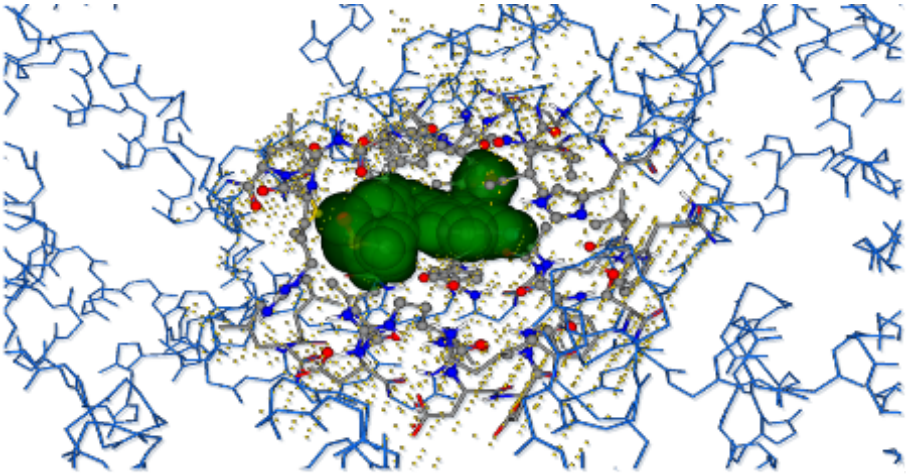
Protein-Ligand Inverse Screening and its Application in Biotechnology and Pharmacology



Dissertation
with the aim of achieving the degree

Dr. rer. nat.

at the Faculty of Mathematics, Computer Science and Natural Sciences
Universität Hamburg

submitted to the
Department of Computer Sciences
of Universität Hamburg

Karen T. Schomburg

born in Braunschweig

Hamburg, May 2014

Gutachter:
Prof. Dr. Matthias Rarey
Prof. Dr. Norbert Ritter
Prof. Dr. Jürgen Pleiss

Tag der Disputation: 14. Juli 2014

Abstract

The thesis at hand presents the development of a new computational target prediction method. Small molecules are rarely only binding to a single protein, but can interact with numerous proteins, their targets. Ignorance of molecule-protein interactions can lead to various problems, wherein the most dangerous probably are side-effects evoked by drugs binding to so-far unknown off-targets. Resolving these problems is the aim of target prediction methods, which try to find all target proteins for small molecules.

The *i*RAISE method developed in this thesis faces the special requirements of structure-based inverse screening which in comparison to normal screening (one protein, many ligands) predicts targets for one small molecule from large protein libraries.

In order to account for the large amounts of protein structural data, *i*RAISE introduces a database representation for efficient and consistent handling and storing of protein data. Further, protein active sites and small molecules are in the first screening step abstracted by a descriptor representation. The chosen descriptor contains features encoding the interaction pattern and the shape of the active site/molecule. Thus, by matching complementary descriptors, the need for sequential protein-ligand matching on atomic level is avoided. Inter-target ranking has been improved compared to standard protein-ligand scoring functions, which mostly contain a bias towards certain protein structures. A multi-step Scoring Cascade considers the reference ligand as well as the coverage of the ligand and pocket, and thus allows the scoring of structurally diverse pockets. Moreover, a Gaussian-based scoring assesses the significance of a score.

Along with the new target prediction method, an evaluation strategy with new data sets has been developed. The evaluation of binding mode prediction, target ranking and running time shows promising results for *i*RAISE.

Further, structure-based computational methods were successfully applied in a biotechnological case study of the development of a synthetic multi-enzyme pathway, highlighting the application potentials of these methods in areas besides drug design where their use is already established.

Kurzfassung

Die vorliegende Arbeit präsentiert die Entwicklung einer neuen, computergestützen Methode zur Zielproteinvorhersage. Kleine organische Moleküle binden nur selten an ein einziges Protein, sondern interagieren mit einer Reihe von verschiedenen Proteinen, sogenannten Zielproteinen oder *Targets*. Dies kann zu verschiedenen Problemen führen, wobei vermutlich die verheerendsten die Nebenwirkungen von Medikamenten sind, welche durch Bindung von Wirkstoffen an Off-Targets entstehen. Aus diesem Grund versuchen Target-Vorhersagemethoden alle Zielproteine für kleine Moleküle zu identifizieren.

Die in dieser Arbeit entwickelte Methode *i*RAISE behandelt die speziellen Anforderungen an strukturbasiertes inverses Screening, bei welchem im Gegensatz zum normalen virtuellen Screening (ein Protein, viele Moleküle) potentielle Targets für ein Molekül in großen Mengen von Proteinen gesucht werden.

Die großen Datenmengen von zehntausenden von Proteinstrukturen werden in *i*RAISE in einer neu entwickelten Datenbank effizient und konsistent gespeichert. Außerdem werden sowohl Protein- als auch Ligandstrukturen im ersten Screening-Schritt durch einen Deskriptor abstrahiert. Der gewählte Deskriptor repräsentiert sowohl das Interaktionsmuster als auch die Form von Protein und Ligand. Indem die Deskriptoren auf Komplementarität getestet werden, erübrigt es sich, jeden Ligand sequenziell gegen jedes Protein auf atomarer Basis zu testen. Des Weiteren wurde das Ranking von Targets durch innovative Maßnahmen verbessert, um einen Bias bezüglich bestimmter Proteinstrukturen zu beheben, der häufig bei Bewertungsfunktionen für 'normales' Protein-Ligand Docking besteht. Die sogenannte Scoring Cascade besteht aus mehreren Bewertungsstufen, welche neben der Berücksichtigung des Referenzliganden die Abdeckung des Liganden und des aktiven Zentrums betrachtet. Dies ermöglicht die Bewertung von strukturell diverser Taschen ohne Bias. Zusätzlich wurde eine Gauß-basierte Bewertung eingeführt, welche die Beurteilung der statistischen Signifikanz eines Scores erlaubt.

Zusammen mit der inversen Screening-Methode wurden eine neue Evaluierungsstrategie mit Evaluierungsdatensätzen entwickelt. Insgesamt erzielt *i*RAISE auf Basis von Bindemodusvorhersagen, Target-Ranking und Laufzeit vielversprechende Ergebnisse.

Darüber hinaus wurden im Rahmen dieser Arbeit strukturbasierte, computergestützte Methoden in einem biotechnologischen Projekt erfolgreich eingesetzt, in welchem ein synthetischer Multienzym-Stoffwechselpfad entwickelt wurde. Dies hebt die Potentiale der Methode in einem weiteren Anwendungsfeld neben dem Feld der Wirkstoffentwicklung, in welchem die Nutzung dieser Methoden schon etabliert ist, hervor.

Danksagung

An dieser Stelle möchte ich mich für alle Unterstützung bedanken, welche ich in den Jahren als Doktorandin erhalten habe.
Als allererstes möchte ich mich bei Prof. Dr. Matthias Rarey bedanken, der mir einerseits ermöglichte, dieses spannende Thema zu bearbeiten und darüber hinaus jederzeit für Diskussionen und fachlichen Rat zur Verfügung stand. Die überaus angenehme Arbeitsatmosphäre in seiner Arbeitsgruppe und die Möglichkeit, mein Projekt auf internationalen Konferenzen zu präsentieren sorgten für viel Spaß und Motivation bei der Arbeit an diesem Doktorarbeitsprojekt.
Mein Projekt konnte sich auf viele von Jochen Schlosser und Ingo Schellhammer entwickelte Prinzipien des Dreiecksdeskriptors stützen. Aus diesem Grund möchte ich Ihnen ebenfalls danken.
Besonderer Dank gebührt auch den Projektpartnern und Unterstützern des LEXI-SynBio Projektes, welche einen Teil dieser Arbeit erst ermöglichten. Spezieller Dank gilt hier Fabian Rieckenberg, Inés Ardao und Katharina Götz für die tolle Kooperation. Nadine Schneider und dem Team der BioSolveIT ist hier zu danken für die Entwicklung und Bereitstellung der HYDE-Bewertungsfunktion und der LeadIT-Docking Suite.
Meinen derzeitigen und ehemaligen Kollegen am ZBH danke ich für viel Unterstützung, rege Diskussionen und die beste Arbeitsatmosphäre die man sich vorstellen kann. Besonders möchte ich meinen Bürokollegen danken, Axel Griewel für seine geduldige Unterstützung zu Beginn meines Projektes und Stefan Bietz für eine tolle Zeit. Melanie Geringhoff und Janna Eich danke ich für die perfekte Organisation aller Verwaltungsangelegenheiten.
Für das tapfere und zuverlässige Korrekturlesen fast jeder dieser Seiten danke ich Janna Eich und meiner Mutter Ida.
Meinen Freunde waren in allen Promotionsjahren immer für mich da, und halfen mir hervorragend beim Abschalten.
Meiner Familie danke ich für die allwährende Unterstützung, ihr Verständnis und ihren Glauben an mich. Meinen Eltern danke ich besonders für die Gewissheit, dass sie immer für mich da sind und wären, egal was passiert. Meiner Mutter danke ich dafür, dass sie fast jedes meiner publizierten Worte Korrektur gelesen hat und meinem Vater für seine Hilfe bei allen technischen Problemen. Meiner Schwester Annika danke ich besonders für ihre

Unterstützung, ihren Rat und ihre Art, irgendwie alles in richtige Licht rücken zu können. Du bist für mich ein großes Vorbild und neben meiner Schwester eine ebenso liebe Freundin. Als letztes möchte ich Martin danken, der das Pech hatte, mich im letzten Jahr meiner Promotionszeit kennen zu lernen. Trotzdem hat er es tapfer mit mir durchgestanden und hat mir nicht zuletzt mit leckersten Mahlzeiten durch alle Problemphasen geholfen. Danke.

Contents

List of Abbreviations

2D	Two Dimensional
3D	Three Dimensional
Å	Ångström
***i*RAISE**	Inverse RApid Index-based Screening Engine
ADME-TOX	Pharmacokinetics: Absorption, Distribution, Metabolism, Excretion and TOXicity
AUC	Area Under the Curve
BEDROC	Boltzman-Enhanced Discrimination of ROC
EC	Enzyme Commission
EF	Enrichment Factor
FDA	US Food and Drug Administration
NMR	Nuclear Magnetic Resonance
NSLR	Normalized Sum of Logarithmic Ranks
PDB	Protein Data Bank
RAISE	RApid Index-based Screening Engine
RMSD	Root Mean Square Deviation
ROC	Receiver Operator Characteristic
sc-PDB	SCreening Protein Data Bank
TP	True Positive
VDW	Van-der-Waals
VS	Virtual Screening

1

Introduction

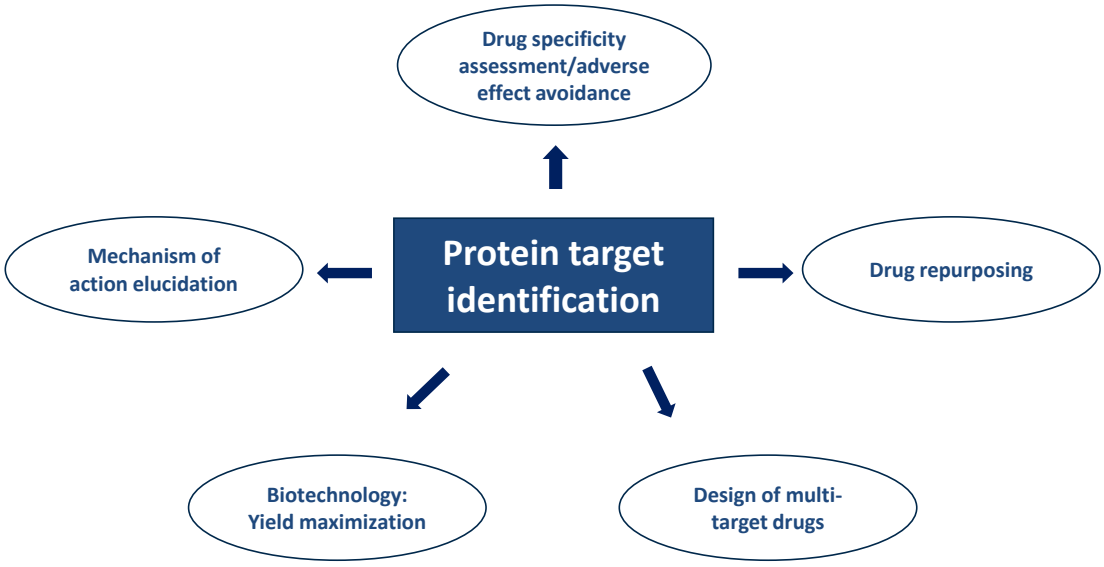


Figure 1.1: Application fields of protein target identification

1.1 Motivation

Life is dependent on proteins. The malfunction of a single protein can lead to disease, disability or death. However, rationally manipulating protein functionality can also cure diseases, alleviate disability and prevent death. Drugs mostly tackle with the functionality of one or several proteins to show their effects. As a consequence, in drug design the knowledge of target proteins and the understanding of their function is crucial for the rational development of new pharmaceutically active compounds.
Proteins are functional units in living organisms, responsible for biological processes like signal transduction, catalysis, molecule transport and gene replication while also having structural functions e. g. building of organic material (e.g. keratine, a protein responsible for the structure of the hair). The 3D structure of the protein determines its function, while the structure of the protein is determined by the sequence of amino acids that it is composed of. Amino acids are the building blocks of proteins, consisting of a carboxy and an amine group and a specific side-chain. The sequence of amino acids is coded in the genes of an organism which encode 20 different natural amino acids.
In order to obtain a certain function, proteins interact with other molecules. These may be other proteins, nucleic acids or small molecules, which bind to the protein. Medicinal chemists exploit this effect by specifically designing small molecules as ligands for a protein target of interest which then alter the natural function of the protein as these ligands compete with the natural ligands for binding to the protein. Small molecules bind to proteins in specific pockets, often located partially buried within the structure of the protein.
A ligand can only bind to the so-called active site or pocket of a protein if the pocket and ligand are chemically complimentary. This means that the ligand not only needs to fit sterically into the binding pocket of the protein, but also needs to form an energetically favorable interaction pattern with the amino acids of the active site. The concept of the chemical complementarity is also called lock-and-key concept, as introduced by Fischer (Fischer [1894]). As the term conveys, there is some specificity involved: Not every key can interact with every lock. Accordingly, specific ligands bind to one protein. The specificity lies in the interaction pattern between ligand and protein.
The interactions responsible for a binding of a small molecule are mainly of non-covalent nature. Hydrogen bonds, ionic interactions and the hydrophobic effect are forces that contribute most to the reversible binding. Only rarely binds a ligand covalently to the protein. In general, a molecule binds to a protein in a biological environment if the interactions between the molecule and the protein are energetically more favorable than the interactions between water or solution molecules filling the binding pockets and the protein. The energy gain of

the protein-ligand binding also has to compensate for the entropic loss by fixation of degrees of freedom of the molecule and the protein (Bissantz et al. [2010]). The strength of the binding of a molecule to a protein varies, depending on the interactions built.
The effects of such a binding are diverse: A small molecule can inhibit or activate a protein. In case of an enzymatic reaction, the protein can change the structure of the small molecule or even change with the binding of a molecule its own tertiary structure, e.g., a membrane protein can open a channel in its midst.
Nature is not very consistent with the selective lock-and-key principle. Rarely does a small molecule bind only to one specific protein just as rarely as a protein interacts only with one ligand. A study examining published in-vitro binding assay data from drug discovery projects even concludes an average of 964 reported active compounds per target (Southan et al. [2011]), while a drug-target network analysis finds an average number of 6 targets per approved drug (Vogt and Mestres [2010]).
The reasons why there is promiscuity in nature are manifold, based on evolution, efficiency and safety. During evolution, mutations of the amino acid sequence of the protein occurred. If the binding of a molecule is too specific, then even minor structural changes of the binding pocket would disrupt the binding. Other reasons why compounds are promiscuous in nature is the catalytic activity of enzymes. A compound may be the substrate for several enzymes with divers catalytic activity, thus the compound may be for one enzyme the building block for a larger compound and another enzyme breaks it down into smaller fragments.
Also, the chemical reaction that is catalyzed by an enzyme is not limited to one special molecule, but many molecules that are chemically similar can function as substrates for the reaction. Thus some enzymes are not selective in their substrates, but can convert many similarly composed molecules (Nobeli et al. [2009]). This way, there does not have to be one enzyme for each substrate, but rather one enzyme for a class of substrates. However, this means that the active pocket of the enzymes has to be capable of holding slightly different molecules. As a consequence, sometimes compounds bind to enzymes although they are no substrates for the catalyzed reaction, but remain in the active site and block it for other compounds, causing inhibition. One example is the compound 6-phospho-D-gluconate, which inhibits the enzyme phosphoglucose isomerase (EC number 5.3.1.9), but is the substrate of the enzyme 6-phosphogluconic dehydrogenase (EC number 1.1.1.44) (Schomburg et al. [2013]).
Furthermore, during evolution, some pathways developed in parallel to guarantee a rescue path in order to account for the blocking of one pathway by disease or a mutation, or react to a change in the nutrient supply. This resulted in compound promiscuity towards different proteins which are capable of fulfilling the same function. Finally, another reason for

promiscuity in nature is efficiency: Often one compound is used for several purposes, e.g., activating one target and inhibiting another. This is needed due to the complex conditions that an organism has to react to, e.g., on environmental impacts, where some processes need to be triggered simultaneously while others need to be inhibited. Using one messenger compound for both tasks is more efficient in that case.

Next to the question of the reasons for promiscuity, also the question how molecular promiscuity is realized on the level of structure and interactions arises. As the lock-and-key principle discussed above states, protein binding pockets are composed like locks and compounds are keys. But how can one lock fit different keys? The answer to this is the composition of binding pockets and the flexibility of the protein and the ligand. Active sites can have very specific binding regions where hydrophilic amino acids are dominating. Here, the ligand has to fit in the sense of providing the matching interaction partners. However, there also can be regions in the binding pockets which are hydrophobic. Here, the ligands which are binding can be very diverse in structure as long as they fit sterically and have hydrophobic regions to place there. Finally, also the binding pocket does not have to be occupied completely by ligands. Therefore, ligands with different size can fit as well.

Next to the composition of the binding pocket, also flexibility of the protein and ligand plays a role in promiscuity. The protein can adjust the binding pocket to the ligand with rotation of amino acid moieties or protein backbone movements, which lead to larger structural changes. Gatti-Lafranconi and Hollfelder find that next to the choice of cofactor or metal involved in the catalysis, the conformational changes of the enzyme are mainly responsible for substrate promiscuity (Gatti-Lafranconi and Hollfelder [2013]). The ligand can change its conformation on binding as well. Therefore, the protein and ligand can adjust to each other, which is a principle known as induced fit (Koshland Jr [1958]). Sturm et al. (Sturm et al. [2012]) evaluate explicitly the structural basis for ligand promiscuity in a study on ligands that are found to bind to multiple proteins. They conclude that, firstly, binding sites of different proteins that accommodate the same ligand can be highly similar in structure, even if the proteins are not related in sequence but also, secondly, that in dissimilar binding sites the ligand binds in various conformations, adjusting to the proteins. Both studies emphasize the role which flexibility plays in promiscuity of protein-ligand complexes.

Having discussed how and why there is promiscuity of a ligand or target in nature, the question arises how these effects can help or hinder natural science like development of drugs or biotechnological process design. Studying which molecules bind to one target is of interest in the design of inhibitors as drugs for disease-related targets or in finding substrate efficiently processed by enzymes in biotechnological questions. However, there are also scientific questions, where the identification of all targets for one molecule is crucial.

The following scientific applications would profit from a reliable target identification method for small organic molecules:

Identification of targets...

- ...next to a primary target for side-effect identification of drugs
- ...for a compound with an effect on an organism, where the mechanism of action is unknown
- ...for a known drug in drug repurposing projects
- ...for the design of multi-target drugs
- ...for substrates in the design of new uses of enzymes in synthetic biotechnological multi-enzyme pathways for maximizing yield
- ...in selectivity studies, where a compound shall be active on a part of a target class only, e. g., only on kinases expressed in certain tissues or on targets of special organisms.

Identifying further targets of a drug is helpful in early design steps of the drug design process such as lead optimization. In the last decades, the number of new drugs introduced to the market does not reflect the rise in investment on the design of new drugs (Kennedy [1997]). One problem of the development process is that many leads are developed with much care and effort but fail in the last steps of the process, the clinical phases. Adverse effects and toxicity of the agents in humans are revealed in these phases. Roughly 10 percent of all projects fail due to adverse effects in humans and roughly 11 percent fail due to animal toxicity. Added together, a total of 21 percent fail due to unwanted and beforehand unknown side-effects (Kennedy [1997]). A study of productivity challenges in the pharmaceutical industry reveals a rise of this number from 11% (percent of reasons for clinical failures) in 1991 to 22% in 2000 (Khanna [2012]).
An early assessment of the specificity would reduce the number of failures in the late development steps and save much money and effort. A study of the strategies that were used to identify potential drug candidates between 1999 and 2008 (Swinney and Anthony [2011]) shows that although in this period the major focus of methods lay on target-based screenings, only 17 drugs were found by this approach, while 28 were found by phenotype screening which was actually applied in fewer projects. This reveals a major drawback of target-based approaches: The concentration on only one target of interest leads to high attrition rates

due to adverse effects. Adding target identification to target-based drug development for early identification of unwanted side-effects shows promise to lower these high attrition rates.

The study of drugs with unknown mechanisms of action can lead to useful insights for the development of new drugs or for the optimization of these drugs. An area of interest in this question is Chinese medicine, where often a phenotypic effect is known, but the mechanism on target level is not clear. The molecular entities responsible for the phenotypic effect are of high interest for pharmaceutical companies since the Chinese medicine extracts have already been applied and tested for a long period of time and, therefore, promise few and non-harmful side effects. One approach of computational target identification for compounds extracted from herbal Chinese medicines is published by Fauzi et al. (Mohd Fauzi et al. [2013]), who in detail look at the targets of extracts from Panax Ginseng. Zhang et al. use a combination of a computational target-identification approach and enzymatic assays to determine the targets of natural products of Chinese Medicine used to treat diabetes and inflammation (Zhang et al. [2011]). Supporting studies like these, Chen compiled a database of ingredients with compound info for traditional Chinese medicines especially for the purpose of drug screening (Chen [2011]).
Next to finding the targets for natural traditional medicines, even today not all mechanisms of action for approved drugs are known in detail. Querying the DrugBank (Wishart et al. [2006]) for drugs with "Mechanism of Action: Unknown" yields 80 hits of a total of about 1500 approved drugs (Date of query: September 2013).

Drug repurposing or drug repositioning is the use of drugs in another disease context than the one where they were developed and approved for (Ashburn and Thor [2004]). The most known case of successful drug repositioning is sildenafil, which activity as an agent for erectile dysfunction was discovered only accidentally while it was primarily of interest in a study concerning blood pressure. In the last years, the number of literature published on drug repurposing exploded, showing the high interest in this field (Andronis et al. [2011], Xu and Cote [2011], Moriaud et al. [2011], Swamidass [2011], Loging et al. [2012], Haupt and Schroeder [2011], Dudley et al. [2011], Muthyala [2012], Reaume [2012], Smith [2012], Oprea et al. [2012]). The approach of drug repositioning has huge advantages, mainly that the drugs already passed several toxicity and selectivity tests, even clinical tests and are less risky to fail due to toxicity issues. Much time and money can be saved, which is of interest especially in so-called orphan disease projects (Ekins et al. [2011]). Orphan diseases affect only a small part of the population and are, therefore, not of great interest to pharmaceutical companies, as the profits of investments into drug development are not expected to be high.

Already, there are successful applications: Azathioprine as drug for rheumatoid arthritis was repurposed for renal transplant, Bleomycin, drug for various cancers was repurposed for pleural effusion, Clycloserine as drug for urinary tract infection was repurposed for Tuberculosis ...among many other examples (FDA [2013]).
Computational methods addressing drug repurposing cover approaches such as literature mining (Andronis et al. [2011], Loging et al. [2012]), target structure based methods like comparison of the 3D structure of protein binding sites (Moriaud et al. [2011], Haupt and Schroeder [2011]) as well as chemical similarity based methods (Dudley et al. [2011]). Recently, databases containing drugs which are interesting for repurposing have been composed. The NCGC Pharmaceutical Collection by the NIH Chemical Genomics Center is a collection of clinically approved drugs in electronic and physical form (Huang et al. [2011a]), created with the purpose of repurposing. The Rare Disease Repurposing Database by Xu et al (Xu and Cote [2011]) matches drugs approved by the FDA (FDA [2013]) to the orphan designation database, which defines diseases with orphan status (these diseases have special conditions concerning tax and market launch). These efforts lead to a rational approach to drug repurposing, which in the past was dominated by serendipity or clinical observation (Liu et al. [2012]).

Multi-target drugs are compounds that affect several targets of one disease. While classical drugs that are designed with regard to one target are called magic bullets, the term magic shotguns has been used for drugs designed for multiple targets (Roth et al. [2004]). A multi-target approach has two main advantages: Firstly, the drug is more effective against the disease if several targets are affected at the same time and, furthermore, the risk of drug resistance is lowered, since if one target mutates successfully, others are still inhibited by the drug. For some diseases, e.g., nervous system disorders, the classical one-drug, one-target approach fails since too many factors are involved (Roth et al. [2004]). However, the design of multi-target drugs is complex, as the designing steps for the drug have to be addressed to all targets that shall be affected. Balancing affinities of one compound for several targets is highly challenging.
Jenwitheesuk et al. (Jenwitheesuk et al. [2008]) published an approach, where they screened computationally drug-like compounds against all targets known to be connected to Malaria (Plasmodium falciparum) and HIV-1 infections. The authors found several compounds which were active against multiple targets. They verified six compounds to be inhibitory on several targets of Plasmodium falciparum experimentally. Bottegoni et al. (Bottegoni et al. [2012]) propose a fragment-based method as a rational approach to computational multi-target design. Next to first methods approaching the design of multi-target drugs rationally,

Medina-Franco et al. state that in the past, many drugs -although initially designed for one target- in reality exhibited their effect by affecting several targets (Medina-Franco et al. [2013]).

In biotechnology, in the last decades, the design of synthetic enzymatic pathways started to show promising results (Cho et al. [2010], Li et al. [2004], Wu et al. [2011], Santacoloma et al. [2010]). These pathways consist of up to dozens of enzymes which transform a substrate to a more profitable product. Here, the challenge lies in finding enzymes for each degradation step that function under the same conditions, and here the secondary activities of enzymes are often exploited, meaning that enzymes are used for reactions for which they were not primarily known. Finding the enzyme that reaches the highest yield among a list of enzymes from various organisms is one challenge. Further challenges in the setup are feedback and buffer inhibitions, which can be avoided by an early identification of all targets for each intermediate in the synthetic pathway.

As the preceding passages show, the reasons for identification of all targets for one compound are manifold. Various names evolved for methods that predict targets for one compound: Target-fishing, off-target prediction, polypharmacology prediction, ligand promiscuity identification, reverse screening, inverse screening. However, all names describe methods which have one aim: Rationally exploiting the advantages of promiscuous compounds and predicting the disadvantages in a reliable manner. In the past, multi-target drugs or repositioned drugs merely occurred due to serendipity. Finding such effects with reason will certainly be a major issue in drug design in the following decades.

1.2 Overview of content

This section shortly summarizes the content of this thesis.
Firstly, the approaches how the scientific field currently identifies or predicts the targets of a ligand are discussed in detail in chapter 2. The top level abstraction of these methods is experimental or computational. As will be discussed, there are many different computational approaches each with its advantages and disadvantages. The description of the state of the art elucidates current challenges and unresolved problems in target prediction, which are the issues that shall be studied in the research project of this thesis. Also, the task of target identification is placed in the huge field of computational structure-based drug design.
Following the state of the art description is a listing and discussion of the aims and objectives

of this thesis in Chapter 3. The second part of Chapter 3 describes the technical conditions of this thesis.
Then, in Chapter 4, the computational methods of the in this thesis developed inverse Screening software *i*RAISE are described.
Chapter 5 focuses on data used in this thesis. For validation and method development, different data sets are needed: Small, diverse data sets during method development and large data sets for statistical performance evaluation. In Chapter 5, it is described which data sets are used for which purpose and where the development of new validation data sets was necessary and how these were composed.
Following is the description and discussion of the evaluation strategy based on the evaluation experiments and performance metrics in Chapter 6.
The results of the target prediction of *i*RAISE on the previously discussed experiments are given in Chapter 7. The overall performance is shown and example cases are discussed to show where *i*RAISE has potentials and limitations.
In Chapter 8, a case study of structure-based cheminformatic methods on biotechnological applications is given. Questions and problems which can be studied with computational methods are elucidated and in an cooperation project, the methods are applied and evaluated experimentally.
Finally, in Chapter 9, the accomplishments of this thesis are summed up, open problems and its applicability is discussed. An outlook on following improvements is given as well.

2

State of the Art

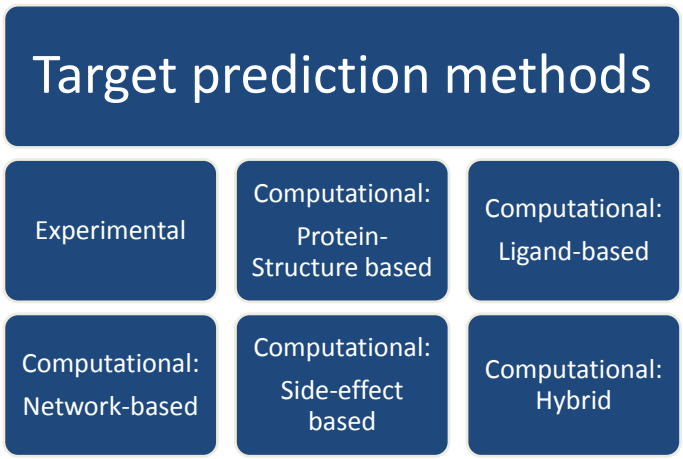


Figure 2.1: Overview of methods available for protein target identification

2.1 Computational structure-based methods

The support that computational structure based cheminformatic methods can supply is currently well-used in drug discovery projects. Whereas in other fields like biotechnology and biochemistry, first applications started to make use of these methods only recently, in the drug discovery pipeline (shown in Figure 2.2), cheminformatic methods are well established. The pipeline consists of the following steps:

- **Target-identification**
  Drugs interact with proteins. Therefore, the target protein has to be identified at the beginning of each drug discovery project, which is mostly done by experiments. Computational methods are rarely used in this step so far. Once a target is identified, it has to be validated whether its modulation changes the disease state and activity assays have to be established. Not each protein is easily crystallizable but if it is possible, a 3D structure of the protein is determined by X-ray crystallization, NMR structure analysis or homology modeling.

- **Lead-identification**
  Once the protein target is identified, a lead, e.g. a substance that binds strongly to the active site of the protein is searched. This task can be done by looking for natural ligands, e.g. substrates of enzymes and understanding which functional groups are essential for binding, i.e., understanding the mechanism of action and the binding mode. Furthermore, if no ligands are yet known from which conclusions can be derived, experimental high throughput screenings can be carried out. If a 3D structure of the target protein is known, molecule libraries can be screened by virtual screening, or rational de-novo design can be conducted.

- **Lead-optimization**
  Once a lead is found, it has to be optimized with regard to administration, distribution, metabolism and toxicity (ADME/Tox properties). In this step, its selectivity is further studied and improved.

- **Clinical trials**
  In animal studies and clinical trials with phases I to III, adverse effects and the efficacy of the proposed drug are studied.

Often, the use of computational methods during the drug-discovery pipeline depends on the available data, determining which methods are applicable. The use of structure-based methods like the inverse screening approach presented in this thesis is reserved to projects where

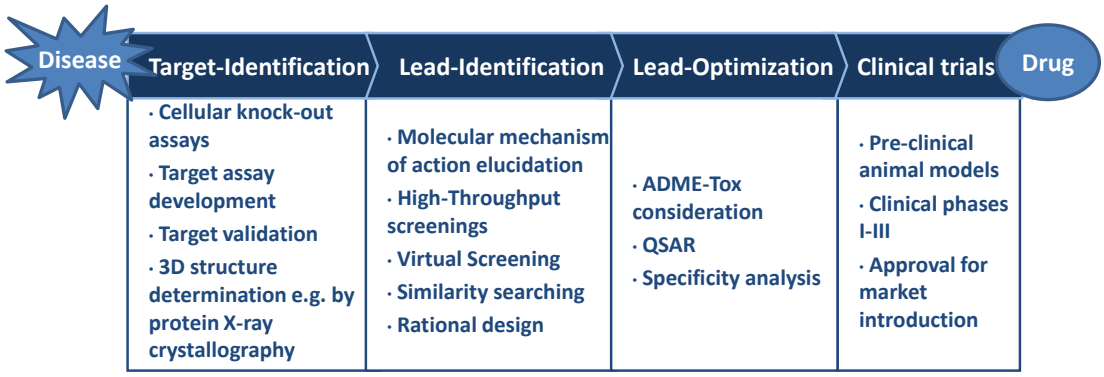


Figure 2.2: Steps of the drug discovery pipeline

the 3D structure of the protein is known and available at a sufficient resolution. If this is the case, structure-based methods can support the drug design pipeline during each step: Elucidation of the binding mode by co-crystallized ligands, analysis of the active site, comparison of the active site to those of other proteins, lead identification by virtual screening, lead optimization by protein-ligand docking, de-novo design, pharmacophore-based approaches and many others.
Inverse virtual screening can be applied mainly in the following steps:
For target-identification: If a drug has an unknown mechanism of action, and for example its efficacy shall be enhanced, first of all, its target needs to be identified. An off-target identification of a drug can also turn it into a lead for another drug discovery project, which is the step lead-identification. In the step lead-optimization the selectivity of the lead can be evaluated by inverse screening through identification of further targets. This can significantly reduce the number of unspecific drugs that go into the last and most expensive step of the drug pipeline, the clinical trials.

As the term 'inverse virtual screening' conveys, it means the reverse setup compared to normal virtual screening, where libraries of millions of substances are screened for one protein target. In structure-based approaches, docking-based virtual screening dominates nowadays. The term docking describes the method of predicting the binding mode of a ligand in the active site of a protein. It consists of the two steps of the placement of a ligand into the active site of a protein followed by a scoring function predicting the binding affinity. While the first docking methods were published in the 1980ies and fast virtual screening applications in the 1990ies (there are numerous recommendable reviews covering the development of docking and virtual screening: Schneider and Böhm [2002], Pujadas et al. [2008], Kitchen et al. [2004], Taylor et al. [2002], Yuriev et al. [2011], Tanrikulu et al. [2013]), the method

of structure-based inverse screening is a decade younger, with the first methods being developed at the beginning of 2000 (see section 2.2.2).
Although the methods used in protein-ligand docking and virtual screening have been developed and further improved for over three decades now, there are still many open challenges. These are found mainly in the aspects of protein flexibility, evaluation of crystal structures, dependence of the methods on expert knowledge of the user for a reliable outcome, ability of the scoring functions to predict binding affinities, and the treatment of active site water (Waszkowycz et al. [2011]). This listing shows that during the development, early problems like ligand flexibility have already been addressed. This is not the case in inverse virtual screening. Being still in its very early development stages, the challenges have not been so clearly defined. So far, the first methods show open problems but very few are yet addressed by improved follow-on approaches. While the knowledge gathered during the docking and virtual screening development phases can also be of use in inverse screening, the open challenges of this field are adopted as well, adding to the new challenges of the reverse setup.

2.1.1 Screening versus docking

For the 'normal' setup of identifying a ligand for a protein, the difference between protein-ligand docking and virtual screening lies in the timescale that is needed to generate a protein-ligand pose. In docking, where one or some molecules are placed into the active site of a protein, the task can take minutes per ligand, while in virtual screening, where up to millions of compounds are screened, the time per ligand should be reduced to seconds or milliseconds (Waszkowycz et al. [2011]). This means that virtual screening methods employ algorithms where the processing of one ligand is accelerated by some abstraction. Until now, this difference is not employed in inverse virtual screening/computational target identification methods. However, as there are currently not many structure-based approaches published, this categorization has probably not been necessary. While the term inverse docking has been used in some publications (Chen and Zhi [2001]), others use the term inverse screening although their methods are based on simple protein-ligand docking and do not employ measures for abstraction and speed-up as in 'normal' virtual screening. As nearly none of the so far published structure-based computational target-identification methods employ measures for speed-up and abstraction, these methods are all described in section 2.2.2.4, without differentiation of inverse docking or inverse screening.

2.2 State of the art of inverse virtual screening

This section summarizes the current state of research concerning target identification, i.e., the methods which are available to determine the target profile for a compound. The aim of this chapter is to give an overview of the current situation in this field of research with current open problems and challenges.
First of all, it is elucidated how experimental methods in laboratories are used to identify the targets a compound binds to. Then, computational methods that are relevant in predicting polypharmacology at the time of writing this thesis are discussed. These methods are divided by type, meaning that ligand-based, network-based, side-effect based and structure-based methods are discussed separately. Representative literature for each method is discussed; however, the discussed literature does not contain each approach of each method, since this would go beyond the scope of this overview. Since the method presented in this thesis is based on the 3D structure of proteins, it is of the type structure-based. Therefore, the methods of this category are directly comparable with the presented method in this thesis. For this reason, the main part of this chapter is focused on methods of this type, which are consequently discussed in detail, while the other methods are mainly discussed in terms of (dis-)advantages in order to give an overview of their basic ideas. The discussion of the type structure-based methods is therefore subdivided into several subsections.

2.2.1 Experimental methods

Experimental methods to determine the targets that a compound or drug has an effect on (activating or inhibiting) are mainly isolated activity assays, cellular screening and pharmacogenetic profiling.
Cell-free activity assays are used for high throughput molecular-target screening. These activity measurements monitor the activity change in the absence or presence of a compound of (ideally) a single protein.
Next to single activity assays, a cell-free screening approach for experimental target identification is affinity chromatography for proteins (Jenkins et al. [2007]). In this approach, the compound of interest is fixed to the stationary phase while the proteins of interest, e.g., in form of a cell extract, are in the mobile phase. This approach is limited concerning the modus of protein-ligand interaction, since if the ligand needs to migrate deeply into the protein into the active site, it would not bind if fixed to the solid phase.
Cellular screenings show the phenotypic effect on the cell level but are more complex and not suitable for high-throughput approaches. Screenings at the whole organism level, like

in vivo screening using, e.g., the zebrafish as model organism (Zimmermann et al. [2007]) are even higher dimensional. Despite the advantage of showing the phenotypic effect on the organism, whole organism screenings are even less suited for target identification on protein level.
So-called high-throughput screening where one target is experimentally screened with thousands of compounds is well established today. In this experimental setup, once an activity assay is available for a target, it is relatively easy to perform activity experiments with a huge number of compounds. However, the reverse setup poses many challenges, which is why there is no high-throughput inverse experimental screening. For this setup, an activity assay for each target that shall be screened needs to be developed. This task can be very difficult for some targets and even if activity assays are available, conducting thousands of different assays is highly complex.

In summary, even today with screening robots and other high-throughput methods, determining the selectivity of a compound experimentally is not trivial. Questions currently discussed in literature are: what is a good selectivity measure, how many assays are needed to determine if a compound is promiscuous and how are the assay results mapped to a numerical number that can be compared across compound assays (Wang and Greene [2012]). The fact that even today experimental methods are expensive and time consuming is also crucial. Furthermore, not each compound or protein is readily available for activity tests, but often have to be obtained lavishly. Another known problem of experimentally determined activity is the often problematic comparability of assay data measured in different laboratories under different conditions. Additionally, another issue is that there are not few compounds which in some way interfere with the activity measuring methods and lead to misinterpreted results due to aggregation or solvation effects (Baell and Holloway [2010]).

2.2.2 Computational methods

Computational methods for target prediction are either built on experimental data from which they derive predictions or use the nowadays established chemical knowledge of protein-ligand interactions. This section is divided into subsections on the basis of the computational model used. Since the results of the computational models have to be evaluated, good data sets are crucial for a reliable evaluation. Therefore, at the end of this section, the data used for evaluation is discussed as well.

2.2.2.1 Ligand-based computational methods

Ligand-based methods use known ligand-target binding information, describe the ligands by descriptors coding, e.g., the shape, topology and functionality and predict with similarity measures unknown ligand-target interactions (Bender and Glen [2004]). The growing amount of available data of ligand-target activities in databases like PubChem (Wang et al. [2009]), DrugBank (Wishart et al. [2006]), ChEMBL (Gaulton et al. [2012]) and others is exploited by these methods. The basic assumption is that similar ligands bind similar targets, whereby 'similar' is defined in various ways. An advantage is the independence from available 3D structure information of the target. Furthermore, descriptor calculations are relatively fast, meaning the results for predictions for a new ligand are rapidly available. A disadvantage is the dependency on the available data: if a ligand-target interaction so far has not been observed, it is not represented in the data and cannot be considered in the predictions. Furthermore, the reduction of the ligands to descriptors can result in misleading predictions.

Ligand descriptors can be divided into one-dimensional (1D), two-dimensional (2D) and three-dimensional (3D) ones. 1D descriptors code molecular properties like molecular weight, number of rotatable bonds, hydrophobicity and others. 2D descriptors describe the topological connectivity of the molecule, one example is the Daylight Fingerprints (James and Weininger [2006]). 3D descriptors consider the shape of the molecule.

Niijima et al. (Niijima et al. [2010]) use a machine-learning method to identify 1D features that allow a prediction of unknown ligand-target interactions. Applying their method to cytochrome P450 enzymes, they rank 1400 features like chemical paths, logP, charge, rings and bond information by their predictive potential. Keiser et al. use in their Similarity Ensemble Approach (SEA) (Keiser et al. [2007]) the 2D Daylight Fingerprint (James and Weininger [2006]) with the Tanimoto Coefficient as similarity measure. With their method, they are able to predict chemogenomic activity classes assigned with a statistical confidence value for new ligands. For evaluation of their method, they confirm in-vitro predictions experimentally (Keiser et al. [2009]).

Koutsoukas et al. use Extended Connectivity Fingerprint (ECFP) descriptors (Rogers and Hahn [2010]) for molecules. They propose a data set as benchmark set for further studies, consisting of about 80000 compounds extracted from the ChEMBL database that show activity against human targets. Based on this data set they compare a Kernel-based and a Naïve-Bayes based approach for prediction of targets for unknown ligands. Their results show for both methods good predictions with 80% true targets in the first 3% of the data. They state that the performance of their methods exhibit large variations depending on the target class (Koutsoukas et al. [2013]).

AbdulHameed et al. (AbdulHameed et al. [2012]) use a shape/chemistry overlap as 3D descriptor by applying the ROCS approach (Rush et al. [2005]) and show a good enrichment of true positives compared to random selections. Another 3D shape overlap method is the Gaussian Ensemble Screening method introduced by Perez et al. (Perez Nueno et al. [2012]).
Nettles et al. (Nettles et al. [2006]) critically examine the use of 2D and 3D descriptors for target-fishing and conclude that 2D descriptors outperform 3D descriptors in correct target prediction. However, in many cases, a complementary use of both descriptor forms is found to be worthwhile, especially if no high 2D similarity to the query compound is found in the available data. Such a complementary 2D/3D approach is applied by Kinnings et al. (Kinnings and Jackson [2011]) as well. They use 2D descriptors (Daylight Fingerprints) as a pre-filter to reduce the number of geometric matchings in the following step. Then, the similarity is calculated considering both, 2D and 3D descriptor-similarity, in combination. The authors also conclude that a complementary approach of both descriptor types is most successful.

2.2.2.2 Network-based computational methods

In polypharmacology, networks of a protein-ligand interaction space or a gene-ligand interaction space are compiled. This method links chemical structures to biological activity data. Using experimentally determined activity data from large screening assays or compilations of data from literature and databases, a network consists of targets or genes as nodes and interactions (e.g. activity with the same compound) as edges (Hopkins [2008]).
The argumentation for using networks for target prediction is as follows: The compiled networks show that there are alternative paths, so-called rescue pathways for all life-essential paths of an organism. Consequently, inhibiting only one node on one of the pathways will not have a phenotypic effect. Hopkins (Hopkins [2008]) argues that a *one-two punch* approach hitting several parallel pathways with polypharmacological drugs or drug combinations is needed for a reliable effect. Only networks reveal the parallel pathways, and show where a target promiscuity of a compound is wanted (when the targets lie on parallel pathways) and where it is unwanted (when the targets are on totally different pathways in the network).
Arguments against this method are similar to those described above for ligand-based methods: The prediction relies on the so far discovered data and predictions can only be derived from the available data. The compiled networks can be used, on the one hand, to identify promiscuous targets, and, on the other hand, to exploit the available data for new predictions as described below.
Networks compiled from experimentally determined activity data can be completed by us-

ing computational prediction methods. As experimental activity determination is expensive and challenging, the completion of the available data with predicted activities is necessary. Strategies for the prediction are diverse: often machine learning methods are trained with the existing data, classifying chemical structures with chemical fingerprints.
Paolini et al. (Paolini et al. [2006]) use data of several sources (in-house screening data, commercially available screening data and data from literature and patents) to compile a pharmacological space in form of a data warehouse. The interaction network compiled consists of 486 (human) targets with more than 3500 edges (interactions). With the network, an assessment of both, compound and target selectivity, is possible. Three methods to calculate a promiscuity index for the targets are introduced. One considers the number of connections in the network, a second the proportion of ligands shared with other targets weighted by the average number of targets the ligands are active against and a third considers the strength of a connection in the network (i.e., the height of an activity measured). Using the data of the network, Paolini et al. trained a Laplacian-modified Bayesian classifier to predict polypharmacology. Using part of the data for training and part as test set, their method shows promising results for new predictions. However, they state that their method is dependent on their data which unfortunately is sparse in some parts.
A ligand descriptor based method is applied by Mestres et al. to introduce their so-called in-silico target profiling. Using ligand-target activity data from seven different databases and in-house screening experiments, ligand activity is mapped with a low-dimensional molecular topology descriptor. In this way, a network of targets connected by ligand descriptors is compiled. For the prediction of the activity of a ligand with unknown activity, the Euclidian distance of this ligand's descriptor to all descriptors in the network is calculated. The network reveals groups of target and cross-pharmacological connections when, e.g., all targets are connected that have ten or more shared ligand descriptors. Mestres et al. show with their target-drug networks, that at the time of writing, the average number of targets per drug is 6. (Gregori-Puigjane and Mestres [2008], Mestres et al. [2009], Nonell-Canals and Mestres [2011]).
Yamanishi et al. (Yamanishi et al. [2010]) connect the chemical space, the pharmacological space and drug-target interaction networks (data collected from various databases) using a two-step approach of firstly using the chemical structure to predict pharmacological effects and secondly predicting drug-target interactions with a supervised bipartite graph inference method. The bipartite graph consists of heterogeneous nodes of drugs or targets and of interactions as edges. Their results show that for their predictions using the pharmacological effects outperforms using chemical similarity alone.

2.2.2.3 Side-effect based computational methods

The effect of a drug on a target, e.g., an activity inhibition does not necessarily lead to the wanted phenotypic effect in the organism. Therefore, in polypharmacology, a significant challenge is not only to predict a potential target, but also to predict a potential phenotypic pharmacological effect such as *abnormal hepatic function* or *cardiac dysrhythmia* for a drug. The number of methods so far available for side-effect prediction is limited, but some methods exploit known side-effects for target predictions. These methods depend on a side-effect description as in package inserts of drugs, and are therefore limited to the data available from marketed drugs.

As pioneers of this method, Campillos et al. predict from side-effects extracted from the package inserts of about 750 drugs, whether two drugs share a target (Campillos et al. [2008]). Applying a weighting scheme, they account for different occurrence frequency of the side-effects as well as for the correlation of effects. The probability of two drugs to share a target is calculated as a combination of a chemical similarity of the drugs (based on a 2D fingerprint) and the similarity of the side-effect listings.

Using the SEA approach (side-effect approach) introduced in the section ligand-based methods (see section 2.2.2.1), Lounkine et al. (Lounkine et al. [2012]) map adverse effects of drugs to targets and are therefore able to predict potential targets together with side-effects for an unknown drug via its chemical similarity.

Combining side-effect prediction with their previously published network-based pharmacological effect prediction method (see above, discussed in section 2.2.2.2), Mizutani et al. (Mizutani et al. [2012], Yamanishi et al. [2012]) correlate drug-protein interaction profiles with side-effect profiles. Constructing a network of proteins and side-effects connected by edges, they are able to predict the phenotypic side effects from a protein interaction profile of a drug.

