the interface of the two lobes. The heterocyclic ring of ATP forms hydrogen bonds with the Hinge-loop (magenta) Activation loop (orange) undergoes a conformation change upon phosphorylation of a conserved residue allowing substrate to bind. Arginine of the HRD motif in the Catalytic loop interacts with phosphate in the activation segment. Aspartate of HRD motif accepts proton from the substrate hydroxyl group during phosphotransfer mechanism.

containing the kinase domain [2a]. The Non-RTK have a kinase domain and often possess several protein-protein interaction domains such as SH2, SH3 and the PH domain [1c]. RTKs activated by growth factors modulate signaling by catalyzing the transfer of gamma-phosphate group from the ATP to target proteins. RTKs regulate key processes in cell proliferation, differentiation, migration, metabolism and programmed cell death [6]. Activation of Non-RTKs involves heterologous protein-protein interactions for enabling transphosphorylation [1c]. Most Non-RTKs couple to receptors including RTKs and those that lack intrinsic enzymatic subunits and relay intracellular signals originating from receptor activation [4, 6]. An example is the recruitment and activation of Src family members by activated platelet derived growth factor receptor (PDGFR) which induces entry into S phase and mitosis [1c].

Many disease states result from disrupted signal transduction pathways [3a]. In particular, dysregulation of TKs is associated with a number of human diseases including diabetes and large range of cancers [2, 6-7]. It is believed that the dysregulation occurs via a gain of mutations, gene rearrangements, gene amplification, and/or over expression or abnormal stimulation of receptors [2a, 6, 8]. For example mutations in epithelial growth factor receptor (EGFR) in

glioblastomas, ovarian tumors and non-small cell lung carcinoma, renders the tyrosine kinase active in the absence of the activating ligand [2b]. Over expression of RTK ERBB2 causes increased kinase activity in human breast cancer[9].

Kinases are pharmacologically targeted by – a) directly targeting the kinase catalytic activity by interfering with phosphorylation mechanism b) inhibiting activation of receptor kinases by blocking their oligomerization c) Antibodies against receptor kinases or their ligands to interrupt signaling through ligand neutralization, blocking ligand interaction or receptor internalization [2b, 9]. Small molecular ATP competitive inhibitors were the first promising therapeutic strategies targeting the catalytic activity of kinases and have been the target of choice in the small molecule space. Most small molecular kinase inhibitors are ATP mimics [7, 10] by presenting one to three hydrogen bonds to residues that normally interact with adenine ring of ATP. The adenine ring forms two key hydrogen bonds at N-1 and N-6 positions with the kinase hinge region – the segment connecting the N-terminal and the C-terminal lobe [11]. The ribose binds in the ribose-binding pocket and the triphosphate groups lie in a channel extending to the substrate binding site. Kinases have a conserved activation loop that assumes a large number of conformations that regulate access to the ATP binding site which allows the enzyme to switch between active and inactive state [3]. In the active state the loop is often phosphorylated while in the inactive state it blocks the substrate binding site.

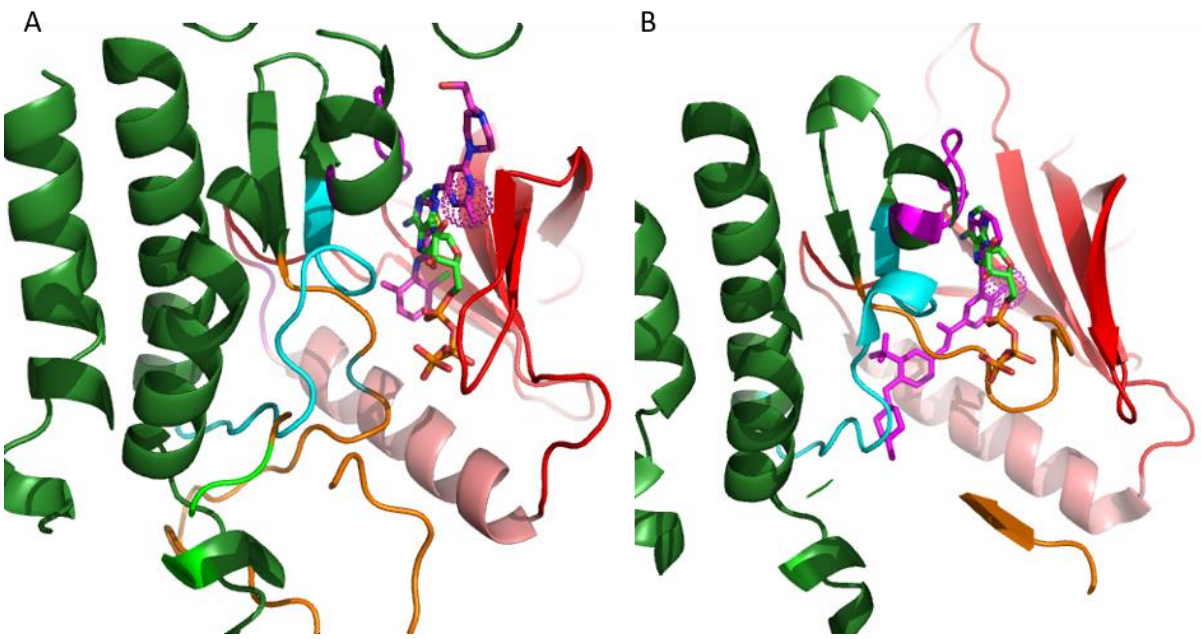


Figure 5-2 Shows the binding of type-1 and type-2 inhibitor in the ATP pocket along-side ATP for comparison. The inhibitors mimic the interactions that the heterocyclic ring of ATP has with the hinge loop (magenta). A) Dasatanib, a type-1 inhibitor is shown in the binding pocket of a kinase domain locked in active state. The activation loop (orange) is positioned such that it gets phosphorylated and is able to recruit substrate proteins. Catalytic loop (cyan) enables transfer of phosphate group to substrate protein. B) Shows Pontatinib bound kinase domain locked in an inactive state. ATP bound state is active state but is shown here for comparison to figure 2A. The conformation of activation loop and the catalytic loop is different from that found in active state.

Several ATP competitive inhibitors have been approved for clinical use or are in clinical trials [10b]. The binding mode of the inhibitors is categorized based on the conformation of a conserved Asp-Phe-Gly (DFG) motif within the activation loop.

The type 1 inhibitors constitute the majority of ATP-competitive inhibitors and block the kinase in the active conformation of the kinase as the DFG motif of activation loop faces into the ATP binding site (Figure 5-2A). The heterocyclic ring of such inhibitors occupies the adenine binding site while the other parts of the molecule occupy the adjacent hydrophobic regions I and II. Examples include FDA approved Dasatinib for CML. Type-1 inhibitors have high cross-reactivity within the kinase family due to a high degree of sequence and structural similarity in ATP binding site. In general, type-1 inhibitors tend to be promiscuous as they target the well-conserved ATP binding sites in the active conformation of kinase enzyme. Figure 5-2A shows superimposed binding poses of Dasatanib (magenta sticks) and ATP (green sticks) into the Abl2 kinase domain. Heterocyclic ring of Dasatanib occupy ATP purine binding site which serve as scaffold for side-chain that occupy hydrophobic site-I near the pocket shown in magenta spherical dots.

Type II inhibitors bind the inactive conformation of the kinase in which the DFG motif is facing outward such that aspartate side chain is facing out to the solvent. The 180-degree rotation opens up an additional hydrophobic pocket, the so-called specificity pocket which is exploited by type II inhibitors. Type-2 inhibitors tend to be more selective because the inactive “DFG-out” kinase conformation allows additional interactions between the inhibitor and specific, not-well-conserved exposed hydrophobic sites within the kinase domain. Examples include FDA approved imanitib and ponatinib against abelson murine leukemia viral oncogene-1 (ABL1) and PDFGR. Figure 5-2B shows ponatinib (magenta) bound to inactive state of DDR1 kinase (PDB: 3ZOS). Allosteric site that the type-II inhibitors target is shown in magenta spherical dots.

As the kinase inhibitors target the orthosteric and well conserved ATP binding pocket, they are multi-targeted and often inhibit a large number of kinases in a non-specific manner [10a]. Improved tyrosine kinase selectivity is a major challenge for developing promising lead compounds into therapeutics due to the side-effects caused by off-target activity. Dasatinib is a potent type-1 kinase inhibitor and is effective in patients with imatinib-resistant chronic myelogenous leukemia. It inhibits several other kinases including C-Kit, PDGFR, Ephrin receptors. Another example is Sunitib approved by the FDA for the treatment of renal cell carcinoma, which inhibits vascular endothelial growth factor receptor (VEGFR), PDGFR and c-kit, also tends to inhibit AMP-activated protein kinase that accounts for some of the cardiovascular toxicity. The degree of cross-reactivity has been determined by a number of studies which report inhibitor activities against a large panel of kinases. Davis et al. screened a total of 70 known inhibitors against a panel of 379 kinases in a competition binding assay [10a].

It is desirable to profile highly potent inhibitors for kinase specificity early in the lead optimization process. We hypothesize that computational models could be used to predict a hit compound’s kinase activity profile early in the

lead optimization process. Further, in a second step, the selectivity profile for to be synthesized derivatives of hit compounds can also be predicted thereby contributing to the prioritization of hit compounds for hit-to-lead optimization. Several computational approaches have been developed for predicting kinase activity profiles. Sheinerman et al. developed a computational approach to design a binding site signature that uses three-dimensional (3D) X-ray structure information of a kinase-inhibitor complex to predict the small-molecule's selectivity profile [12]. Subramanian et al. applied this approach to predict off-target kinase selectivity profile for 15 molecules against 280 members of the human kinome [13]. A co-crystal structure of the ligand of interest is a pre-requisite for this method. The input data includes interacting residues in the binding pocket of the target kinase enzyme. Sciabola et al. used the Free-Wilson approach to build quantitative structure-activity relationship (QSAR) models for a series of chemical analogs [14]. The Free-Wilson concept states that the biological activity of a molecule can be described as sum of activity contributions from specific substructures [15]. A limitation therefore is that it cannot make predictions about functional groups that are not present in the original set of compounds. Subramanium et al reported an average accuracy/sensitivity/specificity of 0.81/0.37/0.93 for 15 kinase inhibitors at an activity cutoff of $K_D \sim 3$ µM against a subset of 280 kinases. Sciabola used an in-house scaffold library for their study reporting a correlation of greater than 0.85 between experimental and predicted IC_{50} values for two series of compounds.

For the present study we developed QSAR models for predicting activity profiles of kinase inhibitors against a panel of kinases using a neural network based methodology. The objective of QSAR modelling is to correlate chemical structure with biological activity in a quantitative way. There are three prerequisites for QSAR modelling: a) a quantitative description of molecular structure (descriptor) b) biological activities of a diverse set of molecules and c) a mathematical technique for correlating descriptors to predict activity. Machine learning techniques are commonly applied to develop non-linear mathematical QSAR models. Here we use Artificial Neural Networks (ANN) as implemented in BCL::CHEMINFO to generate the kinase selectivity models [16].