These methods are useful concerning the prediction of phenotypic pharmacological effects but are limited concerning predictions for new drugs, since they rely on the sparse available data. The so far available methods are not able to make predictions if the drug of interest is very different in structure or in target profile from the drugs that are in the available data set. Another issue of the approach to predict targets via side-effects is the relatively abstract description of phenotypic effects in package inserts, which makes it difficult to map effects to distinct targets. Further, mapping an expressed phenotype to a single target for protein target prediction on target level is hardly possible.

2.2.2.4 Protein-structure based computational methods

In contrast to the above discussed methods, the advantage of protein-structure based methods is that they do not necessarily depend on already available activity data. For example, the methods based on docking use the chemical knowledge of protein-ligand binding for predictions. However, the name protein-structure-based instantly reveals the main obstacle of these methods: They rely on the availability of 3D protein structures. These are determined experimentally by X-ray crystallography or NMR nuclear magnet resonance structure determination as well as computationally by modeling. Being the most established and in comparison with modeled structures more reliable method, many approaches only use structures determined by X-ray crystallography. The public database Protein Data Bank (Berman et al. [2000]), available at www.pdb.org, contains all published 3D protein structures. Additionally, many companies have their own internal collection of 3D structures which are not available to the public. However, the number of publicly available 3D protein structures grows rapidly, with a rise from about 20000 X-ray structures available in the PDB in 2003 to more than 40000 structures in 2008 to more than 82000 structures available in 2013.

The category of structure based methods can be subdivided into protein-ligand docking based, pharmacophore-screening based and binding-site comparison based methods (Rognan [2010]).

Methods based on protein-ligand docking

In protein-ligand docking, small molecules are placed in protein binding pockets; positions where the ligand can bind with reversible interactions to the protein are predicted. Docking is applied traditionally in protein-ligand screening, where a library of small molecules is docked to one protein in order to find molecules which bind to this protein. Often, docking and screening are differentiated, with screening being defined as more coarse but much faster than docking.

For target identification, the inverse method is used: In protein-ligand docking based inverse screening methods, a docking algorithm is employed to place the one query compound into a database of 3D protein structure binding pockets. For identification of binding pockets, either reference ligands bound in protein-ligand complexes or pocket prediction algorithms are used. Molecule poses generated by the docking algorithms are evaluated with diverse scoring functions which measure the steric fitting of the ligand in the binding pocket and the complementarity of possible chemical interactions between protein and ligand. Ranking of the poses by the values of the scoring function gives a ranking of targets for the query compound. For evaluation of the methods, the results can be either verified by experimental

activity assays (prospectively) or if activity data is already available by comparison of the ranking positions of true targets for the query ligand and the activity data (retrospectively). Below, inverse screening approaches based on protein-ligand docking are described in detail. Most of them apply docking algorithms which are not especially adjusted to the *inverse* problem of target rather than ligand identification. Nevertheless, the inverse approach has other requirements than the normal method, which can result in rendering the traditional approaches of docking or scoring approaches inappropriate for the task of inverse screening. These are time issues because handling the data for a database of proteins is more challenging than handling the data for the same number of small molecules, as well as scoring issues, since the scoring functions used for traditional protein-ligand docking are often unsuitable for comparing predicted binding strength on the basis of docking scores among different targets (Kellenberger et al. [2008], Wang et al. [2012]). The discussion of the so far published approaches shows where adjustments crediting these issues are integrated as well as where these issues pose problems on the methods.

INVDOCK Chen et al. (Chen and Zhi [2001] and Chen and Ung [2001]) describe INVDOCK as one of the first structure based-inverse screening approaches. Their method is based on an own implementation of the DOCK protein-ligand docking program (Kuntz et al. [1982]). Based on geometric algorithms, an active site is represented as a cavity defined by overlapping spheres, with their own implementation of the method developed by Wang et al. (Wang et al. [1999]) accounting for ligand flexibility. Ligand poses generated by the docking algorithm are optimized with respect to the torsion angles applying an energy-based optimization strategy.

For the optimization as well as for the evaluation of the obtained poses, a scoring function based on protein-ligand interactions in the AMBER force field is used (Weiner et al. [1984]). The energy score of the docked pose is related to a threshold energy score, which is the number of ligand atoms weighted with a constant factor -α, which is close to 1.0 and is determined by linear regression of the computed energy score of a large set of co-crystallized protein ligand complexes. The score of the docked ligand has to be lower than the threshold score to select a structure as a potential target for the ligand. The authors propose to use more sophisticated weights for the threshold score but do not further evaluate this consideration.

Next to this threshold score, the computed score for the generated pose is also compared with the score of the ligand in the cavity or, if it is empty, with a ligand in a cavity of other related protein cavities. In order to save computation time, a cavity is considered a screening hit, if one pose is found above the threshold scores, and no further search for the

best binding mode is carried out.
In 2002, Chen et al. compiled a database of pharmacologically relevant protein-ligand complexes for inverse screening called TTD (Therapeutic Target Database) (Chen et al. [2002]). The authors searched literature for relevant therapeutic targets, stating that at the time of 2002, medical treatment addressed approximately 500 targets. The first published version of the TTD contained 433 targets, updates in 2010 (Zhu et al. [2010]) and 2012 (Zhu et al. [2012]) led to 932 targets in the currently available version. The targets contained in TTD are cross-linked to other databases and annotated with pathway information, known drugs, disease conditions and related literature and are presented via a web-server to public queries. Validation studies presented with the above described INVDOCK inverse screening engine are performed with a pre-version of TTD consisting of 1040 structures of 38 proteins (2700 cavities) connected with toxicity and side effects as stated by the authors.
In a first study, nine protein structures are screened in a re-docking approach, resulting in RMSD values between 3.65 and 6.55Å (Chen and Zhi [2001]). The relatively high upper limit of observed RMSDs is according to the authors due to the abortion of the search strategy at the first 'good' pose. If this aborting is turned off, and the best pose is searched, the RMSD values range from 0.94 to 2.41Å. For evaluation, INVDOCK is tested with vitamin E and 4H-tamoxifen. For 4H-tamoxifen, INVDOCK lists 20 potential targets, whereof 10 are confirmed in the literature as targets, and for 10 no information exists whether 4H-tamoxifen is binding or not. For vitamin E, two of the 25 predicted targets are confirmed, for four there is experimental evidence of some effect for vitamin E found in literature and for the rest there is no data of possible vitamin E binding.
In a following study (Chen and Ung [2001]), Chen et al. use eight compounds for screening (aspirin, gentamicin, ibuprofen, indinavir, neomycin, penicillin G, 4H-tamoxifen, vitamin C). The search is restricted to 1425 cavities of the database of human or mammalian proteins. The average running time of one compound is at the time of publishing 12 days on a 250 MHz SGI R10000 Octane workstation. The results show that the predictions for each compound contain confirmed and not confirmed targets. For all compounds together, only eight true targets of which there are 3D structures contained in the cavity database are missed. Of the predicted targets, 29 of 68 do not have any experimental data on binding of the compounds, thus cannot be classified as true or false targets.
A third study that applies the INVDOCK inverse screening approach is published by Ji et al. (Ji et al. [2006]) and covers the screening of 11 anti-HIV drugs. The results show that 86-89% of the by INVDOCK identified potential targets are consistent with adverse drug effects reported on the drugs. 67-100% of the reported adverse effects are covered by the predicted targets.

The INVDOCK approach as the first published inverse screening approach reveals the problems of docking based inverse screening without adjustments to the "inverse" problem: The computing time of docking into the targets is relatively high, although in the evaluation experiments only a small number of targets are screened. The computing time saving measures which were taken have the disadvantage of significantly worsening the results. Also, the evaluation of the method is difficult with the data that Chen et al. used since no clear separation of true and negative targets was available.

TarFisDock A unique feature of the TarFisDock approach published by Li et al. (Li et al. [2006]) is its web-platform, which allows any scientist to use the application. The method of Jiang et al. is based on the protein-ligand docking program DOCK (Version 4.0: Ewing et al. [2001]). The scoring function of protein-ligand interactions of the Amber force field is used (Weiner et al. [1984]) to rank the generated docking poses. Neither the docking procedure of DOCK, nor the scoring function are specially adjusted to the challenges of the inverse screening problem.
The web-platform does not only give access to the TarFisDock screening approach, but also provides a protein database which can be screened: Gao et al. compiled a database of 3D protein structures named PDTD (Potential Drug Target Database) (Gao et al. [2008]). Consisting at the time of publishing of 1100 protein entries, the database was compiled of targets gathered from TTD (Chen et al. [2002], DrugBank (Wishart et al. [2006]) and Thomson Pharma (www.thomson-pharma.com), following the aim to compile a database of potential drug targets. The 1100 protein entries represent about 830 potential drug targets with 3D structures. The entries in the database contain active binding sites, annotation of related diseases, biological function and associated pathways of the contained targets.
For evaluation of the TarFisDock approach, Li et al. (Li et al. [2006]) screen the PDTD data set with vitamin E and 4H-tamoxifen and are able to identify 30% and 50% of reported targets, respectively, where for 4H-tamoxifen, 50% of true targets are ranked in the top 5% of the database. In another application, Cai et al. (Cai et al. [2006]) use the TarFisDock program to screen the PDTD database with anti-Helicobacter pylori agents to identify potential targets. They confirm the results with experimental assays and protein crystal structure determination of two discovered targets. The running time depends on the flexibility of the screened compound, ranging from 5 to 20 hours on a single CPU at the time of publishing.
Since no adjustments were implemented to the traditional protein-ligand docking algorithms or scoring functions, this approach is faced with the issues described above of inverse screen-

ing.

sc-PDB Inverse Screening Rognan et al. published several papers (Paul et al. [2004]), (Muller et al. [2006]), (Kellenberger et al. [2008]) on an inverse screening approach which focuses on the compilation on a sensible 3D structure database of proteins for screening. Kellenberger et al. (Kellenberger et al. [2006]) published the first version of the so-called sc-PDB in 2006, and subsequently updated the database, with the currently latest version published in 2011 (Meslamani et al. [2011]). The data set of protein structures contained in the sc-PDB consists of 'druggable' binding sites extracted out of the Protein Data Bank after the following criteria: resolution, type of structure and consistency of annotated data. The binding sites which are represented in the database are annotated with EC number, source organism, name, and cofactors. Only binding sites containing ligands with pharmacological potential are selected for the database, meaning that Kellenberger et al. exclude binding sites with cofactors, sugar-like ligands and crystallization or solution agents as ligands. In the first version, their database consisted of 6415 binding sites, which increased up to 8166 in the version published 2011.
The first inverse screening approach of the sc-PDB is described in Paul et al. [2004], at which time a pre-version of the sc-PDB consisted of 2148 structure entries. Their inverse screening method uses an approach based on the GOLD docking procedure (Verdonk et al. [2003]), using the virtual screening settings, without any further adjustments to the inverse screening setup. Their method used 64 cpu-hours per ligand. For validation, the method is applied for the prediction of targets for four chemically diverse ligands (biotin, 4-hydroxytamoxifen, 6-hydroxy-1,6-dihydropurine ribonucleoside, methotrexate). They compare RMSD values of the binding mode predicted by GOLD to co-crystallized ligands with the result of an average of 0.6 Å. Furthermore, they evaluate at which position the GOLD fitness function ranks known targets for the four ligands, which results in good enrichment rates. However, poor enrichment rates are obtained for the ligand AMP (adenosine 5'-monophosphate), which leads the authors to state that inverse screening should be "reserved to rather selective ligands".
A subsequent target identification study (Muller et al. [2006]) also used the sc-PDB as structure data set. As query five molecules with a 1,3,5-triazepan-2,6-dione scaffold as representatives of a combinatorial library were used. To identify true targets for these molecules, rules were composed based on the GOLD fitness score and the fact that proteins are stored in multiple entities in the sc-PDB data bank: 50% of all target entries had to be among the top 2% of all entries, a minimum of two entries of the same target had to be in the top 2% of all entries and the average fitness score of all entries of the same target

had to be above the value 50. Evaluating their method, experimental inhibition assays of nine enzymes showed mixed results, with three not showing any inhibition, one showing ambiguous results and one assay confirming the prediction of an inhibition.
A following publication (Kellenberger et al. [2008]) addresses the ranking of true targets problem by using the GOLD fitness score and the same four ligands as they were used in Paul et al. [2004]. They find that a combination of the GOLD fitness score with two topological molecular interaction fingerprint (IFP) scores leads to the best enrichment.
Another example where the sc-PDB database was screened inversely by a protein-ligand docking based approach was published by Zahler et al. (Zahler et al. [2007]). The docking tool GlamDock (Tietze and Apostolakis [2007]) was used to generate poses, with its scoring function followed by an energy minimization of the ligand torsion angles of the docked conformation. Zahler et al. aimed to identify targets for four kinase inhibitors. Next to known targets, they identified further structures of the sc-PDB as potential targets and conducted kinase activity assays for evaluation. Besides to finding known targets in the database with an enrichment factor in the first 1% between 16 and 20 for the four inhibitors, one kinase was predicted to be inhibited selectively by one of the four compounds. For this kinase, experimental activity assays confirmed the predictions.

Methods based on pharmacophore-screening

Pharmacophore screening based methods use a known cavity of the compound of interest to derive a 3D pharmacophore describing the chemistry and geometry of the cavity by feature points. With this pharmacophore, databases of cavities, e.g. from the Protein Data Bank, are screened.
As one example, the approach of Campagne-Slater et al. (Campagna-Slater et al. [2010]) is discussed here. They use an algorithm that predicts pockets in human proteins in the Protein Data Bank to compile a cavity database independent from co-crystallized ligands. First of all, the cavity database is filtered for cavities containing the same amino acid types as the query cavity. Then, pharmacophores are used to filter for cavities which also have the correct geometric arrangement of the amino acids. The pharmacophores are derived from known binding pockets, e.g. in case of the in the paper discussed study several methyl-lysine binding cavities are used to compile sets of pharmacophores. Therefore, in this approach, the ligand and known binding pockets are used to derive pharmacophores to search only those cavities from the database, which already have comparable amino acid compositions. This reveals one disadvantage of pharmacophore-screening based methods: The possible hits are limited to very similar binding pockets. Nevertheless, one ligand can bind to very diverse binding pockets considering amino acid composition. These would be missed by the

approach of Campange-Slater et al (Campagna-Slater et al. [2010]).
More prominent in pharmacophore screening approaches is to screen the molecule against a pre-compiled database of pharmacophores of protein binding sites and compare the ligand to these pharmacophores, as shown by Steindl et al. (Steindl et al. [2006]) or Liu et al. with the PharmMapper server (Liu et al. [2010]). PharmMapper is available via a web server and the in-house database of over 7000 pharmacophore-models can be screened within a few hours. The ligand is placed on the protein pharmacophore models and pharmacophore features are compared. A scoring of the feature concurrence allows a ranking of matching pharmocophores, e.g., a ranking of matching protein structures. As an evaluation, tamoxifen was screened against the pharmacophore database. In the top ranked 100 proteins, four are known as true targets. 71% of the known targets appear among the top 300 ranked proteins. Here, as only pharmacophoric feature points are compared, the above mentioned problem is not encountered, diverse pockets can be found with this method. However, the reduction to pharmacophoric points is an abstraction which may lead to false positives.
Meslamani et al. (Meslamani et al. [2012]) compare four different pharmacophore-matching approaches to ligand-similarity and docking-based target prediction. In the applied pharmacophore generation step, only those pharmacophores which are derived form the protein-ligand complex are selected which promise to be selective, based on a statistical estimate. For evaluation, the sc-PDB described above is used as protein structure database. 157 diverse ligands of the reference ligands contained in the sc-PDB are used as queries. The study concludes that ligand-based similarity based on ECFP (2D) and ROCS (3D) outperforms pharmacophore- and docking-based methods. However, their study also reveals cases for which only the pharmacophore or docking-based approaches were successful.

Methods based on binding-site similarity

Another 3D structure-based method is based on the assumption that similar binding pockets bind similar ligands. It is presumed that based on a known binding pocket of a ligand other pockets where the ligand binds can be identified by finding similar pockets of the known binding pocket(s). This method tends to miss potential targets for a ligand since the ligand can change conformation on binding and, thus, the binding sites where it binds to need not necessarily be very similar in shape or interaction spot arrangement.
Exemplarily, the approach of Miletti and Vulpetti (Milletti and Vulpetti [2010]) is discussed here. Their identification of targets for a compound based on pocket similarity consists of the two steps of creating a database of pocket descriptors and then screening the database with descriptors generated of the pocket. Miletti and Vulpetti use a fingerprint consisting of atom types in spheres with increasing radius to describe the pockets. A similarity score

of the descriptors is completed with a score for the superposition of the two binding sites to obtain a ranking of pockets from the database for the query pocket. For evaluation, Miletti and Vulpetti used screening data for kinases from the Ambit panel (Karaman et al. [2008]). For 17 kinase inhibitors, they obtained ROC AUC values between 0.5 and 0.9 with a median of 0.63 (ROC and AUC values are described in detail in Chapter 6).

2.2.2.5 Hybrid computational methods

Combining several of the above described computational methods results in hybrid approaches. Such protocols have the advantage that potential limitations of one method might be straightened out by other methods. In real applications, the method used is dependent on constitution and amount of available data. Hybrid approaches take over the choice of method from a user and combine scores originating from various approaches in one. A disadvantage is that the results are difficult to interpret, since the composition and meaning of a final score is often not clear.

One example that shall be discussed here, is the approach of Simon et al. (Simon et al. [2011], Peragovics et al. [2012]), which combines side-effect analysis with protein-ligand docking to predict new targets for a query compound. In their method called drug effect profiling, Simon et al. firstly compile an interaction pattern matrix consisting of docking scores of each drug entry from the DrugBank (Wishart et al. [2006]) to 149 non-target proteins. Secondly, a so-called binary effect profile matrix is compiled, which lists for each drug from the DrugBank whether a pharmacological effect of a list of 177 is observed or not. These two matrices are combined by canonical correlation analysis and linear discriminant analysis into the effect probability matrix (Simon et al. [2011]). This matrix gives a probability for a drug to have a pharmacological effect.

Next to the prediction of therapeutic effects, Simon et al. extended their method by exchanging the effect profile matrix with a target profile matrix, where one row contains the target profile of a drug as documented in the DrugBank. The statistical combination of both matrices into the target probability matrix gives probabilities for each drug of binding to each of 77 proteins of the target profile matrix (Peragovics et al. [2012]). Of the newly predicted drug-target interactions, the authors focused on antipsychotic drugs for evaluation and searched the literature for conformations of the predictions. Of 84 drug-target interactions that were not recorded in the DrugBank, 39 were found to be reported in literature.

Another example is the approach of Tian et al. (Tian et al. [2013]) who combine inverse docking with a pharmacophore mapping in a Bayesian classifier. They find that their classifier is satisfactory able to classify inhibitors of diabetes type II targets. Subsequently, they apply their method for predicting targets for compounds found in traditional chinese medicine.

The hybrid approach introduced by Meslamani et al. (Meslamani et al. [2013]) automatically combines a series of four ligand and two structure-based target prediction methods. A workflow contains decision conditions for choosing the best method based on the input data. For a ligand test set of 189 ligands, the primary target was identified for 72%.

2.2.3 Data

For the evaluation of predictions resulting from computational target prediction methods, data showing the promiscuity of ligands is needed. As computational methods are merely simplified models of the real situation in nature, evaluation shows to which extend the model is able to mirror the effects in nature. Therefore, as a validation set, data from experimental assays which show the binding affinity of a ligand to as many different targets as possible would be desired. Then, the predictions of a computational model can be divided into true positives, i.e., correct positive predictions, true negatives, i.e., correct negative predictions, and false positives and false negatives. However, so far, there is not much data of this kind available. Often, in experimental setups, one target is screened with thousands of compounds, while the reverse setup where thousands of targets are screened with one compound is not performed due to the problems discussed in section 2.2.1.

Most difficult to obtain is negative data in this scenario, where it is documented for a compound that it is not active on a target. This data is often not published, because if active compounds are wanted in a project only the active compounds found are published. Therefore, the identification of true negative and false positive predictions of computational methods is a hardly solved issue. Furthermore, if an activity of a compound on a target is not documented in literature, it cannot be concluded that the compound is inactive but only that there has not been an activity test of the compound on that target. With this lack of data, the validation of computational methods is difficult.

There are few target classes for which the desired data is available. One target class which is evaluated thoroughly in terms of positive and negative activity of compounds, namely the selectivity of the tested compounds, are kinases. Kinases are a class of enzymes with EC number 2.7.1.* which transfer phosphate-moieties to substrates. They are involved in signal-transduction and are over-expressed in cancer tissues, which is the reason for the interest in selective kinase inhibitors. In the above discussed approaches, Miletti and Vulpetti (Milletti and Vulpetti [2010]) and Zahler et al. (Zahler et al. [2007]) use kinase data sets for evaluation.

Table 2.1 shows an overview of data used in the computational structure-based target prediction approaches discussed above in section 2.2.2.4. The overview demonstrates the lack of negative data points in most of the used data sets.

2.3 Unsolved challenges

The discussion of the state of the art of structure-based inverse screening methods revealed three main challenges:
The first is the adjustment of docking methods to the inverse screening setup. Many methods that use docking in the reverse setup for target identification do not adjust the algorithms in particular for this problem. These methods are more similar to reverse docking than reverse virtual screening. Therefore, for an inverse screening method, many opportunities exist which could speed up the process.
The second main challenge is the inter-target ranking. Scoring functions for protein-ligand complex assessment are in general not suited for comparison of different targets. The absolute values of the scores are only capable to rank several ligands in one target. This is due to the diverse features and structures of each active site and due to the fact that the scoring functions were not designed for this problem yet.
The final challenge lies in the data for validation, as discussed in the preceding section. Unlike the DUD (Huang et al. [2006]), which is an evaluation data set for molecular docking already in the second generation, for inverse screening so far no standard data set which allows statistical evaluation of the performance of target prediction tools is established. While even for ligand-based inverse screening predictions a benchmark data set has been proposed in 2013 (Koutsoukas et al. [2013]), no such data set exists for structure-based inverse screening methods. Therefore, no thorough statistical comparison of different methods is possible. In addition, further improvement of existing methods proves difficult if the only possible validation of predictions is by experimental activity assays. There is a lack of a data set with diverse, high quality protein-ligand complexes and negative data points. A recent review of Kharkar et al. on reverse docking uses the words "Unlike the benchmarks for traditional molecular docking [..], there is scarcity of such sets for reverse docking." (Kharkar et al. [2014]) They conclude that due to the lack of benchmarks, the effectiveness of strategies for multiple target identification is unclear.

Data set	Composition	True positives	True negatives	Number of structures	Availability	Application in inverse screening
sc-PDB (2012 version) [1]	'Druggable' protein-ligand complexes from the PDB meeting quality criteria	yes	no	8077 entries (2377 proteins)	Free access	Screened with 4 chemically diverse compounds [2],[3]
TTD (2012 version) [4]	Therapeutic targets, compiled by search from textbooks, and publications	yes	no	2025 targets	Via web-server	Pre-version of 1040 entries screened with 8 compounds [5],[6]
PDTD [7]	Drug targets from TTD, DrugBank and Thomson Pharma, structures from PDB	yes	no	1207 entries (841 proteins)	Free access	Screened with 2 compounds [8]
Kinase data [9]	PDB structures with Kinase activity data in the Kinase Ambit Panel	yes	yes	189 kinases	Activity data published	Binding site similarity study with 17 kinase inhibitors [9]
Kinase complexes of sc-PDB	Complexes with EC number 2.7.1.*	yes	yes	327 complexes, 84 kinases	yes	screened with 3 indirubin derivatives, true negatives by experimental validation [10]

Table 2.1: Overview of composition and applications of data sets in structure-based inverse screening approaches. ([1] Meslamani et al. [2011],[2] Paul et al. [2004],[3] Kellenberger et al. [2008],[4] Zhu et al. [2012],[5] Chen and Zhi [2001],[6] Chen and Ung [2001],[7] Gao et al. [2008],[8] Li et al. [2006],[9] Milletti and Vulpetti [2010],[10] Zahler et al. [2007]

3

Research Aims and Preconditions

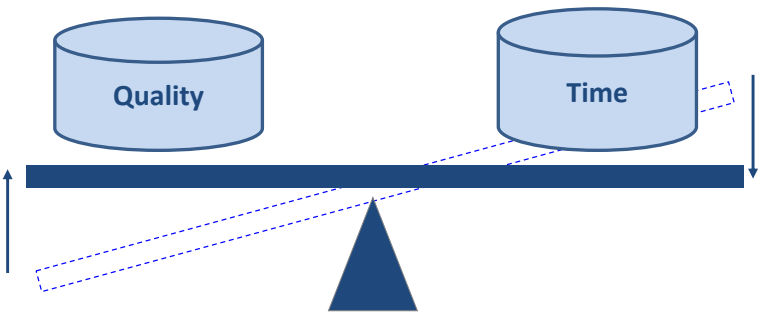


Figure 3.1: The aim of this thesis is a balance between high quality predictions and reasonable running time of the algorithms for the prediction of protein targets for small molecules.

3.1 Aims and objectives

The preceding chapter shows the current state of the art in the field of computational target identification. Many limitations become obvious when looking at these methods and results. This section shall define which obstacles shall be faced during this dissertation, and what the main aim of the project is.
The aim of this dissertation project is the development of a computational structure-based inverse screening method which manages to balance between a good quality of target prediction and a reasonable time frame of the predictions. Therefore, in contrast to most so far published approaches, the to be developed method shall regard the 'inverse' aspect more thoroughly. The aim of the research project of this thesis is summarized as follows:

The aim of this thesis is the development of a structure-based inverse screening method that is able to reliably predict protein targets for small organic compounds in a reasonable amount of time.

The literature analysis of structure based computational target identification methods reveals weak spots mainly in the following aspects: Consistent and appropriate handling of ten thousands of 3D protein structures (concerning time as well), capability of the scoring functions for inter-target ranking and evaluation on suitable and meaningful data.
Thus, these challenges have to be addressed by a new inverse screening method. The following obstacles are defined for this research project:

- Automatic processing of protein structures: For the large scale analysis of thousands of protein-ligand complexes, the protein structures and active sites cannot be prepared one at a time by the user. Therefore, the method has to be able to determine the active site residues, cofactors and metals in an automatic manner.

- Consistent and efficient storage of protein-ligand complexes: Once processed, the protein-ligand complexes and the active site must be stored in a form that allows multiple and consistent access to the data. Since preparation is a time-consuming step, the storage of this data saves much time and holds information about the active site used.

- Abstraction of the active sites: Describing the features of the active site in a form that allows a rapid pre-processing which eliminates protein structures where the query compound is definitely not binding to.

- Inter-target ranking: Scoring the poses of one compound in different targets so far holds a huge challenge since traditional protein-ligand scoring functions are not suited for comparing the absolute values among different targets.
- Evaluation data: So far, the data sets available for evaluation for computational target prediction lack true negatives. No standard has been developed, which allows the comparison of different methods.
- Usability by medicinal chemists: Nowadays, no method should be developed without keeping in mind what a user wants to do with it, what makes it usable for the user and how the usage and the results can be presented to the user, e.g., through a graphical user interface.
- Usability of method in other fields: So far, structure-based methods are mostly used in drug design. An evaluation shall show how and to what extent inverse virtual screening and other structure-based methods can be used in a biotechnological case study where a synthetic enzymatic pathway is developed.

3.2 Preconditions and course of the project

The presented thesis was prepared in the Research Group for Computational Molecular Design at the Center for Bioinformatics of the University of Hamburg (ZBH) from January 2010 to May 2014.
In this dissertation project, an inverse screening method was realized in form of a new software '*i*RAISE'. As a basis for the inverse screening algorithms, an indexing and screening method was used similarly to the Trixx BMI method developed by Dr J. Schlosser in his dissertation (Schlosser and Rarey [2009]). The *i*RAISE software uses the Naomi-Software Library of the ZBH and the BioSolveIT (www.biosolveit.de). The algorithms behind the Trixx BMI software were reimplemented at the beginning of the dissertation, together with S. Urbaczek and A. Henzler, who also developed the 'Trixx' software library used in this project containing functionality for interaction spot assignment, triangle descriptor generation and the clash score grid. The ComplexDB database which was developed during this dissertation was supported by S. Urbaczek and S. Bietz. Further, *i*RAISE uses the software FastBit (Wu [2005]) und Qt (http://qt-project.org/). The GUI ComplexViewer which was also developed in this research project is based on Qt and uses the 3D molecule/protein visualization library developed by the BioSolveIT.
While the *i*RAISE-software development was the main focus of this thesis, a case study in a biotechnology project was conducted: As a part of the SynBio-LEXI project with more

than a dozen project partners from various institutes in Hamburg, molecular docking and inverse screening were applied and experimentally validated. The results of this case study was published in the *Journal of Biotechnology* and presented in form of a talk at two conferences and in form of a poster at two conferences. The *i*RAISE inverse screening approach was published in the *Journal of Chemical Information and Modeling* (at the time of writing this thesis, one publication is accepted and one in reviewing status) and in form of a talk and a poster at two conferences. See Appendix I for a list of the publications, talks and poster presentations.

4
Methods

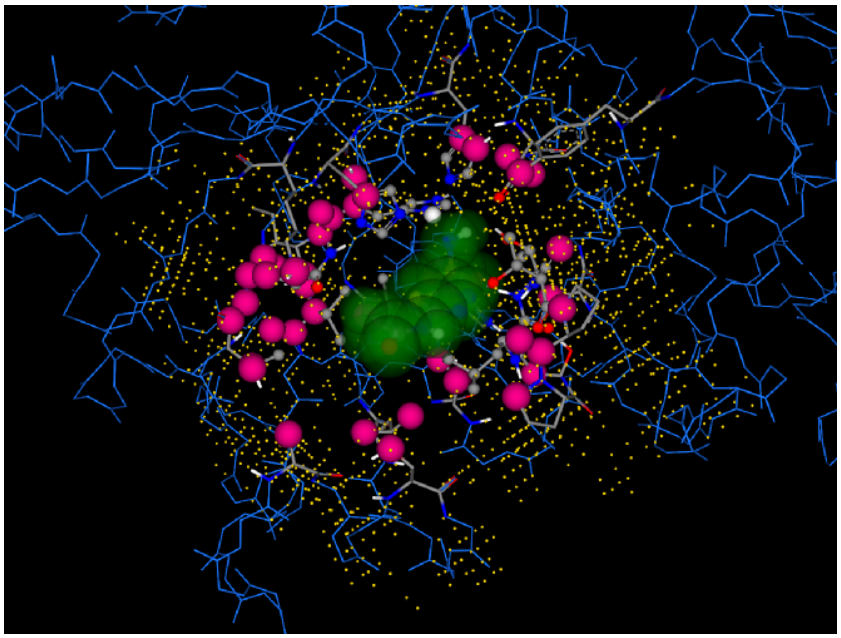


Figure 4.1: An *i*RAISE pose with the grid representation of the active site and the pocket atoms which are not covered by the ligand highlighted with pink spheres. Created with the ComplexViewer.

In this chapter, the details of the *i*RAISE inverse screening method are described. Throughout the chapter, it is focused on those aspects of *i*RAISE, where special measures where taken to account for the inverse setup of a protein-ligand screening. These measures differentiate *i*RAISE from traditional virtual screening methods which search potential ligands for one protein.
First of all, some basic concepts which are used by *i*RAISE are given (section 4.1). The heart of *i*RAISE's method is a triangle descriptor representation of the protein and the ligand. Therefore, in the following, the concept of the triangle descriptor representation is given (section 4.2). The special adjustments of the descriptor concepts to the requirements of *i*RAISE are shown. Subsequently, the complete workflow of *i*RAISE is shown in an overview which demonstrates how the many parts and used technologies are combined into one tool. Each step of the workflow is described (section 4.3).
The following section focuses on the representation of protein structures in form of a database (section 4.4). Next, the scoring scheme used by *i*RAISE is described in a separate section, presenting special measures applied for ranking proteins (section 4.5). Following, a statistical assessment of significant scores is shown (section 4.6). Next to these method details, the parallelization strategy is given (section 4.7). Finally, a graphic user interface for a viewer of the protein database and the screening solutions of *i*RAISE is shown (section 4.8). For a user guide of *i*RAISE see Appendix C and for implementation details see Appendix A. Furthermore, in Appendix B, a list of parameters used in *i*RAISE is given.
The name '*i*RAISE' for the inverse screening method gives credit to the descriptor index-technology used: **i**nverse **RA**pid **I**ndex-based **S**creening **E**ngine.

4.1 Basics concepts

This section gives basic concepts used by *i*RAISE. These concepts are used in the later described approaches and thus shortly explained here.

4.1.1 NAOMI molecule and protein initialization

In the software library used in *i*RAISE (called NAOMI ZBH library), libraries are available for molecule and protein initialization from various file formats. These steps are not discussed in detail here. It shall only be noticed that the NAOMI initialization from file is based on a very strict chemical model. In order to assure correctness, chemically invalid molecules or proteins are not corrected but discarded. A limitation of the library is that up to now, it has no model to handle molecules with covalently bound metals and, therefore, is not able to initialize such molecules (Urbaczek et al. [2011], Urbaczek et al. [2012]).

4.1.2 The MolString

Molecules in NAOMI can be represented by an internal unique string. This string is used for efficient re-initialization and contains all atom and bond information of the chemical model used by NAOMI. Further, it can be used as a unique identifier, similar to a USMILES (James and Weininger [2006]). This string is used exceedingly in the protein structure database developed for efficient and consistent storage of protein structures (see section 4.4).

4.1.3 Abstraction of proteins and ligands by a descriptor

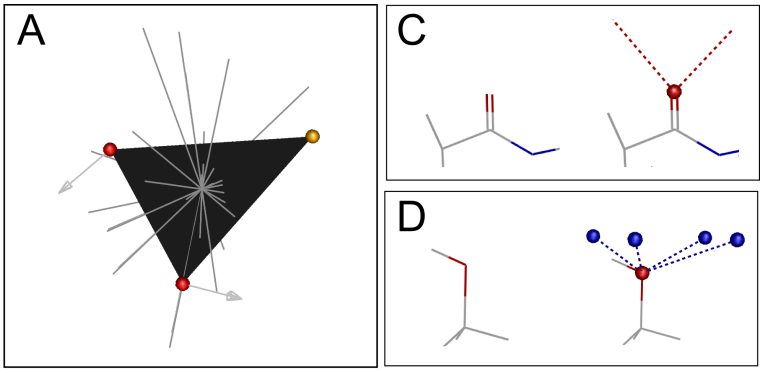
A descriptor representation of molecules and proteins is an abstraction coding the features important for the binding between protein and molecule. Due to the simplification of the molecule and the protein to a few descriptors, huge amounts of protein-ligand combinations can be evaluated rapidly. However, the protein-ligand binding is estimated only coarsely by this abstraction. The protein-ligand complex is reduced to complementary descriptors. Therefore, a post-processing after the abstract match of protein and ligand on descriptor basis is needed. The abstract representation as a first step discards all obviously not matching protein-ligand combinations. Thus, only a small amount of the combinations are sent to a more elaborate and more time-consuming post-processing step which finally decides whether a ligand can be placed in a protein cavity. In this manner, the descriptor representation breaks the sequential examination of each protein-ligand combination in detail. These concepts are valid for both cases, traditional and inverse virtual screening, i.e., if for one protein several compounds are tested or if for one compound several targets are tested.

4.2 TrixX triangle descriptor

*i*RAISE is based on the TrixX triangle descriptor, firstly introduced by Schellhammer and Rarey (Schellhammer and Rarey [2007]) for traditional virtual screening. Schlosser and Rarey then extended the descriptor by a shape component and a bitmap-based index representation allowing an efficient storage and query management (Schlosser and Rarey [2009]).

Properties of the triangle descriptor In Figure 4.2A, an exemplary triangle is shown. The triangle descriptor consists of the following properties for both, ligand and protein representation:

- **Triangle corners**: The triangle corners represent interaction spots. An interaction spot is either a potential hydrogen bond partner (donor or acceptor), a metal inter-



B

Property	Min, Max	Bin-length	Data type	Encoding	Matching tolerance
triangle corners	0,8	-	-	-	-
side length	1Å,9.9Å	0.1Å	unsigned short	range	1.2Å
directions	-	no binning	unsigned long	range	36°
bulk rays	1Å, 15Å	0.5Å	byte	range	0.75Å
coordinates	-	no binning	float	-	-

Figure 4.2: Exemplary triangle descriptor, a list of descriptor properties and hydrophilic interaction assignment. (A) The red corner spots are hydrogen bond acceptor interaction points, the gold spot is a hydrophobic interaction point. The arrows at the acceptor spots indicate the interaction directions. The rays originating from the triangle center map the shape. (B) The descriptor properties with their minimal and maximal values, the size of the bins, the data type used and the matching encoding and tolerances. (C) Assignment of a hydrogen bond acceptor interaction spot with two directions to a carbonyl oxygen. (D) Assignment of rotatable interaction spots at a hydroxyl group. The blue spots are hydrogen bond donor interaction points. The dotted lines are interaction directions.

action spot or a hydrophobic interaction spot. In the Schellhammer and Schlosser version of the TrixX descriptor, the interaction spots were determined by FlexX site interaction centers (SIACs) and compound interaction centers (CIACs). The interaction spots used in *i*RAISE are described in section 4.2.1 and 4.2.2. In Figure 4.2, the triangle has two different interaction types as corners, colored respectively: red spots are hydrogen bond acceptor interactions and the gold spot is a hydrophobic interaction. Furthermore, blue spots represent hydrogen bond donors. Triangles always need to have at least one hydrophilic interaction spot. Triangles consisting of hydrophobic spots at all three corners are not constructed.

- **Interaction directions**: For the hydrophilic interaction spots, interaction directions are part of the descriptor. In Figure 4.2, the interaction directions of the two hydrogen bond acceptors are indicated by arrows at the corners.

- **Triangle side length**: The side lengths of the triangle are used to describe the distances between the interaction spots.
- **Steric bulk**: A canonized representation of 80 bulk rays originating from the center of the triangle represents the steric properties of a compound or an active site. The bulk rays are generated by a once refined icosahedron and go through the face of the 80 sub-triangles of the refined icosahedron.
  For a compound, the length of the bulk rays represent the shape and are restricted by the surface of the ligand. For an active site, the bulk rays represent the volume and reach until the surface of the protein. In Figure 4.2, the rays are shown in gray originating from the center of the triangle.
- **Identifiers**: In the original TrixX descriptor, a molecule identifier and a fragment identifier were part of the descriptor. These identifiers were used to map the descriptors to their respective molecules and fragments in the matching procedure.

Adjustments to the descriptor for the reverse setup In *i*RAISE, the triangle descriptor is used in a reverse fashion: Originally, the descriptors were calculated and stored for a large number of molecules and then queried with descriptors derived from one protein. In *i*RAISE, descriptors are calculated and stored for many proteins and are subsequently queried with descriptors derived from one molecule. Therefore, two adjustments were made:

- **Identifiers**: In *i*RAISE, two identifiers are needed: A pocket and a protein identifier. One protein can have several binding pockets. Therefore, the two identifiers are needed to define which distinct pocket the descriptors were derived from.
- **Coordinates of the triangle corners**: The *i*RAISE descriptor also contains the triangle corner coordinates. This feature is not needed for matching, but for superposition. Triangle matches are handled differently in the original TrixX and in the reverse setup. In the original method, if a descriptor match occurred, the ligand was re-initialized and the descriptors were recalculated to obtain the triangle corner coordinates needed for ligand pose generation in the active site of the protein. Since for proteins, the re-initialization and especially the triangle generation step is much more elaborate and time-consuming, it is avoided by storing the triangle coordinates. Therefore, accepting larger memory demands for storing the coordinates is rewarded with time-saving during screening.

In section 4.2.1 and 4.2.2, the assignment of triangle descriptors to the interaction spots is described.

Representation of the triangle descriptor in a bitmap index The triangle descriptors are stored in a bitmap index for two reasons: Firstly, descriptors once calculated are stored consistently for repeated queries. Secondly, a bitmap-based representation allows efficient queries. Schlosser regarded the FastBit bitmap index developed by Wu et al. (Wu [2005]) best suited for storing the high-cardinality descriptor data which is accessed mostly read-only. The FastBit bitmap index has the advantage of applying a compressing scheme (WAH: Word aligned Hybrid code) exploiting the CPU word size. Defining range and equality encoding for the descriptor properties as needed renders the queries most efficient.
In order to represent the triangle descriptors in the FastBit bitmap index, special measures are taken to prepare the descriptor properties. The combination of interaction types at the triangle corners are mapped into a single number representing the triangle type. The triangle type codes a defined combination: Triangles of type 0 for example have hydrogen bond donors at all three corners, type 1 triangles have two hydrogen bond acceptor and one donor and so on. Since a triangle always possesses at least one hydrophilic interaction spot, i.e., one hydrogen bond donor or acceptor, there are nine different triangle types. These types are used to partition the triangles into folders with the respective number. This reduces the matching queries needed and thus the computing time (see next paragraph).
The side length of the triangles are binned into bins of 0.1 Å length and only triangles with side lengths in the range between 1 Å and 9.9 Å are constructed.
Further, the interaction directions are stored in one number: An icosahedron is used to map the direction vector to a single number. A bit string of size 20 is used to represent the faces of an icosahedron. A bit is set to one if the direction vector passes through the respective face of the icosahedron, and all others are set to zero. This way, also multiple directions of one interaction can be represented in a simple fashion (see section 4.2.1). The next paragraph explains how matching tolerances are stored also in the bit string of the directions.
The bulk rays are binned to bins of 0.5 Å size. Bulk rays are in the range of 1 and 15 Å, therefore 28 bins are needed. The threshold of the maximum ray length was doubled in *i*RAISE compared to TrixX, because in parametrization studies (see section 7.9) it was found that the original length of 7.5 Å was too short and resulted in many clashing poses. This effect can be explained by the fact that the original TrixX fragmented large ligands, which is not done in *i*RAISE.
The coordinates are not binned and stored in float values in the FastBit index.
In the Table shown in Figure 4.2B, the boundaries of the properties and the bin sizes for all properties are summarized.

Matching of protein and compound descriptors In *i*RAISE, protein triangle descriptors are stored in the index and the molecule triangle descriptors are used as queries. In order to reduce the number of queries, *i*RAISE takes advantage of the partitioning of triangles by type, thus only triangles with the correct type are queried. Protein and ligand features have to be complementary to each other to interact. Therefore, the complementary type of triangle is calculated for the query triangles for matching. This means that a hydrogen bond acceptor is turned into a hydrogen bond donor and vice versa. Hydrophobic points are left as they are.
For the other properties like the bulk rays or the triangle side lengths, matching tolerances are applied as listed in the Table shown in Figure 4.2B. These tolerances, however, are the default screening parameters, special parameters can be applied to soften or tighten the matching. The running time is strongly influenced by these parameters, since they determine the number of triangle matches which have to be processed.
Concerning the side lengths, a molecule triangle is allowed to be smaller or larger than that of the protein triangle index within the specified tolerance. Concerning the bulk, however, a molecule triangle bulk ray has to be truly smaller than the protein triangle bulk plus the tolerance. This ensures that the molecule fits into the protein, while it does not have to fill the pocket with its shape completely.
The tolerance of directions is directly encoded in the bit-string coding the faces of the icosahedron. Not only the bit of the face which the direction passes is set to 1, but also the bits of the faces in an angle of $36\,^{\circ}$. This angle results from the level of discretization of rotatable interactions, see section 4.2.1. A bit-wise 'AND' during descriptor matches then gives a match if any of the set faces of the icosahedra of index and query triangles match.

Data partitioning The index is partitioned for two reasons: Firstly, one index partition needs to fit into the main memory and secondly for parallelization (see section 4.7). The partitioning is set up in two levels, equally to the original TrixX setup. The first level is the division of descriptors into types, as described above. The type level can be screened sequentially, therefore only one type level partition has to fit into the main memory. The second level is the total number of descriptors in all type levels. This limits the number of proteins contained in one partition. This level is needed, because during matching, all matched descriptors need to fit into the main memory, which may be all index descriptors and all query descriptors in the worst case.

4.2.1 Ligand triangle descriptor generation

In Figure 4.3, the steps of the molecule triangle descriptor generation are shown. The input is a molecule structure (Figure 4.3A). For this structure, hydrophilic and hydrophobic interaction spots are calculated (Figure 4.3B). Then, triangle descriptors are derived for all combinations of three interaction spots Figure 4.3C) if the following criteria are met: The side lengths are minimally 1 Å and maximally 9.9 Å and at least one interaction spot is hydrophilic. The first criterion discards triangles which are too pointed, while the second criterion discards non-specific triangles with only hydrophobic corners.

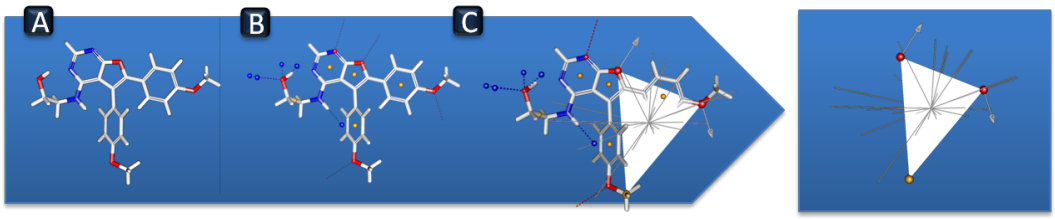


Figure 4.3: Pipeline of the triangle descriptor generation for a molecule A) The input is a molecule structure B) The molecule is reduced to interaction spots (red: hydrogen bond acceptors, blue: hydrogen bond donors, gold: hydrophobic spots) C) For each combination of three interaction spots, triangle descriptors are generated.

Hydrophilic molecule interactions Hydrophilic interaction spots are either hydrogen bond donors or hydrogen bond acceptors. Hydrogen bond acceptors are, e.g., the lone pairs of the oxygen atoms of hydroxyl groups. For these interactions, the spots are placed on the oxygen atom and the direction of the interaction points to the lone pair.
A hydrogen bond donor is, e.g., the hydrogen of a hydroxyl group. In order to facilitate the matching of triangle descriptors and to account for the distance between hydrogen bond donor and acceptor in a hydrogen bond, the spots of a hydrogen bond donor interaction are placed where an acceptor would be in an optimal hydrogen bond. The spot is therefore placed 2.8 Å distant from the oxygen atom in direction of the hydrogen atom.
Sometimes it occurs that an interaction spot has multiple directions. For example, carbonyl oxygen atoms have two free lone pairs and can therefore serve as hydrogen bond acceptors in two directions (see Figure 4.2C). In order to avoid construction of two triangle descriptors with the same interaction spot but differing interactions, both directions are represented in one descriptor. The faces of the icosahedron are set to true for both directions.
In Figure 4.3B, the described interaction spots are shown. The red acceptor spots are placed

directly on the atoms and the blue donor spots distant to the heavy atom.

Representation of flexibility in interaction spots Hydrophilic interaction spots are used to account for one aspect of molecule flexibility, i.e., rotatable groups. One example of such groups are the above mentioned hydroxyl groups. These are freely rotatable and in nature only fixed by an interaction partner. The conformation represented in a protein crystal structure is, thus, only one of many rotational states of those groups. For rotatable groups, the interaction spots are sampled in $72\,^{\circ}$ steps around the rotatable atom, see Figure 4.2D. In this way, more possible matchings of the protein to the ligand can be derived from the descriptor representation, but clearly also the number of descriptors is multiplied. Therefore, only those sampled interactions spots are used, whose directions point into accessible areas. This means that if a direction of, e.g., a molecule interaction points into its own volume, then this interaction spot is discarded. This is shown in Figure 4.2D, where only 4 of the 5 possible interaction spots are sampled.

Hydrophobic molecule interaction spots (Placement points) Hydrophobic spots do not have directions, since chemically hydrophobic contacts are energetically favorable independent from any directions. For molecules, hydrophobic spots are placed in the middle of rings and on the bonds between carbon atoms which are not bound to a hetero atom.

4.2.2 Protein triangle descriptor generation

In Figure 4.4, the steps of the protein triangle descriptor generation are shown. The triangle descriptors are calculated for active sites only, therefore, the first step is to calculate binding pockets of the protein (Figure 4.4A and B). See section 4.3.1.2 for details on this this step. Then, the interaction spots are assigned for the atoms of the active site (Figure 4.4C). This step is divided into two parts: First, hydrophilic interaction spots are assigned and then hydrophobic spots are located (see below for a detailed description). Finally, triangle descriptors are derived from the interaction spots following the same criteria as described above for molecules.

Hydrophilic protein interactions Hydrophilic interaction spots are generated following the procedure as described for molecules in section 4.2.1. Further, protein binding sites may contain metal ions with which a ligand may interact. In such an interaction, metals are interacting with hydrogen bond acceptors, since the metals are positively charged. In order

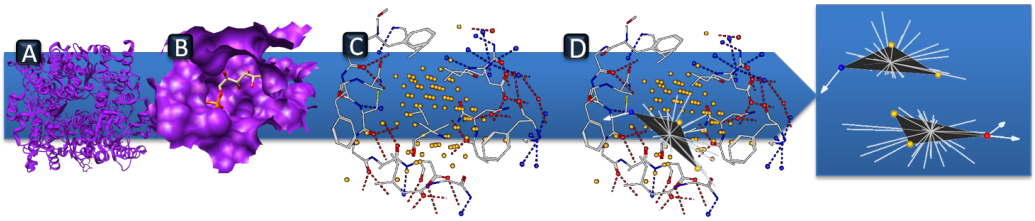


Figure 4.4: Pipeline of the triangle descriptor generation for a protein. A) The input is a protein structure. B) The protein structure is reduced to one or several active sites C) The interaction spots are determined for the active site (red: hydrogen bond acceptors, blue: hydrogen bond donors, gold: hydrophobic spots) D) Triangle descriptors are generated for each combination of three interaction spots.

to avoid a fourth interaction type, metal interaction spots are represented as hydrogen bond donor spots.

Representation of protein flexibility in interaction spots For the protein descriptor, the same measures are applied to account for rotatable groups as discussed for molecules in section 4.2.1.

Hydrophobic protein interaction spots (Placement points) Hydrophobic spots are placed into the binding site based on a grid representation which is shown in Figure 4.1. The grid spans around the complete active site (e.g. if the active site was determined as all atoms around a reference ligand in 6.5 Å, the grid is calculated for 10 Å). All grid points that are within the protein surface or too near to the protein surface are discarded. Then, all grid points in the neighborhood of hydrophilic interaction spots are discarded. In this step, each grid point is assigned with a hydrophobilic score, depending on how many hydrophilic interaction spots are nearby. The score resulting from flexible interaction spots is lower than the one resulting from other hydrophilic interaction spots. Then, the grid points with the highest hydrophilic score are discarded. The remaining grid points are turned into hydrophobic interaction spots. The final number of points is dependent on the total number of hydrophobic atoms in the active site and, therefore, maps the hydrophibicity of the active site.

This procedure leaves a manageable amount of hydrophobic spots which, nevertheless, cover those parts of the binding site which are not represented by the hydrophilic interactions spots.

4.3 *i*RAISE workflow

In Figure 4.5 the total *i*RAISE procedure is shown. The two main parts are the *registration procedure* and the *screening procedure*. While the preceding sections gave detail on some main components of *i*RAISE, the workflow is described now step by step.

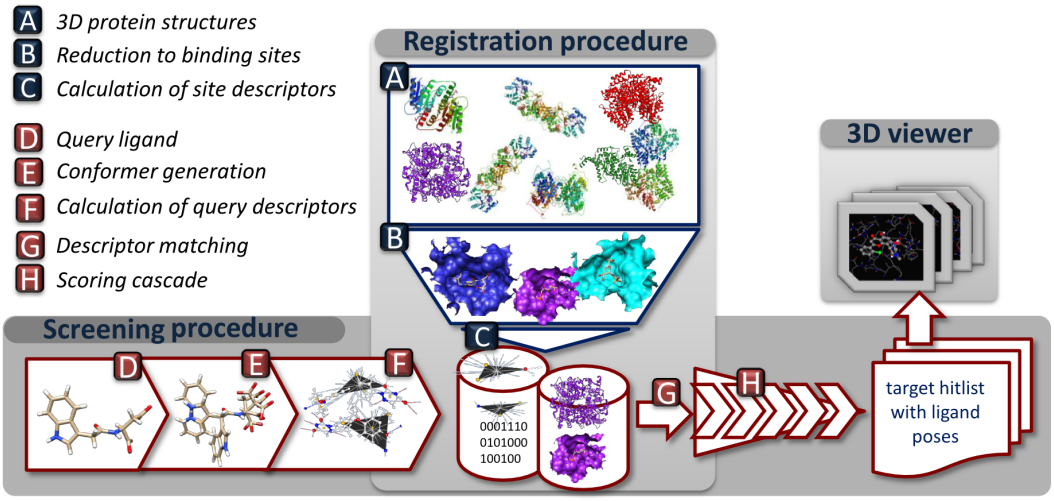


Figure 4.5: Schematic representation of the workflow of *i*RAISE. *(This figure was originally published in Schomburg et al. [2014].)*

4.3.1 Registration procedure

In the registration procedure, the protein data is processed. This step has to be done only once for a set of protein structures since the processed protein data is stored in a database and the generated triangle descriptors in a bitmap index which can then be accessed repeatedly. Therefore, the focus of this step is on a consistent and efficient storage of the data while running times are not overly important.

4.3.1.1 Protein initialization

Proteins have to be provided in PDB file format. The protein is initialized with the NAOMI protein library (ProLib) which automatically derives the protein structure from the heavy atom coordinates stored in the PDB file. All other molecules next to the protein are initialized as well, so that they can be accessed if needed. The result is a complex containing the

protein, ligands, metals and waters, if contained in the PDB file. Additionally, default hydrogen coordinates are assigned automatically.

4.3.1.2 Active site determination

In the next step, the protein is reduced to binding sites. Two approaches are supported for this step:
Either, a reference ligand can be provided by the user which is then used to define the active site. In this case, all amino acids within a user-provided distance around the reference ligand, e.g., 6.5 Å are selected.
The active site can also be annotated automatically using the ligands contained in the PDB file. Each ligand in the PDB file is used to build an active site except ligands which are cofactors, crystallization or solution agents. An exclusion list contains all these unwanted ligands for active site determination in two forms: In the form of the HET PDB code and in the form of a unique SMILES. The HET code is a three letter code used in the PDB file to uniquely annotate molecules. The second representation as a USMILES is used since sometimes HET codes are annotated wrongly in the files or the HET codes are changed by the Protein Data Bank in new releases for some reason. The unique SMILES are used in combination with the HET codes, since the initialization by NAOMI (Urbaczek et al. [2011]) sometimes identifies a different molecule than annotated by the HET code. One example is an aldehyde versus an carbonyl. The respective form is determined by NAOMI by the bond length between oxygen and the carbon atom. Therefore, if the coordinates indicate a different state of the molecule than the HET code, the unique SMILES is failing, but the HET code nevertheless identifies the molecule as unwanted for active site determination.
The exclusion list is compiled by joining HET code lists from Strömbergsson (Strömbergsson and Kleywegt [2009]), Boström (Boström et al. [2006]) and Meslamani (Meslamani et al. [2011]) and adding further HET codes. This list contains in total 1207 PDB HET codes, see Appendix F for a list. Next to the list of HET codes of cofactors and solution agents, a list of HET codes is given as well, which contains ions and one list of ligands with covalently bound metals. Ligands with covalently bound metals are added to the exclusion list since NAOMI does not handle such ligands.

4.3.1.3 Storage in protein database and bitmap index

Once the atoms of the active site are determined, the triangle descriptors are calculated (see section 4.2.2). Then, the protein and the active site information is written to the protein database (see section 4.4 for details on the protein database) and the triangle descriptors are stored in the bitmap index.

4.3.2 Screening procedure

The screening or querying procedure exploits the previously calculated information for fast matching and pose generation. This step needs to be efficient since thousands of proteins shall be screened in a reasonable time. Therefore, the procedure applies a filtering procedure which goes from rough descriptor matching through grid-based pose assessment to detailed atomic analysis. The amount of proteins and molecule poses per protein is reduced in each step, guaranteeing that the most time-consuming steps are only conducted on a subset of reasonable protein-ligand poses. The steps of the screening procedure are now described in detail.

4.3.2.1 Molecule conformation generation

The flexibility of molecules is modeled by creating conformations. The maximal number of conformations created per molecule is set by the user. For conformation generation, the CONFECT module developed by Schärfer (Schärfer et al. [2013]) is used. This conformation generation procedure has the advantage that the number of conformations needed to sample a diverse conformational space is lower in comparison to other tools. It is based on a generalized representation of torsion angles in molecules. For each general torsion pattern, observed frequency distributions of crystallized molecules are generated. These distributions are used to incrementally construct and rank the conformations.

4.3.2.2 Unique molecule triangle descriptor generation

For each of the generated conformations, the triangle descriptors are calculated as described in section 4.2.1. Often, many triangle descriptors are identical for different conformations. This is the case if, e.g., only a terminal group is rotated in one of the conformations. Then, the descriptors covering the remaining part of the molecule are identical. In order to avoid multiple matchings of the same descriptor with the index, duplicate descriptors are eliminated. The descriptors are only eliminated if all properties as well as the corner coordinates are identical. The comparison is done on the basis of the binned descriptors, which is rather time-efficient. For each descriptor a list of conformation identifiers is used for mapping the descriptors to the respective conformations, which is needed in the matching procedure.

4.3.2.3 Matching procedure

In the matching procedure, the descriptors are matched sequentially against the protein descriptor index as described in section 4.2. For each match, the transformation is calculated

which superposes the triangles. This transformation is memorized since it is later used for calculating the transformation of the ligand into the respective protein active site for pose generation. In addition, the identifiers are memorized; further information is not needed. Therefore, a match consists of the transformation, a protein conformation key, a protein pocket key and a ligand conformation key. These matches are gathered until each descriptor has been matched against the bitmap index.
Then, the matches are sorted by protein conformation key and then by pocket key. This allows to sequentially process the matched proteins and pockets. Therefore, each protein of which a descriptor has been matched has to be re-initialized from the protein database only once. For each descriptor match, a pose of the matching molecule conformation is created by applying the transformation of the match to the molecule conformation. Applying the transformation to the ligand atoms is more efficient than transforming all protein atoms.

4.3.2.4 Scoring

In order to evaluate the quality for the generated poses, for each protein, the protein-ligand poses obtained from the matching procedure are scored by a five-step Scoring Cascade, which is in detail described in section 4.5.

4.3.2.5 Solution handling

Once the final protein-ligand poses and their scores are obtained, these results are stored in a solution SQLite-database. It contains two tables with the following attributes:

- Table *Solutionoverview*:
  - Query key
  - Query name
  - Query USMILES
- Table *Solutiondetails*:
  - Query key
  - Protein conformation key
  - Pocket key
  - Conformation key
  - Coordinates

- Score

In parallel screening (see section 4.7), a separate solution database is created for each screened partition to avoid write-locks. In a consolidation step, the databases are then attached and a single solution database with the content of all parallel screenings is created. The solution database can contain poses for multiple ligands if the protein data set was screened with various query molecules. The solution databases contain keys of the protein database, which allows a join of the data.

4.4 ProteinDB and ComplexDB

Since no efficient and consistent handling of protein structures was established before, a protein structure database was compiled. This database avoids the use of files and allows to store information calculated in the registration procedure, which speeds up the re-initialization. A consistent storage of the protein structures exactly as they were used to calculate the triangle descriptors is needed to guarantee a correct descriptor index-protein structure mapping. Further, it has to be memorized, how the active site was determined and which reference ligand was used, since it is needed by *i*RAISE scoring (see section 4.5). SQLite was used for the protein structure database since it has the advantages of not needing a separate server (therefore it is easily portable on different platforms), using the structured query language (SQL) and storing all information, e.g. tables, indices, etc., in one file.
In summary, the database is used as a consistent data container of the protein structures as well as for efficient re-initialization of active sites and proteins.
The scheme of the developed database is sketched in Figure 4.6. This scheme shows two levels of the database: The ProteinDB which stores the residues of a protein and the ComplexDB which consists of the ProteinDB tables and adds further tables to store ligands, waters and metals, thus, everything that is contained in a PDB file of a protein-ligand complex. The division into two databases allows the use of only the ProteinDB if needed, but also to easily add the information of a complex. In SQLite, all the tables are added to a single file, therefore, the separation does not produce any hindrance on using them together. In the database scheme in Figure 4.6, the blue headed tables code the information of the protein. The red tables code the information which is needed to store a complex as it is contained in a PDB file, with ligands and waters. Purple headed tables store the information which is gathered in the pre-calculation step about the active site.

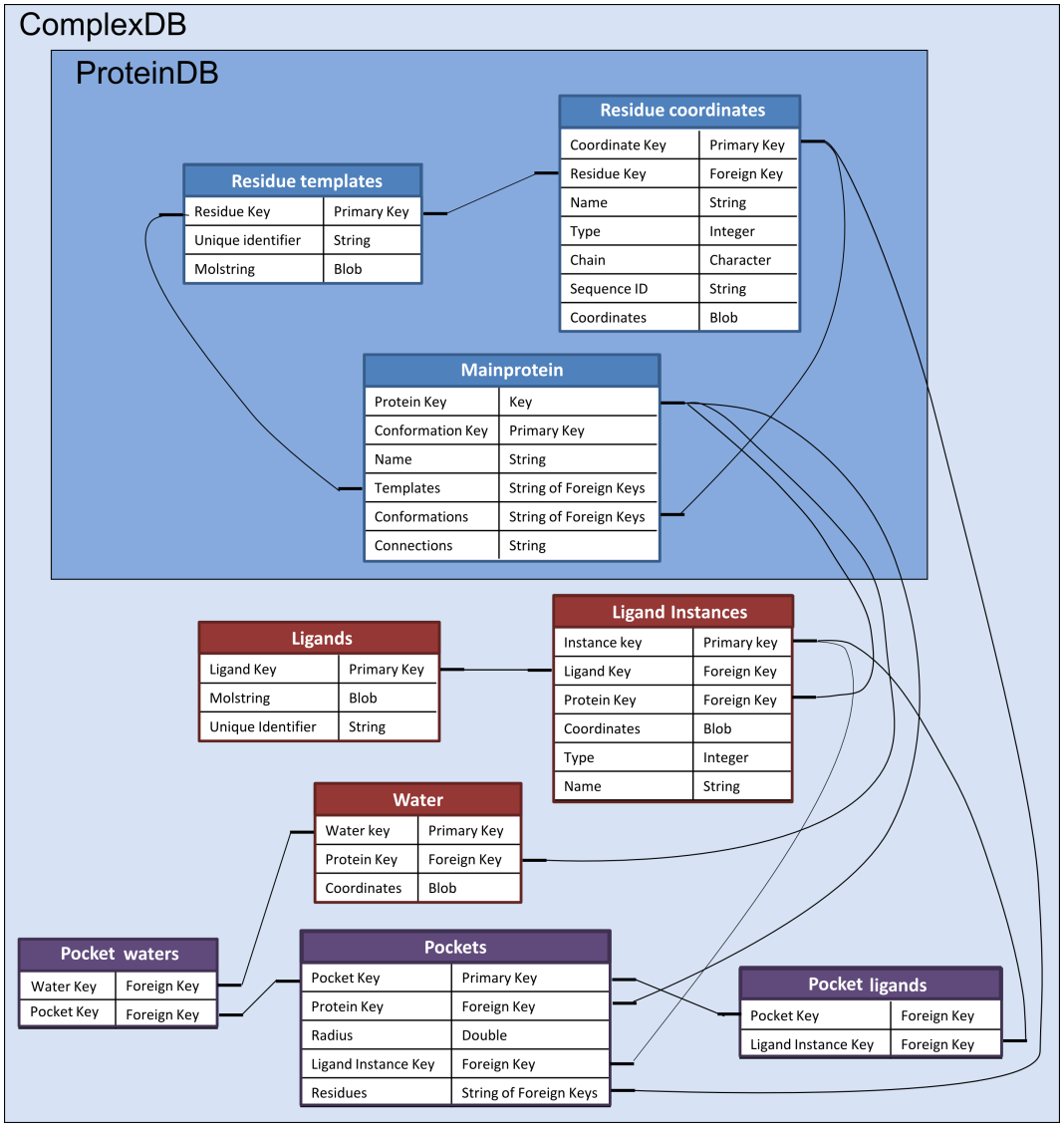


Figure 4.6: Schematic representation of the database tables of the ProteinDB and the ComplexDB. The blue tables code the information of the protein. The red tables code all information of small molecules like ligands, cofactors, metals and water. The purple tables code the information of the active site.

ProteinDB The ProteinDB contains three tables, the *mainprotein* table, the *residue templates* table and the *residue coordinates* table.

The *mainprotein* table contains a protein key and a conformation key, which is the primary key. Therefore, one protein can have several conformations. The protein key can be set from the outside if the user provides several conformations of the same protein. Providing several

conformations -also called *ensembles*- is one way of handling protein flexibility. Next to the keys, the *mainprotein* table contains the protein name as a combination of the name in the PDB file and the file name since a non-speaking name like 'protein' is often used in the files. Further, it contains three strings used for re-initialization: A list of residue template keys, a list of residue conformation keys and a connections string. The residue template in combination with a residue conformation gives an amino acid with coordinates and the connections string stores the information on how the residues are connected.
The *residue templates* table contains unique amino acid templates. Each amino acid is stored only once in this table, the first time it is encountered in a set of proteins added to the database. So for the first protein added to the database, at the beginning many templates are added to this table until no new amino acid is encountered. The unique amino acids contain the 22 natural amino acids, modified amino acids (e.g. with covalently bound ligands) and terminal amino acids. The residue templates table contains two attributes next to the primary key: The unique SMILES which is used as unique identifier to compare if the current amino acid already exists as a template or not. The MolString is a string representation of atoms, bonds and valence states needed to reconstruct the amino acid by NAOMI.
The *residue coordinates* table contains the coordinates of the amino acid templates. It contains as many entries as the sum of all amino acids of the proteins added to the database. Further, it contains a primary key and the residue template key of its template. Additionally, this table stores the name of the residue, the type, the chain character and the sequence ID. This information is added from the PDB file.
With the lists from the *mainprotein* table the residues can be connected easily to re-initialize a protein.

ComplexDB The ComplexDB is composed of all tables of the ProteinDB, which store the protein information. Further, tables are added for small molecule storage (ligands, ions, waters, metals) and for active site information (pockets).
Ligands from a complex are stored in a similar way as amino acids. The *ligands* table contains only a primary key, a USMILES as unique identifier and the MolString. This table serves as a ligand template table, while coordinates matching the templates are stored in the *ligands instances* table. The *ligands* table, therefore, contains only as many rows as there are topological different ligands in the set of complexes added to the database. The *ligand instances* table has as many rows as the sum of all ligands in the set of complexes added. Thus, the storage of duplicate information is avoided.
Next to the coordinates, the *ligands instances* table contains a primary key, the matching

Number of proteins	Size sum of PDB files	Size ComplexDB
100	63 MB	36 MB
1000	558 MB	333 MB
10000	5.7 GB	3.4 GB
56779	32 GB	19 GB

Table 4.1: Comparison of space requirements of proteins stored in PDB files and stored in a ComplexDB.

ligand foreign key, the protein foreign key, the name and a type. The type is a number coding of which type the small molecule is: Ligand, ion, metal or reference ligand. If the molecule is of type 'reference ligand', then this ligand was added only for active site determination and is not part of the protein-ligand complex when re-initialized.
For waters, a separate table was added. This table contains a primary key, a protein key and the coordinates.
With the so far mentioned tables, all information contained in a PDB file needed to reconstruct a protein-ligand complex with all small molecules is stored in the ComplexDB.
The remaining tables contain information of active sites/pockets of the proteins. The *pockets* table contains a primary key, the protein key as foreign key, the radius which was used for calculating the active site around a reference ligand, the ligand instance key of the ligand which was used to determine the active site and a string of the residue coordinate keys which are part of the active site. One protein can have several pockets if several active sites are calculated. The two tables *pocket ligands* and *pockets water* store foreign keys of other ligands and waters which are part of the active site.

Space and time requirements The size of a ComplexDB is about half of the size needed to store the raw PDB files. Table 4.1 compares the space requirements of raw files and the database for random subsets of 100, 1000 and 10.000 PDB files and for all protein-ligand complexes available from the PDB (56779 files). The table shows that the respective databases need only about 60 percent of the space of the files. Reading a protein and the annotated active site from the database only takes 0.06 seconds on average, independent from the database size, as long as the database fits into working memory.

4.5 Scoring Cascade

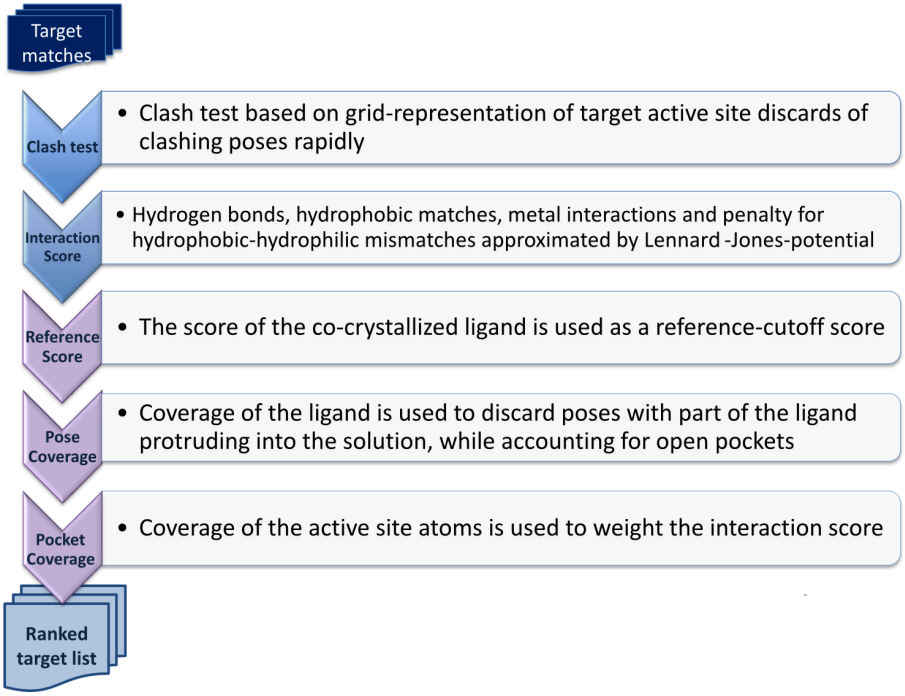


Figure 4.7: The five steps of the Scoring Cascade start with the target matches from the descriptor matching. A clash test discards clashing poses, then a simple interaction score is calculated followed by the reference score cutoff. The pose and pocket coverage account for diverse shapes of protein pockets. The result of the Scoring Cascade is a ranked list of proteins.

Since standard protein-ligand scoring functions are not well suited for inter-target ranking (Kellenberger et al. [2008], Wang et al. [2012]), special measures have to be taken to improve the ranking of proteins. Such measures are applied in the *i*RAISE-Scoring Cascade consisting of five consecutive scoring steps.

The Scoring Cascade is invoked once all descriptor matches are obtained from the index matching. Therefore, it starts with all poses of the ligand in all proteins resulting from the matches. The aim is to further reduce the number of matches with each step until only reasonable poses remain. While the number of poses needed to process is thus reduced in each step, the scoring steps can become more and more elaborate since the most time-consuming steps are applied to the lowest number of poses.

The measures that the Scoring Cascade takes to improve the inter-protein ranking are based on learning from the co-crystallized ligand and accounting for diverse shapes of protein pock-

ets. It is assumed that a ligand which is co-crystallized with a protein is a true binder and, thus, contains information which can be exploited. The shape of pockets may vary in size and buriedness and the scoring must not reward any shape of the pocket, but a good binding of a ligand to that pocket.
In total, the Scoring Cascade consists of five steps, see Figure 4.7 for an overview of the individual steps. The steps are applied to the poses of the query molecule for one protein after another to avoid multiple re-initialization of a protein.
Each step is now explained in detail. The final score resulting from the Scoring Cascade is called the *sc-score*.

4.5.1 1. Clash test

The clash test is used to rapidly discard poses which do not fit sterically into the protein pocket. Reasons why poses are not fitting although the descriptor maps the shape by the bulk are either that the bulk rays are not long enough, or that the sampling of the 80 bulk rays is not fine enough or that the tolerance of the matching of the bulk rays was too soft. The clash test is based on a grid representation of the active site atoms, shown in Figure 4.8 as yellow dots. If an atom of the ligand pose covers a grid point or gets too close to one, the pose is discarded. The clash test is rather soft, nevertheless, about two thirds (64%) of the ligand poses are discarded in this step. The grid calculation is relatively time-consuming (see section 7.10), but once calculated, the clash test on hundreds of poses is performed rapidly. The grid of 10 Å around the active site is calculated with a spacing of 0.8 Å.
Following the coarse-grained grid clash test is an atom-based clash test.

4.5.2 2. Interaction score

The interaction score is relatively simple compared to other elaborate protein-ligand scoring functions. It estimates the binding of protein and ligand on the basis of Lennard-Jones potentials with different parameters for hydrophilic interactions, metal-contacts and hydrophobic contacts. In Figure 4.9, the different potentials are sketched. For scoring mismatches of hydrophobic and hydrophilic atoms, a penalty based on the positive part of a Lennard-Jones potential is used. Each pose passing the grid clash test is scored with the interaction score. Before the score is calculated for a pose, the hydrogen atom positions of the protein-ligand complex are adjusted to form the optimal hydrogen bond network. This is done with PROTOSS (Lippert and Rarey [2009],Bietz et al. [2014]), an algorithm calculating the optimal positions, protonation and tautomeric states of the protein-ligand complex of the active site.

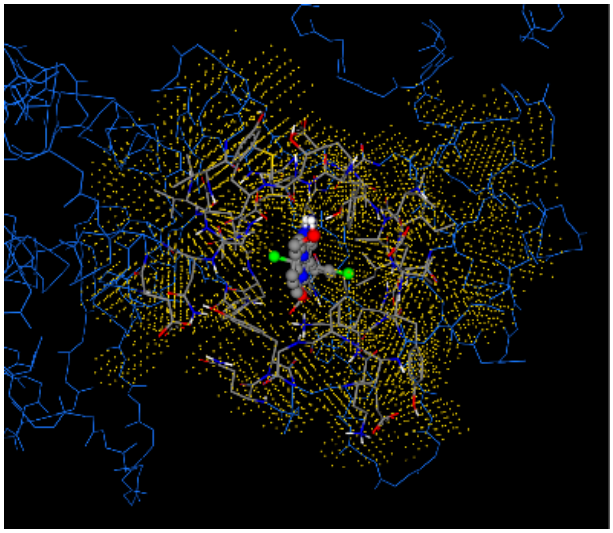


Figure 4.8: Abstraction of the active site with a grid representation for rapid clash tests of ligand poses.

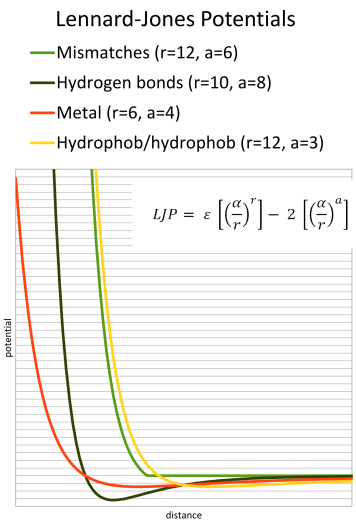


Figure 4.9: Lennard-Jones potentials for scoring hydrophilic interactions, metal- and hydrophobic contacts and penalizing mismatches.

4.5.3 3. Reference score

Besides each pose, the reference ligand used to determine the active site is also scored with the interaction score. The assumption is that being co-crystallized with the protein, the reference ligand is truly binding to the protein and its score is a good estimate of the score a ligand can achieve when binding to this active site. Therefore, the interaction score of each pose is compared to that of the reference ligand. If the score of the pose is less than 75% of the score of the reference ligand, the pose is discarded. This step discards inselective, low-scored poses and on average 50% of all target matches are discarded here.

4.5.4 4. Pose coverage

The fourth step of the Scoring Cascade is the ligand pose coverage score. In this step, the account to which the pose of the ligand is buried in the active site is assessed. Usually, if a ligand is truly binding to a protein, it is well-covered by the protein in the active site. However, often artificial poses at the outer rim of the pocket can also be obtained in the descriptor matching step. Such poses are unwanted and have to be discarded. Nevertheless, these artificial poses protruding into the solvent have to be differentiated from poses in shallow pockets. Therefore, the challenge in comparing the poses among various proteins lies in the diversity of pocket shapes. In Figure 4.10A-D, two scenarios are sketched: In A and B, a buried pocket is shown, whereas in C and D a shallow pocket is shown. In order to be able to compare the coverage in both pockets, again the reference ligand is considered. First, the ligand coverage is calculated for the pose as well as for the reference ligand with the following formulas:

$$LigandCoverage = \frac{1}{|A|} \sum\nolimits_{a \in A} Coverage(a) \tag{4.1}$$

$$Coverage(a) = \begin{cases} 1, \text{if } (nearproteinatoms(a) + \frac{1}{|N(a)|} \sum_{b \in N(a)} nearproteinatoms(b)) > 3 \\ 0, otherwise \end{cases} . \tag{4.2}$$

The ligand coverage is calculated as the part of the atoms which are covered. In the *Ligand Coverage* formula, *A* is the set of heavy atoms of the ligand. In the *Coverage* formula, the *nearproteinatoms*-parameter is the number of all atoms of the active site located in a radius of 4.5 Å around a ligand atom *a*. Further, the coverage of the neighboring atoms is added (*N(a)* = set of bound atoms to atom *a*) divided by the number of neighboring atoms. This sum has to be greater than three, meaning that more than three receptor atoms have to be near the ligand atom and/or each of its neighboring atoms for an atom to count as *covered*.

The neighboring atoms are incorporated into the formula to account for accumulated uncovered atoms. Thus, a molecules of which connected atoms are not covered get a lower pose coverage score than a molecule where the same number of unconnected-atoms are not covered.

The calculated ligand coverage for each pose is compared to the coverage of the reference ligand. If the score of a pose is less than the reference ligand coverage times 1.2, the pose is discarded. Also, a pose is discarded if less than 10% of the ligand atoms are *covered*, independent from the coverage of the reference ligand.

In Figure 4.10A, a buried pocket with its reference ligand is shown. All atoms are covered. In B, a pose on the rim of the pocket is shown, protruding with a large part into the solvent and consequently many uncovered atoms are detected (indicated by pink dots in the sketch). Since the coverage of the pose is much less than that of the reference ligand, this pose would be discarded.

In Figure 4.10C, a shallow pocket with a reference ligand is shown, of which some atoms are not covered. In Figure 4.10D, a pose of a ligand is shown, which also has some uncovered atoms, but fewer compared to the reference ligand. Therefore, this pose is further processed in the next scoring step.

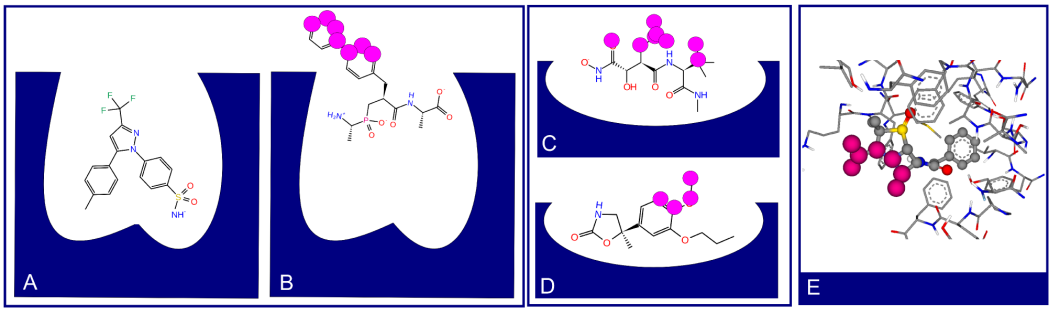


Figure 4.10: Schematic representation of the ligand coverage. Uncovered atoms are highlighted in pink. (A) A reference ligand in a buried pocket with no uncovered atoms. (B) A docking pose protruding into the solvent with seven uncovered atoms. This pose would be discarded by the ligand coverage criterion. (C) A shallow pocket with a reference ligand with seven not covered atoms. (D) A docking pose in the shallow pocket with also four uncovered atoms. This pose would not be discarded by the ligand coverage criterion. (E) A protein-ligand complex where the active site is open to one side and parts of the ligand are not covered.

In Figure 4.10E, the protein-ligand complex of a penicillin acylase with penicillin G is shown as an example. The ligand atoms which are not covered are marked with pink spheres. The

pocket is open to one side and, hence, the ligand is partly uncovered.

4.5.5 5. Pocket coverage

The final step of the Scoring Cascade is the pocket coverage score. The pocket coverage scoring improves the ranking of true targets for small molecules. It assesses how well a pocket is occupied by a ligand. This score is not used as a cutoff like the third and fourth step, but is applied as a weighting function. By down-weighting the interaction score if a pocket is not well-covered by a pose in comparison to the reference ligand, pockets which are better fitting to the ligand are preferred. The pocket coverage is calculated with the following formulas:

$$Pocket\ Coverage\ =\ \frac{1}{|P|}\sum_{a\,\in\,P} Coverage\ (a)$$

$$Coverage\ (a) = \begin{cases} 1, & if\ distance\ to\ any\ ligand\ atom\ < 4.5\ \text{Å} \\ 0, & otherwise \end{cases}$$

The *Pocket Coverage* is calculated as the number of covered pocket atoms divided by the total number of pocket atoms ($|P|$), with P as the pocket atoms. The *Coverage* of a pocket atom (a) is set to 1 if a ligand atom is located within a distance of 4.5 Å to a pocket atom, otherwise to 0. Scores of poses which result in a pocket coverage of less than 80% of the reference ligand pocket coverage are weighted down with a factor of 0.8.
Figure 4.11A and B show a sketch of the pocket coverage as well as protein-ligand complex examples. In Figure 4.11A, a pocket is filled well with a reference ligand. Few pocket atoms are not covered, indicated by pink dots. In comparison, in Figure 4.11B, the pocket is filled with a smaller ligand which occupies only part of the pocket resulting in many not covered pocket atoms.
In Figure 4.11C-E, real screening results are shown. In Figure 4.11C, the small ligand (pantothenoic acid) of the PDB-complex 1SQ5 (pantothenate kinase) is shown in its co-crystallized target with only a few not-covered pocket-atoms shown by pink spheres. A pose of the same ligand is shown in a larger pocket of the PDB-complex 1OWE (urokinase) in Figure 4.11D. Here, only part of the pocket is filled by pantothenoic acid and, thus, a large amount of the pocket atoms are uncovered. In comparison, the pocket coverage of the pose of the ligand of the 1owe PDB-complex in the 1owe-pocket shown in 4.11E is much higher.

Thus, the score of the pose in 4.11D would be weighted down.
The pocket coverage score is only used to weigh the scores and not to discard poses since a small ligand can be able to bind to a spacious pocket, the binding is just not as selective as in the case of a ligand filling a pocket well.

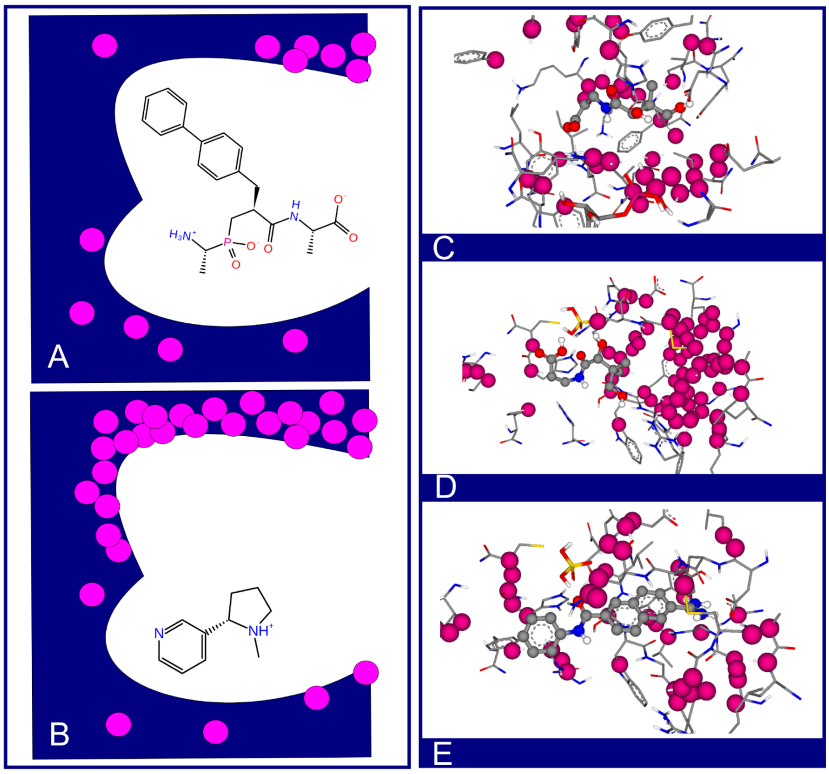


Figure 4.11: Schematic representation of the pocket coverage. Not covered pocket atoms are highlighted in pink. (A) A reference ligand pocket fills the pocket well and only the atoms of the outer rim of the pocket are uncovered. (B) A small ligand does not cover the pocket as well as the reference ligand and many more pocket atoms are uncovered. (C) Pose of a small ligand in its true target (1sq5), with only uncovered atoms at the rim of the pocket. (D) Pose of the small ligand (1sq5) in a spacious pocket (1owe) with many uncovered pocket atoms. (E) The spacious pocket (1owe) with its true ligand and less uncovered atoms in the pocket.

4.5.6 Measures not enhancing ranking of true targets

During the development of the Scoring Cascade, also some approaches were tested which did not result in an improvement of true target ranking. For completeness and as information

source for further developments, a list of these measures is given here:

- Increased weight of repulsive score contribution
- Decreased weight of metal score contribution
- Increased or decreased weight of hydrophobic score contribution
- Selectivity score based on the number of poses per pocket with similar score
- Inclusion of saturation of interactions spots of pocket into score

4.6 Gaussian-based weighting and cutoff

After the five steps of the Scoring Cascade, a Gaussian-based cutoff is used to assess if a score is statistically significant for a protein. The cutoff is based on the average score of the 84 chemically diverse ligands of the Astex Diverse Set (see Chapter 5).
For calculating the cutoff, an *i*RAISE screening project, i.e., each protein, is screened with all 84 ligands. For each of the 84 ligands the sc-score is calculated without using step three of the Scoring Cascade -the reference score cutoff- in order to get a full spectrum of scores, not only those which are in a comparable range to the reference ligand. Therefore, for each protein target up to 84 scores are calculated. If a ligand cannot be placed into a pocket of an active site at all, no score is calculated.
These scores are used to calculate a normal distribution of scores for each target. As the exemplary score distributions of three targets in Figure 4.12 show, the scores indeed are almost normally distributed, allowing this approach.
The cutoff score (=gs-score) is then defined as the average score of all (for which a pose could be generated) 84 sc-scores. A cutoff at the average score plus one standard deviation was found to be too restrictive. This cutoff score is on the one hand used as a weight for the scores assessed in a screening. The *i*RAISE Gauss-weighted score (*gsw-score*) is thus the *i*RAISE Scoring Cascade score (*sc-score*) of a pose of the query ligand in a protein weighted with the Gauss cutoff score (*gs-score*):

$$gsw - score = \frac{sc - score}{gs - score}$$

Subsequently, poses are ranked by the gsw-score. On the other hand, the gsw-score can also be used to decide up to which point of a ranked list of proteins for a query ligand the scores still show true targets. Considering all proteins with gsw-scores greater than 1.0 as potential targets is a better way to assess, e.g., which targets to test in experiments than

using a hard cutoff of testing the first 10% of the ranked proteins.

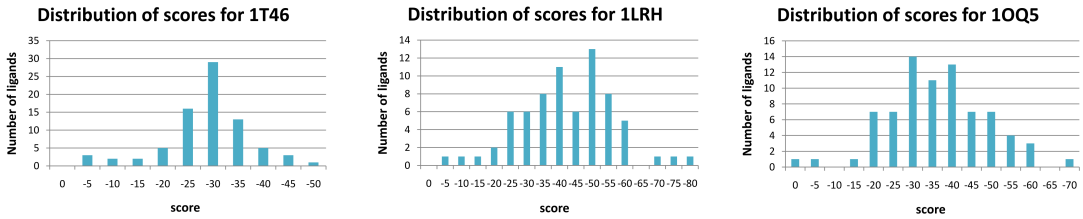


Figure 4.12: Examples for sc-score distributions of three targets screened with all 84 Astex Diverse Set ligands.

In Figure 4.13, the distribution of gs-scores is shown for a large set of nearly 8000 protein structures (sc-PDB, see chapter 5). Also, pockets of gs-scores at the lower bound, the middle and the upper bound of the distribution are shown.

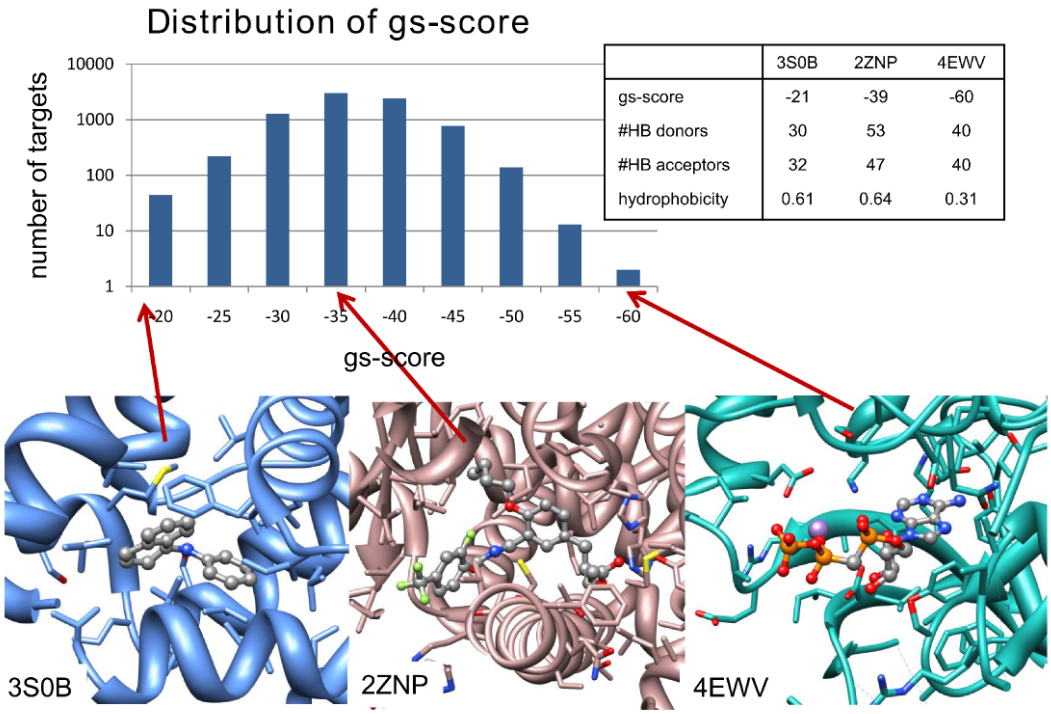


Figure 4.13: Distribution of target-specific gs-scores. The complex 3S0B with a minor gs-score of -21 is highly hydrophobic. The complex 2ZNP with an average gs-score of -39 is a larger pocket containing many hydrogen bond partners. The complex 4EWV with a high gs-score of -60 is hydrophilic and contains a metal ion. (#HB donors = number of hydrogen bond donors, #HB acceptors = number of hydrogen bond acceptors, hydrophobicity= number of hydrophobic amino acids of active site divided by total number of amino acids of active site).*(This figure was originally published in Schomburg et al. [2014]).*

4.7 Parallelization

Parallelization of *i*RAISE is realized on two levels. Data parallelization is in the scenario of *i*RAISE favored over task parallelization since the data (different protein targets and protein descriptors) is independent in itself and easily separable. The gain of parallelization of the computing tasks is not as beneficial.

The first parallelization level is the initial separation of data, i.e., protein structure data separation before the preprocessing procedure. The separated data can then be preprocessed in parallel, and several screening projects with separate descriptor indices and protein databases are created. These can then also be screened in parallel.

The second level of parallelization is realized during the preprocessing step on protein descriptor level. The bitmap index is partitioned during the descriptor calculation phase in order to follow FastBit's requirements of fitting one partition into working memory on the level of descriptor type partitioning and to reduce the number of matches which are kept in memory as well as on the level of protein structures. On average, about 100 proteins are stored in one protein structure partition.

In Figure 4.14, the parallelizations levels are shown in an overview. The screening part can be parallelized by screening several protein structure partitions concurrently. Each process creates a separate solution database in order to avoid write-locks. *i*RAISE therefore provides a method to consolidate all created solutions databases in one after all partitions are screened (see Appendix C).

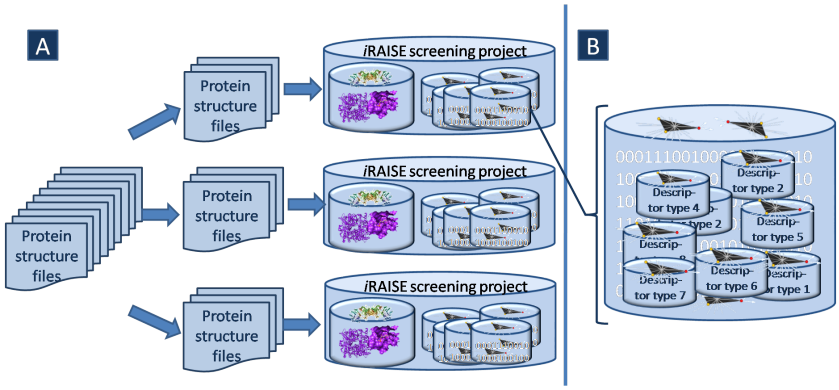


Figure 4.14: Overview of *i*RAISE's parallelization levels: (A) The first level is the separation of protein structure files, which are then preprocessed into separate screening projects. The second level is the separation of the descriptors into index partitions on about 100 protein structures. (B) The index partitions are further separated into descriptor type indices.

4.8 Graphic user interface: The ComplexViewer

*i*RAISE is a command line tool and does not provide a graphic user interface for the tasks of indexing (dataset preprocessing) and screening. See Appendix C for a user guide. However, a graphic interface for viewing the protein database content and browsing the solutions of *i*RAISE, named *ComplexViewer*, is provided (user guide: Appendix D).

A use case of viewing the content of a protein database with the ComplexViewer is shown in Figure 4.15. The content of the database is displayed as a list (Figure 4.15A). The ComplexViewer allows to select individual entries for a detailed view with a 2D structure diagram of the reference ligand and properties of the active site (protein name, reference ligand name, number of amino acids, number of atoms of pocket, number of hydrogen bond acceptors of the pocket, number of hydrogen bond donors of the pocket and the hydrophibicity). The pocket can also be assessed by a double click on the protein entry in a 3D-visualization in a separate window (Figure 4.15B). Per default, the reference ligand is shown with the residues of the active site. The display can be changed to also show water molecules, the protein backbone or the van-der-Waals spheres of the reference ligand.

In Figure 4.16, the second use case of the ComplexViewer, i.e., displaying screening solutions is shown.

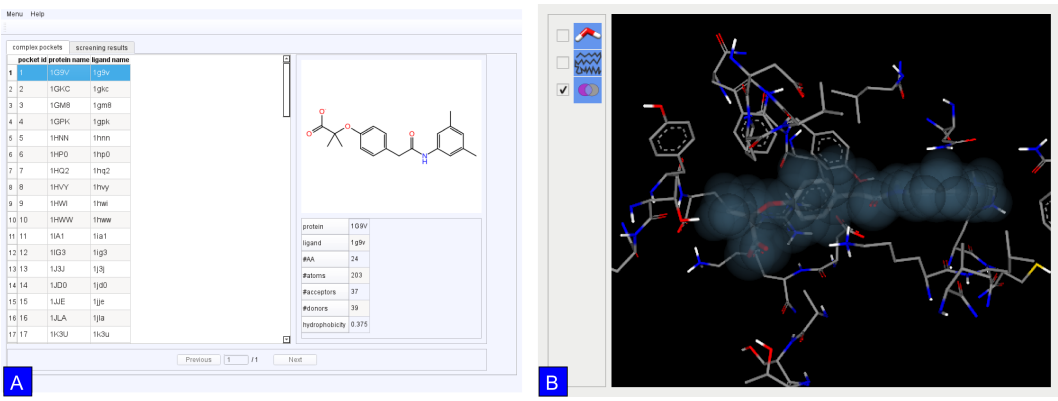


Figure 4.15: The ComplexViewer as a graphic viewer for protein database content. (A) The complex pockets tab shows the content of the protein database. The reference ligand and several pocket properties of a selected pocket are shown on the left. (B) The 3D visualization displays the pocket amino acids with the reference ligand. In the left panel, the options of showing waters, the protein backbone and the van-der-Waals spheres of the ligand are given. Here, the ligand is shown in the van-der-Waals sphere-mode.