Artificial neural networks

Artificial neural network (ANN) models are a type of mathematical model that are inspired by the biological brain and with an adaptive structure that allows for pattern recognition. The basic computational units of ANNs, known as neurons, receive inputs from external sources and combine them in a non-linear manner into an output signal by way of a transfer function, usually given as a sigmoid function:

$$f_j(I) = K\left(\sum_i I_i w_{ij}\right)$$

where $K = \frac{1}{1+e^{-I}}$

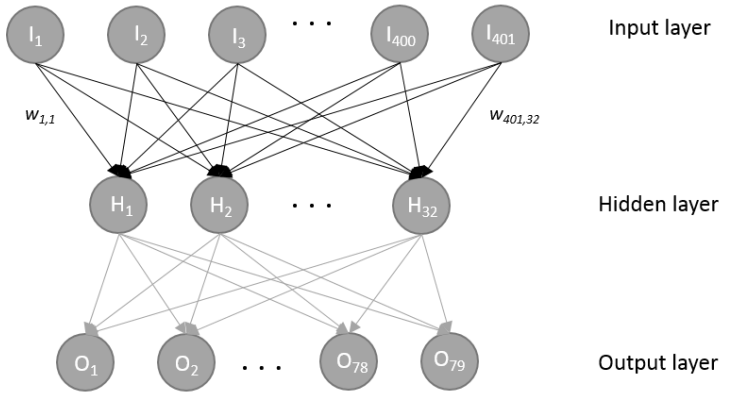


Figure 5-3 Architecture of neural network. Neural networks consist of multiple layers containing basic units of computation known as neurons. As the name suggests, input layer get input data while output layer neurons provide output. A neural network may be composed of one or multiple hidden layers.

Neural networks are composed layers of neurons and include an input layer, one or more hidden layers, and an output layer (Figure 5-3). Each layer of a neural network is connected to a subsequent layer by weighted connections. Neural networks are able to form a high-level mathematical model of a set of data by adjusting these weights, a process known as "training" or "learning". Training works by providing data to the neural network and allowing it to predict an output. The difference between the calculated output and the data point's known target value is used to determine how to change each weight in a process known as backpropagation (reference here). In the case of biological activity predictions, training an ANN consists of iterative weight changes that minimize the error between expected and predicted biological activity in terms of root mean square deviation.

Results and Discussion

Artificial Neural Network (ANN) QSAR models for predicting kinase selectivity profiles were built using the cheminformatics framework implemented in BCL::CHEMINFO. Inhibition data of 70 kinase inhibitors against 379 kinases reported by Davis et al [10a] was used to train the ANNs. The chemical structure of each inhibitor was encoded using molecular descriptors and this numeric description was used as the input to the ANNs, and binary experimental kinase activity was used as output for training. We will first describe the dataset used for building the models followed by molecular descriptors used for numerical encoding.

Training Dataset

ANN QSAR models were trained using kinase inhibitor data published by Davis et al [10a]. Davis et al reported interaction profile of a diverse set of 70 known kinase inhibitors against 379 kinases. The molecules that are tested

represent mature inhibitors optimized against specific kinases of interest. The study was performed using ATP site-dependent competition binding assays. Five models were developed using different K_D cutoffs for specifying active molecules – 0.1 μM, 0.5 μM, 1 μM, 3 μM and 10 μM.

Molecular descriptors

Chemical structures were encoded using a set of molecular descriptors using BCL::CHEMINFO [16-17]. The descriptors are translationally and rotationally invariant geometric functions that describe the distribution of molecular properties in the structure (e.g. mass, volume, surface area, partial charge, electronegativity, polarizability, etc.). Descriptors can be grouped into five categories based on the level of information they provide – 1D descriptors are computed as scalar values derived from a molecular formula, for example molecular weight and total charge. 2D descriptors are calculated using molecular connectivity information and include properties such as hydrogen bond acceptors/donors, number of ring systems, and approximations of surface area and volume. 2.5D descriptors are calculated using information about the molecular configuration (i.e. connectivity and stereochemistry). Conformation-dependent or 3D descriptors encode atomic properties (e.g. partial charge, polarizability) in a three-dimensional fingerprint using radial distribution functions (RDF) and 3D autocorrelations (3DA). Table 5-1 lists all the descriptors used in developing kinase selectivity model.

Table 5-1 List of descriptors used for describing molecules for QSAR models.

1D descriptors	2D autocorrelation descriptors	3D autocorrelation descriptors
Molecular weight	Atom sigma charge	Atom sigma charge
HbondDonor	Atom vcharge	Atom vcharge
HbondAcceptor	Atom in aromatic ring	Atoms in aromatic ring
LogP	Atom in fused aromatic ring	Atoms in fused aromatic ring
Total Charge	Atom signed polarizability	Atom signed polarizability
Number of rotatable bonds	Atom heavy sigma charge	Atom heavy sigma charge
Number of rings	Atom heavy vcharge	Atom heavy vcharge
Topological polar surface area		
Molecular girth		
Maximum ring size		
Bond girth		
Number of atoms in aromatic rings		
Number of atoms in fused aromatic ring		
Number of atoms in fused rings		
Atom Vcharge statistics		
Atom sigma charge statistics		

Artificial neural network model development and validation

Neural networks trained in this study contain 400 inputs (a result of encoding chemical structure with molecular descriptors), 32 hidden neurons, and one output neuron for each of kinase included in the model. The ANNs were trained using simple back propagation and a sigmoid transfer function with a weight update parameters eta=0.1 and alpha=0.5 [16-17].

Metrics to evaluate ANN prediction accuracy

Five models were generated by using different K_D cutoff values for specifying the active molecules. Each model predicts activity of a small molecule in terms of 379 binary outcomes for each of the kinase molecules. The binary predictions fall into the following four categories:

- True Positives (TP) – Experimentally active predicted to be active.
- True Negatives (TN) – Experimentally inactive predicted to be inactive.
- False Positives (FP) – Experimentally inactive predicted to be active.
- False Negatives (FN) – Experimentally active predicted to be inactive.

Table 5-2 Comparison of models developed using different cut-off values for indicating active molecules. The table gives an overall number of true/false-positives and negatives calculated over all 379 kinase molecules. Computed overall accuracy, the Matthew's correlation coefficient, sensitivity and selectivity is reported for each model.

Activity cutoff (μM)	ACC	MCC	SEN	SEL	TP	FP	TN	FN
0.1	68.10	0.14	58.99	68.71	971	7787	17097	675
0.5	78.18	0.26	54.02	81.26	1620	4410	19121	1379
1	78.99	0.31	53.13	83.41	2055	3760	18902	1813
3	78.59	0.37	54.20	84.79	2915	3218	17934	2463
10	78.45	0.42	57.59	85.27	3765	2944	17048	2773

Table 5-2 shows the overall accuracy (ACC), the Matthew's correlation coefficient (MCC), sensitivity (SEN) and specificity/selectivity (SEL) of each model calculated by pooling all true-positives, false-positives, true-negatives and false-negatives, across all kinases and small-molecules. The measures to assess quality of predictive models are defined as follows:

- Sensitivity (SEN) – TP / (TP + FN)
- Selectivity (SEL) – TN / (TN + FP)
- Accuracy (ACC) – (TP + TN) / (TP + TN + FP + FN)
- Matthews correlation coefficient (MCC) –
  $((TP \times TN - FP \times FN)) \,/\, \sqrt{((TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}$)

The three metrics accuracy, sensitivity and specificty are very stable for models using activity cutoff values of greater than 0.5 µM. However the Matthew's correlation coefficient is highest for cutoff value of 10 µM even with higher number of indicated active kinase-inhibitor pair increases. The prediction accuracy of models can also be evaluated using values derived from receiver operator characteristic (ROC) curves. A ROC curve plots the true positive rate (TPR, i.e. active molecules predicted as active) versus the false positives rate (FPR, i.e. inactive molecules predicted as active) as a fraction of the total number of known inactive molecules. A TPR vs. FPR slope of one, which

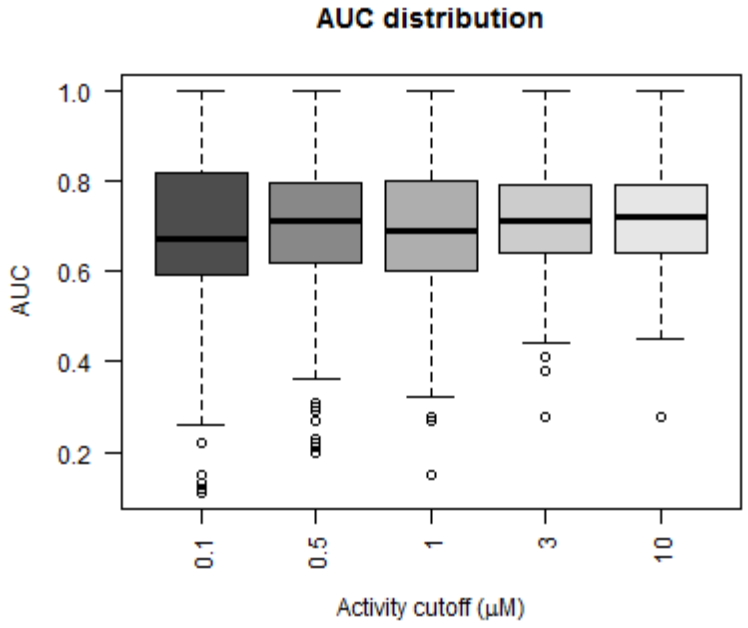


Figure 5-3 Box plot showing the performance of the five models using distribution of AUC values for individual kinases. The five models have been generated using activity values specified at different cutoffs as indicated on the x-axis. The upper and lower edges of box correspond to the first and third quartiles. The horizontal dash in the box represents the median value. The whiskers extend from edge to highest/lowest value that is within 1.5 × Inter-Quartile (IQR) of the box, where IQR is the distance between the first and third quartile.

results in area under the curve (AUC) of 0.5, indicates a model which is no better than random at correctly predicting a compound as active vs inactive. An increase in slope and therefore area under the curve indicates an increase in predictive power over a random guess. Figure 5-3 depicts a box plot showing the performance of the five models in terms of AUC values for each of the 379 kinases. The upper and lower edges of box correspond to the first and third quartiles. The horizontal dash in the box represents the median value. The whiskers extend from edge to highest/lowest value that is within 1.5 × Inter-Quartile (IQR) of the box, where IQR is the distance between the first and third quartile. The AUC value for more than 50% of kinases is above 0.75 for models that consider 3 µM and 10 µM as activity cutoff. Models were compared statistically using Man-Whitney paired test to see which model