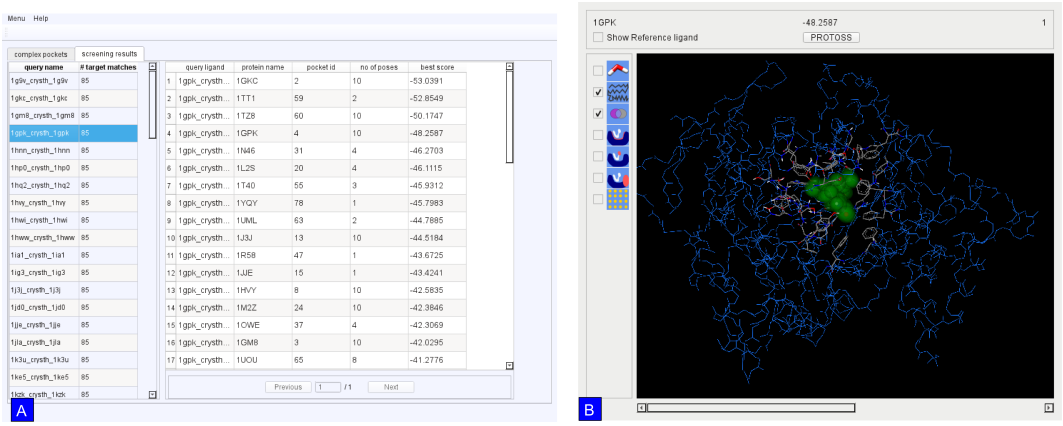


Figure 4.16: The ComplexViewer as a graphic viewer for browsing screening solutions of *i*RAISE. (A) The screening results tab contains two tables: A table containing all ligands with which the screening project was already queried, and on the right, the list of predicted targets of the selected ligand. (B) In the 3D window, the pose of a ligand in a selected target is shown. In the left panel, the options of showing waters, the protein backbone, van-der-Waals spheres for the ligand, the active site atoms, the uncovered ligand atoms, the uncovered pocket atoms and the grid (from top to bottom) are given. The slider at the bottom provides a switch between different ligand poses in the same protein. At the top, the name of the current protein, the score and ranking of the current pose is given. A button labeled PROTOSS allows to align the hydrogen atoms of the current pose optimally. Further the option to simultaneously display the reference ligand is given.

The screening result tab contains two lists: Firstly, the list of ligands with which the *i*RAISE project has already been screened (Figure 4.16A), and secondly for one chosen ligand its list of predicted protein results with the name, IDs and score of the best scored pose. A selection of an entry of this result list with a double click shows a 3D presentation of the selected protein with the ligand pose (Figure 4.16B) in a separate window.

The display per default shows the best pose of the ligand with the pocket atoms. The name of the protein, the score of the pose and the number of the pose are given in the header line. Several options are available for changing the displayed complex: At the top, it can be selected whether the reference ligand is shown as well or not. A PROTOSS button can be used to optimally align the hydrogen atoms of the currently shown complex. At the left, the view can be changed by adding waters, the protein backbone, the van-der-Waals spheres of the ligand, the active site atoms, the uncovered ligand or pocket atoms and the grid. A slider at the bottom can be used to browse through the poses of the ligand in this protein, ordered by score.

5

Data Sets

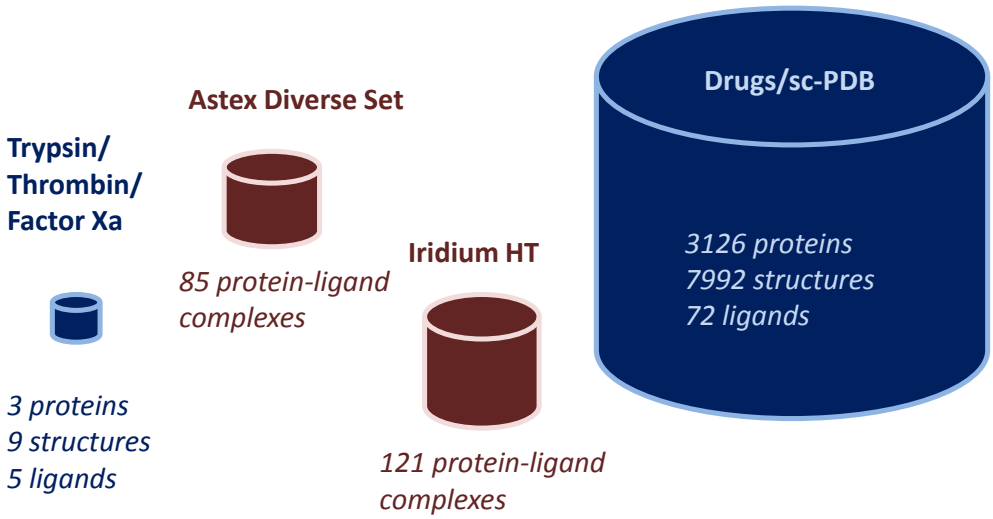


Figure 5.1: Overview of data sets with different sizes used for inverse screening. (Astex Diverse Set: Hartshorn et al. [2007], Iridium-HT: Warren et al. [2012]

The development of computational models depends highly on available experimental data. Data is needed for understanding the concepts of nature. In structure-based computational methods, the binding of a small molecule to the active site of a protein is the center of attention. 3D structures of proteins, ligands and protein-ligand complexes help to gain insights into the mechanism of complex formation. For the development of structure-based target identification methods, ideally, a data set with 3D structures of different proteins bound to the same ligand is available. Furthermore, experimentally determined activity data is required to show which proteins a ligand has an effect on and also on which it has none. 3D structures of protein-ligand complexes are experimentally determined by X-ray crystallography or Nuclear Magnet Resonance (NMR) analysis. In general, it is assumed that if an experimentally determined structure of a protein-ligand complex exists, the ligand binds to this protein. Binding affinity measurements support this thesis.
Since computational concepts are based on the knowledge derived from protein-ligand complexes, this data needs to be reliable. Nowadays, structures determined by X-ray crystallography are still preferred to those solved by NMR. NMR structures of proteins are still limited in the size of proteins that can be solved, but more importantly, the quality of NMR structures still is debatable in some cases as the assessment of the quality of these structures is not easy (Spronk et al. [2003], Nabuurs et al. [2003]).
There are several factors which help to estimate the quality of a 3D protein-ligand complex which was solved by X-ray crystallography. In X-ray crystallography, the electron density of the atoms is experimentally measured. The quality of the electron density varies in the protein-ligand complex, depending on the quality of the protein-ligand crystal, the experimental setup and type of X-ray and internal movements of parts of the protein.
It is often assumed that less ordered parts of the protein, e.g., not stabilized loops, lead to poorly resolved electron density. The less clearly resolved density around some atoms is mapped in a factor for each atom, describing the vibration energy of these atoms. Then there is a factor which rates the agreement of the model of the structure and the measured electron density in total. Another factor that can be used as quality criterion is the resolution of the structure. This factor illustrates up to which distance two points can be distinguished from each other. Therefore, a smaller resolution indicates a better structure; however, since the resolution is averaged over the complete structure, no direct conclusion can be derived for the active site of a protein-ligand complex.
The review of Davis, Teague and Kleywegt discusses these terms and highlights the use of X-ray crystallographic data in structure based computational approaches (Davis et al. [2003]). Errors and challenges of structure determination by X-ray crystallography are discussed with focusing on the ligand by Malde and Mark (Malde and Mark [2011]). Stereochemistry, orien-

tation, tautomeric state, protonation and conformation are not always clearly resolved in the available structures. For proteins, these problems arise in determining the states of amino acids. Also, the positions of water molecules are usually not clear. Another factor that has to be considered concerns the crystallization conditions, which are often very different to the biological conditions, where the protein-ligand complex is formed.
The fact that 3D structures of proteins and ligands are only models themselves and have many limitations has to be considered whenever using these structures. However, keeping the problems in mind, protein-ligand complex data is still the best data available when designing and evaluating computational structure-based methods.

Next to the quality of the data, the composition has to be considered. In structure-based computational approaches, the requirements for data vary with different purposes and steps during the development of a new approach. The needs for data range from data for method development over validation data sets to applications of the method. In computational structure-based inverse screening method development, protein-ligand complex data is needed for the purposes listed below:

- **Method development**
  During method development, concepts and theories are derived from data. Theories on how the formation of protein-ligand complexes can be modeled computationally, how the strength of the binding of a ligand to the active site of a protein can be estimated and where the limitations of the model are found are evaluated on experimental data. For this purpose, a small data set is suitable which is in the best case very diverse in chemical composition. A small data set allows manual evaluation of each case and a rapid testing during method development. However, as the data set has to represent the huge chemical space, it has to be very diverse. Even if the chemical space is sampled diversely in the set, it, though, can never cover each property occurring in the space. Nevertheless, if each data point represents another chemical class, many important features can be assessed with the data set. In this dissertation project for immediate evaluation during method development, the Astex diverse data set was used (see section 5.1).

- **Parametrization**
  Once concepts are developed and implemented, the parameters of the model need to be evaluated. Each computational model has parameters, as it is only a model, and often the parameters determine the level of abstraction. The requirements of

a data set for parametrization are similar to the requirements for data for method development. The data set needs to be chemically diverse to sample the chemical space reasonably. The size of the data set can be larger, since parametrization is done automatically with scripts and not each data point is evaluated manually.
For this research project, parametrization was conducted on the Iridium data set (see section 5.2).

- **Proof of concept**
  The aim of inverse virtual screening is the identification of targets for a compound. As a proof of concept that the computational model is capable of correct predictions, a data set is required, where each protein is annotated as target or no target for a compound. Since true negatives are often problematic in the available data as has been discussed before, this data set is crucial for evaluation if the method is not only capable to predict true positives, but also if it is able to predict true negatives. At the time of this dissertation project, no data set was available for this purpose, which is why the small data set "TTFXa" was composed during the dissertation project. Section 5.4 describes its composition and creation.

- **Validation**
  The validation of inverse screening methods has to show two aspects: Firstly, the performance and quality of the method has to be validated, meaning that the ability of the method to find true targets for a compound in a large protein structure data set has to be evaluated. In order to show statistical relevance, this evaluation has to be done for a suitable amount of diverse ligands.
  Secondly, since inverse screening methods are designed for screening large numbers of proteins in a reasonable amount of time, experiments on large data sets need to show the time requirements of the method. A data set suited for these purposes did not exist prior to this dissertation project, either. Therefore, for validation the "Drugs/sc-PDB data set" was constructed, as will be described in section 5.5.

- **Comparison to other methods**
  A new method should be compared to other methods of the field for showing the impact of the differences in the method on results and performance. Therefore, if other published methods used available data for their validation, new methods should also run the experiments on this data to allow a comparison. The Astex Diverse Set was used for comparison to reverse classic docking. The sc-PDB Diverse Set (see section 5.3) was used for comparing the method developed in this thesis to a pharmacophore-based method by Meslamani et al. (Meslamani et al. [2012]) as well

Name	Number of complexes	Source	Composition
Astex Diverse Set	85	Hartshorn et al. [2007]	diverse high quality complexes
Iridium-HT	121	Warren et al. [2012]	high quality complexes, redefined
Drugs/sc-PDB	7992	own development based on the sc-PDB (Meslamani et al. [2011]) and the DrugBank (Wishart et al. [2006])	large data set for statistical validation of inverse screening on drug-like ligands
Trypsin/Thrombin/ Factor Xa	9	own development	small data set for proof of concept studies

Table 5.1: Overview of data sets

as to two classic docking approaches. The newly developed Drugs/sc-PDB data set was used for comparison to a sequence-based target prediction method.
Other structure-based approaches like TarFisDock or Invdock compiled own structure data sets of which the structures of the proteins were not made publicly available, which renders them infeasible for comparison. In Section 2.2.3 the problems with the so far used data sets are discussed.

- **Application examples of the method**
  Once a new method is developed and validated, it can be applied to 'real world problems'. This dissertation investigates how inverse screening and structure-based computational methods in general are applicable to biotechnology for the construction of a synthetic multi-enzyme pathway. The description of the data used for this study can be found in chapter 8.

Table 5.1 gives an overview of the data sets discussed above. The next sections shortly describes the mentioned data sets and in detail discusses the compilation of the data sets that were newly created during the course of this thesis.

5.1 Astex Diverse Set

This data set was composed by Hartshorn et al. in 2007 (Hartshorn et al. [2007] as an evaluation test set for docking. It consists of 85 high-resolution crystal protein-ligand complexes. Hartshorn et al. used protein-ligand complexes of proteins thought to be relevant in pharmaceutical research and of ligands which are drug-like. They clustered structures by sequence and chose high quality structures as representatives of the clusters. They manually curated the data set, i.e., assigned protonation states. In the 85 protein-ligand complexes, twice the ligand trypsin is contained (in 1OF6 and 1X8X), thus the number of unique ligands is 84.

5.2 Iridium Data Set

The Iridium data set was composed by Warren et al. in 2012 (Warren et al. [2012]). Sources for complexes were four different data sets which are used to validate docking studies, one of which was the above described Astex Diverse Set. The 728 complexes of all four data sets were redefined, i.e., the structure model was newly built based on the electron density. Then the complexes were divided into three categories based on quality criteria of the protein and the ligand structure. Only 121 structures with complete density for the ligand were classified as highly trustworthy. The others were classified as mildly trustworthy or not trustworthy. In this thesis, only those complexes that were classified as highly trustworthy were used.

5.3 sc-PDB Diverse Set

The sc-PDB Diverse Set (Meslamani et al. [2012]) consists of the sc-PDB protein structures data set and 157 diverse ligands. The ligands are a diverse subset of the co-crystallized ligands of the sc-PDB. During the creation of the ligand set, the sc-PDB version 2010 was used. Since this version was no longer available during the time of writing this thesis, the sc-PDB version of 2012 was used instead. Since this version though does not contain all the complexes of all the 157 ligands, a subset of 117 ligands was used in this thesis (see Appendix H for the list of 117 HET codes).
For true positive assignment, two approaches were followed: Firstly, as proposed by Meslamani et al., true positives were assigned via the UniProtID. For this approach, the UniProtIDs were mapped to PDB codes following the list published together with the sc-PDB

2012 (http://cheminfo.u-strasbg.fr:8080/scPDB/2012). Proteins that were co-crystallized with the ligand or had the same UniProtID as a co-crystallized target were defined as true positives. Secondly, additionally to the UniProtID-based true positives, protein structures with the same EC number were defined as true positives. In contrast to the UniProtID, the EC number is not organism specific, but does nevertheless classify the same protein.

5.4 Trypsin/Thrombin/Factor Xa–Data set

As proof of concept data set and for a detailed study of the capabilities of a new method, a tiny data set with defined true negatives and true positives was compiled. For this data set, serine proteases were used as target class. Serine proteases are a protein class functioning as protein-cleavage enzymes. They have a serine amino acid in the catalytic triad of the active site in common, which is essential for the catalytic reaction.
For the data set, three proteins from the sub-class of trypsin-like serine proteases were chosen. Enzymes from this subclass are found in digestive processes, blood coagulation and immune responses. The three enzymes chosen are trypsin (EC 3.4.21.4), thrombin (EC 3.4.21.5) and factor Xa (EC 3.4.21.6). Trypsin is an enzyme found in the digestive system for cleaving proteins which were ingested by food. Thrombin (also called blood-coagulation factor IIa) is a protein of the coagulation cascade and converts fibrinogen to fibrin, next to activating other blood-coagulation factors. Therefore, it is a key player in inflammation reactions and wound healing. Factor Xa (also called Stuart-Prower Factor) plays a major role in the early stages of the coagulation cascade, where it activates thrombin, next to other effects.
These three proteins were chosen for several reasons: First of all, they have the same overall structure. If structurally aligned, the structure of the protein backbone is nearly identical, with only a loop of thrombin burying the active site deeper than in the structures of factor Xa and trypsin. Figure 5.2 shows the alignment of three structures of each protein, with an ellipse highlighting the differing loop region. Sequence-based methods have proven to perform poorly on this protein class (Glinca and Klebe [2013]). However, the amino acid composition of the active sites differ among the three proteins. All have the same amino acids of the catalytic triad (Asp102, His57, Ser195), but of the other amino acids of the active site, about 6 are varying in the three proteins. There are many studies which in detail elucidate the similarities and differences of trypsin, thrombin and factor Xa structures and describe the design of selective inhibitors (Czodrowski et al. [2007], Di Fenza et al. [2007], Nar et al. [2001], Böhm et al. [1999], Stubbs et al. [1995]). The reader is referred to these studies, since the data set is only used for retrospective studies in this context and the details

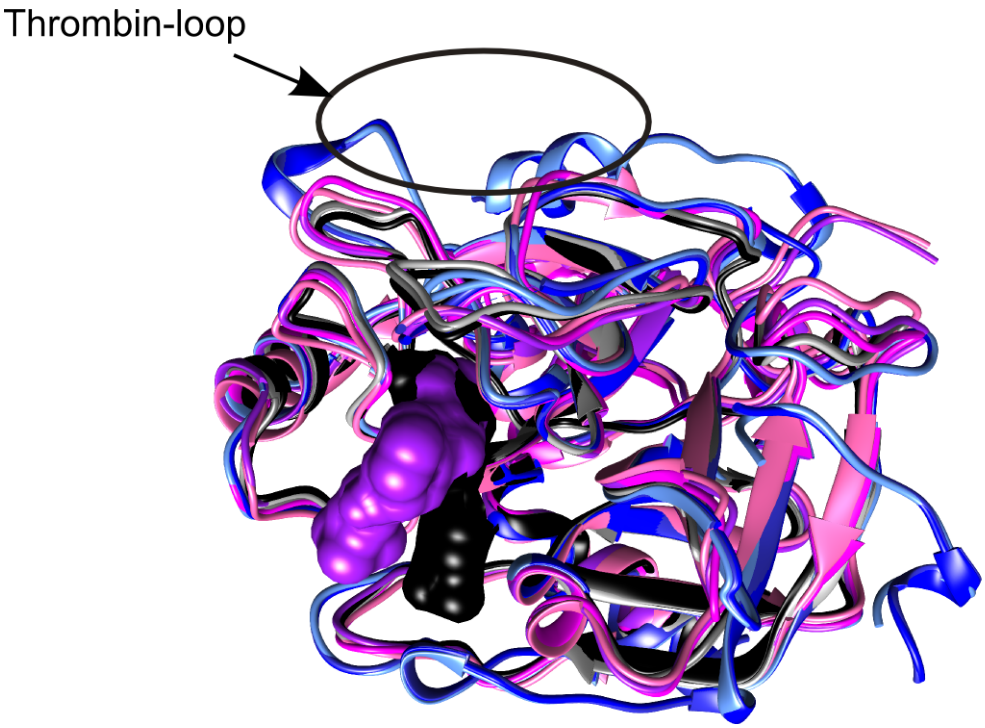


Figure 5.2: Alignment of nine serine protease structures, three structures per protein: Thrombin structures are shown in differing shades of blue (PDB codes: 3RM2, 2BVR, 3SI4), trypsin structures in differing shades of gray (PDB codes: 3GY2, 2G8T, 2G5N) and factor Xa structures in differing shades of pink (PDB codes: 2JKH, 2Y5F, 3KL6). The thrombin-specific loop is highlighted with an ellipse.

for designing selective ligands are not essential. The differences in active site amino acids and general structure is shown by Böhm et al. in a 2D sketch (Böhm et al. [1999]).
With this setup, using the data set in docking studies is not trivial, as the proteins are structurally highly similar. Furthermore, the relevance of these proteins in drug discovery also supports their use in validation studies. The design of selective inhibitors of thrombin or factor Xa which do not target trypsin is of high interest. Thirdly, some inhibitors with well defined activities are known for these proteins, which is needed to complement the data set with ligands.
Five ligands with differing activities were chosen for the data set: Benzamidine (benzenecarboximidamide) and Pefabloc (4-(2-aminoethyl)benzenesulfonyl fluoride) as general serine protease inhibitors, Apixaban (1-(4-methoxyphenyl)-7-oxo-6-[4-(2-oxopiperidin-1-yl)phenyl]-4, 5-dihydropyrazolo[3,4-c]pyridine-3-carboxamide) and Rivaroxaban (5-chloro-N-[[(5S)-2-

oxo-3-[4-(3-oxomorpholin-4-yl)phenyl]-1, 3-oxazolidin-5-yl]methyl]thiophene-2-carboxamide) as specific factor Xa inhibitors and Melagatran (2-[[(1R)-2-[(2S)-2-[(4-carbamimidoylphenyl) methylcarbamoyl]azetidin-1- yl]-1-cyclohexyl-2-oxoethyl]amino]acetic acid) as specific thrombin inhibitors. The inhibitors were obtained by querying the drug bank for inhibitors of the three proteins. The structures of the ligands were collected from PubChem (Bolton et al. [2008]). Figure 5.3 shows the structure diagrams of the five ligands.
The structures for the proteins were collected from the Protein Data Bank. For each protein, three structures were chosen to account for protein flexibility in the data set. The structures were chosen for resolution and co-crystallized ligand: The protein had to be crystallized in a protein-ligand complex, but the ligand was not allowed to be one of the five inhibitors used for screening in the data set. For trypsin, the structures 3GY2, 2G8T, 2G5N, for thrombin the structures 3RM2, 2BVR, 3SI4 and for factor Xa the structures 2JKH, 2Y5F and 3KL6 were chosen.

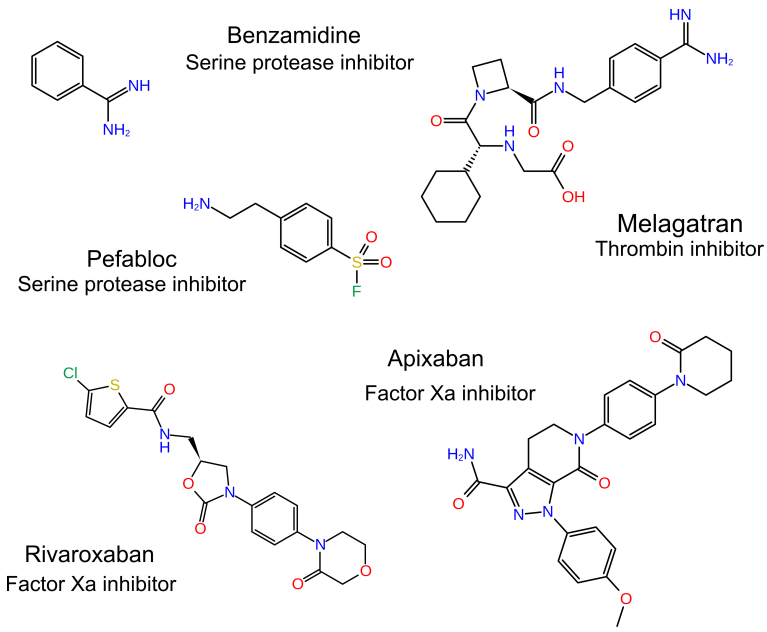


Figure 5.3: Inhibitors of serine proteases used in the TTFXa-Data set as ligands. Two inhibitors are general serine protease inhibitors, two are factor Xa inhibitors and one is a thrombin inhibitor.

5.5 Drugs/sc-PDB

Since no established standard data set was available for large scale statistical validation of structure-based computational target prediction methods, the creation of a new set was necessary. The requirements for a large-scale validation set for target prediction methods are the following:
The data set has to contain complexes covering diverse ligands, diverse targets, high resolution structures, true negatives and true positives. The numbers of targets and ligands has to be sufficient to allow statistical evaluation.
The validation of current state of the art methods was mainly done on data that contains no reliably assigned true negatives since this data is not easily available for most compounds. As a consequence, an expensive experimental validation subsequent to the predictions was necessary. The challenge in compiling a data set with determined true positives and true negatives lies in the determination of true negatives. In many activity studies, only data is reported, on which proteins a compound has an effect, and not, on which it has none. Especially if huge sets of proteins shall be screened with one compound, the information if a protein is a target for that compound is only available for a small set of these proteins. However, using only a few determined data points for the validation of a method is not sufficient. Therefore, the composition of a data set with true negatives was necessary.
For the compilation of such a data set, the fact that drugs are well-studied in selectivity was exploited:
It was assumed that since drugs have to pass several stages of selectivity and toxicity tests, they are rather selective and their true targets are better known than for any other compound class. Clearly, as the drug repositioning projects show, not each target is known for each drug. However, next to the well studied selectivity of drugs, the usage of drugs in a data set also has the advantage that the method is tested on compound classes, which are relevant in pharmaceutical research.
With drugs as compounds, the target data set has to contain proteins which are targeted by the compounds next to not-targeted proteins. Furthermore, the data set shall not be biased towards special target classes. Therefore, instead of compiling a data set of the targets of the drugs and enriching it with target decoys, the sc-PDB was used as target data set (Meslamani et al. [2011]).
This data set has several advantages. Firstly, it represents publicly available structures since it was composed from the Protein Data Bank. Secondly, it contains high quality data, as the authors applied filters for resolution and quality of the complexes. Thirdly, the number of targets is sufficiently high for a validation set with about 8000 or 9000 targets, depending

on the version of the sc-PDB data set. And finally, the targets are, as the authors state, 'druggable' since the authors applied drug-like filters to the co-crystallized ligands. Therefore, it seems reasonable to use this data set as a source for the protein target structures.

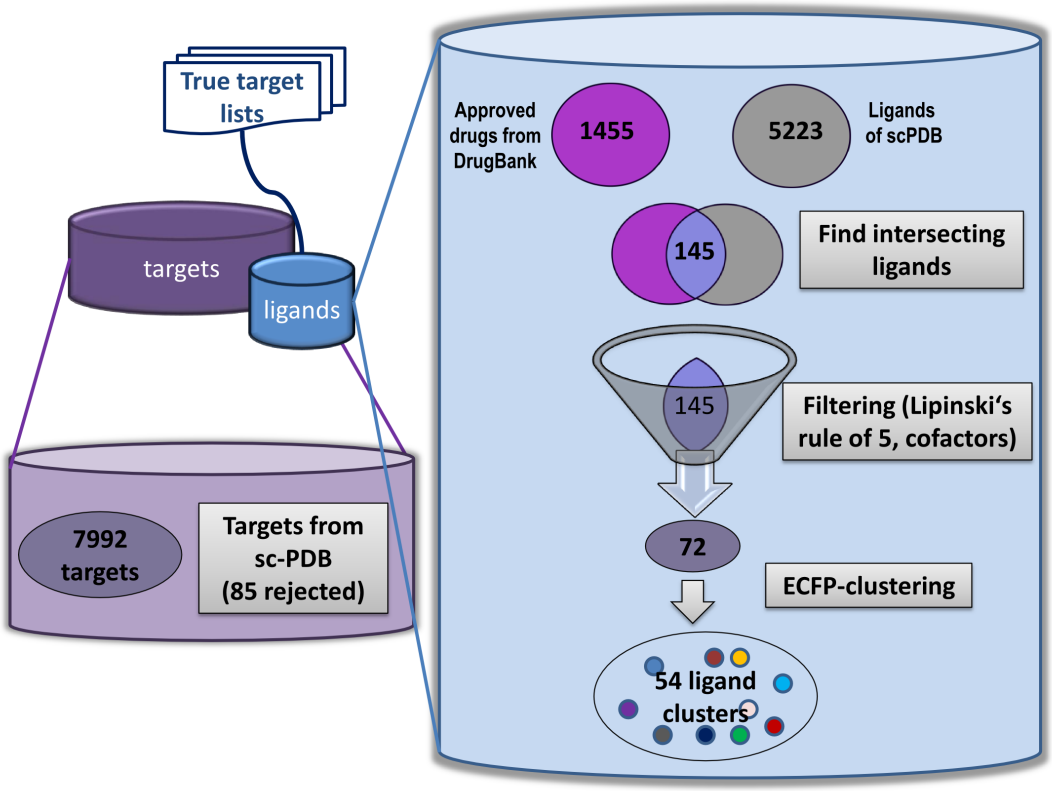


Figure 5.4: Composition and creation of the Drugs/sc-PDB data set. *(This figure was originally published in Schomburg and Rarey [2014].)*

Figure 5.4 shows how the Drugs/sc-PDB data set is composed. The targets' origin is the sc-PDB data set. The version of 2012 originally contained 8077 protein-ligand complexes. The original PDB files were downloaded from the PDB (Berman et al. [2000]) instead of using the prepared sc-PDB files. Of the 8077 complexes, 7992 were chosen for the target database. The remaining 85 were discarded due to obsolete PDB codes in the Protein Data Bank (9), errors in the reference ligand (25) and problems during initialization with NAOMI which can be due to wrong chemical composition of the reference ligand or to rare metals bound to the ligand, which cannot be initialized by NAOMI (51). NAOMI is the chemical library used (see section 4.1 for further details) for compound initialization with strict chemical rules (Urbaczek et al. [2011]).

The ligands are selected by the following steps:

1. Firstly, sets of unique ligands of approved drugs from the DrugBank (Wishart et al. [2006]) and the ligands of the sc-PDB are built. The number of unique approved drugs from the DrugBank is 1455 while there are 5223 unique ligands in the sc-PDB data set.

2. Secondly, the intersection of the compound set of unique approved drugs and the unique ligands of the sc-PDB is built. Ligands are considered equal even if they differ in tautomeric or protonation state. The intersection results in 145 ligands.

3. Thirdly, the resulting ligands are filtered. Lipinski's rule of 5 leaves 81 ligands. Further, thiamine and biotin are excluded since they function as cofactors. Therefore, for many true targets for these compounds, the 'wrong' pocket would be presented in the data set, as not the cofactor, but the substrate binding site was chosen. Therefore, a definition of true targets would be difficult. Furthermore, 6 ligands are excluded since the targets which are listed in the DrugBank are DNA or RNA. This procedure leaves 72 ligands.

4. Finally, since some of the ligands are structurally highly similar and only vary in substituents, the ligands are clustered with ECFP fingerprints. ECFP fingerprints are circular topological fingerprints (Rogers and Hahn [2010]). They describe molecules on the substructural level. Each atom gets a descriptor of its substructural neighborhood, with a given diameter of atoms. For this clustering, a diameter of 3 was chosen. The similarity of two molecules was calculated using the Tanimoto coefficient. As a similarity threshold 0.2 was chosen, since manual evaluation showed that this threshold gave results, which resemble chemical similarity as a chemist would expect. In Figure 5.5, all 38 molecules are shown, which build single clusters. In Figure 5.6 the clusters of molecules with 2 or more members are shown.

For filtering, intersectioning and unifying, MONA (Hilbig et al. [2013]) was used.

Finally, for completion of the data set, for each of the drug ligands, all the 7992 protein structures have to be annotated as true or false targets. A structure of the sc-PDB is classified as a true target if it meets one of the following criteria:

- The protein structure is co-crystallized with the drug compound.

- The protein structure has the EC number of a target of the drug compound listed in the DrugBank
- The protein structure has the name of a target of the drug compound listed in the DrugBank
- The protein structure has the UniProtKB ID of a target of the drug listed in the DrugBank

In order to be able to classify the protein structures via UniProtKB ID, a mapping of the PDB code and UniProtKB ID had to be assigned. This mapping was conducted using the SIFT structure annotation service (Velankar et al. [2013]). The true target lists of PDB codes for each drug can be found in Appendix E. In Figure 5.7 an overview of the number of different proteins and number of true positive structures for each ligand is given.

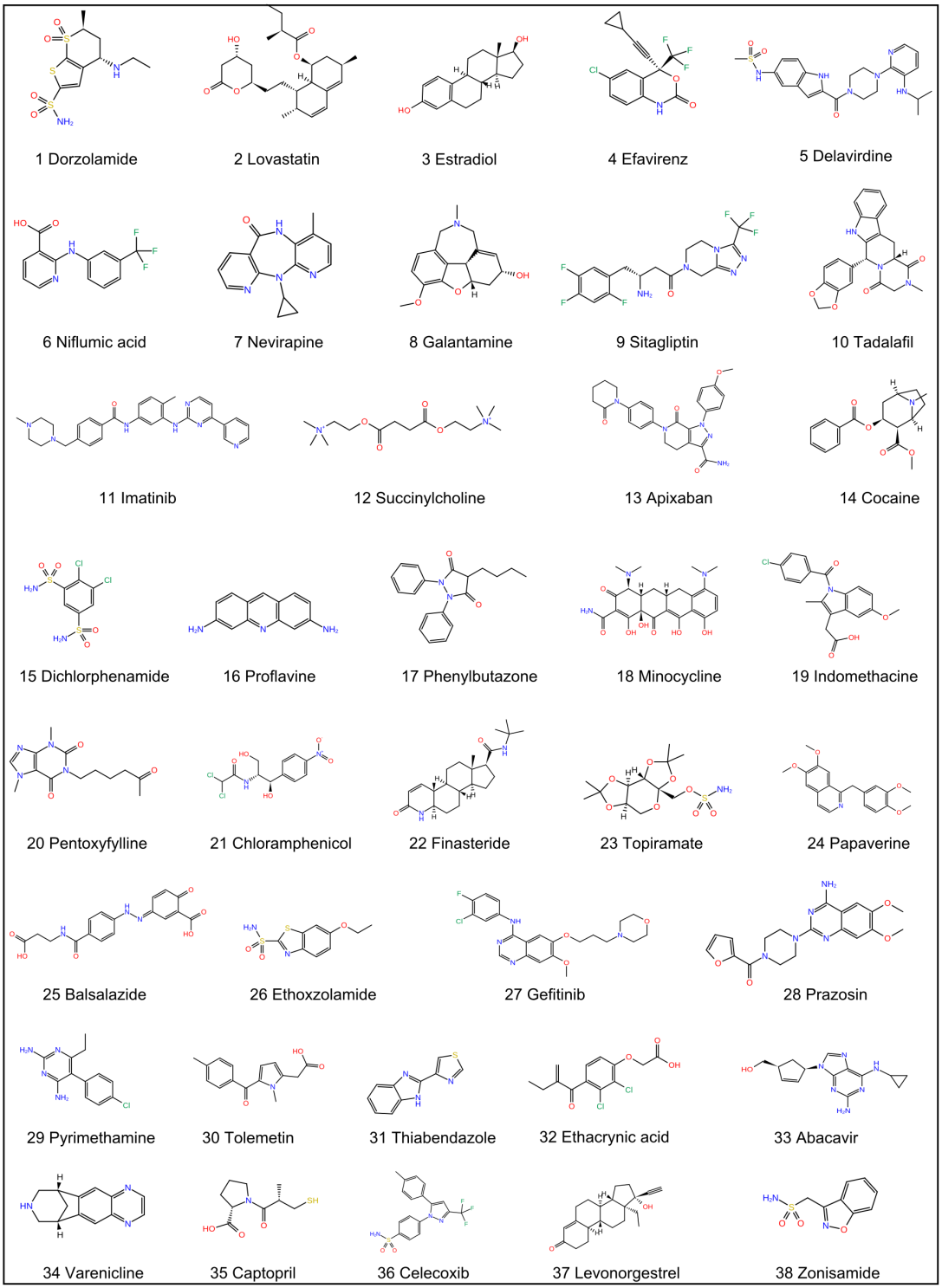


Figure 5.5: Part 1 of the ligands of the Drugs/sc-PDB data set

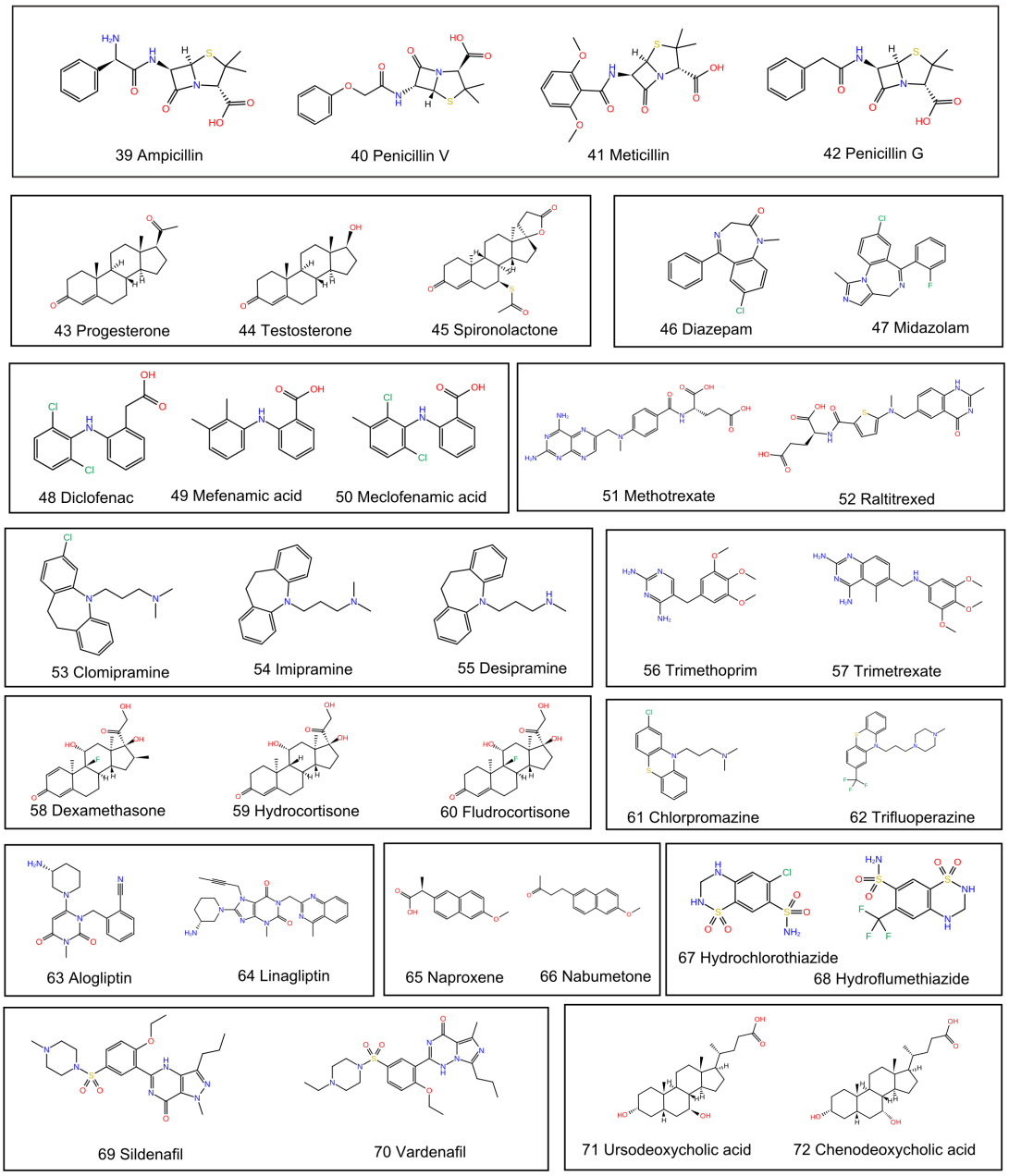


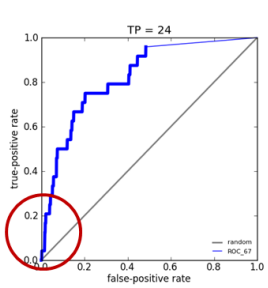
Figure 5.6: Part 2 of the ligands of the Drugs/sc-PDB data set: Clusters of two or more drugs

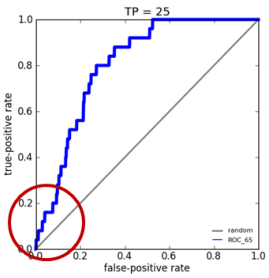
Nr	Drug name	Number of protein targets listed in the DrugBank	Number of true positive protein structures in sc-PDB	Nr	Drug name	Number of protein targets listed in the DrugBank	Number of true positive protein structures in sc-PDB
1	Dorzolamide	3	152	37	Levonorgestrel	4	107
2	Lovastatin	3	29	38	Zonisamide	5	161
3	Estradiol	4	80	39	Ampicillin	2	17
4	Efavirenz	1	114	40	Penicillin V	3	36
5	Delavirdine	1	114	41	Meticillin	1	3
6	Niflumic acid	4	53	42	Penicillin G	1	3
7	Nevirapine	1	114	43	Progesterone	3	90
8	Galantamine	3	37	44	Testosterone	3	54
9	Sitagliptin	1	48	45	Spironolactone	2	31
10	Tadalafil	2	91	46	Diazepam	3	13
11	Imatinib	10	188	47	Midazolam	2	2
12	Succhinylcholine	2	3	48	Diclofenac	6	71
13	Apixaban	1	87	49	Mefenamic acid	2	23
14	Cocaine	4	2	50	Meclofenamic acid	2	22
15	Dichlorphenamide	4	152	51	Methotrexate	8	165
16	Proflavine	3	141	52	Raltitrexed	2	51
17	Phenylbutazone	3	22	53	Clomipramine	7	23
18	Minocycline	7	7	54	Imipramine	7	31
19	Indomethacine	8	153	55	Desipramine	8	31
20	Pentoxifylline	3	98	56	Trimethoprim	2	160
21	Chlormaphenicol	3	51	57	Trimetrexate	1	125
22	Finasteride	2	9	58	Dexamethasome	5	13
23	Topiramate	3	152	59	Hydrocortisone	2	8
24	Papaverine	1	91	60	Fludrocortisone	3	38
25	Balsalazide	4	97	61	Chlorpromazine	6	6
26	Ethoxzolamide	1	152	62	Trifluoperazine	8	42
27	Gefitinib	1	140	63	Alogliptin	1	48
28	Prazosin	1	5	64	Linagliptin	1	48
29	Pyrimethamine	1	125	65	Naproxene	2	25
30	Tolmetin	3	25	66	Nabumetone	2	24
31	Thiabendazole	1	1	67	Hydrochlorothiazide	1	153
32	Ethacrynic acid	2	15	68	Hydroflumethiazide	1	153
33	Abacavir	3	137	69	Sildenafil	1	38
34	Varenicline	1	1	70	Vardenafil	1	39
35	Captopril	3	31	71	Ursodeoxycholic acid	1	22
36	Celecoxib	3	65	72	Chenodeoxycholic acid	1	19

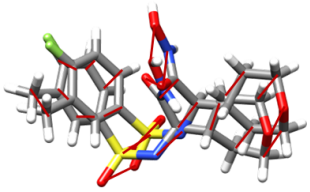
Figure 5.7: List of true protein targets and number of true protein structures for the 72 ligands of the Drugs-sc-PDB data set. In shades of purple, drug clusters are highlighted.

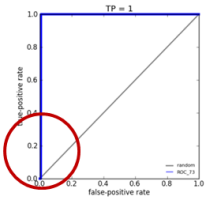
6

Evaluation Strategy and Experiments









Rank	Positive?	code	description
1	yes	1ahg	PHOSPHO-5'-PYRIDOXYL TYROSINE
2	yes	2o2u	MITOGEN-ACTIVATED PROTEIN KINASE 10
3	no	2ztu	D(-)-3-HYDROXYBUTYRATE DEHYDROGENASE
4	no	18gs	GLUTATHIONE S-TRANSFERASE
5	no	3euf	URIDINE PHOSPHORYLASE 1
6	yes	1g9h	ALPHA-AMYLASE
7	no	3it3	ACID PHOSPHATASE
8	no	2i62	NICOTINAMIDE N-METHYLTRANSFERASE
9	no	3dzi	ADP-RIBOSYL CYCLASE 1
10	no	3gqi	PHOSPHOLIPASE C-GAMMA-1
11	no	1ork	TETRACYCLINE REPRESSOR
12	no	2q8s	PEROXISOME PROLIFERATOR-ACTIVATED RECEPTOR
13	no	3r3s	OXIDOREDUCTASE
14	no	5upj	HIV-2 PROTEASE
15	no	1a4r	G25K GTP-BINDING PROTEIN

Figure 6.1: Evaluation strategies cover enrichment metrics, RMSD calculations and rank evaluation.

In this chapter, the evaluation strategy is described. So far, no standard evaluation strategy has been established for inverse virtual screening. Therefore, the evaluation strategy consisting of the evaluation experiments, data sets and choice of performance metrics was newly compiled. While the data has been described in the previous chapter (5), here the experiments and evaluation parameters are discussed.
The aim of the experiments is to thoroughly evaluate the two distinct features of inverse screening: Binding mode prediction and target ranking. In addition, the experiments shall evaluate the performance of *i*RAISE compared to classic docking, pharmacophore-screening and sequence-based target prediction.
In total, the evaluation consists of six experiments:

1. Experiment: Binding mode prediction
2. Experiment: True-target Ranking
3. Experiment: Sensitivity versus selectivity
4. Experiment: Early enrichment
5. Experiment: Comparison of ranking capability with classic docking
6. Experiment: Comparison of enrichment to sequence-based method
7. Experiment: Comparison of enrichment to pharmacophore-based method

In this chapter, initially, the evaluation criteria and performance metrics are introduced and discussed. Then, the evaluation experiments are described. For results and discussion of the experiments, see Chapter 7.

6.1 Evaluation criteria and measures

Evaluation criteria and measures need to assess two points: Firstly, the ability of the method and its limitations need to be monitored in an objective, reproducible way, preferably in comparison to other methods. Secondly, the evaluation has to show to potential users in which case they benefit from using the method but also where the limitations of the model are.
Therefore, the evaluation is conducted on both, on artificial experiments solely usable to compare methods among each other (e.g. re-docking experiments) and on experiments which resemble real use cases (e.g. finding off-targets of approved drugs, see enrichment experiments).

Two capabilities of *i*RAISE need to be covered by experiments: Firstly, the ability of predicting binding modes need to be shown. The simultaneous prediction of protein targets with the binding modes of the query ligand is one great advantage of structure-based inverse screening methods. Secondly, the ability of ranking true targets to the early positions of large protein sets, also called *early recognition* needs to be assessed.

Binding mode evaluation measures For evaluation of the correct binding mode, calculating the RMSDs of re-docking experiments is an established measure. In re-docking experiments, predicted poses are compared concerning the atom coordinates with the co-crystallized position of the ligand in its target. An established cutoff of successful binding mode prediction is an RMSD below 2 Å.

Enrichment evaluation measures For evaluation of the enrichment power of virtual screening methods, several metrics exist. Here, the use of several metrics was chosen for evaluation, since each has its own advantages and disadvantages. Furthermore, recently, concern has been raised in the chemical computing community on the thoroughness of evaluation concerning the use of metrics and data sets (Jain and Nicholls [2008], Kirchmair et al. [2008]), calling for the use of more than one metric.
Established is the use of Enrichment Factors and the AUC (=Area under the ROC curve), which are easily interpretable. For the assessment of the very-early recognition problem, the BEDROC and the NSLR were chosen additionally since they are based on different weighting schemes (exponential versus logarithmic).
Following is a list of all metrics chosen for evaluation of enrichment capability and their definition:
(a = actives found at considered fraction of data set,
A=total number of actives in data set,
n=Number of data points at fraction of data set (actives+inactives) screened,
N=total number of data points,
r_i = rank of ith active.)

- **EF** (Enrichment Factor)

$$EF(\alpha) = \frac{a}{n} \times \frac{N}{A} \tag{6.1}$$

  The Enrichment Factor is a measure of the fraction of actives found in a fraction of the database. The fraction (α) is e.g. 1%, 2%, 10% of the data set. The EF has three main disadvantages: The EF is dependent on the fraction of actives and decoys

and thus cannot be compared for different experiments. Also, the need to set a factor leads to different reported EFs in publications, dependent on the choice of authors. Further, the EF is not bound in its values. An advantage is its straightforward concept. For *i*RAISE's evaluation, the EF at 1%, 2% and 5% is used.

- **AUC** (Area under the ROC curve)
  The ROC curve (=Receiver operator characteristics) is the number of found actives plotted against the number of decoys. The AUC is calculated as the area under this curve.

$$AUC = \frac{1}{nN} \sum_{i=2}^{N} A_i(I_i - I_{i-1}) \tag{6.2}$$

  A_i = found actives at rank i
  I_i = found inactives at rank i
  The AUC has the advantages of being parameter-free and being bound by 0 and 1. An AUC of 0.5 means random performance. The AUC has the disadvantage that it does not assess early enrichment.

- **BEDROC** (Boltzmann-enhanced discrimination of ROC) (Truchon and Bayly [2007])

$$BEDROC(\alpha) = \frac{\sum_{i=1}^{n} e^{(-\alpha r_i/N)}}{\frac{n}{N}(\frac{1-e^{-\alpha}}{e^{\alpha/N}-1})} \frac{\frac{n}{N} e^{\alpha n/N}(e^{\alpha}-1)}{(e^{\alpha}-e^{\alpha n/N})(e^{\alpha n/N}-1)} + \frac{1}{1-e^{\alpha(1-\frac{n}{N})}} \tag{6.3}$$

  The BEDROC metric weights the ranks of true positives exponentially decreasing and therefore can be used to evaluate the early recognition capability of a method. It is bound by 0 and 1 while 0 corresponds to random performance. The disadvantage is its factor α, which determines how much weight is put on the first ranks. Comparing the BEDROC to the EF, Riniker and Landrum found, that the BEDROC-parameter α corresponds reversely to the EF-α, i.e. $\alpha(EF) = \frac{1}{\alpha(BEDROC)}$ (Riniker and Landrum [2013]).
  For evaluation of *i*RAISE, an BEDROC-α of 2 was chosen.

- **NSLR** (Normalized Sum of Logarithmic Ranks) (Venkatraman et al. [2010])
  The Normalized Sum of Logarithmic Ranks is calculated by dividing the Sum of Logarithmic Ranks (SLR) by the maximum SLR:

$$SLR = -\sum_{i=1}^{A} log(\frac{r_i}{N}) \tag{6.4}$$

$$SLR_{max} = -\sum_{i=1}^{A} log(\frac{i}{N}) \tag{6.5}$$

$$NSLR = \frac{SLR}{SLR_{max}} \tag{6.6}$$

The NSLR uses a logarithmic weight to consider the early recognition problem. It has the advantages of being bound by 0 and 1 and being truly parameter-free.

Enrichment classification These metrics were further used to classify enrichments into easily interpretable categories of *excellent*, *good*, *medium* and *bad* enrichment. The following criteria were used for classification:

- **Excellent enrichment**: AUC > 0.7 and BEDROC > 0.6 and EF1% > 3
- **Good enrichment**: AUC > 0.6 and BEDROC > 0.6 and EF1% > 3 (two out of the three conditions have to be fulfilled)
- **Medium enrichment**: BEDROC > 0.4
- **Bad enrichment**: all others

Thus, excellent enrichments show good overall performance as well as early enrichment. Good enrichment still has a good overall performance and at least one good metrics measuring the early enrichment. For medium enrichments, the BEDROC is still significantly better than random, thus the very early enrichment is better than random. For bad enrichments, the predictions failed completely.

6.2 Evaluation experiments

6.2.1 Binding mode prediction

For the binding mode prediction study, the Astex Diverse Set is used (see section 5.1). In this experiment, each Astex ligand is screened against its true target and the pose of this ligand generated by *i*RAISE is compared to the co-crystallized ligand position, evaluated by the RMSD. Standard parameters of *i*RAISE are used and for each Astex ligand maximally 200 conformations are sampled.
The Astex Diverse Set is a set of high-quality protein-ligand complexes. In order to compare the binding mode prediction on this set to a set representing the average quality of complexes from the PDB in a better way, the same experiment is performed on the Drugs/sc-PDB data set (see section 5.5).

6.2.2 True-target ranking

For evaluating *i*RAISE's ability to rank first targets to the beginning of a score-ordered list of a set of targets, also the Astex Diverse Set was used (see section 5.1). Screening each of the 84 ligands against all 85 targets results in a score-ordered list of targets. For this experiment also 200 conformations were maximally sampled for each ligand.
This experiment was performed with two ranking strategies: Firstly, the simple interaction score (see section 4.5) was used for ranking, i. e. only step 1 and 2 of the Scoring Cascade were applied. Secondly the score of the full Scoring Cascade was used.
The number of ligands for which the true target was ranked at position 1 and the first 5%, 10%, 20%, 30% and 50% are used as performance measure.

6.2.3 Sensitivity versus selectivity

The TTFXa data set (section 5.4) is used to study sensitivity versus selectivity since it contains reliable true negative annotation for its five ligands. Therefore, the experiment on this data set is to screen all different serine protease structures with all five ligands and evaluate if the correct proteins were hit, and especially if the correct structures were not hit. For this experiment, also maximally 200 conformations were generated for each ligand.

6.2.4 Early enrichment

For the enrichment experiment, the Drugs/sc-PDB data set was used (see section 5.5). For each of the 72 ligands, the enrichment performance is evaluated using the metrics proposed in section 6.1: The EF1%, EF2%, EF5%, AUC, BEDROC and NSLR.
For this experiment, also maximally 200 conformations were generated for the ligands.

6.2.5 Comparison of ranking capability with classic docking

Experiment number two (section 6.2.2) was also conducted with a classic docking approach for comparison to *i*RAISE. As classic docking approach, the FlexX docking algorithm (Rarey et al. [1996]) together with HYDE scoring (Schneider et al. [2012]) was used. The LeadIt software suite (version 2.1) of the BioSolveIT (www.biosolveit.de) was used for this experiment. The FlexX docking algorithm is an incremental construction docking algorithm. The HYDE scoring function aims at predicting binding affinities by a strict hydrogen bond model and inclusion of solvation and desolvation effects.
FlexX does not need conformations, therefore it was started with a Corina-generated conformation of the Astex ligand (Sadowski et al. [1994]). The HYDE scoring was conducted on maximally 30 FlexX-poses of each ligand in each protein. Each pose was optimized firstly

with Protoss (Bietz et al. [2014]), which optimizes the hydrogen bond network and secondly geometrically for minimizing clashes and conformation strains and optimizing hydrogen bond geometries (Schneider et al. [2012]). For ranking, the best HYDE score was used.

6.2.6 Comparison to sequence-based method

The same enrichment experiment with the Drugs/sc-PDB data set described in section 6.2.4 was used for comparison to the performance of a sequence-based method. As a sequence-based target prediction method, protein-BLAST (Altschul et al. [1997]) from the NCBI/BLAST-server (http://blast.ncbi.nlm.nih.gov/) was used. Default parameter settings were applied for the protein-BLAST algorithm, with exception of the number of results returned, which was set to maximum. The complete PDB was chosen as sequence database. The hits returned were then filtered for those PDB codes which are also contained in the Drugs/sc-PDB data set.
As sequence query, the FASTA-sequence from the PDB was used of those proteins which were co-crystallized with the ligands of the data set: Each of the 72 ligands of the Drugs/sc-PDB data set has at least one co-crystallized protein in the data set. If there are several structures available, the one with the alphabetically first PDB code was chosen. Of the 72 ligands, one had only one true positive, i.e., the protein it was co-crystallized with in the data set. This ligand was omitted from the experiment.
The score returned from the protein-BLAST server was then used to rank the proteins in an ordered list.

6.2.7 Comparison to pharmacophore-based method

For the comparison of *i*RAISE's performance to a pharmacophore-based method, the sc-PDB Diverse Set was used (see section 5.3). This data set was used by Meslamani et al. (Meslamani et al. [2012]) for the evaluation of their pharmacophore-based target prediction method in comparison to two classic docking approaches. The pharmacophores used by Meslamani are generated from co-crystallized protein-ligand complexes. The number is reduced to 10 pharmacophores per pocket with a statistic-based selectivity score.
For comparison to *i*RAISE, the results of Meslamani et al. were extracted from the Supporting Information of Meslamani et al. [2012]. Comparing their approach to classic docking, Meslamani et al. used the Surflex (Jain [2007]) and the Plants (Korb et al. [2009]) docking algorithms. These results were also extracted.
The screening of the sc-PDB Diverse Set with *i*RAISE was conducted in two modi: Firstly, ligand conformations were sampled, with maximally 200 conformers per ligand. Secondly,

no conformations, but the co-crystallized ligand was used for screening.

7

Results and Discussion

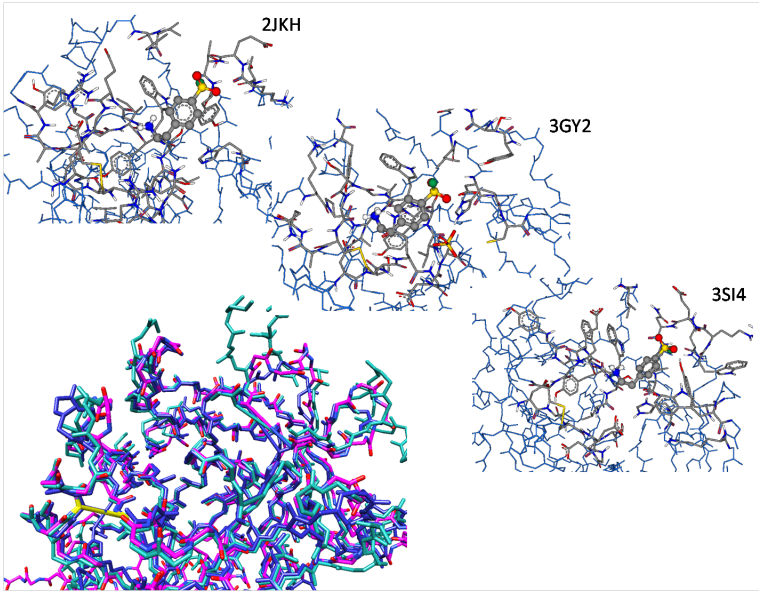


Figure 7.1: Poses of the general serine protease inhibitor pefabloc in the active site of a factor Xa (PDB code 2JKH), a trypsin (3GY2) and a thrombin (3SI4) protein structure. The superposition of the backbone atoms of the three protein structures is shown in the left corner (magenta=2JKH, blue=3GY2, turquoise=3SI4)

This chapter shows and discusses the results of the experiments which were conducted to evaluate *i*RAISE's performance. Firstly, *i*RAISE's capability of binding mode prediction is shown. Then, *i*RAISE's ranking capabilities are evaluated and compared to those of classic docking methods. Next, the enrichment experiments are discussed, and *i*RAISE's performance is compared to a pharmacophore- and a sequence-based target prediction method. A discussion of case studies of unknown target prediction follows. Finally, a short paragraph gives details on parametrization of *i*RAISE and its running time is analyzed.

7.1 Binding mode prediction

The evaluation of *i*RAISE's binding mode prediction capabilities shall display its ability of generating poses resembling the natural bioactive binding mode of the ligand in the active site. The experiment is described in section 6.2.1. RMSD calculations between the poses generated by *i*RAISE and co-crystallized ligands were performed on two data sets, the Astex Diverse Set and the Drugs/sc-PDB data set.
In Table 7.1, the minimal, maximal, average and median RMSDs for both data sets are listed. For the Astex Diverse Set, the numbers were differentiated between the best scored pose, the 30 best scored poses and the 100 best scored poses.

RMSDs	Astex Diverse Set			Drugs/sc-PDB
	best pose	within 30 poses	within 100 poses	within 30 poses
Minimal	0.41 Å	0.26 Å	0.26 Å	0.45 Å
Maximal	8.28 Å	6.8 Å	5.2 Å	7.55 Å
Average	1.2 Å	1.0 Å	0.99 Å	1.57 Å
Median	1.55 Å	0.86 Å	0.79 Å	1.13 Å

Table 7.1: Overview of maximal, minimal, average and median RMSD values for the Astex Diverse Set and the Drugs/sc-PDB data set.

In Figure 7.2, a bar-chart of the sum of ligands with a pose below selected cutoff RMSDs are shown. For the Astex Diverse Set, the RMSDs are given for the best ranked pose, the best RMSD among the first 30 ranked poses and the first 100 ranked poses. For the best RMSD in 30 poses, which is a reasonable amount of poses to generate, more than 80% of the poses get a RMSD lower than 2 Å which is considered as a successful binding mode prediction. This number is comparable to classic docking methods which do not apply post-optimization of the poses, e.g. the Glide docking function was recently evaluated on the same data set

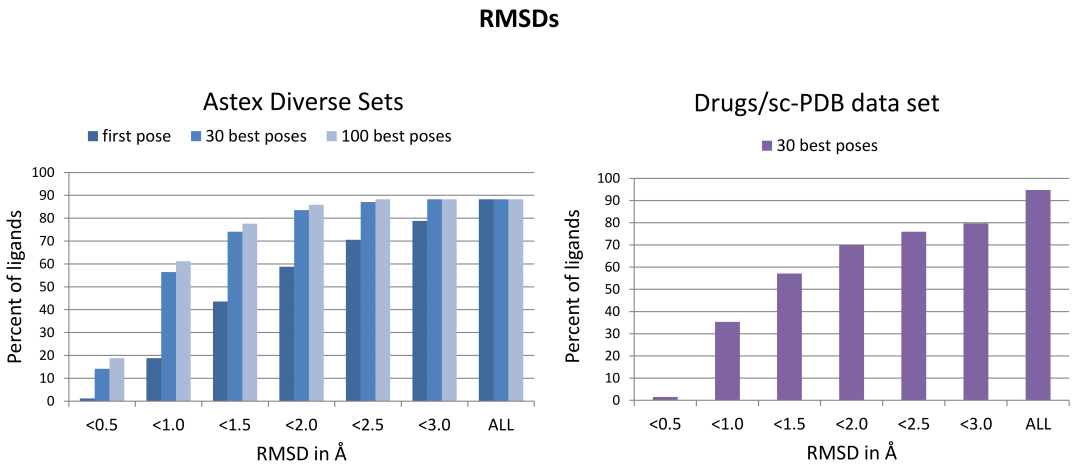


Figure 7.2: Sum of RMSDs of *i*RAISE's re-docked poses of the Astex Diverse Set and the Drugs/sc-PDB data set below the threshold values.

with only 57 of the 85 ligands below 2 Å, i.e. 67% (Wang et al. [2012]).
For the Drugs/sc-PDB data set, which was not manually curated like the Astex Diverse Set, the percentages of low RMSDs are below those of the Astex Diverse Set, hinting that either *i*RAISE predictions are data dependent or that the data set is not as suited for this test as the high-quality one.
The diagrams show that for neither of the data sets, binding modes were predicted for 100% of the ligands. Not all ligands could be placed into their co-crystallized binding pocket. For the Astex Diverse Set, these were about 10%, i.e., 10 of the 85 complexes. These cases were studied further in order to evaluate the reason why *i*RAISE fails here.

In Table 7.2, the ligands are listed, for which no poses were created in their true target applying the default screening settings. Firstly, it was tested, whether a poses was created, if the conformation of the co-crystallized ligand was used. For nine out of the ten ligands, indeed the true target was found in this mode. This means that in the 200 conformations used per default for screening, no conformation close enough to the bioactive one was found and thus the ligand could not be placed in the active site. Next, the maximum number of conformations was set to 500, instead of the per default used 200. With this settings, three out of the ten ligands could be placed in their true targets.

One example where the correct conformation is missing in those generated by default is now shown in detail for PDB complex 1PMN of an imidazole-pyrimidine (see Figure 7.3A) bound to a protein kinase. In Figure 7.3B, the ligand of 1PMN is shown in its co-crystallized

Astex ligand (PDB code of true target)	Pose created with crystal ligand?	Pose created with 500 conformations?
1hvy	yes	yes
1kzk	yes	no
1oyt	yes	yes
1pmn	yes	no
1r1h	yes	no
1t46	no	no
1xoz	yes	no
1y6b	yes	no
1ygc	yes	no
1yqy	yes	yes

Table 7.2: List of Astex Diverse Set cases of which no pose in the co-crystallized pocket was generated by *i*RAISE with the default settings.

conformation (blue) and in the generated conformations (gray). In Figure 7.3C, the active site of 1PMN with the co-crystallized ligand shows that the generated conformations would clash with the protein since the moiety with the three-membered ring is not bent enough.

The ligand of the complex 1t46 is the only one, for which even with the co-crystallized conformation no pose could be created in its true target by *i*RAISE. Further analysis shows that the index matching results in 50 poses in its true target, of which, however, 49 are discarded due to clashes and the last one is discarded since it protrudes with more than two thirds into the solvent.

With one exception, the evaluation thus showed that the reason for failure is mainly due to conformations differing too much from the bioactive conformations. Even if the number of conformations is set to 500, for some ligands the bioactive conformation was not generated. These results clearly show the limitation of *i*RAISE in its dependence on the generated or provided conformations of the query molecule.

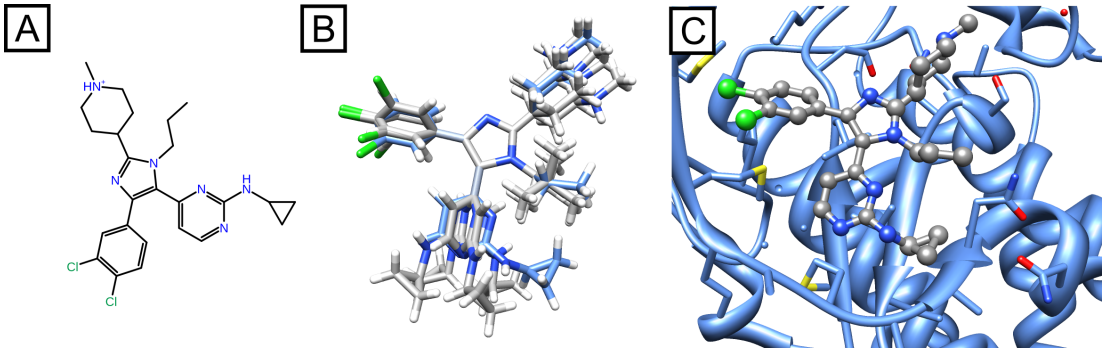


Figure 7.3: (A) 2D structure diagram of ligand 1pmn of the Astex Diverse Set. (B) 3D conformations generated in gray, crystal conformation in blue (C) Active site of 1PMN

7.2 Ranking capability

*i*RAISE's ability to rank true targets to the beginning of a score-ordered list was evaluated on the rank of the co-crystallized targets of the Astex Diverse Set (see section 6.2.2 for the experiment description). The results are summarized in Figure 7.4. This experiment highlights two aspects: Firstly, the measures applied by the Scoring Cascade for selectivity like the reference score cutoff and the ligand and pocket coverage improve the ranking capability significantly. By using the Scoring Cascade, the ranking is improved, e.g., from fewer than a tenth of the ligands at rank 1 with the interaction score to more than one third of the ligands with rank 1 with the Scoring Cascade. The second aspect has already been discussed before: The Scoring Cascade selectivity measures lead to a loss of the total true positives found.

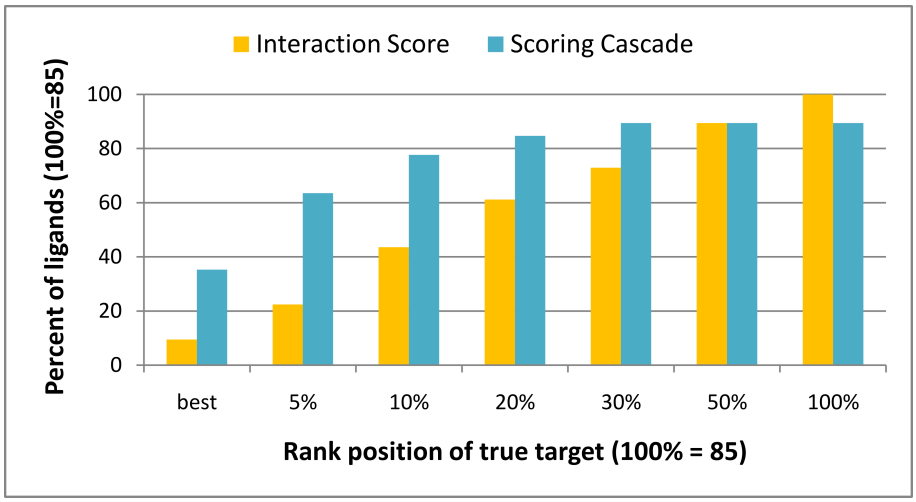


Figure 7.4: Ranking of true target for each of the 85 ligands of the Astex Diverse Set, summed at position 1, the first 5, 10, 20, 30 and 50%. The yellow bars show the ranks for the interaction score and the turquoise bars show the ranks for the full Scoring Cascade

7.3 Comparison to classic docking

The ranking capability of *i*RAISE was also compared to classic docking studies on the same experiment as used in the preceding section for assessing the gain of the Scoring Cascade (see section 6.2.5 for the experiment description). In Figure 7.5, the results of the ranking based on the FlexX/Hyde combination and of *i*RAISE are juxtaposed. The rank sums show that the selectivity measures applied by *i*RAISE in the Scoring Cascade lead to superior

performance in true target ranking at the first percentages of a score-ordered list. While *i*RAISE ranks for about a third of the ligands the true target to the first position of the score-ordered list, the FlexX/Hyde approach only achieves this rank for true targets for a fourth of the ligands, although complex optimization strategies are applied to the poses (see section 6.2.5). The amount of ligands for which the true targets are ranked to the beginning of the score-ordered list is higher for the Scoring Cascade up to a tenth of the list. For later ranks in the target-list, the percent of ligands is higher for the FlexX/Hyde approach. However, in real applications on large datasets, the enrichment of true targets at about 1% is crucial.

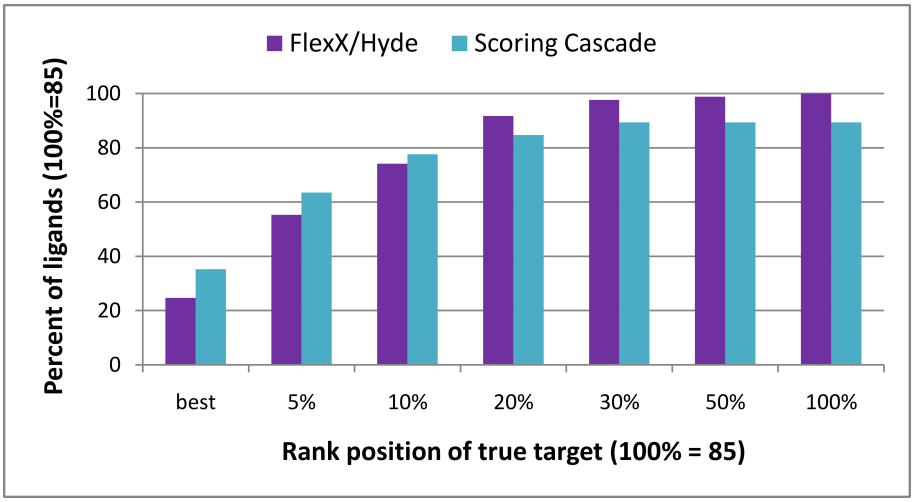


Figure 7.5: Ranking of true target for each of the 85 ligands of the Astex Diverse Set, summed at position 1, the first 5, 10, 20, 30 and 50%. The purple bars show the ranks for docking and scoring with FlexX and Hyde and the turquoise bars show the ranks for the Scoring Cascade of *i*RAISE.

The same experiment has been conducted with the docking program Glide (Friesner et al. [2004]) see Schomburg et al. [2014]. Glide achieves about the same amount of true targets on the first rank. However, only 50% of the true targets are ranked among the first 5%, while *i*RAISE achieves this for more than 60%.

Inhibitor	Thrombin			Factor Xa			Trypsin		
	3RM2	2BVR	3SI4	2JKH	2Y5F	3KL6	3GY2	2G8T	2G5N
Benzamidine									
sc-score	–	–	-30.8	-23.7	-28.5	–	-33.6	-33.4	-34.6
gsw-score	–	–	0.73	0.69	0.80	–	0.84	0.93	0.97
Pefabloc									
sc-score	-51.5	-29.0	-54.3	-40.3	-37.3	-54.3	-34.4	-33.3	-33.9
gsw-score	1.1	0.99	1.28	1.14	0.88	1.20	0.84	0.93	0.97
Apixaban									
sc-score	–	–	–	–	-32.3	–	–	–	–
gsw-score	–	–	–	–	0.88	–	–	–	–
Rivaroxaban									
sc-score	–	–	-28.5	-31.5	39.5	-60.0	–	–	–
gsw-score	–	–	0.69	1.14	0.88	1.33	–	–	–
Melagatran									
sc-score	–	–	–	–	-33.7	–	–	–	–
gsw-score	–	–	–	–	0.93	–	–	–	–

Table 7.3: *i*RAISE scores on the TTFXa data set consisting of three structures for each serine protease thrombin, factor Xa and trypsin. For all structures, true targets of the inhibitors are marked by highlighting the entry in green. A hyphen indicates that *i*RAISE did not predict the structure as a target.

7.4 Sensitivity versus selectivity

The experiment on the TTFXa data set containing nine structures of three serine proteases shall show if the method is able to selectively identify correct targets. Not only true positives have to be found but also true negatives must be classified correctly. A matrix of scores for each ligand of the data set for each structure is shown in Table 7.3. Only the scores for those structures for which *i*RAISE produces a binding pose are listed in the table. Optimally, the *i*RAISE procedure would create only poses for the proteins which are highlighted in green as true targets. Since the protein structures of the same proteins have slightly different conformations because they were co-crystallized with different ligands and in *i*RAISE protein flexibility is not considered internally, it is expected that not each structure of each protein is hit. Therefore, success for a ligand is defined if at least one structure of the correct protein was found.

In Table 7.3, for each inhibitor the sc-score of the Scoring Cascade is given as well as the gsw-score. While the gsw-score is easily interpretable (all scores greater than 1.0 are higher than an average score for that pocket), the absolute numbers of the sc-score are harder to interpret.

Benzamidine and pefabloc are general serine protease inhibitors. For these two inhibitors, *i*RAISE produces correct results with benzamidine-poses for at least one structure of each protein and poses for each protein structures for pefabloc.
Apixaban is a factor Xa inhibitor, for which *i*RAISE also correctly predicts one structure of the correct protein as a target. Rivaroxaban is a factor Xa inhibitor as well. For this inhibitor, the *i*RAISE screening results in poses for all factor Xa structures and one thrombin structure, which is no target of rivaroxaban. However, all factor Xa structures are scored better than the thrombin structure, and the best score for a factor Xa structure is with 1.33 significantly higher than average score of the pocket in contrast to the score of 0.69 for the thrombin structure.
Melagatran is a thrombin inhibitor. For this inhibitor, *i*RAISE only produces a pose for a factor Xa structure, which is not correct. This case was thus analyzed further.
A thrombin structure co-crystallized with melagatran from the PDB (PDB code 4BAH) was superposed with the thrombin structures from the TTFXa data set. Thus, the conformations of the active sites of the structures from the data set could be compared to the protein conformation with bound melagatran. In Figure 7.6, the superposition of 4BAH with 3SI4 is shown as an example. The superposition shows clearly that the binding mode of melagatran from 4BAH would not fit into 3SI4, since the rotated Ile174 would produce a clash. In the 4BAH structure *i*RAISE is able to produce a binding pose with a score of 1.22.
These results show a negative and a positive aspect of the predictions of *i*RAISE: *i*RAISE dependents on conformational protein samples to predict the correct target and produces only poses for correct protein conformations. However, with respect to selectivity, the experiment demonstrates that *i*RAISE is able to predict mostly only true positive targets on the class of serine proteases which are structurally and sequentially highly similar.

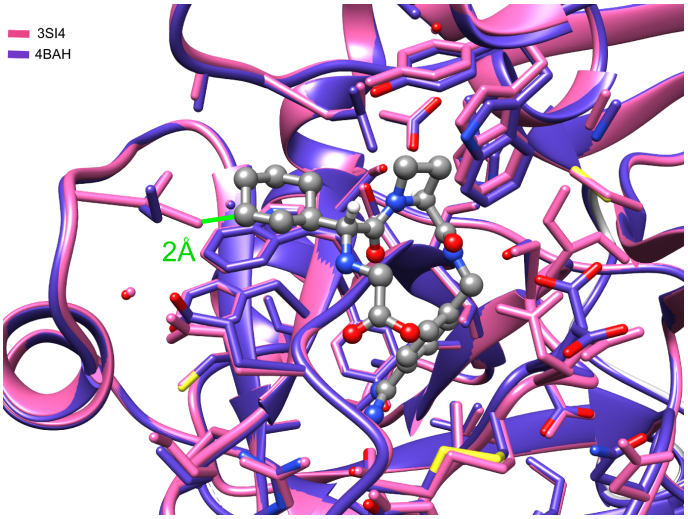


Figure 7.6: Alignment of the two thrombin structures 4BAH in purple and 3SI4 in pink with melagatran as co-crystallized in 4BAH. Melagatran cannot be placed in the active site of 3SI4 in the same binding mode as in 4BAH due to a different conformation of Ile174 which would create a clash with the inhibitor. *(This figure was originally published in Schomburg and Rarey [2014].)*

7.5 Enrichment experiments

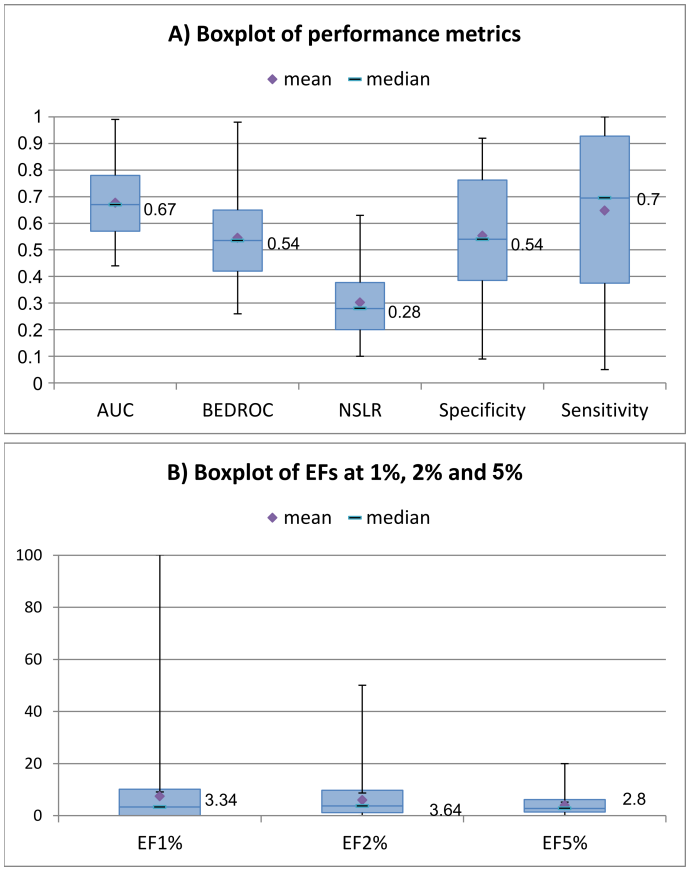


Figure 7.7: Boxplots of (A) the enrichment metrics AUC, BEDROC, NSLR and the specificity and sensitivity (B) the Enrichment Factors at 1%, 2% and 5%. The median is shown with a green line and is printed with numbers into the diagram. The mean is shown in form of a diamond. The blue area shows the area between the first and the third quartile, the lines indicate the minimum and the maximum.

The Drugs/sc-PDB data set was used for enrichment experiments, see section 6.2.4 for the experiment description. The metrics described in section 6.1 were used for evaluation along with the classification scheme of 'excellent', 'good', 'medium' and 'bad' enrichment.
The evaluation metrics map the performance of the inverse screening onto 'one number' and thus allow fast comparison of overall performance among different methods. Since the data set was developed in this thesis, no data of other state-of-the-art methods was available for comparison. Therefore, here the metrics were only be assessed to evaluate the gain of *i*RAISE over random performance.

In Figure 7.7, the metrics *AUC*, *BEDROC*, *NSLR*, *EF1%*, *EF2%*, *EF5%* and the statistical measures *Specificity* and *Sensitivity* are shown in form of a boxplot. The metrics for all 72 ligands are used here.
In Figure 7.7A, the boxplot of the AUC shows that with a value of 0.67 the median is well above random (random=0.5) and that also the maximal 1.0 is reached (which is equal to perfect performance) but also that the minimum is even below random. The median of the BEDROC, which is a better metric for the assessment of the early enrichment (see section 6.1) is with 0.54 well above random (=0.0) and also the maximum almost reaches perfect enrichment (1.0). The NSLR, which also weights the early enrichment is with 0.28 also above random (=0.0) but does not reach perfect enrichment.
The specificity, which is also called true-negative rate, shows which part of the true negatives are truly identified as negatives. The sensitivity, also called true-positive rate, shows which part of the true positives are truly identified as positive. The medians of 0.54 for the specificity versus 0.7 of the sensitivity shows that more positives are identified as true positives than negatives are correctly classified. The low rate of the sensitivity is due to the fact that these numbers are calculated on the level of protein structures and not proteins. As will be shown later in sections 7.5.3 and 7.5.4, often not all structures of a protein in the data set are hit in the *i*RAISE screening, due to e.g., protein flexibility. The low specificity is due to the fact that *i*RAISE merely ranks protein targets. The method was not developed to be able to classify clearly between true negatives and true positives, but the ranking rather says that the first ranked targets bind the query ligand better than the later ranked. Here, true negatives are only those protein structures, in which no binding pose was created and thus no score was obtained as well as protein structures which are scored by the gsw-score lower than 1.0.
In Figure 7.7B, the Enrichment Factors are shown at 1%, 2%, and 5% respectively. The EF1% has a median of 3.34 which means that the enrichment is 3.34 better than random. The medians of EF2% and EF5% show a 3.64 and 2.8 times better enrichment than random.
The discussed metrics and numbers only allow a general assessment of *i*RAISE's performance, but provide no detailed identification of limitations and potentials. Therefore, the metrics of each of the 72 ligands were used to classify the enrichment into excellent, good, medium and bad enrichment following the classification scheme given in section 6.1.
In Figure 7.8, the distribution of the 72 classified enrichments in the four categories are shown. Excellent and good enrichment with a percentage of 28 each cover in total 40 ligands, which is more than half of all ligands. Medium enrichments are achieved for 29% - here improvements are necessary. The part of bad enrichments shows that for 15% of the

ligands, the *i*RAISE target prediction failed.
For further discussion and assessment of limitations, now each category is shown in detail.

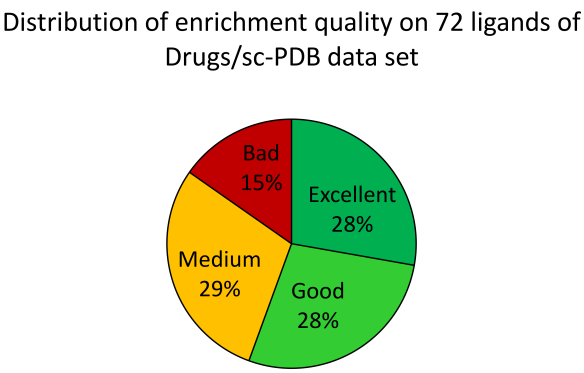


Figure 7.8: Distribution of categorized enrichment for target prediction for the 72 ligands of the Drugs/sc-PDB data set in percent.

7.5.1 Excellent enrichments

The predictions of *i*RAISE are classified as 'excellent' if the following three criteria are fulfilled: AUC > 0.7 and BEDROC > 0.6 and EF1% > 3 (see section 6.1). In Figure 7.9, the metrics for the 20 ligands with excellent enrichments are shown. In the top line, the medians of all 72 ligands are given for comparison. In Figure 7.10, as an example, the three ROC plots for the ligands varenicline, meclofenamic acid and estradiol are shown.
For varenicline, only one true protein structure is contained in the complete data set, which is ranked by *i*RAISE on rank 55 of 7915 and thus an almost perfect ROC curve is reached. Meclofenamic acid binds to prostaglandin G/H synthases 1 and 2, of which 22 structures are contained in the data set. The first structures are ranked at the very high ranks of 5, 19 and 22. The other structures are also all hit consecutively resulting in excellent enrichment. Estradiol binds to the estrogen receptor, to sex-hormone-binding globulin and to beta-hydroxysteroid dehydrogenase. Of these targets, 80 structures are contained in the protein data set. In total, *i*RAISE identifies 74 of these 80 as true targets and the first ranks of true targets are at position 8, 36 and 41. Even if not every structure is hit, the enrichment is excellent.
For the ROC curves of all 20 ligands categorized as 'excellent', see Appendix G, Figure G.1.

		EF1%	EF2%	EF5%	AUC	BEDROC	NSLR
	MEDIAN	**3.34**	**3.64**	**2.8**	**0.67**	**0.54**	**0.28**
3	Estradiol	7.51	9.39	7.75	0.83	0.73	0.46
10	Tadalafil	5.5	3.85	5.05	0.73	0.61	0.38
15	Dichlorphenamide	16.48	12.52	7.63	0.73	0.64	0.51
20	Pentoxifylline	3.07	3.58	4.28	0.77	0.63	0.36
24	Papaverine	15.41	12.11	7.69	0.81	0.72	0.49
26	Ethoxzolamide	3.95	4.61	4.34	0.79	0.66	0.44
30	Tolmetin	4.01	4.01	3.2	0.81	0.66	0.31
34	Varenicline	100.19	50.09	19.99	0.99	0.98	0.55
38	Zonisamide	9.33	7.78	6.08	0.86	0.76	0.53
43	Progesterone	20.04	14.47	8.44	0.71	0.61	0.46
44	Testosterone	16.7	12.99	6.66	0.74	0.64	0.43
49	Mefenamic acid	4.36	8.71	4.35	0.78	0.65	0.32
50	Meclofenamic acid	13.66	9.11	4.35	0.79	0.65	0.35
53	Clomipramine	13.07	10.89	6.95	0.72	0.61	0.34
57	Trimetrexate	9.62	8.82	6.24	0.70	0.60	0.44
58	Dexamethasome	23.12	11.56	6.15	0.85	0.76	0.43
65	Naproxene	12.02	10.02	8.79	0.87	0.77	0.42
66	Nabumetone	4.17	8.35	5.83	0.82	0.71	0.36
67	Hydrochlorothiazide	18.99	16.37	11.1	0.85	0.79	0.63
68	Hydroflumethiazide	18.34	15.39	10.06	0.78	0.71	0.58

Figure 7.9: Metrics for the 20 ligands of the Drugs/sc-PDB data set classified with excellent enrichment.

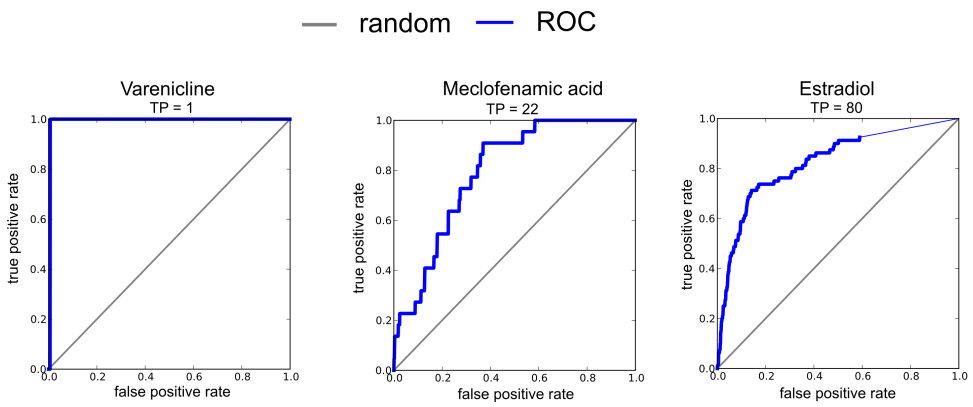


Figure 7.10: ROC plots for 3 ligands categorized as 'excellent' enrichment. Thick blue lines show the true positives found. If not all true targets were identified, a thin line is drawn from the last found positive on assuming random distribution from there on. TP=Number of true positives.

7.5.2 Good enrichments

		EF1%	EF2%	EF5%	AUC	BEDROC	NSLR
	MEDIAN	**3.34**	**3.64**	**2.8**	**0.67**	**0.54**	**0.28**
1	Dorzolamide	17.14	11.54	7.5	0.67	0.58	0.47
2	Lovastatin	3.45	5.18	6.89	0.67	0.55	0.29
4	Efavirenz	29	16.26	8.24	0.69	0.59	0.51
6	Niflumic acid	3.78	2.84	2.64	0.70	0.54	0.28
8	Galantamine	5.42	2.71	2.16	0.73	0.57	0.29
12	Succhinylcholine	33.4	16.7	6.66	0.66	0.54	0.26
14	Cocaine	0	0	0	0.86	0.73	0.24
22	Finasteride	22.26	16.7	6.66	0.69	0.58	0.32
23	Topiramate	8.57	6.26	4.6	0.63	0.52	0.37
28	Prazosin	0	0	11.99	0.95	0.89	0.39
31	Thiabendazole	0	0	0	0.95	0.89	0.33
36	Celexobic	9.25	5.39	2.77	0.63	0.48	0.28
37	Levonorgestrel	10.3	9.83	5.6	0.67	0.55	0.39
42	Penicillin G	0	0	6.66	0.78	0.66	0.26
45	Spiranolactone	12.93	11.31	4.51	0.64	0.52	0.32
46	Diazepam	0	3.85	3.07	0.79	0.65	0.28
48	Diclofenac	4.23	4.23	2.82	0.71	0.55	0.30
59	Hydrocortisone	0	12.52	5	0.74	0.60	0.26
60	Fludrocortisone	13.18	10.55	5.26	0.64	0.52	0.31
72	Chenodeoxycholic acid	0	0	2.1	0.78	0.65	0.28

Figure 7.11: Metrics for the 20 ligands of the Drugs/sc-PDB data set classified with good enrichment.

As criterion for 'good' enrichment, two out of three conditions have to be fulfilled: AUC > 0.6 and BEDROC > 0.6 and EF1% > 3 (see section 6.1).

In Figure 7.11, the metrics for the 20 ligands with good enrichments are shown. For these ligands, the enrichment is not classified as 'excellent', since either the early enrichment or the total enrichment is not perfect.

This becomes obvious in Figure 7.12, where the ROC plots for the ligands cocaine, galantamine and efavirenz are shown (for ROC plots of all ligands of this category, see Appendix G, Figure G.2).

Several transporter and receptor proteins are targets for cocaine. In the protein structure data set, two structures of the target protein muscarinic acetylcholine receptor are present. These are ranked by *i*RAISE at positions 577 and 1624 of the 7915 targets. Thus, no true target is found until more than 7 percent of the data has been screened leading to improvable early enrichment.

For galantamine, the data set contains 37 protein structures of its true targets acetylcholinesterase, cholinesterase and acetylcholine receptor. The first structure of a true target is ranked by *i*RAISE to position 6. Of the 37 true targets, 3 are not identified by *i*RAISE. In total, the enrichment is rather good, only the BEDROC with 0.57 is slightly below the

threshold for excellent enrichment.
In the case of efavirenz, *i*RAISE shows excellent early enrichment but hits less than half of the true target structures in the database. Of 114 protein structures of reverse transcriptase of the human immunodefficiency virus in the structure data set, 60 are not identified as true targets. The early enrichment is very high, with the first 4 ranks occupied with structures of true targets and in the first 50 ranks 25 true target structures. Thus, the enrichment would be usable in real applications, where the first percentages of the rank ordered list are tested experimentally, but the overall performance needs to be improved.
An evaluation of those true target structures to which *i*RAISE could not create a binding mode for efavirenz showed various reasons for the failure: Mostly, the protein conformation was different compared to a structure co-crystallized with efavirenz with respect to the backbone and/or the side chains, highlighting the methods incapability of internally handling flexibility. Another aspect was the mislabeling of the structures as 'true targets': In the true target assignment stage, proteins with the same name, as e.g. a co-crystallized protein are considered as true positives. In this case, a protein labeled with 'pol polyprotein' is considered as true target, which is a collective term for HIV reverse transcriptase and other enzymes like the polymerase, integrase or protease. Thus, here a weakness of the automatic true positive assignment is revealed. A third influencing effect arises from the frequent mutations found in reverse transcriptases. Although mutations are not considered during the true-target assignment method, they certainly have an influence on the binding of the compound. Thus, for some of the mutated structures it cannot be concluded if they still bind efavirenz or not.

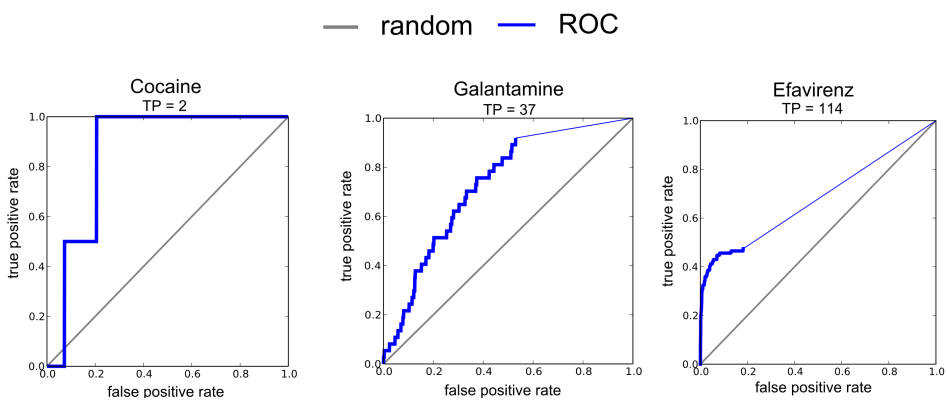


Figure 7.12: ROC plots for 3 ligands categorized as 'good' enrichment. Thick blue lines show the true positives found. If not all true targets were identified, a thin line is drawn from the last found positive on assuming random distribution from there on. TP=Number of true positives.

7.5.3 Medium enrichments

		EF1%	EF2%	EF5%	AUC	BEDROC	NSLR
	MEDIAN	**3.34**	**3.64**	**2.8**	**0.67**	**0.54**	**0.28**
7	Nevirapine	2.64	4.39	2.63	0.57	0.42	0.26
11	Imatinib	3.2	2.13	2.23	0.56	0.40	0.26
13	Apixaban	0	1.15	0.92	0.56	0.41	0.22
16	Proflavine	0	0	0.57	0.60	0.42	0.22
18	Minocycline	0	7.16	2.86	0.59	0.45	0.19
19	Indomethacin	1.96	2.62	2.09	0.58	0.42	0.26
21	Chloramphenicol	0	2.95	1.57	0.60	0.43	0.21
27	Gefitinib	0.72	2.15	1.14	0.57	0.41	0.24
29	Pyrimethamine	0.8	1.6	1.92	0.55	0.40	0.23
32	Ethacrynic acid	0	0	0	0.70	0.50	0.19
33	Abacavir	0.73	0.73	1.46	0.66	0.51	0.29
39	Ampicillin	0	2.95	2.35	0.67	0.49	0.20
40	Penicillin V	2.78	2.78	2.78	0.61	0.47	0.24
41	Meticillin	0	0	0	0.60	0.45	0.15
47	Midazolam	0	0	0	0.66	0.49	0.16
51	Methotrexate	3.64	3.64	2.66	0.58	0.43	0.29
52	Raltitrexed	0	1.96	1.57	0.59	0.44	0.22
54	Imipramine	6.46	4.85	3.22	0.57	0.43	0.23
56	Trimethoprim	2.5	2.5	2.25	0.60	0.45	0.28
61	Chlorpromazine	0	0	0	0.67	0.53	0.19
69	Sildenafil	2.64	1.32	1.58	0.55	0.40	0.19

Figure 7.13: Metrics for the 21 ligands of the Drugs/sc-PDB data set classified with medium enrichment.

If an enrichment was not classified with 'good' or 'excellent' and still has a BEDROC greater 4.0, then it is classified as medium (see section 6.1).

In Figure 7.13, the metrics for the 21 ligands with medium enrichments are shown. As examples, in Figure 7.14, the ROC plots for ethacrynic acid, penicillin V and imipramine are shown (for ROC plots of all ligands of this category, see Appendix G, Figure G.3).

Ethacrynic acid binds to gluthathione S-transferase A2 and serum albumin, for which there are 15 true target structures in the data set. All of these are recovered by *i*RAISE screening, but the first is not found before more than 10% of the data has been screened. Thus the early enrichment is poor while the total AUC is nevertheless good with 0.7. For penicillin V, the structure data set contains 36 true target structures of penicillin acylase and beta-lactamase. Of these, 11 are not identified as true targets, but the early enrichment is good with an EF1% of 2.87. Imipramine binds to the androgen receptor, the adrenergic receptor and the muscarinic acetylcholine receptor of which there are 31 structures in the data set. Only 7 of these are recognized as true targets but with high early enrichment as the ranks of the first true targets are 21, 43 and 95 of the 7915 target structures, respectively. The AUC is low with 0.54 but the EF1% is 6.46.