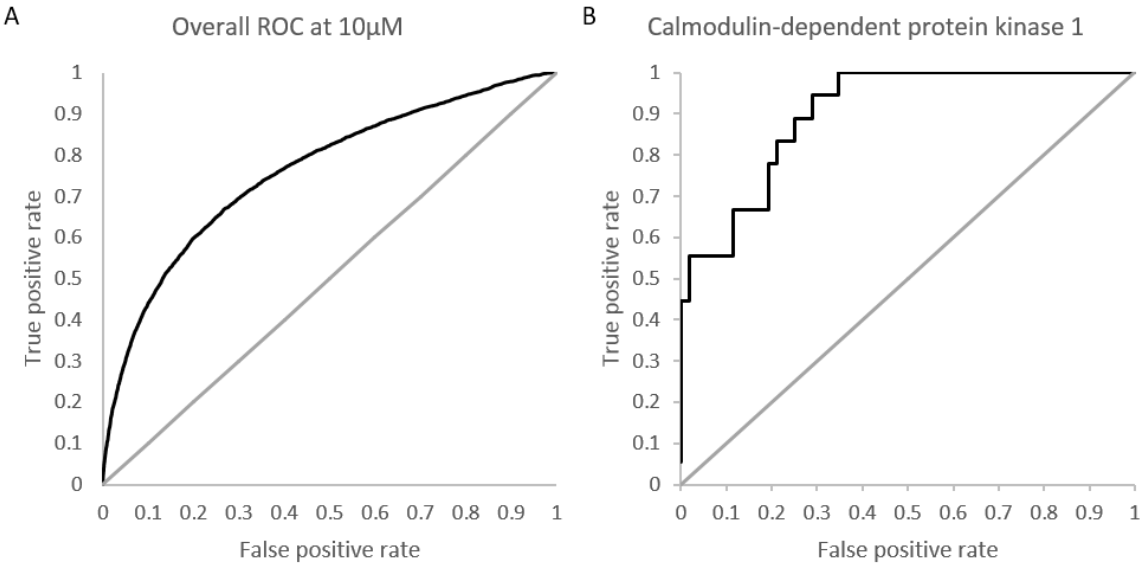


Figure 5-4 A) Overall Receiver Operating Curve computed for model developed at activity cutoff of KD = 10 µM. B) Example ROC curve for calmodulin-dependent protein kinase -1 which has 18 active molecules at KD < 10 µM with an AUC of 0.91.

performs better in terms of higher AUC values. Model built using activity cutoff value of 10 µM performs better than all other models at confidence interval of 95%. Fishers test showed that the 10 µM model is statistically significantly better than a random model which predicts 50% of cases as positive.

Figure 5-4A shows the overall ROC curve for model developed using activity specified at K_D < 10 µM with computed area under the curve of 0.76. Figure 5-2B is an example ROC curve for a kinase with 18 active molecules and high prediction accuracy (88%), and specificity (98%). The calculated AUC for this kinase, Calmodulin-dependent protein kinase-1, is 0.91. Figure 5-5A and B respectively show the heat maps of experimental and predicted activity for model developed using activity specified at K_D < 10 µM.

In the current approach neural network based QSAR models were trained to predict activity of small molecules against a panel of 379 kinases. The MCC for model developed using activity specified at K_D < 3 µM is 0.48 compared to 0.37 for structure based models developed by Subramanium et al[13]. Subramanium et al developed a computational model to predict activity of 15 kinase inhibitors against 280 kinase molecules by designing binding site signatures that use three-dimensional (3D) X-ray structure information of kinase-inhibitor complexes. Davis et al screened all these inhibitors against a panel of kinases except one, Roscovitine. Table 5-3 compares the performance of models developed by Subramanium et al for 14 investigated kinase-inhibitors to models developed in this study. The table shows overall true positive, false positive, true negative and false negative kinase-inhibitor pairs Two models were reported by Subramanium et al for 15 kinase inhibitors at activity cutoffs specified at K_D values of 0.1 µM and 3 µM. The method involves computing the binding site signature computed for each inhibitor

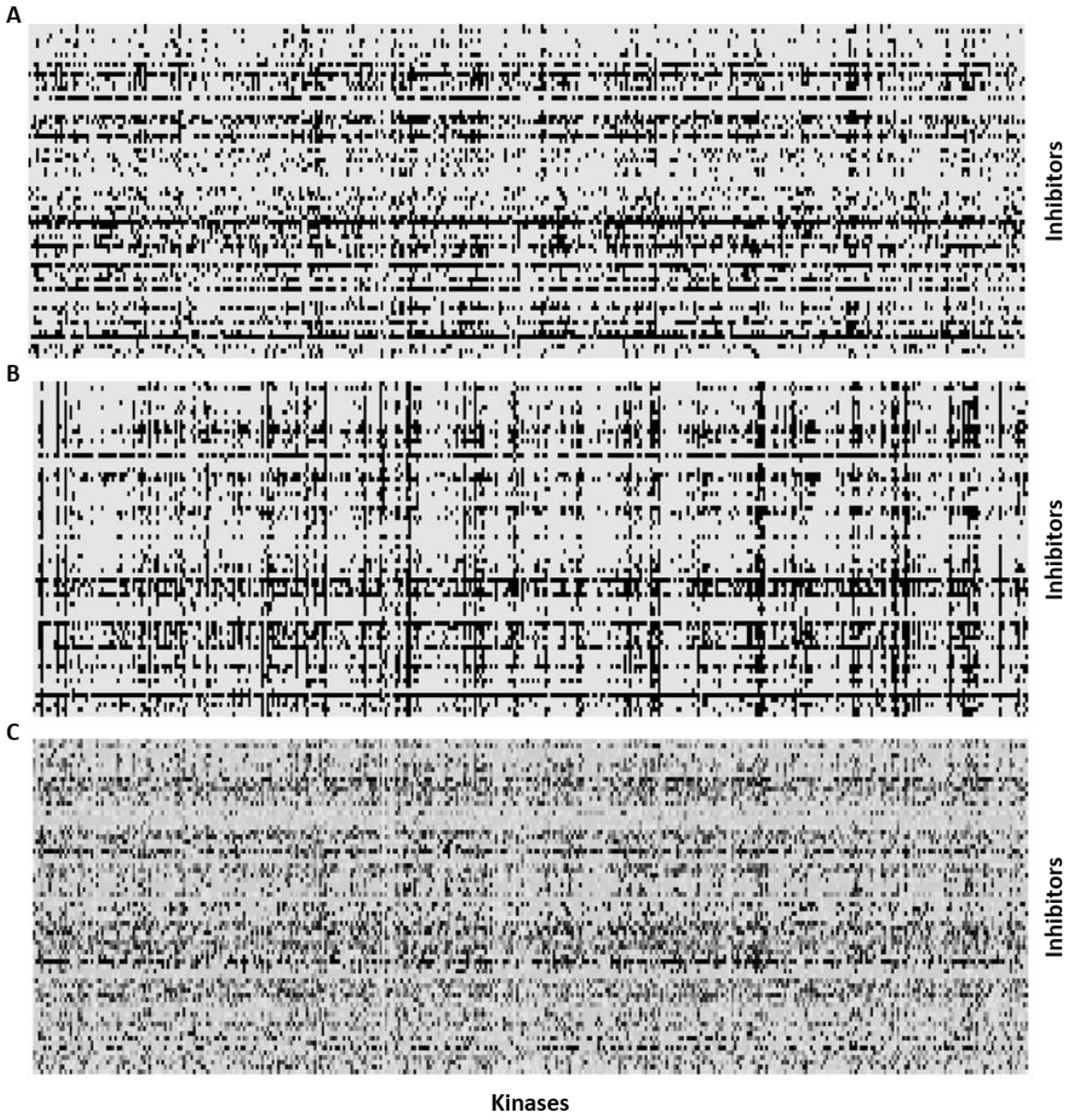


Figure 5-27 Activity matrix of all inhibitors versus all kinases. The pixels in black correspond to active molecules at 10 µM. A) Corresponds to experimental activity matrix as reported by Davis et al. B) Corresponds to predictions made by QSAR model. C) Difference of experimental activity matrix and prediction activity matrix.

using a co- crystal structure of kinase-inhibitor complex. Based on the similarity of binding site signature, the method predicts which other kinases the small-molecule inhibitor can bind. The models developed in this study predict the activity of small molecules against kinases. Here, the neural network predicts activity of each inhibitor against 379 kinases based on the chemical structure of the inhibitor. For each kinase, a different threshold of predicted activity is chosen for specifying activity of small-molecules. The models generated in this study using 0.1 µM cutoff performs worse compared to those reported by Subramanium et al. This is possibly because of sparsity of data as there are

very few kinase-inhibitor pairs with K_D less than 0.1 μM. However, our model generated at cutoff values of 3 μM is better than those reported by Subramanium et al in terms of high values for MCC, accuracy, and sensitivity. Models reported by Subramanium et al have higher specificity but very low sensitivity compared to models reported in this study. Appendix Table 3 shows the activity prediction for 15 molecules using models developed here and those by Subramanium et al at 3 μM. Our model performs better for highly cross reactive inhibitors like Staurosporine and VX-680. In general, since the QSAR models have only been trained on type-1 and type-2 inhibitors of kinase, its utility is limited for molecules that show inhibitory activity against at least one kinase molecule and target the ATP binding site.

Table 5-3 Performance comparison of kinase activity models developed by Subramanium et al and those developed here. Models developed with activities specified at different cutoffs are reported in the table. Reported is the accuracy, sensitivity and specificity of models developed for 15 kinase inhibitors studied by Subramanium et al. Models developed in this study were used for predicting activity of these 15 kinase inhibitors and results are tabulated.

	Cutoff (μM)	ACC	MCC	SEN	SEL	TP	FP	TN	FN
Subramanium et al	0.1	87.30	0.35	52.26	90.49	185	370	3521	169
	3	81.04	0.37	36.62	93.57	342	213	3098	592
this study	0.1	67.67	0.24	75.39	67.00	340	1727	3507	111
	3	78.63	0.48	74.45	79.83	944	891	3526	324
	10	79.26	0.52	73.79	81.39	1174	762	3332	417

Conclusions

In this study, QSAR models were developed for predicting activity of kinase inhibitors against a panel of 379 kinase enzymes. Kinase activity data was reported by Davis et al for 70 inhibitors in terms of K_D values obtained using ATP site-dependent competition binding assays. Five models were developed using activities specified at different cutoffs of K_D values. Statistical tests suggest that model using 10 μM as cutoff for activity has better predictive ability compared to other models. This model allows prediction of kinase specificity profile for weak binders. A pre-requisite to using this model is that a given small molecule to be tested should be active against at least one of the 379 tyrosine kinases present in the dataset as the neural network has been trained only on a small chemical space of known inhibitors. The predictive ability of the model varies significantly with AUC values for 75% of kinases ranging from 0.5 to 0.1. This model is a good starting point for predicting the selectivity profile of a new molecular entities against different kinases. This is especially useful after a computational high throughput screening of a virtual compound library when compounds need to be prioritized for experimental testing. Ideally a diverse set of drug-like molecules would be ordered and tested. The selectivity QSAR model developed here could be used for short listing compounds by scanning for molecules that are predicted to be selective.