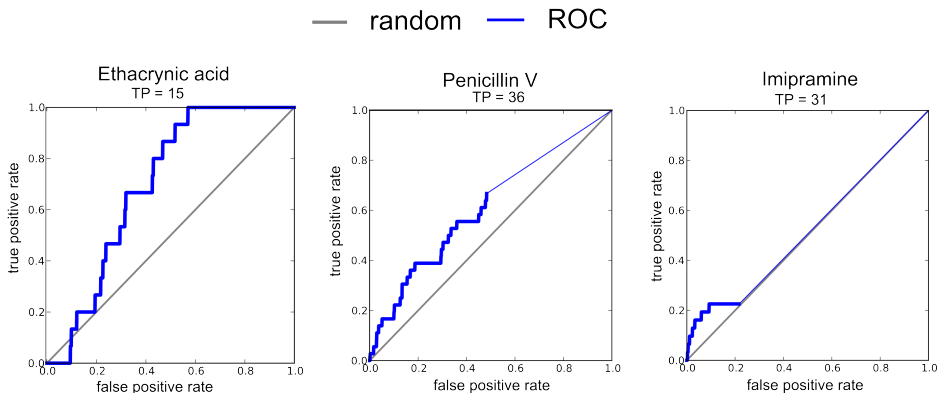


Figure 7.14: ROC plots for 3 ligands categorized as 'medium' enrichment. Thick blue lines show the true positives found. If not all true targets were identified, a thin line is drawn from the last found positive on assuming random distribution from there on. TP=Number of true positives.

7.5.4 Bad enrichments

		EF1%	EF2%	EF5%	AUC	BEDROC	NSLR
	MEDIAN	**3.34**	**3.64**	**2.8**	**0.67**	**0.54**	**0.28**
5	Delavirdine	0	0.44	1.23	0.52	0.36	0.20
9	Sitagliptin	0	1.04	0.42	0.47	0.30	0.14
17	Phenylbutazone	0	0	0	0.44	0.26	0.10
25	Balsalazide	0	0.52	0.62	0.48	0.32	0.16
35	Captopril	0	0	0.64	0.53	0.36	0.15
55	Desipramine	3.23	4.85	1.93	0.52	0.38	0.19
62	Trifluoperazine	4.77	2.39	1.43	0.54	0.38	0.19
63	Alogliptin	2.09	2.09	0.83	0.56	0.39	0.18
64	Linagliptin	0	0	0	0.49	0.33	0.14
70	Vardenafil	2.57	2.57	1.02	0.53	0.38	0.18
71	Ursodeoxycholic acid	4.55	2.28	2.73	0.51	0.37	0.18

Figure 7.15: Metrics for the 11 ligands of the Drugs/sc-PDB data set classified with bad enrichment.

In total, eleven of the 72 ligands of the data set cannot be classified with excellent, good or medium, but are classified as 'bad'. For these ligands, the target predictions by *i*RAISE failed.

In Figure 7.15 the metrics for these ligands are listed. In Figure 7.16, the example ROC plots of phenylbutazone, captopril and trifluoperazine are shown (for all ROC plots of ligands of the category 'bad' enrichment, see Appendix G, Figure G.4).

Phenylbutazone is the ligand for which *i*RAISE performs worst. Of 22 true positive structures of its targets prostaglandin synthase, only one is identified as true positive at rank 1352.

Captopril binds to angiotensin converting enzyme, collagenase 3 and matrix-metalloprotease, of which 31 structures are contained in the data set. The *i*RAISE screening finds 25 of the 31 true targets structures, but the ranks of these structures are almost randomly distributed among the data set, no enrichment can be observed. In the true target structures found as well as in those not found, all three target classes are present, thus no bias of *i*RAISE towards one of the classes is observed. As targets for trifluoperazine 42 structures for troponin C, androgen receptor, calmodulin, xanthine dehydrogenase and adrenergic receptor are contained in the data set. Of these, less than a fifth are identified correctly as true targets and the ranking of the ones found is almost randomly distributed among the first ten percent of the structure data.

As *i*RAISE performs worst on phenylbutazone, this case was further studied. Firstly, literature and database study revealed serum albumin and lipoxygenase as further targets of phenylbutazone (Günther et al. [2008]). For serum albumin, the *i*RAISE screening finds 9 of 10 structures at ranks 139, 182, 536, 574, 627, 924, 978, 1099, 1476, 1531, 1626, and 1740. Thus the first structure is found before 2% of the data has been screened. For the enzyme lipoxygenase, one structure is found at rank 270 and two structures are not identified as true targets. Thus, firstly it can be concluded that the ligand in general is not the problem, since for serum albumin the predictions of the screening were correct.

Secondly, it was studied, at which step of *i*RAISE the true targets are discarded. In total, there are 22 true target structures of prostaglandin synthase, i.e., cyclooxygenase. In the default *i*RAISE screening setup, only one was found, at rank 1352 (PDB code 3B99). As a first step, the screening was repeated with 500 conformations of phenylbutazone instead of 200. With this setup, only one further structure (of a cyclooxygenase 2) was identified as true target (PDB code 3MQE). Therefore, it does not seem to be only a problem of the limitation of the conformations. Thus the *i*RAISE procedure was studied step by step. In the index matching step (where descriptors of the ligand and the targets are matched), all true targets get a match. The following step is a coarse (grid-based) clash test. Here, for five of the 22 structures all matches are discarded. Following is a detailed atom-based clash test. In this step, of the remaining 18 structures 14 are discarded. Two more structures are then discarded due to insufficient ligand coverage. Of the remaining two, one is discarded by the reference score cutoff. Thus only one structure remains.

Therefore, in this case, probably not enough poses are generated by *i*RAISE, or the clash tests are too strict or not enough ligand conformations sampled. Also, it is possible that protein flexibility is the problem.

Another aspect might also be the reason for failure: Cyclooxygenases have two binding sites, a substrate-binding and an allosteric site, where the second part of the natural reaction takes

place. Literature study did not reveal whether the inhibition of cyclooxygenases by phenylbutazone is due to competitive inhibition of the first reactive center, or the second. In the structure data set, only the first binding site is represented. Although the missing of the true targets is probably due to a limitation of *i*RAISE and its inability to induce conformational changes which may be necessary for the binding of phenylbutazone, it is also possible that the inhibition is allosteric and thus cannot be found in this screening setup.

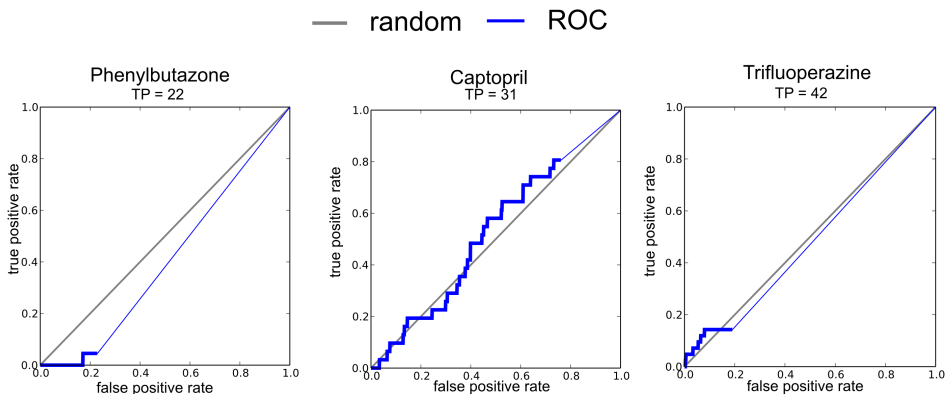


Figure 7.16: ROC plots for 3 ligands categorized as 'bad' enrichment. Thick blue lines show the true positives found. If not all true targets were identified, a thin line is drawn from the last found positive on assuming random distribution from there on. TP=Number of true positives.

	median AUC	median NSLR	median BEDROC
pBLAST	0.56	0.14	0
*i*RAISE	0.67	0.28	0.54

Table 7.4: Median metrics for sequence-based target prediction by pBLAST and for *i*RAISE on 71 ligands of the Drugs/sc-PDB data set.

7.6 Comparison to sequence-based target prediction

Sequence-based methods are rapid and easy to perform. A starting sequence of a protein to which the ligand of interest is known to bind to is compared against the sequences of other proteins. The advantage is that since only sequences of proteins are needed and not structures, much more proteins can be assessed than in structure-based methods. A disadvantage is that a starting sequence is needed, thus if no protein is known to which the ligand binds, this method cannot be used.
The experiment of sequence-based target prediction is described in section 6.2.6. The same data has been used as in the enrichment experiments described in the preceding sections. The sequence-based experiment was conducted to compare *i*RAISE's performance to this method and to see if the Drugs/sc-PDB data sets contains examples which are not predictable with straight forward sequence comparisons. In real applications, sequence-based methods will always be used if possible. Using structure-based methods needs vindication since they are much more time consuming.
For the evaluation of the sequence-based target prediction capability, the same metrics were calculated as used for assessing *i*RAISE's performance on this data set. For comparison of the overall performance, in Table 7.4, the medians for pBLAST and *i*RAISE are shown opposite to each other. These metrics are the medians for 71 of the 72 ligands of the Drugs/sc-PDB because one ligand was excluded from the sequence-experiment (see section 6.2.6).

The comparison of the median metrics shows that *i*RAISE outperforms pBLAST concerning the total performance. However, for a more detailed evaluation, these metrics are plotted for pBLAST and *i*RAISE for all 71 ligands in Figure 7.17. The plots of the AUC (Figure 7.17A), the BEDROC metric (Figure 7.17B) and the NSLR (Figure 7.17C) for all ligands show two aspects: For a part of the ligands, the metric values of pBLAST are very high, corresponding to perfect enrichment and target predictions. *i*RAISE does not reach as high values in the performance metrics. However, for more than half of the 71 ligands, the sequence-prediction

fails while *i*RAISE's performs well on most of those cases. Two conclusions can be drawn from these observations: Firstly, the data set indeed contains entries which are not solvable with sequence-based target prediction and is thus appropriate for evaluating structure-based target prediction methods. Secondly, *i*RAISE is able to predict targets which are not found by sequence-based target predictions methods and are therefore not trivial.
In Figure 7.17D, the number of different protein targets for each ligand is plotted to assess if the sequence-based methods or *i*RAISE's performance is correlated with this number. Most of the ligands for which pBLAST's performance is best have only one true protein target. Nevertheless, there is also a ligand (dichlorphenamide) with 4 different protein targets and another (imitinib) with 10 different targets on which pBLAST performs well. However, the targets for dichlorphenamide are 4 different carbonic anhydrases and the targets for imatinib belong to the kinase enzyme family. Both protein sets exhibit high sequence similarity.
Clearly, the sequence-based methods performance is highly dependent on the input sequence. If other input sequences had been chosen, the performance would be different. One example, where *i*RAISE performs very well and pBLAST fails is prazosin (fifth data point from the left). Prazosin only has adrenergic receptors annotated in the DrugBank as true targets, however, in the data set, it was co-crystallized with a quinone reductase and thus the correct targets could not be found. One example where pBLAST has a perfect AUC of 1.0 and *i*RAISE's predictions fail with an AUC of 0.5 is sitagliptin, which only has dipeptidyl peptidase 4 as a target in the structure data set and the sequence prediction was also started with the sequence of this enzyme.

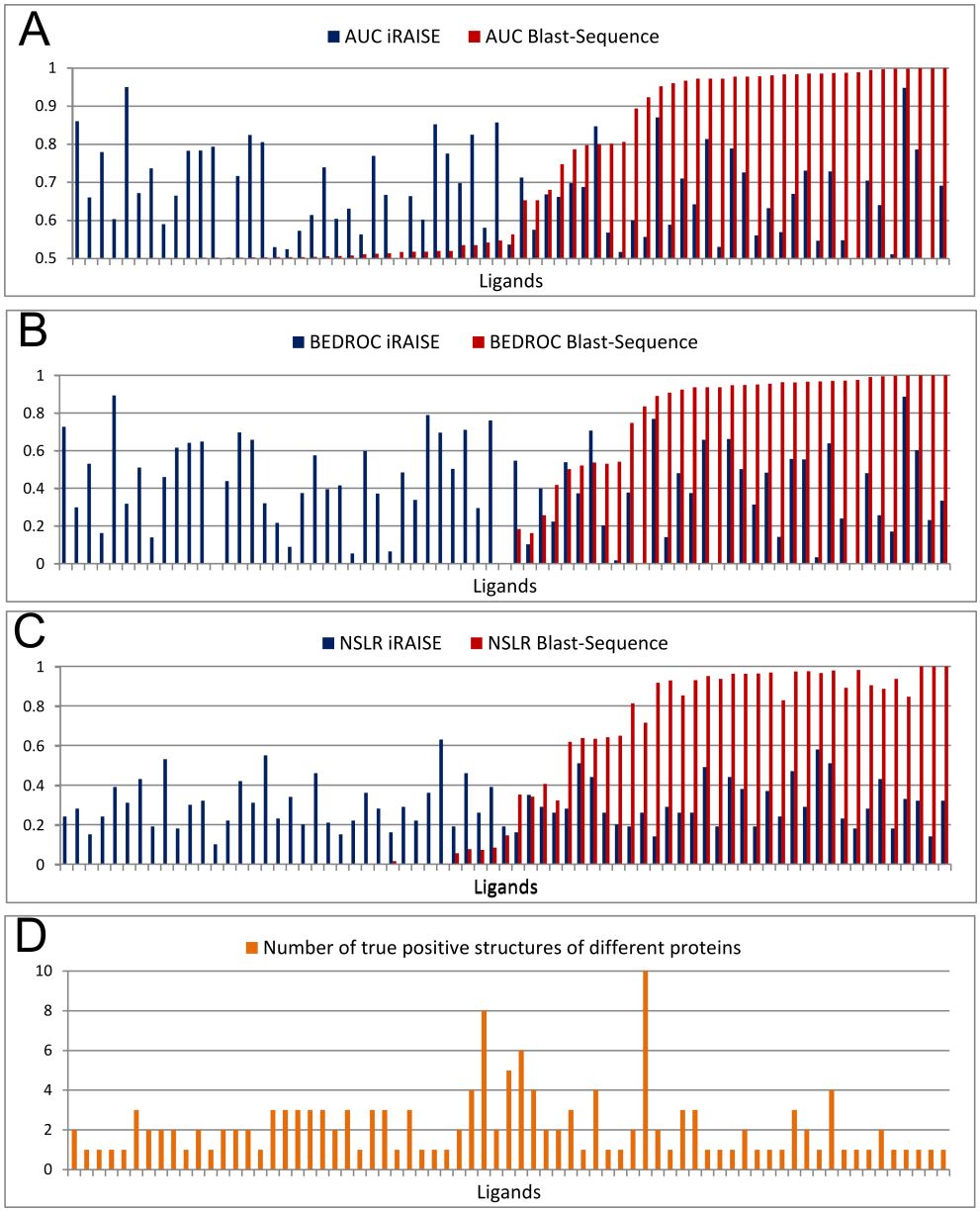


Figure 7.17: Results of pBLAST sequence-based target prediction in comparison to *i*RAISE's plotted for 71 ligands of the Drugs/sc-PDB data set. All plots are sorted by AUC values of pBLAST predictions. A) AUC values . B) BEDROC values. C) NSLR values D) Number of different true targets for each ligand contained in the structure data set. *(This figure was originally published in Schomburg and Rarey [2014].)*

Method	Median rank of first TP on 2556 distinct proteins (percent)	Median rank of first TP on 2879 distinct proteins (percent)	Median rank first of TP on 2879 distinct proteins with EC-TP (percent)
pharm-rigid1	4 (0.16%)		
pharm-rigid2	4 (0.16%)		
pharm-flex1	6 (0.23%)		
pharm-flex2	4 (0.16%)		
Surflex1	65 (2.5%)		
Plants1	113 (4.4%)		
*i*RAISE-flex		33 (1.15%)	8 (0.28%)
*i*RAISE-crystal		2 (0.07%)	2 (0.07%)

Table 7.5: Medians of ranks of first true target of 117 ligands of the sc-PDB Diverse Set for pharmacophore-based methods (rigid1, rigid2, flex1, flex2), two docking methods (Surflex and Plants), for *i*RAISE with conformations (*i*RAISE flex) and for *i*RAISE with the crystallized ligand (*i*RAISE crystal). EC-TP = annotation of true positives (TP) with EC numbers.

7.7 Comparison to pharmacophore-based target prediction

For the comparison of *i*RAISE to a pharmacophore-based target prediction method, the results and evaluation experiment of Meslamani et al. were used (see section 6.2.7 for the experiment). For the 117 ligands of the sc-PDB Diverse Set, the rank of the first true positive target (=FTP) in the score-ordered list was evaluated, based on four varying pharmacophore-based approaches of Meslamani, two classic docking approaches (Surflex and Plants) and two *i*RAISE approaches, one with ligand conformations (*i*RAISE flex) and one with the crystal ligand conformation (*i*RAISE crystal). The ranks for the four pharmacophore-based approaches and of the two classic docking approaches were extracted from the Supporting Information of Meslamani et al. [2012]. For the ranks of *i*RAISE for each of the 117 ligands, see Appendix H.

In Table 7.5, the medians of the first true positives of all methods are listed. Since the target prediction with *i*RAISE was done on a more recent version of the sc-PDB with a higher number of distinct proteins (2556 versus 2879), the absolute numbers are not comparable. Therefore, the percentages are given in the table as well. Next to the true-positive assignment after Meslamani, true positives were also assigned via the EC number. Median FTPs for *i*RAISE for both approaches - screening with up to 200 conformations and screening with the crystal conformation - are listed. The median rank of the FTP of *i*RAISE-flex is with 1.15% superior to both docking methods, and with true positive assignment via EC

numbers with 0.28% comparable to the pharmacophore-based methods.
The median FTP of *i*RAISE-crystal is only comparable to the pharmacophore-based methods. This experiment is artificial, since the binding conformation of a ligand would in a real scenario not be known beforehand. Since in the pharmacophore-based methods, though, the pharmacophores are derived from the co-crystallized complexes, preliminary information is exploited here as well. For the docking-based methods, however, the results were not available for comparison. The median FTP of *i*RAISE-crystal is at 0.07% of the target list even superior than all pharmacophore-based methods.
Next to the total performance, the performance at several percentages of the data set is assessed. In Figure 7.18, the FTP ranks for the four pharmacophore-based approaches, the two classic docking approaches, *i*RAISE-flex and *i*RAISE-crystal are summed for position 1 and the first 1%, 2%, 5%, 10%, 20%, 30% and 50% of the data set. Here again the percentages are calculated for the two *i*RAISE approaches on 2879 distinct proteins and for the other methods on 2556 distinct proteins. *i*RAISE-crystal outperforms all other methods till 5% of the database and *i*RAISE-flex outperforms the classic docking methods at all percentages. Since in experiments only the first percentages of the ranked lists can be screened, the performance of the methods at the first percentages is the most important.

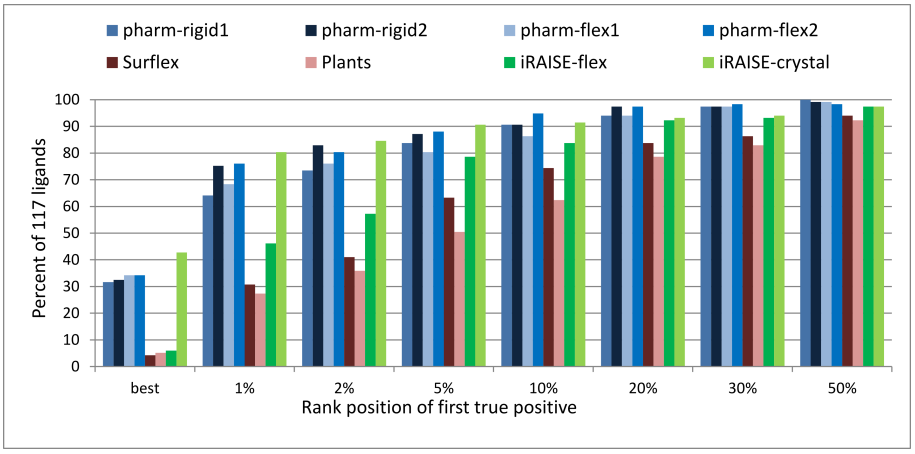


Figure 7.18: Ranking of true target for each of the 117 ligands of the sc-PDB Diverse Set, summed at position 1, the first 5, 10, 20, 30 and 50%. The blue bars show the ranks of four pharmacophore-based methods, the red bars the ranks of the two classic docking approaches and the green bars the ranks of *i*RAISE with flexible ligands and the co-crystallized ligand.

In Figure 7.19, for all 117 ligands, the FTP is plotted on logarithmic scale for one pharmacophore-based approach, the two classic docking approaches, *i*RAISE-flex and *i*RAISE-crystal. This

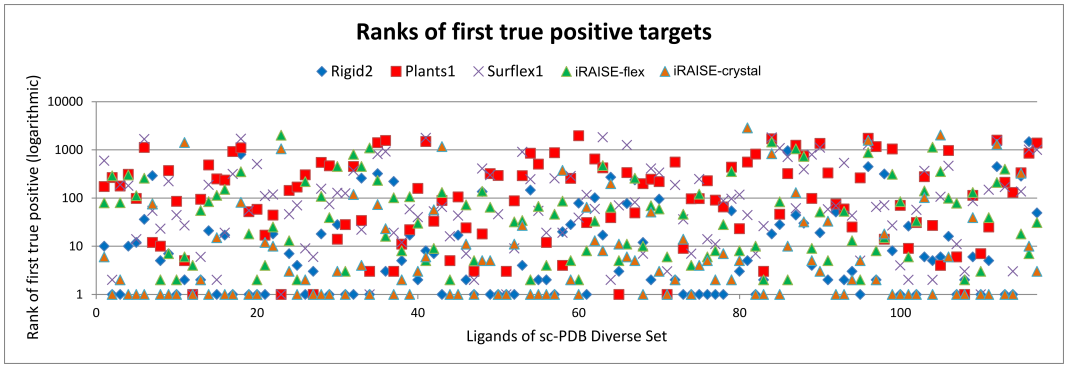


Figure 7.19: Distribution of ranks of first true positives for the 117 sc-PDB Diverse Set ligands by a pharmacophore-based method (Rigid2), two classic docking approaches (Plants1 and Surflex1) and the *i*RAISE-flex and *i*RAISE-crystal screening.

plot shall show, if the predictions of the four methods are correlated, i.e., if all approaches perform well or poor on the same ligands. The vast scattering of the FTP ranks for the five methods indicates that this is not the case. Each method has obviously difficulties with different ligands.

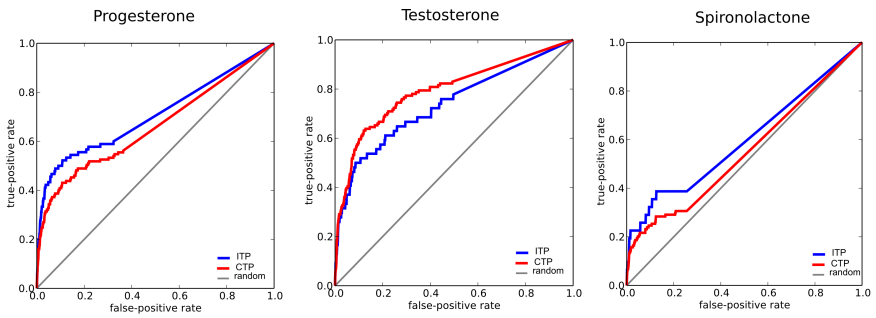
7.8 Prediction of unknown targets

In this section, *i*RAISE's ability of predicting unknown and off-targets for a compound is evaluated. Firstly, the clustering of drugs by ECFP of the Drugs/sc-PDB data set is studied for finding so far not-annotated targets. Secondly, the capability of *i*RAISE to predict different binding modes for one drug in several targets is evaluated. Thirdly, *i*RAISE is used to study drugs with unknown mechanism-of-action.

7.8.1 Analysis of drug clusters

During development of the Drugs/sc-PDB data set, the drugs were clustered with a similarity measure (see section 5.5) for two reasons: Firstly, the number of different molecule clusters (38 singletons and 14 clusters of two to four molecules, Figure 5.5 and Figure 5.6) shows how diverse the set is. Secondly, the clusters can be exploited for target analysis.
The target annotations by the DrugBank are only for three of the 14 clusters for all drug members identical. These are the clusters of hydrochlorothiazide and hydroflumethiazide, of linagliptin and alogliptin and of vardenafil and sildenafil. For the other 11 clusters the target annotations from the DrugBank are differing among the members of a cluster.
There are two reasons why target annotations may differ for structurally highly similar molecules: Firstly, they bind to different targets in spite of their similarity, which is an not uncommon effect called an activity cliff (see Stumpfe and Bajorath [2012]). Secondly, the annotation of the compounds in the DrugBank may be incomplete. The molecules of a cluster might bind to the same targets while not all interactions are yet observed or reported. With *i*RAISE, these effects can be studied and it can be hypothesized which case holds true for the clusters of this data set. As an experiment, the target annotations for all molecules of one cluster were combined, and the enrichment was calculated as if these combined target structures were all true positives. The ROC plots of this experiment for the clusters are shown in Appendix G, Figure G.5 for the clusters with three and four molecules and in Appendix G Figure G.6 for the clusters with two drugs.

As an example, the cluster of progesterone, testosterone and spironolactone is discussed. For the structure diagrams of these molecules see Figure 5.6 on page 83. In Figure 7.20, the ROC curves for the three drugs with both true target annotation strategies are shown and the targets are listed.
For progesterone, the DrugBank annotation labels the progesterone receptor, the estrogen receptor and the mineralcorticoid receptor as true tagets. Of these, there are 90 structures in the data set, of which 36 structures (of all three targets) are not found. Combining the true targets with the others of this cluster adds the structures of the androgen receptor,



Individual true positive targets	Progesterone receptor Estrogen receptor Mineralcorticoid receptor	Androgen receptor Amino oxidase Serum albumin	Mineralcorticoid receptor Androgen receptor
Individual true positive structures	90	54	31
Combined true positive structures	137	141	134

Figure 7.20: ROC curves of the three drugs progesterone, testosterone and spironolactone which are members of the same cluster. Comparison of ROC curves for true positive targets based on the annotation by DrugBank (blue lines) and by combination of the true targets for all three drugs (red lines). ITP = individual true positive annotation, CTP= combined true positive annotation

amino oxidase and serum albumin to the total of 137 true target structures. The ROC curves show that the total enrichment gets worse with the added true positives. Looking into details shows that *i*RAISE suggests that indeed, the androgen receptor and serum albumin are targets for progesterone while amine oxidase is not. Of the androgen receptor, nearly each structure is hit in the *i*RAISE screening and the first ranks are at positions 2 and 4 of the *i*RAISE-score ordered list of all 7915 structures. All serum albumin structures are also identified as true targets, but no amine oxidase structures is hit.

For testosterone, the added true positives are the progesterone receptor, the estrogen receptor and the mineralcorticoid receptor. The screening by *i*RAISE suggests that all three receptors are also targets for testosterone: The first estrogen receptor structure is hit at rank 9, the first progesterone structure at rank 61 and the first mineralcorticoid receptor structure at rank 82.

For spironolactone only the mineralcorticoid and the androgen receptor are annotated as targets in the DrugBank. The *i*RAISE screening predicts for this drug the estrogen receptor as true target with the first structure ranked to position 10. Serum albumin structures and progestrone structures are also identified as true targets but not ranked very high. Amino

oxidase is not predicted as target for spironolactone.
Of these predictions, literature search confirms the following interactions: spironolactone-progesterone receptor (Fernandez et al. [1983]), spironolactone-estrogen receptor (Meyers et al. [2010]), progesterone-serum albumin (Andre et al. [2003]), testosterone-estrogen receptor (Huang et al. [2011b]), testosterone-progesterone receptor (ChEMBL Gaulton et al. [2012]).
For those interactions of which no literature reporting could be found, the question whether the prediction is correct remains unanswered. However, with this experiment, it could be shown that the predictions of *i*RAISE on the Drugs/sc-PDB data set contain true targets which are yet not listed in the DrugBank. By joining ligand-based similarity with structure-based target prediction the suggestions based on ligand similarity can be further studied and examined.

7.8.2 Analysis of capability to predict diverse binding modes

For a structure-based target prediction method it is essential that it is capable of predicting ligands in various binding modes. Molecules may bind to structurally diverse proteins and exhibit different binding modes. Thus, the method should not be biased towards any binding mode. The Drugs/sc-PDB data set contains several ligands which bind to as much as 5 to 10 different proteins. In order to evaluate *i*RAISE's capability to predict diverse binding modes for one ligand, the example of diclofenac was consulted (see Figure 5.6 on page 83 for the 2D structure diagram of diclofenac).
Diclofenac is a drug used as anti-inflammatory and pain suppressor. According to DrugBank, it binds to six different targets: The primary targets are prostaglandin G/H synthases 1 and 2 (also known as cyclooxygenases 1 and 2). Next to these, diclofenac binds to phospholipase A2, transthyretin and serum albumin and is a substrate of UDP glucuronosyltransferase. Of these proteins, there are 71 structures in the structure data set of the Drugs/sc-PDB data set. The *i*RAISE screening successfully identifies all proteins as true targets.
In Figure 7.21 in the left column the best *i*RAISE-scored poses in the active sites of all five targets are shown. In the right column, the complete proteins are shown in ribbon style. This column shows how structurally diverse the proteins are, with prostaglandin G/H synthase containing a heme, phospholipase A2 as a rather small protein, transthyretin dominated by β-sheets and serum albumin with mainly α-helices.
Representatively for the binding to prostaglandin G/H synthases, in Figure 7.21A, the binding to a prostaglandin G/H synthase 1 is shown. The carboxy-group of diclofenac is contributing most to its hydrophilic interactions. In 1HT5, *i*RAISE predicts a pose where this group interacts with a serine and a tyrosine. In Figure 7.21B, the binding to UDP glucurono-

syltransferase is shown for which diclofenac is a substrate. Here, *i*RAISE predicts a pose where the carboxy-group interacts with a tyrosine and a backbone amide. The binding of diclofenac to a phospholipase structure is shown in Figure 7.21C where the carboxy-group gets protonated and interacts with an aspartate and the calcium ion of the active site. In Figure 7.21D the binding of diclofenac to transthyretin, which is a hormone carrier protein, is shown. Here the active site is located between two protein domains and the carboxy-group interacts with several serine amino acids. Finally, in Figure 7.21E, the pose predicted by *i*RAISE in serum album is drawn, where the carboxy-group interacts with arginine.
In this study no conclusion can be drawn if the by *i*RAISE predicted binding modes are correct, since the structures do not contain diclofenac as co-crystallized ligand. However, it can be concluded that *i*RAISE is able to predict diverse binding modes and is able to identify diverse protein structures as true targets for a ligand.

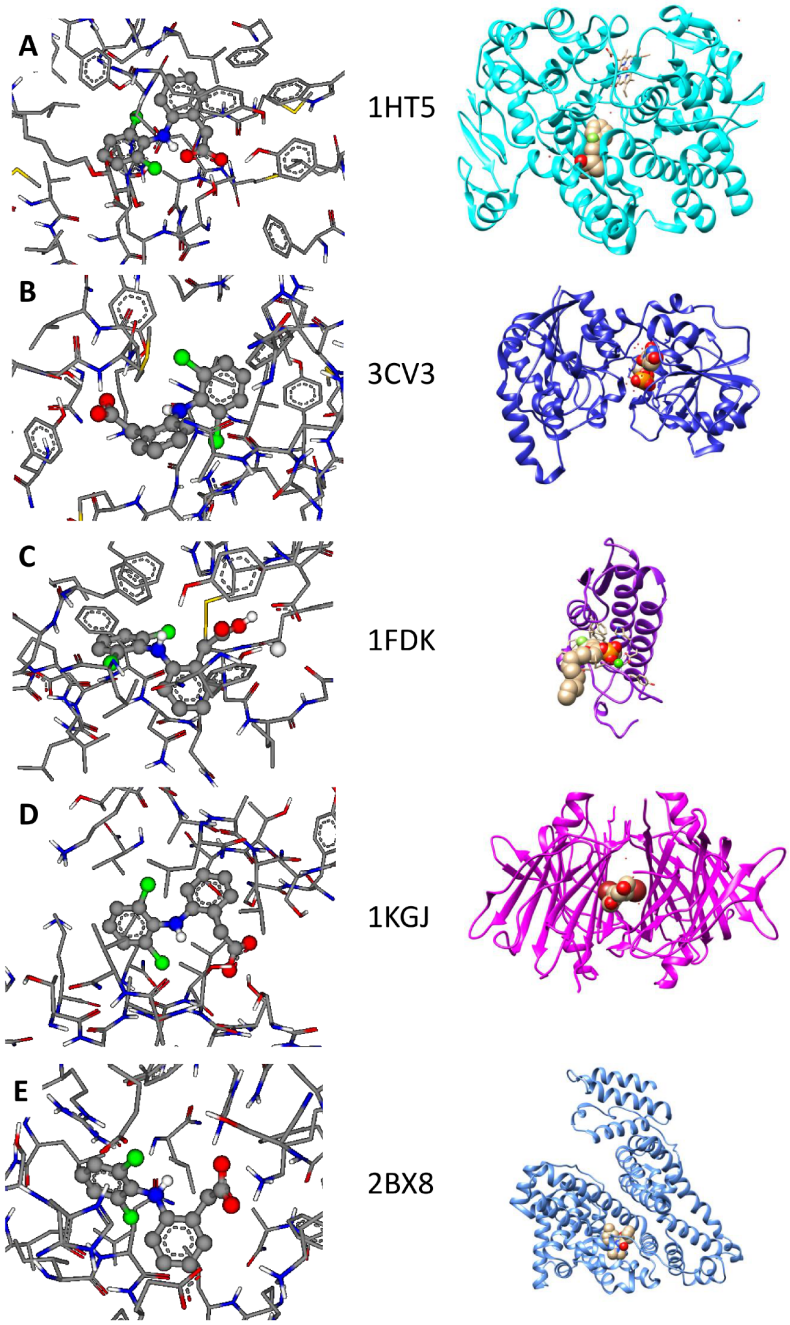


Figure 7.21: Poses of diclofenac in five of its target proteins. A) prostaglandin G/H synthase 1 (PDB code 1HT5) B) UDP glucuronosyltransferase (PDB code 3CV3) C) phosholipase (PDB code 1FDK) D) transthyretin (PDB code 1KGJ) E) serum albumin (PDB code 2BX8)

7.8.3 Unknown mechanism-of-action

The prediction of targets for drugs with unknown mechanism of action is another application of inverse virtual screening. For such compounds, the phenotypic reaction is known but not the responsible mechanism on protein level. Inverse virtual screening can aid in identification of the targets on protein level. For many drugs with unknown mechanism of action registered in the DrugBank, the mechanism of action is completely unknown yet. One example is Clioquinol, which is anti-fungal but was removed from the market in 1983 due to neurotoxicity. For this compound, DrugBank states "Clioquinol is bacteriostatic, however, the precise mechanism of its action is unknown." (Wishart et al. [2006], accession number DB04815). Target predictions for such drugs could aid tremendously in drug design, since if the primary target of the drug would be found, rational design on basis of the target could be conducted. Further, by predicting the side-effect causing targets, the selectivity of the drug could be increased with rational drug design and thus new powerful drugs developed.
For this thesis, however, the evaluation of predicted targets is difficult, since the results would need to be tested experimentally to prove the predictions of *i*RAISE. Therefore, the DrugBank was searched for drugs of which the mechanism of action is yet not unambiguously clarified, but for which already targets are suspected to be the cause of the observed phenotype. It was then evaluated if the result of an *i*RAISE screening on basis of the structures of the sc-PDB supports the hypothesis stated in the DrugBank or not.
Of this type, sulindac (see Figure 7.22 for the 2D structure diagram) was chosen as an example, since it fulfills the above described condition and the sc-PDB structure data contains the necessary structures.
Sulindac is an anti-inflammatory drug, for which is stated in the DrugBank "The exact mechanism of its NSAIA [nonsteroidal anti-inflammatory agent] properties is unknown, but it [sulindac] is thought to act on enzymes COX-1 and COX-2, inhibiting prostaglandin synthesis." (Wishart et al. [2006], accession number DB00605). Therefore, it was evaluated if the *i*RAISE screening finds COX-1 and COX-2 (=prostaglandin G/H synthases) as true targets from the sc-PDB for this drug.
Both targets were indeed identified by *i*RAISE as true targets. The best ranked structure of COX-2 (3LN1) is scored with -1.36 (gsw-score) and the best ranked COX-1 structure (2OYE) with -1.11, which is clearly above average and thus considered as a target by *i*RAISE.
In Figure 7.22, the poses are shown for both cyclooxygenases. The co-crystallized ligand of COX-2 (shown on the left side) is rather similar to sulindac and the pose predicted for sulindac covers the same part of the active site. The ligand of the COX-1 complex is much smaller than sulindac and thus the pose of sulindac protrudes into another subpocket of the active site.

In summary, the *i*RAISE screening supports the hypothesis that COX-1 and COX-2 are targets for sulindac.

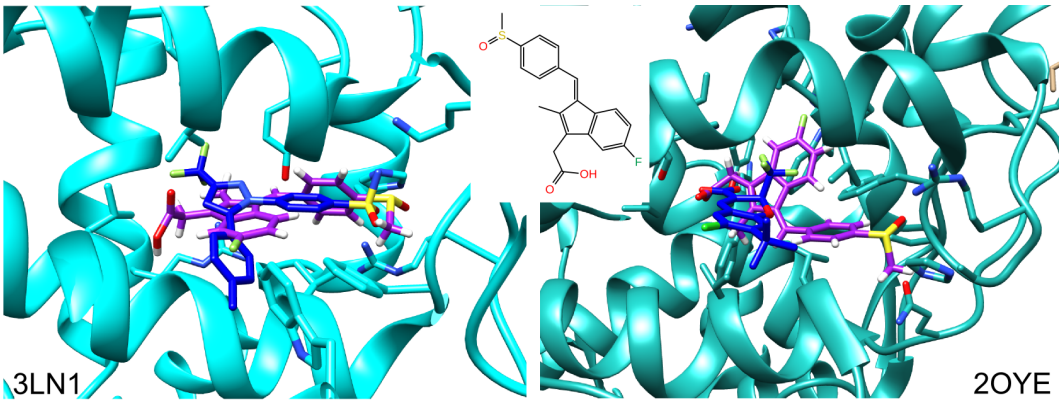


Figure 7.22: Poses of sulindac (pink) and 2D structure diagram A) prostaglandin G/H synthase 2 (PDB code 3LN1) B) prostaglandin G/H synthase 1 (PDB code 2OYE). Co-crystallized ligand of the shown complexes colored in blue.

7.9 Parametrization

In Appendix B, a list of *i*RAISE's parameters is given. For parameter setting studies, the Iridium-HT data set (see Chapter 5) was used. The running time and the rank of the true target for the 121 ligands were monitored in the parametrization experiment. The parametrization was a balancing act between improving the ranking of the true target, i.e. increasing the specificity, versus finding for each ligand still its true target at all, i.e. not decreasing the sensitivity.

The selectivity/sensitivity balance is mainly influenced by scoring parameters. Next to the ranking of true targets, the binding mode can be assessed via RMSD calculation of the ligand to the crystal pose in its true target of the Iridium targets. The parametrization of tuning *i*RAISE to produce good binding modes was not on the level of scoring parameters, but on the level of data amount. The more conformations are created per ligand and the softer the triangle matching parameters are set and the higher the number of poses allowed to pass the Scoring Cascade, the better the binding mode prediction, i.e. the smaller the RMSDs. The compromise between running time and precision of binding mode can be dependent on the number of proteins to be screened in a project. Best parameters for a good binding mode/running-time compromise are set as defaults.

7.10 Running time evaluation

The running time of *i*RAISE is evaluated for *i*RAISE's registration procedure and for the screening procedure separately. The running time was averaged over the Astex Diverse Set 84 times 85 (84 ligands against 85 targets). Here, the Iridium-HT set was not used on purpose, since in the Iridium data set several pockets per protein are used. In this evaluation, however, the running time per protein structure was assessed in the default case that for each pocket a new protein has to be evaluated. This way, the worst case that for each pocket a separate protein has to be loaded from the database was assessed. The Astex Diverse Set is also especially suited for this evaluation, since the protein triangle descriptors of the Astex proteins nearly fill one partition and are therefore a good estimation of the running time if each partition is screened in parallel.
The running time evaluations were performed on a Suse 12.2 workstation with Intel® Core™ i5/3570 CPU@3.4GHz, 4 cores and 8GB RAM, single threaded.

Registration procedure Like discussed before, the running time for creating a *i*RAISE project with indexed descriptors and the protein database is not as critical as the running time for screening this project, since the registration procedure needs to be executed only once for a set of proteins. In Figure 7.23, a distribution of the steps of the registration procedure is shown. The highest running time is required by the triangle descriptor calculation and only a fraction is used by binning the triangle descriptors and writing them to the bitmap index (=Indexing descriptors). Initialization of the protein-ligand complex from a PDB file and calculation of the active site (=Initialization from file) requires the second-least time while writing the protein to the database barely contributes to the total time. Averaging the 622 seconds needed in total over the 85 proteins results in averagely 7.3 seconds for one protein.

Screening In Figure 7.24 A, the overall time distribution of the screening procedure and in B in detail the time distribution for the most time-consuming steps of the Scoring Cascade is shown. The first steps of molecule conformation generation, triangle descriptor generation and unique descriptor calculation (=Descriptor generation) all do not contribute much to the overall running time. The most time-consuming step is the querying of the FastBit descriptor index with all query triangle descriptors. The following steps then only have to be executed for targets, for which a match occurred. Reading the protein from the database

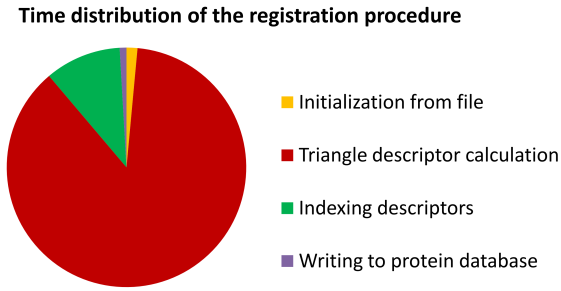


Figure 7.23: Distribution of running time of *i*RAISE's registration procedure on the Astex Diverse Set.

is comparably fast, while the re-initialization of the protein with the grid and further score information needed later takes a greater part of the running time. The next step, the grid-based clash-test and prescoring is the second-largest part of the running time and the last step, the Scoring Cascade the third largest parts.
This last step, the Scoring Cascade is further evaluated in Figure 7.24 B. The most time-consuming step is Protoss, followed by the atom-based clash test, the interaction score and the pose coverage. The pocket coverage as last step is applied to the fewest poses, since the Scoring Cascade acts as a filter reducing the number of poses from step to step. Therefore this step is not significantly contributing to the overall running time.
In total, averaged over all 84 Astex ligands and averaged over the 85 targets, the screening procedure takes 7.1 (median 5.1) seconds for screening one protein structure. The time needed is highly dependent on the ligand: descriptors of small, hydrophilic ligands match many target descriptors and almost each protein has to be re-initialized, while for large ligands only a small number of proteins is matched in the descriptor step and thus the following steps after matching need only be applied to few proteins. In the 84 ligands of the Astex Diverse Set for example, the time to screen all 85 targets ranges between 82.4 seconds (indirubin-3'-monooxmine, ligand of the complex 1Q41) and 3259.1 seconds (pantoate, ligand of the complex 1N2J).

Comparison to inverse docking The running time of *i*RAISE on the Astex Diverse Set with an average of 7.3 seconds for the preprocessing procedure and 7.1 (median 5.1) seconds for the screening procedure were compared to the running time of FlexX-docking and HYDE scoring in the LeadIT-Suite of the BioSolveIT (www.biosolveit.de).
The LeadIT-suite also has a mode of scripting protein preparation, therefore, the initializa-

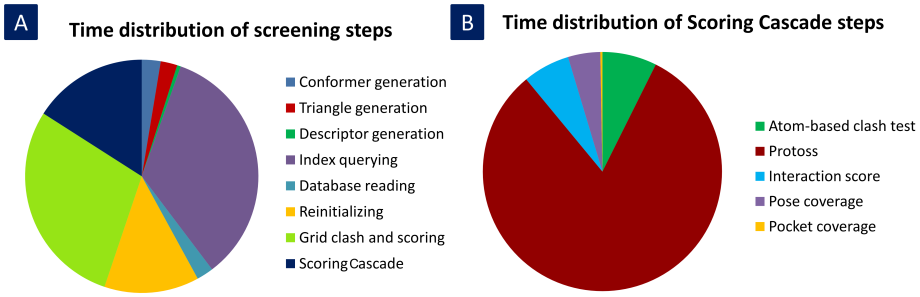


Figure 7.24: Distribution of running times of *i*RAISE screening (A) and in detail of the most time-consuming steps of the Scoring Cascade without the grid-clash test (B) on the Astex Diverse Set.

tion did not have to be done manually with the GUI. On average, this protein preparation took 26 seconds per protein. The screening then took on average 113 seconds for pose generation with FlexX and scoring with HYDE.

8

Biotechnological Application Case Study

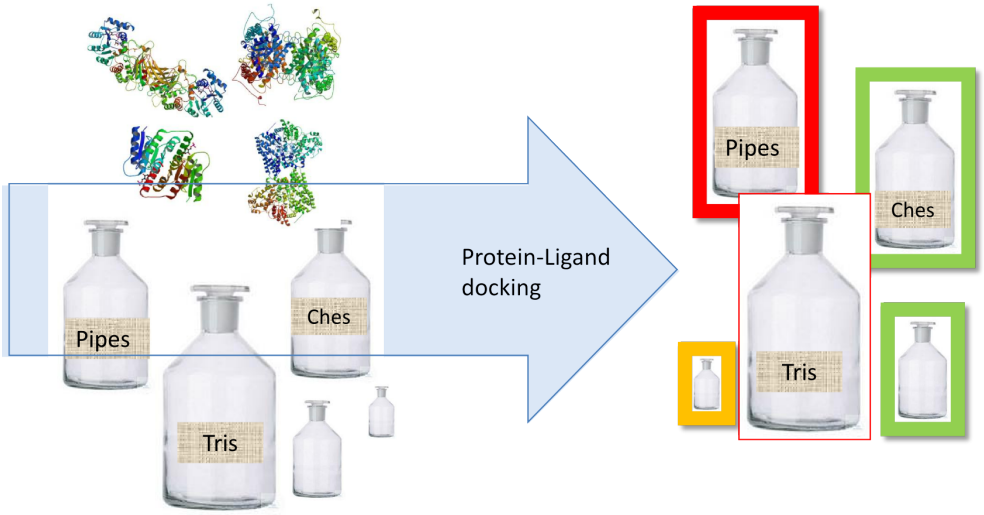


Figure 8.1: Computational methods like protein-ligand docking can support biotechnological experiment setup, e.g., the choice of buffers.

In order to understand the nature of questions and problems in biotechnological contexts, a case study in this field was conducted. The aim was to gather an understanding of how computational structure-based methods may support biotechnological method development, of what requirements computational methods have to fulfill to be able to answer biotechnological questions and last but not least, of how biotechnological scientists see and use structure-based computational methods to support their work.
In this chapter, first the biotechnological project is described, then the identification of questions where computational methods may support the project is discussed, and finally, the computational assessment of the identified questions is described.

8.1 Project description

The project on which the case study was conducted was part of one of Hamburg's Excellence Cluster projects. Its title is 'Fundamentals for Synthetic Biological Systems (SynBio)'. Twelve academic groups from various institutes and universities in Hamburg joined their expert knowledge to work on a specific synthetic multi-enzyme pathway for learning and developing methods which support similar approaches in future.
Synthetic multi-enzyme pathways are used for high-yield biotechnological productions of chemicals which have higher economic value than the substrates. Since enzymes catalyze chemical reactions and, thus, the energy needed to turnover a substrate in a product is rather low, they are used systematically in biotechnology. Relatively new, however, is the concept of joining several enzyme reactions subsequently in a reaction chain which is not found in nature. This way, the turnover of substrates to products is possible in a way that does not occur in nature. Such reaction chains can consist of more than a dozen different enzymes.
Multi-enzyme pathways show much promise as they allow new chemical reactions to be processed entirely by enzymes, however, they also pose many challenges. Enzymes have different reaction optima with respect to reaction solution, temperature and pH, may need cofactors and can be inhibited by intermediates of the reaction pathway. These factors have to be evaluated during the setup of a multi-enzyme pathway. Optimally, if all enzymes are active and stable at the same conditions, the complete reaction pathway may be carried out in one single reaction container, called one-pot reaction.
The subject of the SynBio project was the synthetic multi-enzyme reaction pathway for H_2 production from starch presented by Zhang et al. (Zhang et al. [2007]). The synthetic pathway consists of 13 different enzymes, partly from the pentose-phosphate pathway. On the

basis of this pathway, the partners of the SynBio project studied the design of multi-enzyme pathways.
There are already some computational approaches supporting the design of multi-enzyme pathways: Prediction of possible enzymatic pathways (Arita [2000], Cho et al. [2010], Li et al. [2004], McShan et al. [2003], Wu et al. [2011]), assessment of enzyme stability in non-aqueous solutions with Molecular Dynamic studies (Lousa et al. [2012]) or identification of enzymes by reaction intermediate-docking (Hermann et al. [2006], Hermann et al. [2007]).
The following tasks were identified to be addressed computationally for yield-optimization in the multi-enzyme pathway of H_2-production:

- Prediction of inhibitory potential of buffer agents
- Analysis of feedback-inhibition
- Identification of enzyme structure with highest activity

The first two tasks were addressed with protein-ligand docking and the third was addressed with the inverse screening tool *i*RAISE developed during this thesis. All three tasks are discussed in detail in the following sections. The assessment and results of the first task have been published in Schomburg et al. [2012] in collaboration with partners from the SynBio project.

8.2 Prediction of inhibitory potential of buffer agents

If several enzyme reactions shall be highly efficient in one reaction pot, the enzymes all have to be reactive at the same reaction conditions. One factor of the reaction conditions are buffer agents. These agents keep the pH value of the reaction medium stable, even if hydroxyl or oxonium ions are released during the reaction. Buffering agents are chosen after the pH range which they are able to buffer. However, for a pH range several possible buffer agents exist. Often, the buffer is chosen after habit, although rational selection promises higher enzyme activity. A buffer agent can compete with reaction substrates for the active site and therefore competitively inhibit the enzyme reaction. Such effects lower or even stop the substrate turnover completely. Therefore, a rational selection is needed to avoid such effects.
With a standard protein-ligand docking approach adapted to this problem, a prediction of the inhibitory potential of a buffer compound is possible. Here, the classic protein-ligand docking approach was adapted to be able to to classify a buffers as *inhibiting*, *potentially inhibiting* and *not inhibiting* for one enzyme. In the evaluation, the approach was tested

on buffer inhibition examples reported in literature (retrospective validation) as well as on five of the 13 enzymes from the multi-enzyme pathway of the SynBio project (prospective validation).

8.2.1 Method

The workflow of the method is shown in Figure 8.2. Firstly, the 3D structures of the proteins were collected from the Protein Data Bank. The protein structures were chosen with respect to resolution, organism (the organism that was used in the collaborative working groups of the SynBio project was preferred), and co-crystallized ligand. See Figure 8.4 for the PDB codes used.

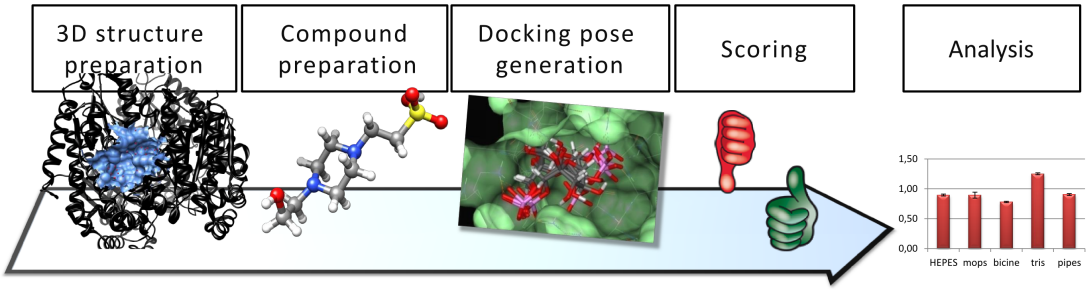


Figure 8.2: The workflow of the computational classification of the inhibitory potential of buffer compounds is divided into the five steps 3D structure preparation, compound preparation, docking pose generation, scoring and analysis.

Then, the 3D structures of the buffer agents which should be tested were collected. Buffer agents were chosen after the buffering range which had to cover the pH optimum of the enzymes and for the literature cases the buffer agents for which an inhibition was reported were used. For the prospective study, 14 buffers were chosen; a 2D visualization is shown in Figure 8.3. The 3D structures of the buffers were collected from the PubChem database (Bolton et al. [2008]). Each protonation state was sampled with the Naomi software library of the ZBH (Urbaczek et al. [2011]). Further, the 3D structures of the substrates and products were collected and processed the same way.
Then, the proteins were prepared for docking by active site identification with a co-crystallized reference ligand or by catalytic residues described in the literature.
For protein preparation and the succeeding docking procedure, the LeadIT software suite from the BioSolveIT was used (www.biosolveit.de/leadit).

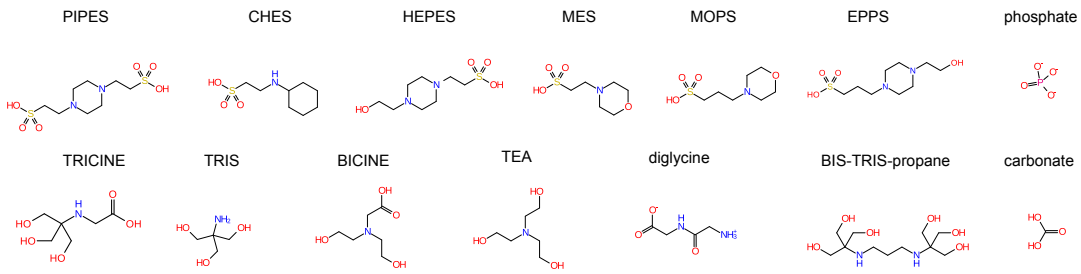


Figure 8.3: Structure diagrams of the buffers compounds. Abbreviations: PIPES (1,4-piperazinediethanesulfonic acid), CHES (2-(cyclohexylamino)ethanesulfonic acid), HEPES (2-[4-(2-hydroxyethyl)piperazin-1-yl]ethanesulfonic acid), MES (2-(N-morpholino)ethanesulfonic acid), MOPS (3-morpholin-4-ylpropane-1-sulfonic acid), EPPS (3-[4-(2-hydroxyethyl)piperazin-1-yl]propane-1-sulfonic acid), TRICINE (2-[[1,3-dihydroxy-2-(hydroxymethyl)propan-2-yl]amino]acetic acid), TRIS (2-amino-2-(hydroxymethyl)propane-1,3-diol), BICINE (2-[bis(2-hydroxyethyl)amino]acetic acid), TEA (Triethanolamine), BIS-TRIS-propane (2-[3-[[1,3-dihydroxy-2-(hydroxymethyl)propan-2-yl]amino]propylamino]-2-(hydroxymethyl)propane-1,3-diol)

The LeadIT suite uses the FlexX docking algorithm (Rarey et al. [1996]) for pose generation. The integrated Protoss tool (Lippert and Rarey [2009]) automatically optimizes the hydrogen network of the protein-ligand pose. For each docking, 100 poses of the ligand were generated. For the prospective experiments, poses for each of the five enzymes of a library of all protonation states of the buffers and the substrate and the product were generated. For the retrospective cases, the reported inhibitory buffers, not inhibitory buffers and the substrates and products were docked.

For scoring, the HYDE scoring function was used (Schneider et al. [2012]). It is suited for this problem for two reasons. Firstly, it focuses on hydrogen bond formation and dehydration. Since buffer compounds are highly hydrophilic and can therefore form many hydrophilic interactions in a protein active site, they might be over-scored by some scoring functions. HYDE, however, penalizes dehydration and non-perfect hydrogen bonds and is thus capable of generating sensible scores even for very hydrophilic compounds as buffers. Secondly, the HYDE scoring function is not biased to scoring large compounds higher, as many scoring functions tend to (Pan et al. [2003]). As buffer compounds highly differ in size as well as compared to substrates and products of the enzymatic reaction, this feature is essential for the experiments conducted here.

For classification of the buffer compounds into the categories *inhibiting*, *potentially inhibiting* and *not inhibiting*, the scores of the buffer compounds were compared to that of the reaction substrate. If the score of a buffer was higher than 90% of the substrate score, the

buffer was categorized as *inhibiting*. A score between 90% and 75% of the substrate was categorized as *potentially inhibiting* and all lower scores as *not inhibiting*.
For the prospective experiments, enzyme activity assays were conducted for validation of the predictions by the collaborators of the SynBio project. For details on the experiment setup, see Schomburg et al. [2012].

8.2.2 Retrospective experiments

The retrospective experiments cover four cases from the literature where an effect on an enzymatic reaction by a buffer compound was reported. Of the four evaluated cases, three inhibitions are due to competitive inhibition of the active site, while in the fourth case the enzyme reaction is enhanced by the buffer due to transphosphorylation by the buffer agents. In the following, the results are discussed for each enzyme separately.

Amylosucrase, EC 2.4.1.4 MacKenzie et al. (MacKenzie et al. [1977]) report an inhibition of amylosucrase by TRIS buffer. Fortunately, crystal structures of both, the protein in complex with the substrate sucrose (PDB code 1JGI) and in complex with the TRIS buffer compound (PDB code 1G5A) are available, rendering these complexes suitable for scoring. Further, the cacodylate buffer used by MacKenzie was docked. The sucrose complex is scored with -26kJ/mol, the TRIS complex with -28kJ/mol and the docked cacodylate with -17kJ/mol. Docking TRIS even results in a higher score of -31kJ/mol. As the score for the TRIS buffer is higher than the substrate score, the inhibition is successfully recovered by the docking approach. Furthermore, the *not inhibiting* buffer cacodylate is also correctly classified.

Exopolyphosphatase, EC 3.6.1.11 For the enzyme exopolyphosphatase, Wurst and Kornberg showed an inhibition by the buffer compounds CHES and MES (Wurst and Kornberg [1994]). The substrates of the ester hydrolysis are poly-phosphates of variable length. For docking, triphosphate was used. The score for the substrate triphosphate is -19 kJ/mol, for the CHES buffer compound the score is -39 kJ/mol and for the MES buffer compound -30 kJ/mol. Therefore, these buffer compounds are classified as *inhibiting* by our method, in agreement with the experimental observation by Wurst and Kornberg.

Creatine Kinase, EC 2.7.3.2 Several buffer compounds were reported to have an inhibitory potential on the enzyme creatine kinase. Gerhardt shows inhibition by phosphate buffer, PIPES buffer, sulfate buffer, MOPS buffer and BIS-TRIS-propane buffer (Gerhardt

Buffer compound	Active site		Cofactor binding site	
	HYDE score in kJ/mol	Classification	HYDE score in kJ/mol	Classification
BIS-TRIS-propane	-47	inhibiting	-45	inhibiting
MOPS	-32	inhibiting	-24	potentially inhibiting
PIPES	-31	inhibiting	-25	potentially inhibiting
TRIS	-28	inhibiting	-29	inhibiting
phosphate	-19	potentially inhibiting	-21	not inhibiting
sulfate	-19	potentially inhibiting	-23	potentially inhibiting
creatine (substrate)	-24			
ATP (cofactor)			-29	

Table 8.1: HYDE scores and categorization of buffer compounds docked into the active site and cofactor binding site of creatine kinase

[1983]). The enzyme creatine kinase catalyzes the reaction from creatine to phospho-creatine utilizing ATP as cofactor. Therefore, the inhibition can be caused by either occupation of the cofactor binding site or the substrate binding site. Consequently, we docked all the buffer compounds into the substrate binding site as well as into the cofactor binding site. In addition, we docked an imidazole buffer since this buffer was recommended by Gerhardt. Table 8.1 shows the scores and the categorization of all compounds. For the co-factor binding site, only BIS-TRIS-propane gets a significantly higher score than the substrate ATP. For the active site, however, phosphate and sulfate buffer are categorized as *potentially inhibiting* and all other buffers are categorized as *inhibiting*. Only the recommended imidazole buffer is categorized as *not inhibiting*.

Alkaline phosphatase, EC 3.1.3.1 For alkaline phosphatase, buffers can enhance the activity by acting as transphosphorylating agents. McComb and Bowers studied the enzyme's activity at differing conditions and found an activity increase in buffers able to transphosphorylate (McComb and Bowers [1972]). Fortunately, a structure of a transition state of the enzyme is available in the Protein Data Bank (PDB code 3MK0). In this

Buffer compound	HYDE score in kJ/mol	Experimental activity increase (in percent)
triethanolamine	-39	51
D-mannitol	-39	12
TRIS	-27	55
diethanolamine	-26	62
2-amino-2-methyl-1,3-propandiol	-26	21
2-isopropylaminoethanol	-26	16
2-methylaminoethanol	-25	42
glycerol (substrate)	-24	

Table 8.2: HYDE scores of buffer compounds docked into a transition state structure and into a basic state structure of the enzyme alkaline phosphatase. The percental activity increase was reported by McComb and Bowers (McComb and Bowers [1972])

structure, a serine amino acid in the active site is phosphorylated. This phosphate group will in the following step be accepted by a buffer compound. Therefore, docking to this structure with buffers able to transphosphorylate should reveal buffers increasing the activity. On the other side, docking to the active site of a structure in the basic state of the enzyme should reveal inhibitors. Table 8.2 shows the scores for buffers with higher scores than the substrate glycerol docked to the transition state enzyme and the basic state. All buffer compounds that are scored higher than the substrate in the transition state enzyme structure can act as transphosphorylators and have a phosphate-accepting hydroxyl group near the covalently bound phosphate of the active site in the docking pose.
In Table 8.2 the percental activity increase reported by McComb and Bowers is shown. It does not correlate completely with the ranks of the buffers by the score. The rank of the score of D-mannitol does not correlate with the percental activity increase. However, D-mannitol is the only buffer scored higher than the substrate in the basic state enzyme structure. Therefore, it may also inhibit the enzyme, which could be the reason why the activity is lower than in the other buffers, which are not predicted to bind to the active site.

8.2.3 Prospective experiments

For five of the 13 enzymes of the SynBio project which were available in the laboratories of the collaborators, predictions were made for a library of 14 buffers. In enzyme activity assays then the buffers available in the laboratories were tested by always choosing a low-scored, a high-scored and a middle-scored buffer for the experiments, if possible.

In Figure 8.4 the results of the study are shown. In A to D, each enzyme is listed with its systematic name, the PDB code of the protein structure used and a structure diagram sketch of the catalyzed reaction. The bar diagrams show the relative enzyme activity measured in different buffer compounds. The relative activity is calculated by normalizing all activities by the highest measured activity. The color of the bars shows the category into which the buffers where categorized: Green is classified as *not inhibiting*, yellow as *potentially inhibiting* and red as *inhibiting*. The plots on the left show the relative score correlated with the relative activity. The relative score is the score of the buffer compound normalized with the score of the substrate (therefore, the relative score can be greater than 1).
The bar diagrams show that the experimental measurements confirm the trend of the predictions for propanediol-oxidoreductase isoenzyme, phosphoglucose isomerase and alcohol dehydrogenase A. For these enzymes, the activity trend follows the ranking of the buffers. For fructose 1,6-bisphosphate aldolase, two buffers are predicted to be inhibiting, but almost no activity decrease is observed in the experiments. Here, it is possible that an inhibitory effect occurs only at higher buffer concentrations. For glucose-6-phosphate dehydrogenase, all buffer compounds are predicted to be *inhibiting*, and except the MOPS buffer, the activity trend follows the ranking of the predictions. As no buffer of the buffer library is classified as *not inhibiting*, the validation of the predictions is difficult. Furthermore, this enzyme is known to be highly flexible and the protein structures available are all in the 'open' form of the enzyme, while a hinge movement on substrate binding closes the active site. Therefore, the score of the substrate might be underestimated since the correct structure of the bound substrate cannot be assessed with the score.

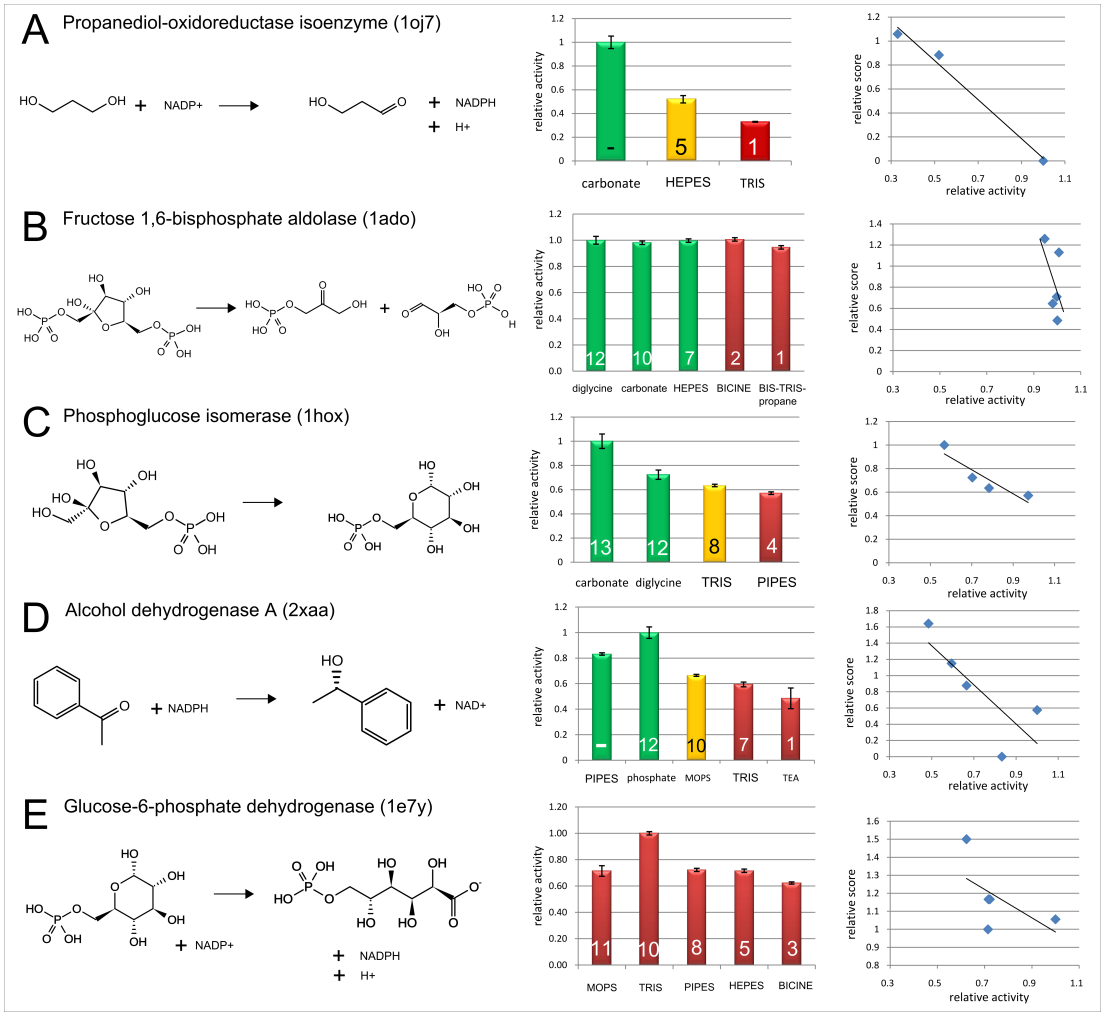


Figure 8.4: Results of the buffer activity screenings. (A) 1,3-Propanediol oxidoreductase isoenzyme (B) Fructose 1,6-bisphosphate aldolase (C) Phosphoglucose isomerase (D) Alcohol dehydrogenase A (E) Glucose-6-phosphate dehydrogenase. Enzymatic reaction diagrams of the reactions used to monitor activity and results are shown. The histograms in the middle show the relative activity of the enzyme normalized with the highest activity in different buffers. Red bars show *inhibiting* compounds, yellow bars *potentially inhibiting* compounds and green bars *not inhibiting* compounds. The bars show the relative activity of the enzyme in the different buffers (normalized by the highest activity in the experiment) and are sorted on the abscissa by HYDE score. The numbers written on the bars indicate the rank of the buffer in the buffer library of 14 buffers. The correlation diagrams on the right show a correlation between the relative activity and the relative docking score. These contain only the buffers that could be docked and therefore get a score.

8.3 Prediction of feedback inhibition

The method described in the previous section was also used to predict feedback or crossover inhibitions. This term is used for describing the inhibition of an enzyme in the pathway caused by a product or substrate of another enzyme in the pathway. The detection of feedback inhibitions is important for the pathway reaction setup. Enzymes whose substrates or products inhibit others should be placed in a separate reaction container to avoid a reduction of activity.

In the SynBio project, a sub-pathway of four of the thirteen enzymes was studies for feedback inhibitions. The collaborators studied the setup of the pathway from enzyme 9 to 12 (triose-phosphate isomerase, aldolase, fructose-bisphosphatase, phosphoglucose-isomerase) by evaluating if a one-pot approach is feasible for these four enzymes.

By predicting inhibitions of these four enzymes by substances present in the total pathway, the evaluation can be supported by protein-ligand docking.

8.3.1 Methods

The protein-ligand docking approach discussed in the previous section could be applied for this problem as well. A ligand library consisting of all substrates and products and cofactors of the four enzymes was compiled (see Zhang et al. [2007] for a list of the enzymes and the reaction description with substrates and products), following the protocol described above. For collecting the protein structures of these four enzymes, the same protocol as described in 8.2.1 was used. See Tables 8.3 to 8.6 for a list of the substances contained in the compound library and the PDB codes of the enzymes used for docking.

8.3.2 Results

The scores of all the ligands of the library docked into the four enzymes are shown in Tables 8.3 to 8.6.

For the enzyme triose-phosphate isomerase, only 6-phospho-D-gluconate is scored higher than the substrate. However, the score is much higher, which means that inhibition can already occur at low concentration. For the enzyme fructose-bisphosphatase, more compounds are scored better than the substrate. Among the potentially inhibiting compounds, the cofactor NAD+ is found as well. For aldolase, also many compounds are scored higher than the substrates. However, since both substrates get connected during enzyme reaction, calculating the score by docking them one at a time might underestimate the binding affinity. For phosphoglucose isomerase, three compounds are scored better than the substrate.

Compound	HYDE score in kJ/mol for docking into 1HTI
6-phospho-D-gluconate	-45
glyceraldehyde 3-phosphate (substrate)	-36
D-ribulose 5-phosphate	-35
D-xylulose 5-phosphate	-34
sedoheptulose 7-phosphate	-34
dihydroxyacetone phosphate (product)	-33
D-fructose 6-phosphate	-33
D-fructose 1,6-diphosphate	-31
erythrose 4-phosphate	-31
D-glucose 6-phosphate	-30
NAD+	-29
D-ribose 5-phosphate	-28
D-glucose 1-phosphate	-27
NADP+	-26
phosphate	-12

Table 8.3: HYDE scores of library compounds docked into triose-phosphate isomerase, sorted by best score.

Compound	HYDE score in kJ/mol for docking into 3DR1
NAD+	-52
D-fructose 6-phosphate (product)	-51
D-glucose 1-phosphate	-50
6-phospho-D-gluconate	-48
D-xylulose 5-phosphate	-45
D-fructose 1,6-diphosphate (substrate)	-44
sedoheptulose 7-phosphate	-42
D-ribulose 5-phosphate	-39
D-ribose 5-phosphate	-38
D-erythrose 4-phosphate	-38
D-glucose 6-phosphate	-38
NADP+	-35
glyceraldehyde 3-phosphate	-33
dihydroxyacetone phosphate	-27
phosphate	-20

Table 8.4: HYDE scores of library compounds docked into fructose-bisphosphatase, sorted by best score.

Compound	HYDE score in kJ/mol for docking into 1ADO
6-phospho-D-gluconate	-39
D-ribulose 5-phosphate	-36
sedoheptulose 7-phosphate	-35
D-xylulose 5-phosphate	-33
D-glucose 6-phosphate	-32
dihydroxyacetone phosphate (substrate)	-32
glyceraldehyde 3-phosphate (substrate)	-31
NAD+	-26
D-ribose 5-phosphate	-26
D-glucose 1-phosphate	-23
NADP+	-22
D-fructose 1,6-diphosphate (product)	-22
D-fructose 6-phosphate	-21
D-erythrose 4-phosphate	-19
phosphate	-11

Table 8.5: HYDE scores of library compounds docked into aldolase, sorted by best score.

Compound	HYDE score in kJ/mol for docking into 1HOX
sedoheptulose 7-phosphate	-39
D-fructose 1,6-diphosphate	-38
6-phospho-D-gluconate	-37
D-fructose 6-phosphate (substrate)	-37
D-glucose 1-phosphate	-33
D-erythrose 4-phosphate	-32
D-ribulose 5-phosphate	-32
D-glucose 6-phosphate (product)	-32
D-xylulose 5-phosphate	-31
glyceraldehyde 3-phosphate	-31
D-ribose 5-phosphate	-30
dihydroxyacetone phosphate	-28
phosphate	-15
NAD+	0
NADP+	0

Table 8.6: HYDE scores of library compounds docked into phosphoglucose isomerase, sorted by best score.

Protein-ligand docking is known to predict more false positives than false negatives. Therefore, the results can be used to test only the predicted inhibitors in the experiments and thus reduce the number of experiments in comparison to testing all compounds.

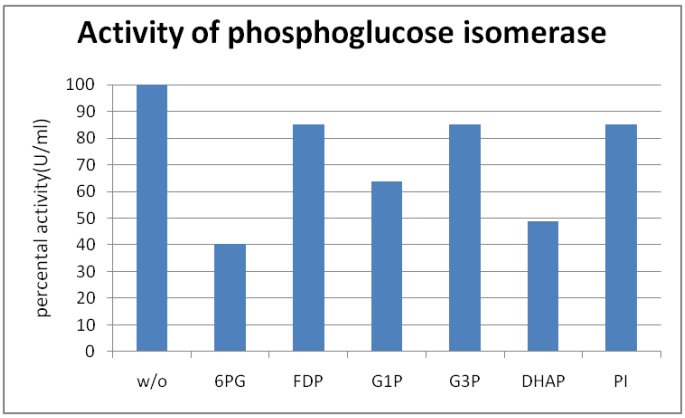


Figure 8.5: Experimental activity test of the enzyme phosphoglucose isomerase in presence of other substances from the multi-enzyme pathway. Abbreviations= w/o: activity without other substance, 6PG: 6-phosphogluconate, FDP: fructose-1,6-disphosphate, G1P: glucose-1-phosphate, G3P: glyceraldehyde-3-phosphate, DHAP: dihydroxyacetone phosphate, PI: phosphate

In the SynBio project, the inhibition prediction could only be tested experimentally for phosphoglucose isomerase by the collaborators. The result of an activity assay (assay for this enzyme as described by Sigma-Aldrich Co. LLC.) where the inhibitor was added in the same concentration as the substrate are shown in Figure 8.5. For 6-phospho-D-gluconate, which is scored as high as the substrate, the experiment indeed shows an inhibition. For D-fructose 1,6-disphosphate, this inhibition is not as high, although it gets a similar high score. Glucose-1-phosphate also scored lower than the substrate, but still gets a good score. This compound also leads to an inhibition. Glyceraldehyde 3-phosphate and dihydroxyacetone phosphate get comparable scores, but dihydroxyacetone phosphate is the second highest inhibition, which is not represented by the scores. The inhibition by phosphate is not predicted and not high with an activity of about 85%.

8.4 Identification of enzyme structure with highest activity

The third task that was studied in the biotechnological case study was how computational methods may support the choice of an enzyme with highest activity. Often several enzymes

are known to catalyze a reaction with varying efficiency. The same enzymes originating from different organisms show high activity differences and often have different reaction optima. Also, they may be more or less selective, depending on the evolutionary path the enzyme followed. Therefore, it may be crucial for yield optimization in a multi-enzyme pathway to choose the best available enzyme. Further, it may be helpful to mutate amino acids of an enzyme for, e.g., increasing selectivity.

The structures of enzymes contained in the Protein Data Bank can be supportive for the choice of an enzyme. Often enzyme structures of different organisms and also mutated structures are available. With molecular modeling, mutated structures can be created in-silico. For a single enzyme, a large number of structures may be available for studying. Therefore for a complete multi-enzyme pathway, it is infeasible to manually study all structures and an automatic, computational process is needed.

Docking one substrate into many different enzyme structures is the reverse problem to the normal protein-ligand docking which could be used in the previous two tasks. Therefore, for assessing this task the inverse screening software *i*RAISE was used which was developed in this thesis. See Chapter 4 for the *i*RAISE method description.

As an exemplary use case, we studied structures of the enzyme phosphoglucose isomerase (enzyme number 12 of the multi-enzyme pathway studied in the SynBio-project) for their selectivity for the substrate D-fructose 6-phosphate over the known inhibitor 6-phospho-D-gluconate.

For the screening experiment, all enzyme structures from the Protein Data Bank with the EC number 5.3.1.9 which were co-crystallized with a ligand were collected. 36 structures from 14 different organisms were available. An analysis of the active site properties showed that for this enzyme, the active site composition is in fact varying: The number of amino acids varies between 19 and 34 and the number of hydrogen bond donors and acceptors varies as well although the co-crystallized ligands are highly similar. In the 36 structures, only 14 different ligands are found.

All 36 protein structures were screened with *i*RAISE with the substrate D-fructose 6-phosphate and the known inhibitor 6-phospho-D-gluconate. In Table 8.7, the resulting scores are listed. The results can now be interpreted by looking for structures which have a high score for the substrate and a low score for the inhibitor.

The substrate and inhibitor can be docked to all structures, therefore both fit in all active sites. The score of the substrate is a little bit higher for most organisms than that of the inhibitor. The scores of the substrate in the structures of Trypanosoma brucei, Thermococcus litoralis and Staphylococcus aureus are only slightly higher, therefore enzymes of these organisms should not be chosen. Many structures of the enzyme from Mus musculus are

available. In most of them, the substrate is scored better than the inhibitor. Since the structures are conformations of the same enzyme, the next step would be the experimental evaluation, whether certain reaction conditions can stabilize an enzyme conformation which prefers binding of the substrate.

8.5 Summary

The biotechnological case study revealed several aspects in the design of synthetic multi-enzyme pathways which can be supported by structure-based computational methods. All of them concerned the activity increase of enzymes. With protein-ligand docking, assessing the negative effects of buffer compounds supports the compilation of optimal reaction solutions and assessing feedback inhibitions help in the reaction container setup. With inverse screening, proteins with best substrate specificity can be predicted from all available structures. The weakness of these methods is that they are dependent on the available enzyme structures. As more and more enzyme structures become available, these methods are even more promising in future, though. In the SynBio project, the predictions of the first and part of the second tasks were evaluated experimentally. The third task was so far not evaluated but shows a potential application for inverse screening in the biotechnological design of multi-enzyme pathways.

Organism	PDB code	Score for fructose-6-phosphate	Score for 6-phosphogluconate
FRANCISELLA TULARENSIS	3M5P	-93	-82
	3Q7I	-101	-88
	3Q88	-83	-73
HOMO SAPIENS	1IRI	-98	-84
	1NUH	-111	-97
LEISHMANIA MEXICANA	1T10	-76	-62
PLASMODIUM FALCIPARUM	3PR3	-92	-80
PYROBACULUM AEROPHILUM	1TZC	-107	-95
	1X9H	-118	-102
	1X9I	-92	-93
TRYPANOSOMA BRUCEI	2O2C	-96	-91
THERMOCOCCUS LITORALIS	1J3R	-79	-74
MUS MUSCULUS	1U0F	-123	-87
	1U0G	-107	-70
	2CXO	-95	-70
	2CXP	-97	-89
	2CXQ	-114	-84
	2CXR	-119	-118
	2CXS	-98	-99
	2CXT	-102	-105
STAPHYLOCOCCUS AUREUS	3FF1	-55	-52
SUS SCROFA	1GZV	-123	-111
TOXOPLASMA GONDII	3UJH	-91	-77
ORYCTOLAGUS CUNICULUS	1DQR	-90	-106
	1G98	-84	-100
	1HOX	-94	-82
	1KOJ	-124	-106
	1XTB	-97	-93
PYROCOCCUS FURIOSUS	1QXR	-79	-86
	1QY4	-87	-87
	1X7N	-68	-58
	1X82	-62	-56
	2GC0	-81	-71
	2GC2	-72	-73

Table 8.7: Results of screening various structures of phosphoglucose isomerase from different organisms with *i*RAISE for the substrate fructose-6-phosphate and the inhibitor 6-phosphogluconate.

9
Conclusion

9.1 Overview

In this thesis, a new approach for inverse virtual screening has been realized in the software *i*RAISE. The method development was focused in large parts on handling large amounts of protein structures efficiently and addressing the problem of inter-protein scoring. Next to the method development, a main part of this thesis was the evaluation of *i*RAISE which included the development of new data sets and an evaluation strategy. Along with the main command line tool *i*RAISE, a viewer of the solutions in form of a GUI was created. In the following, the achievements and limitations of the developed methods are discussed. Further, areas of possible improvements are highlighted.

9.2 Achievements

The achievements are discussed on the basis of the in Chapter 3 defined aims and objectives. In the field of structure-based target prediction the method is outstanding compared to current state of the art methods concerning measures applied to handle large numbers of protein targets, to overcome the linear one ligand-one protein matching by using descriptor representations and to improve inter-target ranking.

Automatic processing of protein structures

*i*RAISE handles protein structures fully automated. On the basis of PDB protein files the method can automatically detect active sites based on co-crystallized ligands. Thus, as input solely protein-ligand complexes in form of the prevalent PDB file format are needed. This overcomes the need of manual protein preparation steps as they are common in various protein-ligand docking tools. Nevertheless, on small amounts of proteins, using the knowledge of a user as assistance in active site determination is reasonable, which is why in *i*RAISE it is possible to provide reference ligands by the user. For large amounts of protein structures, however, the automatic mode evades the need for manual annotation.

Consistent and efficient storage of protein-ligand complexes

The consistent and efficient storage of protein-ligand complexes as well as of the annotation of active sites is crucial in inverse screening approaches. Active site definition has to be stored consistently to allow multiple screening under the same conditions. The efficient handling of protein-ligand complexes requires a memory-saving representation which can be accessed rapidly. With the ComplexDB and the ProteinDB, a database representation has been developed in *i*RAISE which stores the protein-ligand complexes with all information necessary

for screening as well as the active site definition. Thus, time-consuming initializaton steps from the PDB file have to be done only once in the preprocessing step, which stores the information readily accessible in the database. The choice of an SQLite database combines the advantages of using database technology with easy portability and no need of setting up a server.

Abstraction of active sites

A descriptor representation of the active sites as well as of the query ligand breaks the linear screening sequence, since no one to one matching of ligand and protein active site on atomic basis is required. Testing complementarity on descriptor level is much faster and only those structures which match on descriptor basis then need to be processed on atomic level. The triangle descriptor used here was already successfully applied in classic virtual screening. Some adjustments were necessary for the inverse setup like storing further information which is needed to avoid re-calculation of the descriptors during screening. Using a bitmap-index for storing the descriptors allows fast querying and due to binning of the descriptor values efficient memory usage.
Since the descriptor contains information about interactions, their geometric arrangement, their directions and a shape representation, the protein-ligand binding is abstracted on a high level. Thereby, a reasonable estimation of a binding is possible and allows immediate discarding of proteins as targets for a ligand if no match occurred.

Inter-target ranking

In general, inter-target ranking of proteins poses a problem to 'normal' protein-ligand scoring functions used in docking contexts. Often certain proteins are biased with higher scores than others, an effect called inter-protein scoring noise (Wang et al. [2012]). These scoring functions were not developed for inter-target but for inter-molecule ranking, i.e., for ranking different molecules for one protein. Thus, measures have to be taken to avoid this bias. Further, in inverse screening, false positives have a worse effect than in normal docking, since nowadays it is still more expensive to test a long list of proteins for one ligand than to test many ligands for one protein in activity assays. For each predicted target more, a separate activity assay has to be conducted which is time- and cost-consuming.
Facing these challenges, a five-step Scoring Cascade has been developed which step-wise applies more detailed scoring measures. The steps contain coarse and detailed clash tests and exploit information available by evaluation of the co-crystallized ligand. By considering the ligand coverage as well as the pocket coverage on the basis of a co-crystallized ligand

the scoring is capable of dealing with pockets with diverse shapes. Large or small and buried or open pockets can be scored without preferring any shape. Next to the Scoring Cascade, as a further measure, an average score for each pocket was calculated and successfully used to assess which scores are statistically significant for a pocket and which are not.

Performance evaluation

The field of inverse virtual screening method-development is still in its fledgling stage, compared to the about one decade older normal virtual screening. Far fewer evaluation data sets and methods are yet established for inverse virtual screening. Since no standard evaluation strategy existed, a new strategy has been developed based on data sets used commonly in evaluation of 'normal' docking. Further, two new data sets have been created for evaluation purposes, a small one for studying selectivity/sensitivity in detail and a large one which allows statistical evaluation.
The performance of *i*RAISE was compared to state of the art docking, to sequence-based and pharmacophore-based target prediction methods. Concerning the target ranking, *i*RAISE has shown good performance, superior to classic docking. Its performance was comparable to the pharmacophore-based method and if taking into account the same amount of input information even superior. Concerning the comparison to sequence-based target prediction, it was shown that *i*RAISE is able to perform better on those cases which are not trivial, i.e., where several diverse targets existed for one ligand.
With on average a screening time of 7.1 seconds per protein and available parallelization on data level, large amounts of targets can be screened rapidly.

Usability by medicinal chemists

The *i*RAISE inverse screening tool has been thoroughly tested on pharmacological data with drugs as compounds and drug targets as proteins. Successful predictions on this data could be observed, with only few examples where *i*RAISE failed in predicting correct targets completely.
Next to the performance in target prediction, however, the usability by medicinal chemists which are no computer scientists is important for a software to be of use in real world applications. Since the *i*RAISE software is a bare command line tool, also a graphic interface for assessing the screening solutions in form of the ComplexViewer has been developed. This viewer allows the browsing of proteins contained in a screening project and a 3D visualization of the pockets. Further, it allows to browse through screening solutions and provides a 3D graphic representation of the predicted binding poses. For medicinal chemists it may be

crucial to investigate not only the ranked list and scores of predicted targets, but also the predicted binding mode which they are able to revalue with their expert knowledge.

Usability of method in other fields

A thorough evaluation of the potentials and limitations of structure-based computational methods on the basis of classic docking and inverse virtual screening has been conducted in a synthetic multi-enzyme pathway development project. It has been found that docking methods can aid in predicting competitive inhibition of enzymes by buffer solution agents and thus support the composition of buffer solutions supporting enzyme activity and maximizing yield. An application of inverse virtual screening is the prediction of feedback inhibition and selecting the best source organism of an enzyme.

9.3 Limitations

Although several of the open challenges of inverse virtual screening could be faced in the work of this thesis, still areas of improvements and limitations of its usability exist. First of all, the limitations of the evaluation strategy are highlighted followed by the areas of improvements for the inverse screening approach *i*RAISE.

In the development of the Drugs/sc-PDB data set focus has been laid on automatic true-positive annotation of structures based on the target information of drugs contained in the DrugBank. By the automatic true positive annotation based on EC number, protein name and UniProtID, however, errors are unintentionally introduced.
Reasons are misleading names of proteins in the PDB-file header or multi-enzyme constructs which are then labeled with several EC numbers of which it is not automatically possible to decide which is the correct one for the calculated active site. Further, source organism or mutations are not taken into account. Nevertheless, same proteins from different organisms or mutated proteins not necessarily bind the same compounds and thus the assumption that they are targets for a compound might be false. However, the automated annotation is human bias-free and allows to easily create newer versions of the data set on the basis of newer versions of the DrugBank and the sc-PDB.
Another limitation is that definitely not all true targets for the drugs are yet known/annotated in the DrugBank. Therefore, the sensitivity/specificity assessment of a screening tool on this data is not totally correct.

Concerning the *i*RAISE screening approach, its main limitation is its dependence on the

input data which influences several layers of its performance:
Firstly, the performance evaluation has shown that *i*RAISE is highly dependent on the input ligand conformations. Often the bioactive conformation has internal strains and is energetically not favored and thus occurs only by interacting with the protein. Such energetically unfavored conformations are not taken into account by conformational sampling tools – at least if more favored conformations can be build. Therefore, a post-optimization step or a different ligand conformation generation strategy (like the fragment based conformation construction used in FlexX (Rarey et al. [1996]) would be needed to overcome this problem. However, if the final ligand conformation is not known from the start of the screening, then the matching of descriptors would need to be much more tolerant or the descriptor would have to be calculated based on ligand fragments. In both cases, the method would have to deal with many more protein matchings from the descriptor index and possibly more false positives.
Secondly, the method does not handle protein flexibility internally - except for rotatable terminal groups. Thus it is dependent on the input of protein conformations. The performance evaluation has shown that due to the strict scoring in *i*RAISE, not all true targets are hit even if only small conformational changes of the protein would be necessary. One possibility would be to sample the protein conformations beforehand, which however is a very difficult task and nowadays is yet not feasible for large numbers of proteins. While amino acid conformations might be sampled, taking into account backbone flexibility such as hinge movements is still not predictable. Another possibility would be to post-optimize the protein together with generated ligand poses to remove clashes. In this case, also the descriptor matching and clash tests would need to be less strict and more protein-ligand combinations would need to be evaluated. Post-optimization based, e.g., on force fields is computing time expensive. Thus, either the time for screening would increase immensely or a pre-scoring able to reduce the large number of matches to a sensible amount for post-optimization would be needed.
Thirdly, the Scoring Cascade is in large parts dependent on a co-crystallized ligand of the screened protein. It is assumed that a co-crystallized ligand binds to a protein. However, in the Protein Data Bank also protein-ligand complexes exist where the binding is artificial, if, e.g., the experimental method *soaking* has been used, it may occur that a ligand is only *placed* in a protein pocket but would not bind in solution. Using such ligands as references falsifies the scoring result.
In summary, it would be desirable for *i*RAISE to be less dependent on the input data and that the flexibility (protein and ligand) would be handled better internally.
A further area of improvement is the handling of apo protein structures, i.e., structures

not co-crystallized with a ligand. These structures are also of high interest since for many proteins, no crystallized complexes exist. Using a pocket detection algorithm like DogSite (Volkamer et al. [2010]) for active site definition would be easily integrable, but the loss of information derived from the co-crystallized ligand for scoring would need to be compensated.
Concerning the design of the method, one limitation are its many parameters. Many hard cutoff values are used to make decisions, like do descriptors match, when is a pose clashing, when is a pose discarded due to insufficient coverage and others. Here more dynamic decisions would be desirable, reacting to the current screening situation instead of hard-coded cutoffs.
So far, the parallelization takes place solely on the data level. Thus, the running time could be improved by threaded parallel computing exploiting the nowadays standardly available multi-core architectures.

In the application study of structure-based computational prediction methods in a biotechnological context, it became clear that the computational methods yet are limited concerning the integration of external experiment parameters like temperature and pH value. The models of the computational methods are build on physiological data. However, in biotechnological experiments changes of pH and temperature are important issues. However, the underlying theories of the computational models do not hold true at pH value-changes or extreme temperatures.
Lastly, the applicability of the approach could be improved, since so far only a command line interface exists and the user has to use scripts to regulate parallel screening. The ComplexViewer-GUI provides only basic utility of assessing screening solutions and could be extended significantly.

9.4 Outlook

The limitations discussed in the preceding section lead to ideas on next steps of improving the method:

- Integration of sampling of amino acid side chains. This step would be a first assessment of protein flexibility and could be easily integrated. The developed ProteinDB is setup in a way that several conformations for one residue could be stored without duplicating the complete protein data for each sampled amino acid.

- Task parallelization next to data parallelization for further running time improvement.

- Joining the screening with a pose post-optimization strategy. Currently, only hydrogen atoms are relocated for each pose. Optimization of protein and ligand poses before final scoring would certainly improve the results especially in those cases for which the true targets could not be identified. This step would be mandatory if amino acid conformations are introduced, since the active site would need to be optimized with the respective conformations of the amino acids.
- Combination of *i*RAISE screening with other target-prediction approaches like ligand-similarity in a so called hybrid target prediction tool. Recently a method by Meslamani et al. (Meslamani et al. [2013]) was found to be quite successful by combining four ligand-based and two structure-based (docking and pharmacophore-matching) approaches. The predictions of two off-targets were confirmed experimentally. Hybrid approaches have the advantage that the limitations of the methods are often straightened out by other approaches since they are not necessarily overlapping.
- Inclusion of the possibility to use apo protein structures by integration of a pocket detection algorithm, and re-structuring of the Scoring Cascade to render the use of a reference ligand optional.
- Integration of user-defined filters like properties of active sites or user generated pharmacophores. Such constraints are of help if the user looks for a certain class of proteins, e.g., if an enzyme shall be found to which a compound functions as a substrate. In those cases, not only the binding of the ligand to the protein is important, but also the location of certain ligand moieties in vicinity to catalytic amino acids. Such filtering steps currently have to be done manually on the generated results.
- Experimental validation. For many predictions in the evaluation studies it was not possible to conclude if they were correct or false positives. Therefore studying *i*RAISE's predictions by experiments would be a next step to give further insights.

Finally, there are many interesting application scenarios of structure-based target identification, in which it would be interesting to apply *i*RAISE screening:

- Identification of targets for metabolites with unknown function.
- Identification of enzymes catalyzing given substrates.
- Identification of off-targets for so-called orphaned drugs. These drugs have already passed several tests but did not exhibit high enough efficacy against the targeted proteins. Thus, already much money has been send on these drugs without any positive venue.

- Identification of side-effect causing targets for withdrawn drugs.
- Target identification for natural compounds used in medicinal applications like traditional chinese medicine.

References

M.D.M. AbdulHameed, S. Chaudhury, N. Singh, H. Sun, A. Wallqvist, and G.J. Tawa. Exploring Polypharmacology Using a ROCS-Based Target Fishing Approach. *Journal of chemical information and modeling*, 52(2):492–505, 2012. 18

Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997. 91

C Andre, Y Jacquot, TT Truong, M Thomassin, JF Robert, and YC Guillaume. Analysis of the progesterone displacement of its human serum albumin binding site by β-estradiol using biochromatographic approaches: effect of two salt modifiers. *Journal of Chromatography B*, 796(2): 267–281, 2003. 121

Christos Andronis, Anuj Sharma, Vassilis Virvilis, Spyros Deftereos, and Aris Persidis. Literature mining, ontologies and information visualization for drug repurposing. *Briefings in bioinformatics*, 12(4):357–368, 2011. 6, 7

Masanori Arita. Metabolic reconstruction using shortest paths. *Simulation Practice and Theory*, 8 (1):109–125, 2000. 131

Ted T Ashburn and Karl B Thor. Drug repositioning: identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery*, 3(8):673–683, 2004. 6

Jonathan B Baell and Georgina A Holloway. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *Journal of medicinal chemistry*, 53(7):2719–2740, 2010. 16

A. Bender and R.C. Glen. Molecular similarity: a key technique in molecular informatics. *Organic & biomolecular chemistry*, 2(22):3204–3218, 2004. 17

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, TN Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000. 21, 79

Stefan Bietz, Sascha Urbaczek, Benjamin Schulz, and Matthias Rarey. Protoss: a holistic approach to predict tautomers and protonation states in protein-ligand complexes. *Journal of cheminformatics*, 6(1):12, 2014. 56, 91

Caterina Bissantz, Bernd Kuhn, and Martin Stahl. A medicinal chemist's guide to molecular interactions. *Journal of medicinal chemistry*, 53(14):5061–5084, 2010. 3

Markus Böhm, Jörg Stürzebecher, and Gerhard Klebe. Three-dimensional quantitative structure-activity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin, and factor Xa. *Journal of medicinal chemistry*, 42(3):458–477, 1999. 75, 76

Evan E Bolton, Yanli Wang, Paul A Thiessen, and Stephen H Bryant. PubChem: integrated platform of small molecules and biological activities. *Annual reports in computational chemistry*, 4:217–241, 2008. 77, 132

Jonas Boström, Anders Hogner, and Stefan Schmitt. Do structurally similar ligands bind in a similar fashion? *Journal of medicinal chemistry*, 49(23):6716–6725, 2006. 48

Giovanni Bottegoni, Angelo D Favia, Maurizio Recanatini, and Andrea Cavalli. The role of fragment-based and computational methods in polypharmacology. *Drug discovery today*, 17(1):23–34, 2012. 7

Jianhua Cai, Cong Han, Tiancen Hu, Jian Zhang, Dalei Wu, Fangdao Wang, Yunqing Liu, Jianping Ding, Kaixian Chen, Jianmin Yue, et al. Peptide deformylase is a potential target for anti-Helicobacter pylori drugs: reverse docking, enzymatic assay, and X-ray crystallography validation. *Protein science*, 15(9):2071–2081, 2006. 24

Valerie Campagna-Slater, Andrew G Arrowsmith, Yong Zhao, and Matthieu Schapira. Pharmacophore screening of the protein data bank for specific binding site chemistry. *Journal of chemical information and modeling*, 50(3):358–367, 2010. 26, 27

Monica Campillos, Michael Kuhn, Anne-Claude Gavin, Lars Juhl Jensen, and Peer Bork. Drug target identification using side-effect similarity. *Science*, 321(5886):263–266, 2008. 20

Adrià Cereto-Massagué, María José Ojeda, Robbie P Joosten, Cristina Valls, Miquel Mulero, M Josepa Salvado, Anna Arola-Arnal, Lluís Arola, Santiago Garcia-Vallvé, and Gerard Pujadas. The good, the bad and the dubious: VHELIBS, a validation helper for ligands and binding sites. *Journal of cheminformatics*, 5(1):1–9, 2013.