References

1. (a) Edelman, A. M.; Blumenthal, D. K.; Krebs, E. G., Protein serine/threonine kinases. *Annu Rev Biochem* 1987, *56*, 567-613; (b) Hubbard, S. R.; Till, J. H., Protein tyrosine kinase structure and function. *Annu Rev Biochem* 2000, *69*, 373-98; (c) Schenk, P. W.; Snaar-Jagalska, B. E., Signal perception and transduction: the role of protein kinases. *Biochim Biophys Acta* 1999, *1449* (1), 1-24.
2. (a) Lemmon, M. A.; Schlessinger, J., Cell signaling by receptor tyrosine kinases. *Cell* 2010, *141* (7), 1117-34; (b) Paul, M. K.; Mukhopadhyay, A. K., Tyrosine kinase - Role and significance in Cancer. *Int J Med Sci* 2004, *1* (2), 101-115.
3. (a) Taylor, S. S.; Kornev, A. P., Protein kinases: evolution of dynamic regulatory proteins. *Trends Biochem Sci* 2011, *36* (2), 65-77; (b) Steichen, J. M.; Kuchinskas, M.; Keshwani, M. M.; Yang, J.; Adams, J. A.; Taylor, S. S., Structural basis for the regulation of protein kinase A by activation loop phosphorylation. *J Biol Chem* 2012, *287* (18), 14672-80.
4. Hanks, S. K.; Hunter, T., Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB J* 1995, *9* (8), 576-96.
5. Goldsmith, E. J.; Akella, R.; Min, X.; Zhou, T.; Humphreys, J. M., Substrate and docking interactions in serine/threonine protein kinases. *Chem Rev* 2007, *107* (11), 5065-81.
6. Vlahovic, G.; Crawford, J., Activation of tyrosine kinases in cancer. *Oncologist* 2003, *8* (6), 531-8.
7. Eglen, R. M.; Reisine, T., The current status of drug discovery against the human kinome. *Assay Drug Dev Technol* 2009, *7* (1), 22-43.
8. Schlessinger, J., Receptor tyrosine kinases: legacy of the first two decades. *Cold Spring Harb Perspect Biol* 2014, *6* (3).
9. Krause, D. S.; Van Etten, R. A., Tyrosine kinases as targets for cancer therapy. *N Engl J Med* 2005, *353* (2), 172-87.
10. (a) Davis, M. I.; Hunt, J. P.; Herrgard, S.; Ciceri, P.; Wodicka, L. M.; Pallares, G.; Hocker, M.; Treiber, D. K.; Zarrinkar, P. P., Comprehensive analysis of kinase inhibitor selectivity. *Nat Biotechnol* 2011, *29* (11), 1046-51; (b) Liu, Y.; Gray, N. S., Rational design of inhibitors that bind to inactive kinase conformations. *Nat Chem Biol* 2006, *2* (7), 358-64.
11. (a) Toledo, L. M.; Lydon, N. B.; Elbaum, D., The structure-based design of ATP-site directed protein kinase inhibitors. *Curr Med Chem* 1999, *6* (9), 775-805; (b) Zhang, J.; Yang, P. L.; Gray, N. S., Targeting cancer with small molecule kinase inhibitors. *Nat Rev Cancer* 2009, *9* (1), 28-39.
12. Sheinerman, F. B.; Giraud, E.; Laoui, A., High affinity targets of protein kinase inhibitors have similar residues at the positions energetically important for binding. *J Mol Biol* 2005, *352* (5), 1134-56.
13. Subramanian, G.; Sud, M., Computational Modeling of Kinase Inhibitor Selectivity. *ACS Med Chem Lett* 2010, *1* (8), 395-9.
14. Sciabola, S.; Stanton, R. V.; Wittkopp, S.; Wildman, S.; Moshinsky, D.; Potluri, S.; Xi, H., Predicting kinase selectivity profiles using Free-Wilson QSAR analysis. *J Chem Inf Model* 2008, *48* (9), 1851-67.
15. Free, S. M., Jr.; Wilson, J. W., A Mathematical Contribution to Structure-Activity Studies. *Journal of Medicinal Chemistry* 1964, *7*, 395-9.
16. Butkiewicz, M.; Lowe, E. W., Jr.; Mueller, R.; Mendenhall, J. L.; Teixeira, P. L.; Weaver, C. D.; Meiler, J., Benchmarking ligand-based virtual High-Throughput Screening with the PubChem database. *Molecules* 2013, *18* (1), 735-56.
17. Mendenhall, J.; Meiler, J., Improving quantitative structure-activity relationship models using Artificial Neural Networks trained with dropout. *J Comput Aided Mol Des* 2016.

SUMMARY

Drug discovery is a lengthy process that begins with target identification, lead compound discovery, and lead compound identification followed by pre-clinical studies. Traditionally the drug discovery process has been performed experimentally, which is expensive, time consuming, labor intensive and often meets with a low rate of success. Complementary computational technologies have been developed that reduce experimental work. These technologies enable identification of a small set of compounds that need experimental verification.

Once a target protein has been identified, lead compound discovery is carried out by high-throughput screening of a library of compounds. This requires the development of a high-throughput assay for screening against a target of interest. Once a lead compound is identified, optimization studies are carried out to improve its binding affinity to the target of interest. Optimization can be performed through medicinal chemistry studies either with or without the knowledge about of co-crystal structure or the binding pose. In the presence of a crystal structure, the lead compound is designed by analyzing its interaction with the protein target of interest. Derivatives are synthesized and their activities correlated with the structure of the molecule and the putative interactions they can have with the target of interest. In the absence of co-crystal structure of the target and lead compound, structure activity relationship studies are carried out with known binders and non-binders. Optimized compounds are then tested in-vivo for their absorption, distribution, metabolism, excretion and toxicological studies (ADMET). Following these studies, the optimized compounds may have to go through another round of optimization for desired ADMET properties.

Computational tools have been developed which complement each of the described experimental drug discovery tools. Virtual screening (*in-silico*) technologies have been developed to prioritize molecules for experimental testing. These include structure-based methods like docking and ligand-based methods like shape matching. For accurate modelling of molecular structures and their interactions, these drug discovery tools need to sample 3D conformations that include the conformation that binds to the target protein. Small molecules exist in multiple different conformations in solution and may bind to the target protein in one of those solution conformations or in an entirely novel conformation dictated by the target. Computational modelling requires representative low-energy molecular conformations for making reasonable predictions.

The most comprehensive method for conformation sampling is systematic or deterministic sampling where dihedral bonds are rotated by N degrees one by one to sample all possible combinations. However, this method quickly becomes intractable with increase in number of dihedral bonds. For example, if bonds are rotated by 30°, for a molecule containing four bonds the total number of conformations to be sampled is 4^{12}. Physics based or knowledge based methods have been developed that efficiently search the conformational space. Several commercially available software packages are available that use physics-based methods or knowledge-based methods for sampling conformations. Physics-based methods requires free energy computation which is resource

intensive and are often not suitable for high-throughput experiments. Knowledge based methods have been developed which are used widely in both the academia and industry. These methods use existing information about small molecule conformations to sample low energy ligand conformations. This information is derived from the crystallographic structure databases like the CSD or the PDB. These methods analyze the torsional profile of different dihedral bonds from these databases and apply this information during conformation sampling.

We have developed a novel knowledge-based conformation sampling algorithm, BCL::CONF, which derives fragment conformations from the CSD/PDB and applies this information to sample conformations. While not all small-molecules have been crystallized, the hypothesis is that conformational space available to fragments of small molecules has adequately been represented in the crystallographic databases. Brameld, et al have shown that fragment conformations seen in the CSD are accurate representations of those seen in ligands found in complex with proteins in the PDB. BCL::CONF uses frequently observed fragment conformations in the CSD/PDB to sample small molecular conformations. Use of fragment conformations takes into account torsional profile of multiple dihedral bonds, which compose distinct chemical environments, at the same time. This is clearly an advantage over other knowledge-based conformation sampling algorithms, which generally treat dihedral bonds in a disconnected manner. BCL::CONF is thereby able to leverage correlations between dihedral bonds of fragments and apply those during conformational sampling. In addition, use of extended fragment-conformations captures the effect of substituents on the torsional propensities of dihedral bonds given the surrounding chemical environment.

The algorithm performs better than the most popular methods in the field in terms of accuracy and speed of computation. Knowledge-based dihedral bond torsion profiles from the CSD have been used for more than two decades for small molecule conformation sampling. Most of the successful algorithms derive torsion angle preferences from structural databases like the CSD or PDB, or from molecular mechanics simulations of small molecules. A few methods use small fragment conformations generated through molecular mechanics, or perform a stochastic search using molecular mechanics force field for energy calculations.

BCL::CONF uses torsional profiles of frequently occurring fragments in the CSD and the PDB, and applies them to molecules of interest. BCL::CONF performs better because it captures torsional correlation across multiple bonds instead of sampling rotatable bonds one at a time. Exhaustive sampling ultimately finds same conformations as correlated torsion sampling but will need to produce much larger conformation ensemble, and is not generally capable of identifying the lowest energy conformations. BCL::CONF is novel in that it uses conformations of large fragments from structure databases. The algorithm uses a scoring function that favors those molecular conformations whose sub-fragment conformations exist in the CSD or PDB. The scoring function favors conformations of larger fragment over smaller fragments. This algorithm ranks among the fastest available while performing high-accuracy conformation sampling, thanks to a fast look up scoring function that precludes the need

to do expensive physics-based calculations to rank conformations. Use of extended fragment conformations also allows BCL::CONF to take into account substituent effects that affect the local torsion profile of dihedral bonds.

Fragment conformations from the CSD or PDB need to be used carefully as they are often distorted due to crystal packing effect. A fragment conformation must be seen enough times to be included in the fragment library used during conformation sampling, to mitigate the impact of crystallographic errors and the incomplete sampling seen in particularly rare fragments. Rigorous benchmarking yielded four as the minimum number of times a fragment conformation should be seen in the structure database to be used during sampling. The scoring function also helps to control for packing effects. Smaller counts would exist for an anomalous conformation of large fragments caused by packing effects when compared to smaller fragments. If fragments exist in the same conformation multiple times in a structural database, it is more likely that the conformation is not an artifact of packing effect. Larger fragments have greater intra-molecular forces per atom accounted for in the torsional profile, while intermolecular forces (due to crystal packing) per atom remain a constant. As the scoring function gives a higher weightage to conformations of larger fragments, it helps in controlling for crystal packing effects to some extent.

The PDB contains crystal structure entries of a number of proteins complexed with 7K different small molecules. Molecular conformations are often perturbed due to interactions with a protein. Again, the strict conformation count criteria and the scoring function help in controlling the impact of these perturbations. Perturbation free small molecular conformations can be obtained from Nuclear Magnetic Resonance (NMR) studies. NMR studies are usually performed in solution and can be done at room temperature. Conformational ensembles from NMR studies are likely more accurate compared to those obtained using X-ray crystallographic studies.

Fragment conformation concept for sampling small molecules is analogous to amino-acid rotamer sampling technique used by ROSETTA macromolecular software. BCL::CONF, using a library of fragment rotamers, can be used by any external software to sample conformations. This enables the use of BCL::CONF from within ROSETTA for drug discovery applications. Incorporation of BCL::CONF into ROSETTA allows on-the-fly conformation sampling during ligand docking and implementation of structure based *de novo* drug-design algorithms. BCL::CONF is also used to sample ligand conformations in the drug-design module of the online scientific game FOLDIT. In FOLDIT, the structure-based drug design problem is crowdsourced to a large community of game players who, via interactive gameplay, come up with solutions to scientific problems presented as puzzles. The players are provided with basic fragments with which they build ligand molecules in the binding site of a target molecule of interest. BCL::CONF provides an efficient algorithm for online conformation sampling required for fast-paced gameplay.

On the ligand-based drug discovery side, BCL::CONF is being used with QSAR based drug design and evolutionary algorithms. A number of approaches for training QSAR models using multiple conformations generated by BCL::CONF were implemented and tested. Currently in the BCL, 3D QSAR models are developed using a single conformation generated using an external program (CORINA). 2D and 3D molecular descriptors are calculated and used for training

neural network to correlate molecular structure to activity. The hypothesis for using multiple conformations is that ligands that bind the target of interest bind in poses with similar 3D distribution of electrochemical properties. The neural network will learn 3D property distributions that are common and unique to active molecules, thus improving the predictive power.

QSAR models are trained using datasets containing active and inactive molecules against a target molecule of interest. In the Meiler lab, eight datasets derived from the PUBCHEM database have been cleaned and prepared for QSAR modelling. These datasets were the outcome of high throughput screening efforts performed at various academic institutions against targets like serine/threonine kinase, glutamate receptors, etc. Single low energy conformation for both actives and inactive molecules are generated using a conformer generator method. Molecules are numerically represented and used to train neural networks to correlate structure to activity. Current state of the art method in the Meiler lab uses CORINA generated conformations for training QSAR models (MC).

In the work described in this thesis, multiple conformations were used to train the neural network so as to identify 3D conformations that correspond to active conformations for a target of interest. Several schemes for training neural networks with multiple conformations were tried but with limited success. This is possibly due to the parameters used for training ANNs, which were optimized for single CORINA conformation of each molecule. A full benchmark study to optimize parameters for multiple conformations may produce better results.

The neural network learning algorithm was modified to train using multiple conformations. In every cycle of training, the neural network keeps track of all the conformations of a molecule and predicts activity of each. Only the conformation with the highest-predicted activity is used to tune the network. In one experiment, five conformations of both the active and inactive molecules were used to train the network. In another approach, instead of using explicit conformations, average descriptor values calculated over a diverse set of conformations were used. Descriptor values were averaged over ~50 conformations of both active and inactive molecules. The descriptor average values were then used to train a traditional neural network.

A straightforward experiment to test whether neural network parameters need to be tuned is described next. This experiment can be performed by using a dataset of small molecules containing the actual binding conformation against a particular target. For example, a number of kinase inhibitors have been co-crystallized with different kinase enzymes and reported in the PDB. Most kinase inhibitors target the ATP binding site and possibly an allosteric site within the kinase domain. Due to high degree of homology between kinase enzymes, kinase inhibitors often show activity against multiple kinase enzymes and possibly interact in a similar binding conformations across the different kinases. Not all inhibitors have been crystallized against their entire set of possible kinase targets. For example, a kinase enzyme like ABL has close to 30 kinase inhibitors molecules but only about 15 have been co-crystallized with c-ABL. The QSAR model will be trained using only the binding conformation of the active molecules and a single conformation of the inactive ligands generated using BCL::CONF (MC1) or CORINA (MC2). We expect MC1 and MC2 to

perform equally well if there is no bias in neural network tuning. If performance is substantially different, neural network parameters will be needed to be tuned for using BCL conformations. The tuned neural networks can then be used for the next set of experiments to validate the hypothesis that they can learn the correct binding conformation during training. For this experiment, QSAR models (M1) will be trained with multiple conformations of the inhibitors including the binding conformation. Multiple models will be trained using the binding conformation of active ligands and up to five conformations generated using BCL::CONF. If the hypothesis is correct and M1 model is trained correctly, the model should be able to identify the correct binding conformation of molecules and the model performance should be comparable to that of MC1/MC2. If M1 model perform worse than MC1 and MC2, parameter tuning of neural networks is required for M1 model. The final model can then be developed by training without the active ligand binding conformations and varying the number of ligand conformation to benchmark the best parameters. If such a dataset is unavailable where binding conformations of active molecules are known, alternatively a dataset containing conformationally rigid active molecules can be used. Due to limited conformation flexibility of active molecules, correct active conformation is expected to be fed during training.

Here our hypothesis is that active molecules bind a target by adopting similar 3D shapes. This requires that all the active molecules are able conform to the binding pocket. According to the hypothesis, neural network model may perform better if trained using conformations that are common to most active molecules in a dataset. In the preprocessing step of such an experiment, all the conformations of active molecules will be generated and aligned. Each conformation of one active molecule will be aligned with another in a pair-wise manner. Conformations that allow good alignment of most of the active molecules will be used for model training.

In a second project, computational drug discovery techniques were applied to find novel small molecules that can selectively bind DDR1 for therapeutic purpose or for use as probes to explore its biological role. Discoidin domain receptors have been implicated in osteoporosis, cystic fibrosis, fibrosis of kidneys etc. Homology models were developed using crystal structures of similar receptor tyrosine kinase domains. At the time the homology models were developed, DDR1 kinase domain had not been crystallized. DDR1 kinase domain is highly homologous to kinase domains of other receptor tyrosine kinase receptors, many of which have been co-crystallized with ligands. Homology models were developed using homologous receptor kinase domains crystallized with different ligands to model the perturbations in the ATP binding pocket and important loops like activation and catalytic loops. Homology models were selected based on the ability to dock kinase inhibitors and recovering the binding pose observed in homologous RTKs. The docking poses of Dasatinib and Imanitib were later found to be in agreement with poses in DDR1 co-crystal structures.

Recently several novel inhibitors of DDR1 have been reported increasing the number to greater than 100. This has allowed development of QSAR models for prediction of inhibitors against DDR1. Since the number of known inactive molecules is sparse, a generic kinase enzyme inactive dataset was used for developing the QSAR models.

The generic set of inactive molecules reported in PUBCHEM had been identified in a high-throughput screening assay against serine-threonine kinase. Three QSAR models were developed through an iterative process of model development and experimental testing. The QSAR model was used to prioritize small molecules for DDR1 experimental screening from the in-house Vanderbilt high-throughput library. Ten molecules were prioritized for DDR1 kinase activity screening, all of which were found to be inactive against DDR1. The QSAR inactive dataset was updated with the addition of these ten inactive molecules, and the QSAR model was retrained. A second round of experimental screening was performed and 25 molecules were prioritized for testing. Two molecules were found to be active and 23 inactive. The QSAR models were trained again with the dataset after updating the datasets. For the next round of experimental testing molecules were prioritized from the eMolecules database, a commercially available library of molecules. Fifty diverse set of molecules were prioritized based on the predicted activity for experimental testing. Out of these, four molecules were found to have inhibitory DDR1 kinase activity. Two inhibitors showed more than 80% inhibition of DDR1 kinase. These molecules are reported kinase inhibitors but have not been identified as DDR1 inhibitors in literature. Two other inhibitors, which inhibit 60% of DDR1 kinase activity, are novel inhibitors with no known activity against any kinases.

Further experimental validation and categorization are being carried out in the Pozzi lab at Vanderbilt University. Lead optimization studies can be performed at Vanderbilt University through computational design followed by synthesis and testing. The newly identified molecules can be docked into DDR1 homology models using ROSETTA followed by structure-based design. DDR1 kinase is gaining lot of interest in the scientific community as a potential therapeutic target. In the last two years, several new selective inhibitors have been reported. In our studies, we have found at least two novel scaffolds that have inhibitory activity against DDR1 kinase. Further characterization, modifications and optimizations are needed to develop highly active and selective molecules.

APPENDIX

List of abbreviations

3D	3-Dimensional
3D-QSAR	Three-Dimensional Quantitative Structure Activity Relationship
Abl	Abelson Murine Leukemia
ADMET	Absorption, Distribution, Metabolism and Excretion - Toxicity
ANN	Artificial Neural Networks
ATP	Adenosine Tri-Phosphate
CADD	Computer Aided Drug Discovery/Design
CASP	Critical Assesment of Techniques for protein prediction
CHARMm	Chemistry at HARvard Molecular Mechanics
CSD	Cambridge Structure Database
DDR	Discoidin Domain Receptors
DFG	ASP-PHE-GLY motif
DMPK	drug metabolism and pharmacokinetics
DS	Discodin
GPCR	G protein-coupled receptor
HRD	HIS-ARG-ASP motif
HTS	High Througput Screening
LB-CADD	Ligand-Based CADD
MC	Monte Carlo
MD	Molecular Dynamics
MM	Molecular Mechanics
MMFF	Merck Molecular Mechanics Force Field
MOE	Molecular Operating Environment Conformation Import Routine
MOE-SS	Molecular Operating Environment Stochastic Search Routine
NMR	Nuclear Magnetic Resonance
PDB	Protein Data Bank
QM	Quantum Mechanics
QSAR	Quantitative Structure Activity Relationship
RMS	Root Mean Square

RMSD	Root Mean Squared Deviation
ROC	Reciever Operating Characteristic
RTK	Receptor Tyrosine Kinase
SAR	Structure Activity Relationship
SB-CADD	Structure-Based CADD
SVM	Support Vector Machine
TK	Tyrosine Kinases
TYR	Tyrosine
vHTS	virtual-HTS
YRD	TYR-ARG-D

Chapter 2

Appendix Table 1 Optimization of BCL::CONF parameters using different number of iterations and temperature values. Optimization was done for better recovery of native conformations, fewer average number of conformations per molecule and computation time.

Iterations	T	Recovery%										Average number of confo-rmations	Time
		0.25	0.5	0.75	1	1.25	1.5	1.75	2	2.25	2.5		
200	1	11.46	36.36	62.45	76.68	86.17	91.30	96.44	99.21	100.0	100.0	52.40	1.6 s/mol
	2	11.46	36.76	59.68	78.26	87.35	92.49	98.02	99.60	99.60	100.0	57.28	
	3	10.67	37.15	61.66	79.45	88.93	93.28	97.63	99.60	99.60	100.0	60.27	
	4	11.07	37.94	61.66	77.08	86.17	93.28	96.84	100.0	100.0	100.0	61.75	
250	1	9.88	37.55	63.24	76.68	86.17	92.49	95.26	98.02	98.81	99.21	58.00	1.9 s/mol
	2	10.28	35.57	61.66	76.68	85.38	91.70	96.05	98.81	98.81	99.21	64.81	
	3	11.86	36.36	62.45	79.05	88.54	91.70	96.05	98.42	98.81	99.21	66.83	
	4	11.86	39.92	61.66	78.26	88.14	92.49	96.44	98.81	99.21	99.21	66.43	
300	1	10.67	37.15	63.64	79.05	88.93	94.07	97.23	99.21	99.60	99.60	63.23	2.2 s/mol
	2	11.46	37.55	64.82	79.05	87.35	91.70	96.84	98.42	98.81	99.21	68.36	
	3	11.07	37.94	67.59	79.05	88.54	91.70	97.23	98.42	98.42	98.81	70.75	
	4	12.65	39.92	65.22	80.24	87.75	92.89	96.84	100.0	100.0	100.0	71.75	

Protocol capture

The protocol capture contains steps necessary to generate molecular conformations using BCL::CONF. The input parameter files and computational steps are necessary to make fragment library, rotamer library and using the rotamer library for conformational sampling. The final rotamer library and BCL::CONF executable can be downloaded at http://www.meilerlab.org. The commands required for generating rotamer library are provided in scripts which are kept in the Thesis folder in the sub-directory Chapter2. This path Thesis/Chapter2 is referred to $PATH from here on.

Step	Text	Commands	Comment
1.Setup for running protocol capture	Folder $PATH has three folders bin, input and config in it.	Download the BCL::CONF executable at http://www.meilerlab.org and put it in the bin folder. Get bcl_license.txt file and put it in the bin folder.	
1. Prepare the rotamer library from a given structure database.	If the structure database is large, jobs provided in the script will have to be split up.	Run the create_rotamer_library.sh script and provide the database as first parameter by using the following command – /bin/bash $PATH/config/generate_rotamers.sh [your database] You can download the rotamer library obtained from CSD from http://www.meilerlab.org and keep it in **$PATH**/bin to use it.	Input: The structure database using which rotamer library will be created. Output: Rotamer library in the $PATH/input directory is composed of three files and a directory : rotlib.constitutions.txt.gz rotlib.substructure.txt.gz rotlib.configuration_mapping.txt.gz directory - rotlib_conformations
2. Generate conformation data for publication	Steps: 1. Generate conformations using methods of interest. 2. For each method, create a file containing rmsd of generated conformations to native conformation. Each line contains rmsd-to-native for conformations of a single molecule of the benchmark dataset. 3. Name the above file as vernalis_{method}_R.txt. An example file is vernalis_bcl_R.txt which contains rmsd-to-native values for the vernalis dataset.	BCL conformations were generated using – $PATH/bin/bcl-apps-static.exe molecule:ConformerGenerator -rotamer_library 'File(prefix=$PATH/input/rotlib) –ensemble_filenames INPUT -top_models 100 -conformers_single_file OUTPUT –native_ensemble NATIVE –remove_h	Input: - INPUT : $PATH/input/{zeroed_vernalis.sdf} - NATIVE : $PATH/input/{native_vernalis.sdf} Output: - OUTPUT : $PATH/input/{vernalis_bcl_R.txt}
3. Generate publication figures.	Steps: 1. Generate files containing rmsd-to-native data for each method and dataset as mentioned in step 2.	Execute script in $PATH/config to generate plots : $PATH/config/generate_publication_figures.sh	Input: - $PATH/input/{all files listed below} vernalis_bcl_R.txt,vernalis_confimport_R.txt,vernalis_confgen_R.txt,vernalis_dihedral_R.txt,vernalis_omega_R.tx,vernalis_rdkit_R.txt, Output: Image files in $PATH/input

			Comparison of closest to native conformer generated for each molecule in the dataset – Files (example) : vernalis_bcl_moe_comparison.txt (for all molecules), vernalis_bcl_moe_comparison1.txt (molecules with rotatable bonds >0 and <4), vernalis_bcl_moe_comparison1.txt (molecules with rotatable bonds >3 and <6), and so on
3. Generate conformations by user defined parameters	An example command line to demonstrate user defined parameters that can be modified for conformational sampling	$PATH/bin/bcl-apps-static.exe molecule:ConformerGenerator -rotamer_library 'File(prefix=rotlib)' –ensemble_filenames INPUT -temperature 3 -max_iterations 200 -conformation_comparer SymmetryRMSD 0.25 - top_models 100 -conformers_single_file OUTPUT	

Chapter 3

The goal of this project was implementation of BCL::CHEMINFO neural networks that could be trained with multiple molecular conformations. Multiple implementations were tested in increasing order of complexity. The protocol capture will provide step by step directions to perform these experiments. Conformations were generated using BCL::CONF. Appropriate control experiments were performed to evaluate performance of new implementations. The various implementations are tabulated below

Appendix Table 2 List of all the experiments performed toward developing QSAR models using multiple conformations.

Experiments	Description
CORINA	Standard neural network trained using CORINA generated conformations; Control experiment
BCL_ONE	Standard neural network trained using single conformation generated using BCL::CHEMINFO; Control experiment
BCL_CONFORMATION_AVG	Standard neural network trained using average descriptors calculated over multiple conformations.
BCL_CONFORMATION_AVGStdDev	Standard neural network trained using average and standard deviations of descriptors calculated over multiple conformations.
BCL_FIVE_PREVROUND	Trained using five conformations. Modified neural networks that backpropagate errors from the first conformation that has higher predicted activity compared to previous round prediction.
BCL_FIVE_NODP	Trained using five conformations. Neural network modified such that it forward propagates all five conformations during training cycle and backpropagates errors from only the best. No dropout was implemented.
BCL_ONE_NODP	Standard neural network trained using single BCL::CONF generated conformation. No dropout; Control experiment
BCL_FIVE	Same as BCL_Five_Nodp but with drop out
BCL_TWOSTAGE	BCL_CONFORMATION_AVG followed by BCL_FIVE TRAINING

Protocol capture

This protocol provides a step by step process of reproducing all the experiments tabulated in Appendix Table and using them to predict activity for molecules in screening libraries. QSAR models were built using BCL::CHEMINFO which is available at http://www.meilerlab.org and is free for academic use. All the input files are provided in

directory Thesis_directory/Chapter3/ henceforth abbreviated as $PATH. Experiment is present in a separate directory with its own set of scripts and input files. Each directory contains a bin file containing BCL executable, config file containing scripts and parameter files used for training QSAR models, and input file containing all the input and final output files generated during the experiment. The dataset files are kept in $PATH/Datasets

1. Common steps for performing all the experiments –

Step	Text	Commands	Comment
1. File locations used for the experiment	The directory is located at $PATH/Experiment_name		
2. Generate all the input files to carry out the experiment	Run script jobs_setup_script.sh on piranha cluster Conformations are sampled and desired number of conformations are stored.	Run the command from the input directory Rscript $PATH/config/jobs_setup_script.sh	Output: A directory is generated for each dataset. Conformations are stored in a subdirectory for each of the datasets in active_conformation* and inactive_conformation* directories.
3. Prepare descriptor files and setup for submitting jobs to train qsar models	Run script prepare_datasets.sh from the config directory. It will generate descriptors and randomize each dataset.	Run the command from the input directory /bin/bash $PATH/config/prepare_descriptors.sh	Output: Descriptor files are generated in directories named *_conformation
4. Train QSAR models	Run script train_qsar.sh from the config directory on the piranha cluster	Run the command from the input directory /bin/bash $PATH/config/train_qsar.sh	Generates QSAR models for each dataset. Models are stored in models directory under *_conformation directories.
5. Calculate area under curve	Results are stored in $PATH/input/dataset_id/results directory. When multiple conformations are used, blind dataset prediction is done. Then use script $PATH/config/calculate_results.sh	If blind dataset prediction is done as in the case of multiple conformations, execute the script from $PATH/input directory using command /bin/bash $PATH/config/calculate_resluts.sh	

2. Output files specific for each experiment.

Experiments	Outputs

BCL_ONE	• Conformations
	  ○ Active – $PATH/BCL_One/input/dataset_id/active_conformation_1/*_actives_clean_conformations.1.sdf.gz
	  ○ Inactive - $PATH/BCL_One/input/dataset_id/inactive_conformation_1/*_inactives_clean_conformations.1.sdf.gz
	• Models - $PATH/BCL_One/input/dataset_id/one_conformation/models/dataset_id
	• Results – $PATH/BCL_One/input/dataset_id/one_conformation/results/dataset_id/final_objective.ind.merged.txt

BCL_CONFORMATION_AVG	• Conformations – Conformations are generated on the fly and descriptors averaged over them are output in the bin file. Bin file $PATH – $PATH/input/dataset_id/twelve.bin • Models - $PATH/BCL_One/input/dataset_id/models/6_0_hd0.05_vd0.25 • Results – $PATH/BCL_One/input/dataset_id/results/6_0_hd0.05_vd0.25/final_objective.ind_merged.txt
BCL_CONFORMATION_AVGStd Dev	• Conformations – Conformations are generated on the fly and descriptors averaged over them are output in the bin file. Bin file $PATH – $PATH/input/dataset_id/twelve.bin • Models - $PATH/BCL_One/input/dataset_id/models/6_1_hd0.05_vd0.25 Results – $PATH/BCL_One/input/dataset_id/results/6_1_hd0.05_vd0.25/final_objective.ind_merged.txt

BCL_FIVE_PREVROUND	• Conformations –
	  ○ Active – $PATH/BCL_One/input/dataset_id/active_conformations_5/*_actives_clean_conformations.5.sdf.gz
	  ○ Inactive - $PATH/BCL_One/input/dataset_id/inactive_conformations_5/*_inactives_clean_conformations.5.sdf.gz
	• Models – $PATH/BCL_One/input/dataset_id/five_conformations/models/blind*
	• Results – Run $PATH/config/calculate_results.sh$PATH/BCL_One/input/dataset_id/five_conformations/final_results.txt

BCL_FIVE_NODP	• Conformations – ○ Active – $PATH/BCL_One/input/dataset_id/active_conformations_5/*_actives_clean_conformations.5.sdf.gz

	○ Inactive - $PATH/BCL_One/input/dataset_id/inactive_conformations_5/*_inactives_clean_conformations.5.sdf.gz • Models – $PATH/BCL_One/input/dataset_id/five_conformations/models/nodropout_resilient* • Results – Run $PATH/config/calculate_results.sh$PATH/BCL_One/input/dataset_id/five_conformations/nodropout_resilient_final_result.txt
BCL_ONE_NODP	• Conformations – ○ Active – $PATH/BCL_One/input/dataset_id/active_conformation_1/*_actives_clean_conformations.1.sdf.gz ○ Inactive – $PATH/BCL_One/input/dataset_id/inactive_conformation_1/*_inactives_clean_conformations.1.sdf.gz • Models – $PATH/BCL_One/input/dataset_id/one_conformation/models/nodropout_resilient • Results – $PATH/BCL_One/input/dataset_id/one_conformation/results/nodropout_resilient/final_objective.ind.merged.txt
BCL_FIVE	• Conformations – ○ Active – $PATH/BCL_One/input/dataset_id/active_conformations_5/*_actives_clean_conformations.5.sdf.gz ○ Inactive - $PATH/BCL_One/input/dataset_id/inactive_conformations_5/*_inactives_clean_conformations.5.sdf.gz • Models – $PATH/BCL_One/input/dataset_id/five_conformations/models/nodropout_resilient* Results – Run $PATH/config/calculate_results.sh$PATH/BCL_One/input/dataset_id/five_conformations/nodropout_resilient_final_result.txt
BCL_TWOSTAGE	BCL_CONFORMATION_AVG followed by BCL_FIVE TRAINING

Chapter 4

Protocol Capture – Homology modelling and Docking

Homology models were created for DDR1 and DDR2 using ROSETTA [34] for both the active (DFG-in) and in-active (DFG-out) conformations. DDR1 active conformation was built from templates – PDB: 2PVF, PDB: 2X2L, and PDB: 3C4F. The templates are all tyrosine kinase domains in the active conformation in complex with inhibitors and have at least 40% similarity to DDR1 sequence. Structural alignment was done using MUSTANG [35] followed by sequence alignment of DDR1 sequence to the structure alignment. ROSETTA was then used for threading and generating homology models. Hierarchical docking protocol (Figure 4-5) was used to select homology models that would be used for docking studies. Native conformation of dasatinib, ligand that binds active state, was docked into top 50 homology models using a flexible backbone protocol generating 2000 docking per homology model. The ligand was docked in the same relative spot as seen in its co-crystal structure with Abl kinase (PDB: 2GQG). Initial docking protocol involved 2 Å translation and 180° rotation. Top 10% of all models were clustered on the basis of ligand rmsd from the native ligand position in PDB: 2GQG. Top 1% of each cluster was used in the next round of docking. A second subsequent docking round involved smaller ligand perturbations. Finally top scoring models from different clusters were used for docking studies.

Dasatinib and imatinib were docked into DDR1 active and inactive state homology models respectively. Multiple conformations of ligands were docked into homology models. The ligands were docked into the ATP binding pocket using a 5 Å translations and 360° degree rotation. Figure 4-8A and B respectively show score versus rmsd of plot of imatinib and dasatinib docked into DDR inactive and active-state homology models. The rmsd in the plot is the atom-pair root mean square deviation of docked pose of ligand versus the native pose of dasatinib in PDB: 2GQG. The docking funnels give hope for the success of ROSETTA in homology modeling and docking in DDRs.

This protocol capture contains the steps necessary to obtain homology models and perform the docking calculations reported in the manuscript. The input parameter files and representative models of steps necessary to carry out the steps outlined in this protocol relating to the results found in the manuscript are provided in the attached supplementary information. Multiple templates were used to generate both the active and inactive DDR1 kinase models. For the sake of simplicity the protocol capture describes modeling of DDR1 onto the template PDB: 2PVF. The ROSETTA 3.4 software suite is publically available and the license is free for non-commercial users at http://www.rosettacommons.org/. The directory $PATH are as found in downloaded ROSETTA source. Final homology models used for docking studies are provided.

1. Structural alignment of kinase domain templates

Step	Text	Commands	Comment
1A. Prepare kinase crystal structures from the Protein Data Bank.	Homology models for DDR1-in (active) conformation were created using four templates - PDB: 2PVF, PDB: 2X2L, PDB: and 3C4F. DDR-out (inactive) conformation was modeled based of PDB: 3BEA, PDB: 4AT5 and PDB: 4HVS.	Obtain PDB files: Download PDB: 2PVF from the Protein Data Bank at http://www.rcsb.org. Clean the PDB file using the following script rosetta_tools/protein_tools/scripts/clean_pdb.py 2pvf.pdb A > 2PVF.PDB_A.pdb	Input: Kinase domain crystal structures from the Protein Data Bank at http://www.rcsb.org. Output: Active conformation Inactive conformation 2PVF.PDB_A.pdb
1B. Perform a structural alignment of kinase domains using crystal structures from the Protein Data Bank.	Structural alignment was performed separately for active and inactive state using MUSTANG (Konagurthu et al., 2006), as seen in Fig S1.	mustang -p . -i 1OPK.PDB_A.pdb 2PVF.PDB_A.pdb -o results -F fasta -D 2.5	Input: 2PVF.PDB_A.pdb Output: allkinases.afasta allkinases.pdb

2. Sequence alignment of the DDR1 to template sequences

Step	Text	Commands	Comment
2B. Sequence alignment of DDR1 sequence	The sequence of the DDR1 was aligned with the profile of structurally aligned templates using CLUSTALW (Thompson et al., 1994).	Input target sequence ddr1_VAseq.fasta and profile alignment allkinases.afasta to http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::clustalO-profile. Default settings were used. The alignment was modified manually to get good alignment in conserved regions like the DFG and HRD sequences.	Input: ddr1_VAseq.fasta, allkinases.fasta Output: ddr_fasta.aln modified_ddr_fasta.aln

3. Thread target sequence onto template backbone coordinates

Step	Text	Commands	Comment

3. Thread target sequence DDR1 onto template backbone coordinates.	The sequence of the target DDR1 was then placed onto the backbone coordinates of each template structure.	rosetta_tools/protein_tools/scripts/thread_pdb_from_alignment.py --template=2pvf.pdb --target=ddr1_VAseq --chain=A --align_format=clustal modified_ddr_fasta.aln 2PVF.PDB_A.pdb ddr1_on_2pvf.pdb	Input: modified_ddr_fasta.aln 2PVF.PDB_A.pdb Output: ddr1_on_2pvf.pdb

4. Rebuild missing density

Step	Text	Commands	Comment
4. Rebuild missing density caused by gaps in the sequence alignment.	Any missing density and variable loop regions were constructed using the ab initio cyclic coordinate descent protocol in Rosetta. Top 50 models by score were selected and used for iterative docking step which was used for selecting the final homology models used for docking.	Generate loops file: In this case, the loop definitions will span regions where gaps were located in the sequence alignment. List the residue numbers in the loop file as shown in 2pvf.loops. Generate options file: List the desired options for rebuilding loop regions in an options file as shown in ccd.options. Run loop building: rosetta_source/bin/loopmodel.linuxgccrelease @ccd.options –loops:input_pdb ddr1_on_2pvf.pdb –loop:loop_file 2pvf.loops –out:pdb_gz -database rosetta_database Get top 50 models by score	Input: ccd.options, ddr1_on_2pvf.pdb 2pvf.loops, aaddr1A03_05.200_v1_3, aaddr1A09_05.200_v1_3 (fragment files not supplied due to their size) Output: 500 models of ddr1 from 2pvf template with missing density rebuilt: Top 50 models by score were selected for the next step.

5. Hierarchical docking protocol to select homology models

Step	Text	Commands	Comment
5A. Generate input files necessary for docking with Rosetta Scripts	Dock native dasatinib into the ddr1 DFG-in conformation to select models. Multiple docking iterations were performed with smaller ligand perturbations in subsequent rounds.	Prepare input pdb file: Align the top models obtained in step 4 to PDB:2GQG using pymol. Prepare options file: List the desired options for docking in an options file as shown in dock.options Prepare XML file for docking: List the desired specifications for docking in an options file as shown in round1_dock.xml	Input: Top 50 models from the previous step Output: models aligned to PDB:2GQG eg: 18_2pvf_0003.pdb

5B. Dock native ligand conformation (1N1 from PDB: 2GQG for active conformation) into models (Round 1)	Ligand was allowed to sample pocket in a 2 Å radius from the crystallized binding pose. After a rigid body orientation of the ligand centroid is performed through translation and 1000 cycles of 180 degree rotation, varying conformations of the ligand were tested within the site. During high resolution refinement, six cycles of side-chain rotamer sampling around the ligand were coupled with 0.1 angstrom, 0.05 radian ligand movements simultaneously in a Monte Carlo simulated annealing algorithm. A final minimization combines side-chain rotamer sampling with backbone torsion angle minimization with harmonic constraints on the C-alpha atoms.	Dock ligand generating 2000 models for each of the input homology model: rosetta_source/bin/rosettascripts.linuxgccrelease @dock.options –database rosetta_database –in:file:s "$MODEL 1N1_0001.pdb" –in:file:extra_res_fa 1N1.params –out:pdb_gz –out:nstruct 2000 –relax:thorough –parser:protocol round1_dock.xml Filter for the top ten percent of models by interface energy: Ligand interface energy score was used getting top 10% of models.	Input: Each of the 50 models from the previous step(each separately supplied to command at $MODEL), dock.options, round1_dock.xml Output: 2000 models each for each model in for example : 5_74_2pvf_0001_1N1_0001_0348.pdb
5C. Cluster models by ligand RMSD	To select for comparative models that can recuperate the native binding pose. RMSD of docked poses was calculated w.r.t native ligand pose and 10% of top-scoring poses that were within 1Å of the native were chosen.	Get top docked models and cluster models based on their similarity to the native pose. Download the bcl software suite at (the license is free for non-commercial users). http://www.meilerlab.org/index.php/bclcommons/show/b_apps_id/12 Bcl is required for running the script that does all the analysis. The script is not part of rosetta and is provided with the supplementary information cluster_poses.sh –p <$PATH to directory containing models> -t <number of models desired for clustering> -n –c 1.0 –b <$PATH to bcl executible>	Input: Models from previous step (5B) script - cluster_poses.sh Output: Top 10% of docked poses eg:1_18_2pvf_0003_1N1_0001_0367.pdb
5D. Round2 docking followed by cluster analysis	Ligand was allowed to sample pocket in a 0.6 Å radius from the crystallized binding pose. After a rigid body orientation of the ligand centroid is performed through translation and 1000 cycles of 60 degree rotation.	Docking rosetta_source/bin/rosettascripts.linuxgccrelease @dock.options –database rosetta_database –in:file:s "$MODEL 1N1_0001.pdb" –in:file:extra_res_fa 1N1.params	Input: Output models obtained from the previous round Output: Top 10% of docked poses

	RMSD of docked poses was calculated w.r.t native ligand pose and 10% of top-scoring poses that were within 0.3Å of the native were chosen.	–out:pdb_gz –out:nstruct 2000 –relax:thorough –parser:protocol round2_dock.xml Analysis cluster_poses.sh –p <$PATH to directory containing models> -t <number of models desired for clustering> -n –c 0.3 –b <$PATH to bcl executible>	eg:1_1_25_2pvf_0006_1N1_0001_0022_0015.pdb.gz

6. Dock imatinib and dasatanib into homology models

Step	Text	Commands	Comment
6A Generate ligand conformations in MOE	Ligand conformations were generated by MOE (Molecular Operating Environment, Chemical Computing Group, Ontario, Canada) with the MMFF94x force field and Generalized Born solvation model. Energy cutoffs for ligand conformations were dependent on the number of rotatable bonds: 3 kcal/mol for 1-6 rotatable bonds, 5 kcal/mol for 7-9 rotatable bonds and 7 kcal/mol for 10-12 rotatable bonds (Perola and Charifson, 2004) Ligand names dasatanib (1N1) imatinib (STI)	Generate ligand conformations in MOE: See MOE operating guide. Stochastic search with the MMFFx94 force field and Generalized Born solvation model was used to generate conformations within the specified energy cutoff. The ligand conformations were then saved as 1N1.sdf file for conversion to .pdb and .params files for Rosetta. Convert .sdf file of ligand conformations to .pdb and .params file for Rosetta input: rosetta_source/src/python/apps/public/molfile_to_params.py -n 1N1 -p 1N1 1N1.sdf Combine all individual ligand conformations in pdb format to a file called 1N1_confs.pdb. Add the line “PDB_ROTAMERS 1N1_confs.pdb” to the bottom of the 1N1.params file.	Input: ligand coordinates in mol format: 1N1.sdf and STI.sdf Output: 1N1.params, 1N1_confs.pdb STI.params, STI_confs.pdb
6B. Generate input files necessary for docking with Rosetta Scripts.	Ligand was allowed to sample docking poses in a 5 Å radius from the crystallized binding pose. After a rigid body orientation of the ligand centroid is performed through translation and 1000 cycles of 360 degree rotation, varying conformations of the ligand were tested within the site. During high resolution refinement, six cycles of side-	Prepare input pdb files: Top 1% of models output from step 5D were used for docking. Prepare options file for docking: List the desired options for docking in an options file as shown in dock.options. Prepare XML file for docking: List the desired specifications for docking in an options file as shown in dock.xml.	Input: comparative modes from step 5D Output: Top models from among all the starting templates – ddr1_HM1_in.pdb, ddr1_HM2_in .pdb, ddr1_HM3_in.pdb, ddr1_HM4_in.pdb,

	chain rotamer sampling around the ligand were coupled with 0.1 angstrom, 0.05 radian ligand movements simultaneously in a Monte Carlo simulated annealing algorithm. A final minimization combines side-chain rotamer sampling with backbone torsion angle minimization with harmonic constraints on the C-alpha atoms.		ddr1_HM5_in.pdb ddr1_HM1_out.pdb, ddr1_HM2_ out.pdb, ddr1_HM3_out.pdb, ddr1_HM4_out.pdb dock.options, dock.xml
6C. Dock dasatanib (1N1) and imatinib (STI) into DDR1-active and DDR1-inactive comparative models.	For each ligands 5,000 docked complexes were generated.	rosetta_source/bin/rosettascripts.linuxgccrelease @dock.options -database rosetta_database	Input: 1N1.pdb, 1N1.params, dock.xml, dock.options Output: 10% of models were plotted and are shown in figure S3

Protocol Capture – Ligand based vHTS

QSAR models were developed using BCL::CHEMINFO using DDR1 active molecules reported in PUBCHEM and inactives from dataset AID 2689 which contains molecules screened against Serine-Threonine kinases. An iterative approach was used where feedback from experimental studies was used to update computational models. Four such rounds of computational screening were performed followed by experimental testing. In the first round QSAR model was developed using only the molecules reported in PUBCHEM. The Vanderbilt virtual screening compound library (VICB library) was screened to prioritize 10 molecules. No hits were found during experimental screening and these molecules were fed back into the QSAR models. Round two QSAR model was developed using feedback from round1. Computational screening of VICB library was performed to prioritize 50 molecules. Experimental testing found that two compounds had mild inhibitory activity. Round three QSAR model was updated with two actives and 48 inactive molecules identified from round two experimental screening. New molecules reported in the literature were added to the dataset and a third QSAR model were developed. This model was used to predict and prioritize 50 molecules for testing from the eMolecules database.

This protocol provides a step by step process of QSAR modelling approaches used to train QSAR models and using them to predict activity for molecules in screening libraries. QSAR models were built using BCL::CHEMINFO which is available at http://www.meilerlab.org and is free for academic use. All the input files are provided in directory Thesis_directory/Chapter4/Ligand_vHTS/, henceforth abbreviated as $PATH.

Step	Text	Commands	Comment
1. File locations used for running protocol capture	Four round of QSAR training and predictions were carried out. Directories associated with each round are located under the parent directory $PATH. These directories contain files necessary to perform experiments. The $PATH/bin folder contains the BCL executable.	Download the BCL::CONF executable at http://www.meilerlab.org and put it in the bin folder. Get bcl_license.txt file and put it in the bin folder.	
2. Datasets	Each directory contains two raw files containing raw dataset – inactives.sdf actives.sdf VICB library – $PATH/vicb_library_indexed.sdf.gz eMolecules – $PATH/emolecules_indexed.sdf.gz		
3. Preparing datasets	Cleaning molecules and generating 3D conformations using CORINA. Use script $PATH/molecule_pipeline_light.sh	$PATH/molecule_pipeline_light.sh <DATASET.sdf>	Output: inactive_clean.sdf.gz active_clean.sdf.gz
4. Adding property to molecules	Adding 1 or 0 under property name “IsActive” to indicate active or inactive respectively.	bcl.exe molecule:Properties –input_filenames inactive.sdf.gz -add ‘Constant(0)’ –rename ‘Constant(0)’ “IsActive” –output <OUTPUT> bcl.exe molecule:Properties –input_filenames active.sdf.gz -add ‘Constant(1)’ –rename ‘Constant(1)’ “IsActive” –output <OUTPUT>	Rename files so that at the end we have files inactive_clean.sdf.gz active_clean.sdf.gz with property strings.

5. Generate feature files	Generate feature files from clean actives and inactive file. Script – $PATH/generate_sdf_dataset.sh	$PATH/generate_sdf_dataset.sh <INPUT.sdf>	Input: inactive_clean.sdf.gz active_clean.sdf.gz Output: Inactive_clean.bin, active_clean.bin
6. Randomize datasets	Combine and randomize datasets. Script – $PATH/combine_randomize_dataset.sh	$PATH/combine_randomize_dataset.sh active_clean.bin inactive_clean.bin	Input: Inactive_clean.bin, active_clean.bin Output: actives_inactives.randomized.bin
7. Train QSAR models	Train QSAR models	/home/kothiwsk/workspace_molecule/bcl/scripts/machine_learning/launch.py -t cross_validation --config-file $PATH/config.best.ini --id ddr1 --datasets actives_inactives.randomized.bin --max-minutes 600 --cutoff 0.5 --pbs --objective-function 'AucRocCurve(cutoff=%(cutoff)s,parity=%(parity)s,x_axis_log=1,min fpr=0.001,max fpr=0.1)	Input: $PATH/config.best.ini actives_inactives.randomized.bin Output: Models are located at $PATH/round*/models Results are located at $PATH/round*/results Log files are located at $PATH/round*/log_files
8. Predictions	Predictions activities of molecules in a library of molecule	bcl.exe GenerateDataset -source 'SdfFile(filename=<Input>)' -feature_labels $PATH/code_input_prediction.obj -result_labels "Combine(0)" -scheduler PThread 24 -output predictions.csv	Input: $PATH/vicb_library_indexed.sdf.gz or $PATH/emolecules_indexed.sdf.gz $PATH/code_input_prediction.obj (modify file according to change ROUND_NUMBER. Modified files are present in round* directories) Output: predictions.csv
9. Retrieving molecules with high prediction from dataset.	Sort molecules based on activity value and chooseN= n*2 molecules, where n is the desired number of molecules to be screened experimentally. Retrieve molecules	Sorting and choosing top N predicted molecules sort --field-separator=',' --key=3 predictions.csv |tail –N | awk –F, '{print $2}' > indices.txt	Input predictions.csv, screening library, retrieval indices extracted from indices.txt Output: predicted_active_topN.sdf

		$PATH/RetrieveMoleculesByIndex.py <INPUT> <OUTPUT> indices	
10. Cluster molecules to choose diverse a diverse set.	Cluster molecules to choose diverse molecules that should be tested. copy $PATH/clustering directory in “round” directory Output will be dendogram.py and different scaffolds in the cluster_sdf directory.	cp predicted_active_topN.sdf clustering/ cd clustering clean_up_molecules.sh predicted_active_topN.sdf cluster_molecules.py –m predicted_active_topN_clean.sdf.gz –l 0.3 –s 0.1 –c 5 cluster_scaffold.sh cd cluster_sdf	Input: predicted_active_topN.sdf Output: scaffold*.sdf files in cluster_sdf directory that contain unique scaffolds identified at sampling factor (-s) of 0.1

Chapter 5

Appendix Table 3 Molecule wise comparison of predictive ability of models developed by Subramanium et al. and developed in this study. The active molecules are indicated at cutoff value of KD = 3μM.

Inhibitor type	Molecule	Subramanium et al (3μ models)							3 μM QSAR model						
		ACC	SEN	SPE	TP	FP	TN	FN	ACC	SEN	SPE	TP	FP	TN	FN
Type-1	Dasatinib	81	44	96	36	9	193	45	78	47	89	48	30	247	54
	Erlotinib	85	19	97	8	8	233	34	93	NaN	93	0	25	354	0
	Gefitinib	80	52	82	11	47	215	10	79	66	81	29	65	270	15
	LY-333531	77	43	83	18	42	199	24	77	63	80	46	60	246	27
	Roscovitne	96	0	99	0	2	271	10	-	-	-	-	-	-	-
	SB-203580	89	10	99	3	3	249	28	88	18	94	6	19	326	28
	Staurosporine	46	39	97	96	1	33	153	83	87	54	289	21	25	44
	VX-680	64	3	100	3	0	178	102	68	76	64	100	88	159	32
	VX-745	97	20	100	2	0	273	8	73	89	73	8	100	270	1
Type-2	BIRB-796	83	10	98	5	5	230	43	67	84	64	42	118	211	8
	Flavopiridol	78	2	97	1	7	221	54	71	78	69	75	89	194	21
	Imatinib	94	68	96	13	11	253	6	86	68	87	15	46	311	7
	Lapatinib	100	100	100	3	0	280	0	84	67	84	4	60	313	2
	Sorafenib	84	35	96	19	9	219	36	78	81	77	55	72	239	13
	Sunitinib	63	76	45	124	66	54	39	78	80	75	181	38	115	45

Protocol capture

Protocol capture is a step by step guide to the process of building kinase selectivity models. All the files are kept in the zipped directory , Thesis/Chapter5 here on called $PATH. This $PATH contains three directories – a) bin b) config c) input. The config directory contains all the scripts that are required for building models and for data analysis. The input directory contains all the input files that are created during the study. The models are not provided due to the size they occupy on hard disk (each model occupies ~2.3 G). However, the models can be easily created by following the directions given below. Kinase selectivity QSAR models were built using BCL::CHEMINFO which is available at http://www.meilerlab.org and is free for academic use.

Step	Text	Commands	Comment
1. File locations used for running protocol capture	The file containing kinase list is present in input directory – all_list.txt The file contains six digit Uniprot alphanumeric codes of	Download the BCL::Conf executable at http://www.meilerlab.org and put it in the bin folder.	

	kinases reported in the Zarrinkar dataset.	Get bcl_license.txt file and put it in the bin folder.	
2. Generate QSAR selectivity models.	Run script build_selectivity_model.sh from the $PATH/config directory. It will generate selectivity model at the specified cutoff in $PATH/input directory	Run the command from the input directory /bin/bash $PATH/build_selectivity_model.sh <cutoff value>	Output: Generates a directory containing QSAR model for kinase selectivity for kinases listed in *_list.txt at the desired cutoff value $PATH/input/all_$cutoff
3. Calculating area under the curve for each kinase, overall area under the curve for a model (Figure 4A) and heat maps (Figure 5)	Run script calculate_auc.sh from the $PATH/input directory	Run the command from the input directory /bin/bash $PATH/calculate_auc.sh <cutoff>	Output: Files containing AUC value for each kinase of a particular kinase family is output - area_under_curve.txt Figure 5-4A - overall_auc.png Figure 5-5 – A)heatmap_experimental.png B)heatmap_predictions.png C)heatmaps_absdifference.png