Calvin Yu-Chian Chen. TCM Database@ Taiwan: the worldś largest traditional Chinese medicine database for drug screening in silico. *PloS one*, 6(1):e15939, 2011. 6

Xin Chen, ZL Ji, and Yuzong Z Chen. TTD: Therapeutic Target Database. *Nucleic acids research*, 30(1):412–415, 2002. 23, 24

YZ Chen and CY Ung. Prediction of potential toxicity and side effect protein targets of a small molecule by a ligand-protein inverse docking approach. *Journal of molecular graphics & modelling*, 20(3):199, 2001. 22, 23, 31

YZ Chen and DG Zhi. Ligand–protein inverse docking and its potential use in the computer search of protein targets of a small molecule. *Proteins: Structure, Function, and Bioinformatics*, 43(2): 217–226, 2001. 14, 22, 23, 31

Ayoun Cho, Hongseok Yun, Jin H Park, Sang Y Lee, and Sunwon Park. Prediction of novel synthetic pathways for the production of desired chemicals. *BMC Systems Biology*, 4(1):35, 2010. 8, 131

Paul Czodrowski, Christoph A Sotriffer, and Gerhard Klebe. Protonation Changes upon Ligand Binding to Trypsin and Thrombin: Structural Interpretation Based on pKa Calculations and ITC Experiments. *Journal of molecular biology*, 367(5):1347–1356, 2007. 75

Andrew M Davis, Simon J Teague, and Gerard J Kleywegt. Application and Limitations of X-ray Crystallographic Data in Structure-Based Ligand and Drug Design. *Angewandte Chemie International Edition*, 42(24):2718–2736, 2003. 70

Armida Di Fenza, Andreas Heine, Ulrich Koert, and Gerhard Klebe. Understanding binding selectivity toward trypsin and factor Xa: the role of aromatic interactions. *ChemMedChem*, 2(3):297–308, 2007. 75

Joel T Dudley, Tarangini Deshpande, and Atul J Butte. Exploiting drug–disease relationships for computational drug repositioning. *Briefings in bioinformatics*, 12(4):303–311, 2011. 6, 7

S Ekins, J Mestres, and B Testa. In silico pharmacology for drug discovery: applications to targets and beyond. *British journal of pharmacology*, 152(1):21–37, 2007.

Sean Ekins, Antony J Williams, Matthew D Krasowski, and Joel S Freundlich. In silico repositioning of approved drugs for rare and neglected diseases. *Drug discovery today*, 16(7):298–310, 2011. 6

Todd JA Ewing, Shingo Makino, A Geoffrey Skillman, and Irwin D Kuntz. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *Journal of computer-aided molecular design*, 15(5):411–428, 2001. 24

FDA. US Food and Drug Administration, June 2013. URL `http://www.fda.gov/`. 7

M DOLORES Fernandez, GRAHAM D Carter, and TN Palmer. The interaction of canrenone with oestrogen and progesterone receptors in human uterine cytosol. *British journal of clinical pharmacology*, 15(1):95–101, 1983. 121

Emil Fischer. Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der deutschen chemischen Gesellschaft*, 27(3):2985–2993, 1894. 2

Richard A Friesner, Jay L Banks, Robert B Murphy, Thomas A Halgren, Jasna J Klicic, Daniel T Mainz, Matthew P Repasky, Eric H Knoll, Mee Shelley, Jason K Perry, et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of medicinal chemistry*, 47(7):1739–1749, 2004. 99

Zhenting Gao, Honglin Li, Hailei Zhang, Xiaofeng Liu, Ling Kang, Xiaomin Luo, Weiliang Zhu, Kaixian Chen, Xicheng Wang, and Hualiang Jiang. PDTD: a web-accessible protein database for drug target identification. *BMC bioinformatics*, 9(1):104, 2008. 24, 31

Pietro Gatti-Lafranconi and Florian Hollfelder. Flexibility and Reactivity in Promiscuous Enzymes. *ChemBioChem*, 14(3):285–292, 2013. 4

Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107, 2012. 17, 121

W. Gerhardt. *Creatine kinase*, volume 3 of *Methods of Enzymatic Analysis*. Verlag Chemie, Weinheim/Deerfield Beach, 1983. 134

Serghei Glinca and Gerhard Klebe. Cavities tell more than sequences: Exploring functional relationships of proteases via binding pockets. *Journal of chemical information and modeling*, 53(8): 2082–2092, 2013. 75

Elisabet Gregori-Puigjane and Jordi Mestres. A ligand-based approach to mining the chemogenomic space of drugs. *Combinatorial chemistry & high throughput screening*, 11(8):669–676, 2008. 19

Stefan Günther, Michael Kuhn, Mathias Dunkel, Monica Campillos, Christian Senger, Evangelia Petsalaki, Jessica Ahmed, Eduardo Garcia Urdiales, Andreas Gewiess, Lars Juhl Jensen, et al. SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic acids research*, 36(suppl 1):D919–D922, 2008. 111

Michael J Hartshorn, Marcel L Verdonk, Gianni Chessari, Suzanne C Brewerton, Wijnand TM Mooij, Paul N Mortenson, and Christopher W Murray. Diverse, high-quality test set for the validation of protein-ligand docking performance. *Journal of medicinal chemistry*, 50(4):726–741, 2007. 69, 73, 74

V Joachim Haupt and Michael Schroeder. Old friends in new guise: repositioning of known drugs with structural bioinformatics. *Briefings in bioinformatics*, 12(4):312–326, 2011. 6, 7

Johannes C Hermann, Eman Ghanem, Yingchun Li, Frank M Raushel, John J Irwin, and Brian K Shoichet. Predicting substrates by docking high-energy intermediates to enzyme structures. *Journal of the American Chemical Society*, 128(49):15882–15891, 2006. 131

Johannes C Hermann, Ricardo Marti-Arbona, Alexander A Fedorov, Elena Fedorov, Steven C Almo, Brian K Shoichet, and Frank M Raushel. Structure-based activity prediction for an enzyme of unknown function. *Nature*, 448(7155):775–779, 2007. 131

Matthias Hilbig, Sascha Urbaczek, Inken Groth, Stefan Heuser, and Matthias Rarey. MONA–Interactive manipulation of molecule collections. *Journal of cheminformatics*, 5:38, 2013. 80

Andrew L Hopkins. Network pharmacology: the next paradigm in drug discovery. *Nature chemical biology*, 4(11):682–690, 2008. 18

Niu Huang, Brian K Shoichet, and John J Irwin. Benchmarking sets for molecular docking. *Journal of medicinal chemistry*, 49(23):6789–6801, 2006. 30

Ruili Huang, Noel Southall, Yuhong Wang, Adam Yasgar, Paul Shinn, Ajit Jadhav, Dac-Trung Nguyen, and Christopher P Austin. The NCGC pharmaceutical collection: a comprehensive resource of clinically approved drugs enabling repurposing and chemical genomics. *Science translational medicine*, 3(80):80ps16, 2011a. 7

Ruili Huang, Menghang Xia, Ming-Hsuang Cho, Srilatha Sakamuru, Paul Shinn, Keith A Houck, David J Dix, Richard S Judson, Kristine L Witt, Robert J Kavlock, et al. Chemical genomics profiling of environmental chemical modulation of human nuclear receptors. *Environmental health perspectives*, 119(8):1142, 2011b. 121

Ajay N Jain. Surflex-Dock 2.1: robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *Journal of computer-aided molecular design*, 21(5):281–306, 2007. 91

Ajay N Jain and Anthony Nicholls. Recommendations for evaluation of computational methods. *Journal of computer-aided molecular design*, 22(3-4):133–139, 2008. 87

C. A. James and D. Weininger. *Daylight Theory Manual*. Daylight Chemical Information Systems, Inc: 27401 Los Altos, 2006. 17, 39

Jeremy L Jenkins, Andreas Bender, and John W Davies. In silico target fishing: Predicting biological targets from chemical structure. *Drug Discovery Today: Technologies*, 3(4):413–421, 2007. 15

Ekachai Jenwitheesuk, Jeremy A Horst, Kasey L Rivas, Wesley C Van Voorhis, and Ram Samudrala. Novel paradigms for drug discovery: computational multitarget screening. *Trends in pharmacological sciences*, 29(2):62–71, 2008. 7

Zhi Liang Ji, Yi Wang, Lin Yu, Lian Yi Han, Chan Juan Zheng, and Yu Zong Chen. In silico search of putative adverse drug reaction related proteins as a potential tool for facilitating drug adverse effect prediction. *Toxicology letters*, 164(2):104–112, 2006. 23

Mazen W Karaman, Sanna Herrgard, Daniel K Treiber, Paul Gallant, Corey E Atteridge, Brian T Campbell, Katrina W Chan, Pietro Ciceri, Mindy I Davis, Philip T Edeen, et al. A quantitative analysis of kinase inhibitor selectivity. *Nature biotechnology*, 26(1):127–132, 2008. 28

M.J. Keiser, B.L. Roth, B.N. Armbruster, P. Ernsberger, J.J. Irwin, and B.K. Shoichet. Relating protein pharmacology by ligand chemistry. *Nature biotechnology*, 25(2):197–206, 2007. 17

M.J. Keiser, V. Setola, J.J. Irwin, C. Laggner, A.I. Abbas, S.J. Hufeisen, N.H. Jensen, M.B. Kuijer, R.C. Matos, T.B. Tran, et al. Predicting new molecular targets for known drugs. *Nature*, 462 (7270):175–181, 2009. 17

Esther Kellenberger, Pascal Muller, Claire Schalon, Guillaume Bret, Nicolas Foata, and Didier Rognan. sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank. *Journal of chemical information and modeling*, 46(2):717–727, 2006. 25

Esther Kellenberger, Nicolas Foata, and Didier Rognan. Ranking targets in structure-based virtual screening of three-dimensional protein libraries: methods and problems. *Journal of chemical information and modeling*, 48(5):1014–1025, 2008. 22, 25, 26, 31, 55

Tony Kennedy. Managing the drug discovery/development interface. *Drug discovery today*, 2(10): 436–444, 1997. 5

Ish Khanna. Drug discovery in pharmaceutical industry: productivity challenges and trends. *Drug Discovery Today*, 2012. 5

Prashant S Kharkar, Sona Warrier, and Ram S Gaud. Reverse docking: a powerful tool for drug repositioning and drug rescue. *Future medicinal chemistry*, 6(3):333–342, 2014. 30

S.L. Kinnings and R.M. Jackson. ReverseScreen3D: a structure-based ligand matching method to identify protein targets. *Journal of chemical information and modeling*, 51(3):624–634, 2011. 18

Johannes Kirchmair, Patrick Markt, Simona Distinto, Gerhard Wolber, and Thierry Langer. Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection—What can we learn from earlier mistakes? *Journal of computer-aided molecular design*, 22(3-4):213–228, 2008. 87

Douglas B Kitchen, Hélène Decornez, John R Furr, and Jürgen Bajorath. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature reviews Drug discovery*, 3(11):935–949, 2004. 13

Oliver Korb, Thomas Stutzle, and Thomas E Exner. Empirical scoring functions for advanced protein-ligand docking with PLANTS. *Journal of chemical information and modeling*, 49(1):84–96, 2009. 91

DE Koshland Jr. Application of a theory of enzyme specificity to protein synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 44(2):98, 1958. 4

Alexios Koutsoukas, Robert Lowe, Yasaman Kalantar, Motamedi, Hamse Y Mussa, Werner Klaffke, John BO Mitchell, Robert C Glen, and Andreas Bender. In Silico Target Predictions: Defining a Benchmarking Data Set and Comparison of Performance of the Multiclass Naïve Bayes and Parzen-Rosenblatt Window. *Journal of chemical information and modeling*, 53(8):1957–1966, 2013. 17, 30

Irwin D. Kuntz, Jeffrey M. Blaney, Stuart J. Oatley, Robert Langridge, and Thomas E. Ferrin. A geometric approach to macromolecule-ligand interactions. *Journal of Molecular Biology*, 161(2): 269–288, 1982. 22

Chunhui Li, Christopher S Henry, Matthew D Jankowski, Justin A Ionita, Vassily Hatzimanikatis, and Linda J Broadbelt. Computational discovery of biochemical routes to specialty chemicals. *Chemical engineering science*, 59(22):5051–5060, 2004. 8, 131

Honglin Li, Zhenting Gao, Ling Kang, Hailei Zhang, Kun Yang, Kunqian Yu, Xiaomin Luo, Weiliang Zhu, Kaixian Chen, Jianhua Shen, et al. TarFisDock: a web server for identifying drug targets with docking approach. *Nucleic acids research*, 34:W219–W224, 2006. 24, 31

Tobias Lippert and Matthias Rarey. Fast automated placement of polar hydrogen atoms in protein-ligand complexes. *Journal of cheminformatics*, 1(1):1–12, 2009. 56, 133

Xiaofeng Liu, Sisheng Ouyang, Biao Yu, Yabo Liu, Kai Huang, Jiayu Gong, Siyuan Zheng, Zhihua Li, Honglin Li, and Hualiang Jiang. PharmMapper server: a web server for potential drug target identification using pharmacophore mapping approach. *Nucleic acids research*, 38(suppl 2):W609–W614, 2010. 27

Zhichao Liu, Hong Fang, Kelly Reagan, Xiaowei Xu, Donna L Mendrick, Weida Tong, et al. In silico drug repositioning–what we need to know. *Drug discovery today*, 2012. 7

William Loging, Raul Rodriguez-Esteban, Jon Hill, Tom Freeman, and John Miglietta. Cheminformatic/bioinformatic analysis of large corporate databases: Application to drug repurposing. *Drug Discovery Today: Therapeutic Strategies*, 8(3):109–116, 2012. 6, 7

Eugen Lounkine, Michael J Keiser, Steven Whitebread, Dmitri Mikhailov, Jacques Hamon, Jeremy L Jenkins, Paul Lavan, Eckhard Weber, Allison K Doak, Serge Cote, et al. Large-scale prediction and testing of drug activity on side-effect targets. *Nature*, 486(7403):361–367, 2012. 20

Diana Lousa, Antonio M Baptista, and Claudio M Soares. Analyzing the molecular basis of enzyme stability in ethanol/water mixtures using molecular dynamics simulations. *Journal of chemical information and modeling*, 52(2):465–473, 2012. 131

CR MacKenzie, KG Johnson, and IJ McDonald. Glycogen synthesis by amylosucrase from Neisseria perflava. *Canadian Journal of Microbiology*, 23(9):1303–1307, 1977. 134

Alpeshkumar K Malde and Alan E Mark. Challenges in the determination of the binding modes of non-standard ligands in X-ray crystal complexes. *Journal of computer-aided molecular design*, 25 (1):1–12, 2011. 70

Robert B McComb and George N Bowers. Study of optimum buffer conditions for measuring alkaline phosphatase activity in human serum. *Clinical chemistry*, 18(2):97–104, 1972. 135, 136

Daniel C McShan, S Rao, and Imran Shah. PathMiner: predicting metabolic pathways by heuristic search. *Bioinformatics*, 19(13):1692–1698, 2003. 131

Jose L Medina-Franco, Marc A Giulianotti, Gregory S Welmaker, and Richard A Houghten. Shifting from the single to the multitarget paradigm in drug discovery. *Drug discovery today*, 2013. 8

Jamel Meslamani, Didier Rognan, and Esther Kellenberger. sc-PDB: a database for identifying variations and multiplicity of 'druggable' binding sites in proteins. *Bioinformatics*, 27(9):1324–1326, 2011. 25, 31, 48, 73, 78

Jamel Meslamani, Jiabo Li, Jon Sutter, Adrian Stevens, Hugues-Olivier Bertrand, and Didier Rognan. Protein–ligand-based pharmacophores: generation and utility assessment in computational ligand profiling. *Journal of chemical information and modeling*, 52(4):943–955, 2012. 27, 72, 74, 91, 116

Jamel Meslamani, Ricky Bhajun, Francois Martz, and Didier Rognan. Computational profiling of bioactive compounds using a target-dependent composite workflow. *Journal of chemical information and modeling*, 53(9):2322–2333, 2013. 29, 154

Jordi Mestres, Elisabet Gregori-Puigjane, Sergi Valverde, and Ricard V Sole. The topology of drug–target interaction networks: implicit dependence on drug properties and target families. *Mol. BioSyst.*, 5(9):1051–1057, 2009. 19

Marvin J Meyers, Graciela B Arhancet, Susan L Hockerman, Xiangyang Chen, Scott A Long, Matthew W Mahoney, Joseph R Rico, Danny J Garland, James R Blinn, Joe T Collins, et al. Discovery of (3 S, 3a R)-2-(3-Chloro-4-cyanophenyl)-3-cyclopentyl-3, 3a, 4, 5-tetrahydro-2 H-benzo [g] indazole-7-carboxylic Acid (PF-3882845), an Orally Efficacious Mineralocorticoid Receptor (MR) Antagonist for Hypertension and Nephropathy. *Journal of medicinal chemistry*, 53(16): 5979–6002, 2010. 121

Francesca Milletti and Anna Vulpetti. Predicting polypharmacology by binding site similarity: from kinases to the protein universe. *Journal of chemical information and modeling*, 50(8):1418–1431, 2010. 27, 29, 31

Sayaka Mizutani, Edouard Pauwels, Veronique Stoven, Susumu Goto, and Yoshihiro Yamanishi. Relating drug–protein interaction network with drug side effects. *Bioinformatics*, 28(18):i522–i528, 2012. 20

Fazlin Mohd Fauzi, Alexios Koutsoukas, Rob Lowe, Joshi Kalpana, Tai-Ping Fan, Robert Glen, and Andreas Bender. Chemogenomics Approaches to Rationalizing the Mode-of-Action of Traditional Chinese and Ayurvedic Medicines. *Journal of chemical information and modeling*, 53(3):661–673, 2013. 6

Fabrice Moriaud, Stephane B Richard, Stewart A Adcock, Laetitia Chanas-Martin, Jean-Sebastien Surgand, Marouane Ben Jelloul, and Francois Delfaud. Identify drug repurposing candidates by mining the Protein Data Bank. *Briefings in bioinformatics*, 12(4):336–340, 2011. 6, 7

Pascal Muller, Gersande Lena, Eric Boilard, Sofiane Bezzine, Gerard Lambeau, Gilles Guichard, and Didier Rognan. In Silico-Guided Target Identification of a Scaffold-Focused Library: 1, 3, 5-Triazepan-2, 6-diones as Novel Phospholipase A2 Inhibitors. *Journal of medicinal chemistry*, 49 (23):6768–6778, 2006. 25

Ramaiah Muthyala. Orphan/rare drug discovery through drug repositioning. *Drug Discovery Today: Therapeutic Strategies*, 8(3):71–76, 2012. 6

Michael M Mysinger, Michael Carchia, John J Irwin, and Brian K Shoichet. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *Journal of medicinal chemistry*, 55(14):6582–6594, 2012.

Sander B Nabuurs, Chris AEM Spronk, Elmar Krieger, Hans Maassen, Gert Vriend, and Geerten W Vuister. Quantitative evaluation of experimental NMR restraints. *Journal of the American Chemical Society*, 125(39):12026–12034, 2003. 70

Herbert Nar, Margit Bauer, Angela Schmid, Jean-Marie Stassen, Wolfgang Wienen, Henning WM Priepke, Iris K Kauffmann, Uwe J Ries, and Norbert H Hauel. Structural basis for inhibition promiscuity of dual specific thrombin and factor Xa blood coagulation inhibitors. *Structure*, 9(1): 29–37, 2001. 75

J.H. Nettles, J.L. Jenkins, A. Bender, Z. Deng, J.W. Davies, and M. Glick. Bridging chemical and biological space: "target fishing" using 2D and 3D molecular descriptors. *Journal of medicinal chemistry*, 49(23):6802–6810, 2006. 18

S. Niijima, H. Yabuuchi, and Y. Okuno. Cross-target view to feature selection: identification of molecular interaction features in ligand- target space. *Journal of chemical information and modeling*, 51(1):15–24, 2010. 17

Irene Nobeli, Angelo D Favia, and Janet M Thornton. Protein promiscuity and its implications for biotechnology. *Nature biotechnology*, 27(2):157–167, 2009. 3

Alfons Nonell-Canals and Jordi Mestres. In silico target profiling of one billion molecules. *Molecular Informatics*, 30(5):405–409, 2011. 19

Tudor I Oprea, Julie E Bauman, Cristian G Bologa, Tione Buranda, Alexandre Chigaev, Bruce S Edwards, Jonathan W Jarvik, Hattie D Gresham, Mark K Haynes, Brian Hjelle, et al. Drug repurposing from an academic perspective. *Drug Discovery Today: Therapeutic Strategies*, 8(3): 61–69, 2012. 6

Yongping Pan, Niu Huang, Sam Cho, and Alexander D MacKerell. Consideration of molecular weight during compound selection in virtual target-based database screening. *Journal of chemical information and computer sciences*, 43(1):267–272, 2003. 133

Gaia V Paolini, Richard HB Shapland, Willem P Van Hoorn, Jonathan S Mason, and Andrew L Hopkins. Global mapping of pharmacological space. *Nature biotechnology*, 24(7):805–815, 2006. 19

Nicodeme Paul, Esther Kellenberger, Guillaume Bret, Pascal Müller, and Didier Rognan. Recovering the true targets of specific ligands by virtual screening of the protein data bank. *Proteins: Structure, Function, and Bioinformatics*, 54(4):671–680, 2004. 25, 26, 31

Agnes Peragovics, Zoltan Simon, Laszlo Tombor, Balazs Jelinek, Peter Hari, Pal Czobor, and Andras Malnasi-Csizmadia. Virtual Affinity Fingerprints for Target Fishing: A New Application of Drug Profile Matching. *Journal of chemical information and modeling*, 53(1):103–113, 2012. 28

V.I. Perez Nueno, V. Venkatraman, L. Mavridis, and D.W. Ritchie. Detecting Drug Promiscuity using Gaussian Ensemble Screening. *Journal of Chemical Information and Modeling*, 2012. 18

Gerard Pujadas, Montserrat Vaque, Anna Ardevol, Cinta Blade, MJ Salvado, Mayte Blay, Juan Fernandez-Larrea, and Lluis Arola. Protein-ligand docking: A review of recent advances and future perspectives. *Current Pharmaceutical Analysis*, 4(1):1–19, 2008. 13

Matthias Rarey, Bernd Kramer, Thomas Lengauer, and Gerhard Klebe. A fast flexible docking method using an incremental construction algorithm. *Journal of molecular biology*, 261(3):470–489, 1996. 90, 133, 152

Andrew G Reaume. Drug repurposing through nonhypothesis driven phenotypic screening. *Drug Discovery Today: Therapeutic Strategies*, 8(3):85–88, 2012. 6

Sereina Riniker and Gregory A Landrum. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of cheminformatics*, 5:26, 2013. 88

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010. 17, 80

Didier Rognan. Structure-based approaches to target fishing and ligand profiling. *Molecular Informatics*, 29(3):176–187, 2010. 21

Bryan L Roth, Douglas J Sheffler, and Wesley K Kroeze. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nature Reviews Drug Discovery*, 3(4):353–359, 2004. 7

Thomas S Rush, J Andrew Grant, Lidia Mosyak, and Anthony Nicholls. A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *Journal of medicinal chemistry*, 48(5):1489–1495, 2005. 18

Jens Sadowski, Johann Gasteiger, and Gerhard Klebe. Comparison of automatic three-dimensional model builders using 639 X-ray structures. *Journal of chemical information and computer sciences*, 34(4):1000–1008, 1994. 90

Paloma A Santacoloma, Gurkan Sin, Krist V Gernaey, and John M Woodley. Multienzyme-catalyzed processes: next-generation biocatalysis. *Organic Process Research & Development*, 15(1):203–212, 2010. 8

Christin Schärfer, Tanja Schulz-Gasch, Jérôme Hert, Lennart Heinzerling, Benjamin Schulz, Therese Inhester, Martin Stahl, and Matthias Rarey. CONFECT: Conformations from an Expert Collection of Torsion Patterns. *ChemMedChem*, 8(10):1690–1700, 2013. 49

Ingo Schellhammer and Matthias Rarey. TrixX: structure-based molecule indexing for large-scale virtual screening in sublinear time. *Journal of Computer-Aided Molecular Design*, 21(5):223–238, 2007. 39

Jochen Schlosser and Matthias Rarey. Beyond the virtual screening paradigm: Structure-based searching for new lead compounds. *Journal of Chemical Information and Modeling*, 49(4):800–809, 2009. 35, 39

Gisbert Schneider and Hans-Joachim Böhm. Virtual screening and fast automated docking methods. *Drug Discovery Today*, 7(1):64–70, 2002. 13

Nadine Schneider, Sally Hindle, Gudrun Lange, Robert Klein, Jürgen Albrecht, Hans Briem, Kristin Beyer, Holger Claußen, Marcus Gastreich, Christian Lemmen, et al. Substantial improvements in large-scale redocking and screening using the novel HYDE scoring function. *Journal of computer-aided molecular design*, 26(6):701–723, 2012. 90, 91, 133

Ida Schomburg, Antje Chang, Sandra Placzek, Carola Söhngen, Michael Rother, Maren Lang, Cornelia Munaretto, Susanne Ulas, Michael Stelzer, Andreas Grote, et al. BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic acids research*, 41(D1):D764–D772, 2013. 3

Karen T Schomburg and Matthias Rarey. Benchmark Datasets for Structure-based Computational Target Prediction. *Journal of chemical information and modeling*, DOI: 10.1021/ci500131x, 2014. 79, 102, 115

Karen T Schomburg, Inés Ardao, Katharina Götz, Fabian Rieckenberg, Andreas Liese, An-Ping Zeng, and Matthias Rarey. Computational Biotechnology: Prediction of competitive substrate inhibition of enzymes by buffer compounds with protein-ligand docking. *Journal of Biotechnology*, 2012. 131, 134

Karen T Schomburg, Stefan Bietz, Hans Briem, Angela M Henzler, Sascha Urbaczek, and Matthias Rarey. Facing the Challenges of Structure-based Target Prediction by Inverse Virtual Screening. *Journal of chemical information and modeling*, 54(6):1676–1686, 2014. 47, 64, 99

Zoltan Simon, Agnes Peragovics, Margit Vigh-Smeller, Gabor Csukly, Laszlo Tombor, Zhenhui Yang, Gergely Zahornszky-Kohalmi, Laszlo Vegner, Balazs Jelinek, Peter Hari, et al. Drug effect prediction by polypharmacology-based interaction profiling. *Journal of chemical information and modeling*, 52(1):134–145, 2011. 28

Richard B Smith. Repositioned drugs: integrating intellectual property and regulatory strategies. *Drug Discovery Today: Therapeutic Strategies*, 8(3):131–137, 2012. 6

Christopher Southan, Kiran Boppana, Sarma ARP Jagarlapudi, and Sorel Muresan. Analysis of in vitro bioactivity data extracted from drug discovery literature and patents: ranking 1654 human protein targets by assayed compounds and molecular scaffolds. *Journal of cheminformatics*, 3(1): 1–11, 2011. 3

Chris AEM Spronk, Sander B Nabuurs, Alexandre MJJ Bonvin, Elmar Krieger, Geerten W Vuister, and Gert Vriend. The precision of NMR structure ensembles revisited. *Journal of biomolecular NMR*, 25(3):225–234, 2003. 70

Theodora M Steindl, Daniela Schuster, Christian Laggner, and Thierry Langer. Parallel screening: a novel concept in pharmacophore modeling and virtual screening. *Journal of chemical information and modeling*, 46(5):2146–2157, 2006. 27

Helena Strömbergsson and Gerard J Kleywegt. A chemogenomics view on protein-ligand spaces. *BMC bioinformatics*, 10(Suppl 6):S13, 2009. 48

Milton T Stubbs, Robert Huber, and Wolfram Bode. Crystal structures of factor Xa specific inhibitors in complex with trypsin: structural grounds for inhibition of factor Xa and selectivity against thrombin. *FEBS letters*, 375(1):103–107, 1995. 75

Dagmar Stumpfe and Jürgen Bajorath. Exploring Activity Cliffs in Medicinal Chemistry: Miniperspective. *Journal of medicinal chemistry*, 55(7):2932–2942, 2012. 119

Noe Sturm, Jeremy Desaphy, Ronald J Quinn, Didier Rognan, and Esther Kellenberger. Structural insights into the molecular basis of the ligand promiscuity. *Journal of chemical information and modeling*, 52(9):2410–2421, 2012. 4

S Joshua Swamidass. Mining small-molecule screens to repurpose drugs. *Briefings in bioinformatics*, 12(4):327–335, 2011. 6

David C Swinney and Jason Anthony. How were new medicines discovered? *Nature reviews Drug discovery*, 10(7):507–519, 2011. 5

Yusuf Tanrikulu, Björn Krüger, and Ewgenij Proschak. The holistic integration of virtual screening in drug discovery. *Drug discovery today*, 18(7):358–364, 2013. 13

Richard D. Taylor, Philip J. Jewsbury, and Jonathan W. Essex. A review of protein-small molecule docking methods. *Journal of computer-aided molecular design*, 16(3):151–166, 2002. 13

Sheng Tian, Youyong Li, Dan Li, Xiaojie Xu, Junmei Wang, Qian Zhang, and Tingjun Hou. Modeling compound–target interaction network of traditional chinese medicines for type ii diabetes mellitus: Insight for polypharmacology and drug design. *Journal of chemical information and modeling*, 53 (7):1787–1803, 2013. 28

Simon Tietze and Joannis Apostolakis. GlamDock: development and validation of a new docking tool on several thousand protein-ligand complexes. *Journal of chemical information and modeling*, 47(4):1657–1672, 2007. 26

Jean-François Truchon and Christopher I Bayly. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *Journal of chemical information and modeling*, 47(2):488–508, 2007. 88

Sascha Urbaczek, Adrian Kolodzik, J Robert Fischer, Tobias Lippert, Stefan Heuser, Inken Groth, Tanja Schulz-Gasch, and Matthias Rarey. NAOMI: On the Almost Trivial Task of Reading Molecules from Different File formats. *Journal of chemical information and modeling*, 51(12): 3199–3207, 2011. 38, 48, 79, 132

Sascha Urbaczek, Adrian Kolodzik, Inken Groth, Stefan Heuser, and Matthias Rarey. Reading PDB: Perception of Molecules from 3D Atomic Coordinates. *Journal of chemical information and modeling*, 53(1):76–87, 2012. 38

Sameer Velankar, Jose M Dana, Julius Jacobsen, Glen van Ginkel, Paul J Gane, Jie Luo, Thomas J Oldfield, Claire O'Donovan, Maria-Jesus Martin, and Gerard J Kleywegt. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic acids research*, 41(D1): D483–D489, 2013. 81

Vishwesh Venkatraman, Violeta I Pérez-Nueno, Lazaros Mavridis, and David W Ritchie. Comprehensive comparison of ligand-based virtual screening tools against the DUD data set reveals limitations of current 3D methods. *Journal of chemical information and modeling*, 50(12):2079–2093, 2010. 88

Marcel L Verdonk, Jason C Cole, Michael J Hartshorn, Christopher W Murray, and Richard D Taylor. Improved protein–ligand docking using GOLD. *Proteins: Structure, Function, and Bioinformatics*, 52(4):609–623, 2003. 25

Ingo Vogt and Jordi Mestres. Drug-Target Networks. *Molecular Informatics*, 29(1-2):10–14, 2010. 3

Andrea Volkamer, Axel Griewel, Thomas Grombacher, and Matthias Rarey. Analyzing the topology of active sites: on the prediction of pockets and subpockets. *Journal of chemical information and modeling*, 50(11):2041–2052, 2010. 153

Jian Wang, Peter A Kollman, and Irwin D Kuntz. Flexible ligand docking: a multistep strategy approach. *Proteins: Structure, Function, and Bioinformatics*, 36(1):1–19, 1999. 22

Wei Wang, Xi Zhou, Wanlin He, Yi Fan, Yuzong Chen, and Xin Chen. The interprotein scoring noises in glide docking scores. *Proteins: Structure, Function, and Bioinformatics*, 80(1):169–183, 2012. 22, 55, 95, 149

Xiangyun Wang and Nigel Greene. Comparing Measures of Promiscuity and Exploring Their Relationship to Toxicity. *Molecular Informatics*, 31(2):145–159, 2012. 16

Yanli Wang, Jewen Xiao, Tugba O Suzek, Jian Zhang, Jiyao Wang, and Stephen H Bryant. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic acids research*, 37(suppl 2):W623–W633, 2009. 17

Gregory L Warren, Thanh D Do, Brian P Kelley, Anthony Nicholls, and Stephen D Warren. Essential considerations for using protein–ligand structures in drug discovery. *Drug discovery today*, 17 (23):1270–1281, 2012. 69, 73, 74

Bohdan Waszkowycz, David E Clark, and Emanuela Gancia. Outstanding challenges in protein–ligand docking and structure-based virtual screening. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(2):229–259, 2011. 14

Scott J Weiner, Peter A Kollman, David A Case, U Chandra Singh, Caterina Ghio, Guliano Alagona, Salvatore Profeta, and Paul Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society*, 106(3):765–784, 1984. 22, 24

David S Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, 34(suppl 1):D668–D672, 2006. 6, 17, 24, 28, 73, 80, 124

Di Wu, Qin Wang, Rajeev S Assary, Linda J Broadbelt, and Goran Krilov. A computational approach to design and evaluate enzymatic reaction pathways: application to 1-butanol production from pyruvate. *Journal of chemical information and modeling*, 51(7):1634–1647, 2011. 8, 131

Kesheng Wu. FastBit: an efficient indexing technology for accelerating data-intensive science. In *Journal of Physics: Conference Series*, volume 16, page 556. IOP Publishing, 2005. 35, 42

Helmut Wurst and Arthur Kornberg. A soluble exopolyphosphatase of Saccharomyces cerevisiae. Purification and characterization. *Journal of Biological Chemistry*, 269(15):10996–11001, 1994. 134

Kui Xu and Timothy R Cote. Database identifies FDA-approved drugs with potential to be repurposed for treatment of orphan diseases. *Briefings in bioinformatics*, 12(4):341–345, 2011. 6, 7

Yoshihiro Yamanishi, Masaaki Kotera, Minoru Kanehisa, and Susumu Goto. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics*, 26(12):i246–i254, 2010. 19

Yoshihiro Yamanishi, Edouard Pauwels, and Masaaki Kotera. Drug Side-Effect Prediction Based on the Integration of Chemical and Biological Spaces. *Journal of Chemical Information and Modeling*, 52(12):3284–3292, 2012. 20

Yongliang Yang, S James Adelstein, Amin I Kassis, et al. Target discovery from data mining approaches. *Drug discovery today*, 14(3-4):147–154, 2009.

Elizabeth Yuriev, Mark Agostino, and Paul A Ramsland. Challenges and advances in computational docking: 2009 in review. *Journal of Molecular Recognition*, 24(2):149–164, 2011. 13

Stefan Zahler, Simon Tietze, Frank Totzke, Michael Kubbutat, Laurent Meijer, Angelika M Vollmar, and Joannis Apostolakis. Inverse in silico screening for identification of kinase inhibitor targets. *Chemistry & biology*, 14(11):1207–1214, 2007. 26, 29, 31

Shoude Zhang, Weiqiang Lu, Xiaofeng Liu, Yanyan Diao, Fang Bai, Liyan Wang, Lei Shan, Jin Huang, Honglin Li, and Weidong Zhang. Fast and effective identification of the bioactive compounds and their targets from medicinal plants via computational chemical biology approach. *MedChemComm*, 2(6):471–477, 2011. 6

Y-H Percival Zhang, Barbara R Evans, Jonathan R Mielenz, Robert C Hopkins, and Michael WW Adams. High-yield hydrogen production from starch and water by a synthetic enzymatic pathway. *PLoS One*, 2(5):e456, 2007. 130, 139

Feng Zhu, BuCong Han, Pankaj Kumar, XiangHui Liu, XiaoHua Ma, XiaoNa Wei, Lu Huang, YangFan Guo, LianYi Han, ChanJuan Zheng, et al. Update of TTD: therapeutic target database. *Nucleic acids research*, 38(suppl 1):D787–D791, 2010. 23

Feng Zhu, Zhe Shi, Chu Qin, Lin Tao, Xin Liu, Feng Xu, Li Zhang, Yang Song, Xianghui Liu, Jingxian Zhang, et al. Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic acids research*, 40(D1):D1128–D1136, 2012. 23, 31

Grant R Zimmermann, Joseph Lehar, and Curtis T Keith. Multi-target therapeutics: when the whole is greater than the sum of the parts. *Drug discovery today*, 12(1):34–42, 2007. 16

A
Implementation

During the course of this dissertation project, the software tools *i*RAISE and ComplexViewer were developed. This Appendix gives an overview of the implementation details.
All software was implemented in C++ and next to some external software libraries, the libraries of the NAOMI software library developed at the Center for Bioinformatics in the Working group of Algorithmic Molecular Design were used. In Figure A.1, the dependencies of *i*RAISE and the ComplexViewer are shown. The coloring scheme highlights the origin of the libraries: Red libraries are external libraries, yellow libraries are internal NAOMI-libraries and blue libraries were programmed during this dissertation project.
The ComplexViewer uses Qt, boost and SQLite as external libraries and of the interal libraries ComplexDB, ProteinDB, Visualization3dLib, sdf, MolLib, ProLib, ComplexLib and ProtossLib. *i*RAISE uses of the external libraries additionally FastBit. Of the internal libraries it used the ComplexDB, ProteinDB, MolLib, ProLib, ComplexLib. ProtossLib, Geometry3d, Trixx, Interactions, Conformations and the *i*RAISE-Index.

External libraries

- **boost** (www.boost.org) is a C++ library with efficient algorithms, data structures and program options

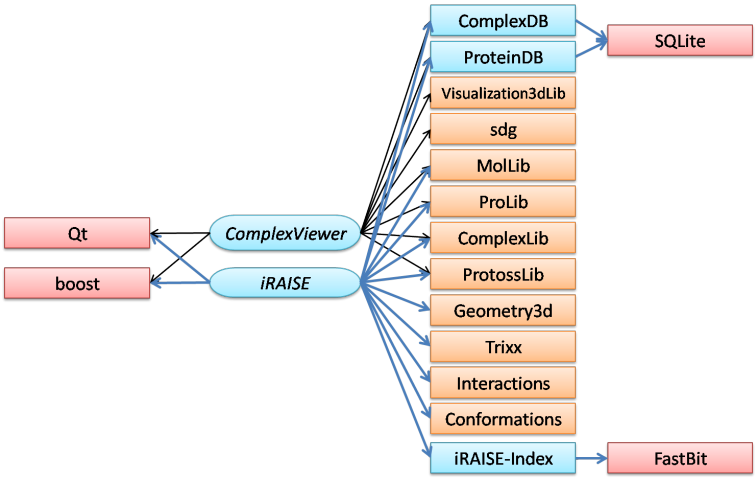


Figure A.1: Dependencies of the ComplexViewer and *i*RAISE. Libraries highlighted in red are external libraries, in yellow are internal Naomi-libraries and in blue are libraries developed in this dissertation project.

- **Qt** (qt-project.org) is used mainly for file-handling and GUI-design
- **SQLite** (www.sqlite.org) is used as SQL database engine.
- **FastBit** (https://sdm.lbl.gov/fastbit/) is used for storage and querying of the descriptors in a bitmap index.

Developed libraries

- **ComplexDB** contains functions for storing and accessing a protein-ligand complex in a ComplexDB.
- **ProteinDB** contains functions for storing and accessing a protein in a ProteinDB.
- **iRAISE-Index** contains interface functions to the FastBit external library.

Internal NAOMI libraries

- **Visualization3dLib** contains functions for displaying proteins and ligands in 3D.
- **sdg** contains functions for 2D coordinate generation for ligands and displaying structure diagrams.
- **MolLib** contains functionality for reading/writing molecules from/to files, initializing and handling of molecules.

- **ProLib** contains functionality for reading and initializing proteins from files.
- **ComplexLib** contains the classes Complex and ActiveSite and functionality to handle a protein-ligand complex.
- **ProtossLib** is the library of Protoss, which automatically places hydrogen coordinates in a protein-ligand complex for an optimal hydrogen-bond network.
- **Geometry3d** contains utilities and functions for objects in 3-dimensional space.
- **Trixx** contains functionality for triangle descriptor calculation and for the grid scoring.
- **Interactions** is used to assign interaction spots to a protein or molecule.
- **Conformations** contains the utilities to create conformations for a small molecule.

Since FastBit is only available for Linux-based operating systems, therefore *i*RAISE also is only available for Linux. Additionally, *i*RAISE is only available as a command-line tool. The ComplexViewer is a viewer for the content of the protein database of a *i*RAISE project and of the solutions. It cannot be used for starting the pre-processing or the screening of *i*RAISE.

B

Parameters

Parameter	**Default value**	**User setable?**
Triangle descriptor parameters		
Descriptor side length minimum (maximum)	1Å(9.9Å)	no
Number of descriptor bulk rays	80	no
Descriptor bulk ray minimum (maximum)	1Å(15.0Å)	no
Descriptor side length matching tolerance	1.2Å	no
Descriptor bulk matching tolerance	0.75Å	no
Descriptor direction matching tolerance in degree	36	no
Index partition maximum number of type descriptors	400.000	no
Index partition maximum number of total descriptors	1.800.000	no
Scoring parameters		
Lennard-Jones terms repulsion (r) and attraction parameters (a)		
-Hydrogen bonds	a=8, r=10	no
-Metal contacts	a=4, r=6	no
-Hydrophobic contacts	a=3, r=12	no
-Mismatches	a=6, r=12	no
Reference score cutoff	75%	no
Pose coverage reference cutoff	80%	no
Pocket coverage reference cutoff	80%	no
Pocket coverage weighting factor	0.8	no
Contact distance atom-atom	4.5Å	no
Screening parameters		
Size of active site around reference	6.5Å	yes
Grid spacing active site grid	0.8Å	no
Number of conformations of molecule	200	yes
Number of poses passing clash grid filter	500	no
Number of poses written to solution databse	10	no

Table B.1: List of parameters used in *i*RAISE and their default-values

C

*i*RAISE Userguide

C.1 About *i*RAISE

*i*RAISE is a structure-based inverse screening tool which predicts targets for small molecules. It is developed at the Center for Bioinformatics of the University of Hamburg (www.zbh.uni-hamburg.de). The software is still in a development status. For conditions of use please see www.zbh.uni-hamburg.de/raise.

This User Guide contains instructions on how to use the software. *i*RAISE is a command line tool restricted to the use on Linux operating systems. It is recommended to be used on a computing cluster in parallel for large amounts of target structures. This User Guide is structured as follows: First of all, the command line options and the folder structure are explained. Then the usage of *i*RAISE is shown for the most common use cases. Finally, the limitations of the software are listed.

C.2 Using *i*RAISE

C.2.1 Folder of the *i*RAISE-package

After extracting the files from the tar-Archive with
`tar xfvz iRAISE.tar.gz`
the *i*RAISE folder with several subfolders is available. Table C.1 gives an overview of the subfolders and their content.

Folder	Content
bin	iRAISE executable and needed software libraries
Scripts	Example shell scripts for usage of the *i*RAISE-executable
ExampleData	Some example protein structures and ligands

Table C.1: Overview of the folders of the *i*RAISE-package

The example protein structures and ligand files found in the ExampleData are taken from the Astex Diverse Set developed by Hartshorn et al. (Hartshorn, MJ, Verdonk, ML, Chessari, G, Brewerton SC, Mooij WT, Mortenson PN, Murray CW. (2007) Diverse, high-quality test set for the validation of protein-ligand docking performance. J Med Chem, 50(4), 726-741).

C.2.2 Starting *i*RAISE

For starting the *i*RAISE software, change into the *i*RAISE directory and type

```
./bin/iRAISE_release --help
```

into your command line. This will start the program and list the command line options. These are also listed in this User Guide in Table C.2. In this table, the first column shows the available command line options with the respective arguments, the second column specifies in which use case an option may be used and the right column gives a more detailed explanation of the option. The following use cases of the *i*RAISE program are possible (abbreviations as used in the table are given in brackets):

- **Create a project (C)**. This use case is needed to set up the screening project for all other use cases. The given proteins are processed and stored in a database and the descriptor index is calculated.
- **Screening with a molecule (S)**. This use case screens the generated structure database for potential targets for a query molecule. It can be carried out repeatedly on a screening project.

- **Evaluation of solutions (E)**. This use case processes generated screening results. The results of several screening runs are gathered and predicted poses can be written to an output file.

Table C.2: Overview of command line options for the iRAISE_release executable

Command line option	Use case	Explanation
-h,- -help	all	Shows a list of command line options with explanations
-v,- -version	all	Print the version
-o,- -output-level <number>	all	Level of information printed during execution. 0=only errors are printed, 1=warnings and errors are printed, 2=Next to warnings and errors, the workflow information is printed.
-n,- -name <string>	all	Name of the screening project. The name can be chosen by the user, but the name must only contain characters which can be used for a folder name. The name is then used to name the folder of the screening project, with the prefix "raise_-resources_". As an example, if you type `-n Trypsin`, then the project data will be put in a folder with the name raise_resources_Trypsin. If this is option is omitted, an automatic name is generated. For screening, use this option to give the project that shall be screened.
-I, - -Index <string>	all	Path, where the screening project will be created/is located. If no path is given, the screening project is created at the location where the *i*RAISE-executable has been called.
-C,- -Create <file-name>	C	First use case of *i*RAISE: Create a screening project from a protein list. This option cannot be combined with the `-S` use case. The argument of `-C` needs to be a file which contains a list of protein pdb files with full path. See `-r` for how to provide the reference ligands. If no reference ligands are given, the active sites are determined automatically from the ligands in the pdb file.

...continues below

Table C.2: Overview of command line options for the *i*RAISE_release executable (continued)

Command line option	Use case	Explanation
-r, - -refs <file-name>	C	File that contains the reference ligands as list of molecule files with full path. This list has to be complementary to the list given to the `-C` option, in a way that the line number of the refs-file contains the reference ligand for the protein of the same line number as in the protein file provided.
-a,- -active-site <double>	C	Radius for active site determination, if this argument is not given, a radius of 6.5 Å is used.
-S, - -Screen <file-name>	S	Second use case of *i*RAISE: Screening of the target database with a small molecule. As input this option requires a molecule file in .sdf, .mol2 or .pdb format.
-c, - -conformations <number/level>	S	Generate conformations for the input molecule for screening. Either maximal number of conformations or three different levels, Q1, Q2, or Q3 can be given. If this option is not given, only the input conformation is used.
-w,- -write-confs	S	Write conformations. If this option is given, the generated conformations of the screening molecule are written in an sdf file to the solution folder in the project folder.
-E,- -Evaluate	E	Third use case of *i*RAISE: Use this option to consolidate solution databases, for example if the screening was executed in parallel.
-e,- -export <number>	E	Export predicted ligand poses as sdf file of the ligand. The number argument specifies, for how many targets (starting with the best ranked target) poses of the ligand are written to the solutions folder of the project folder. If a screening project has been screened with several ligands, poses are written for each ligand. .
- -screen-start <number>	S	Only for expert users for parallel screening: If partitions shall be screened in parallel, this option specifies, which partition is screened with the query molecule

...continues below

Table C.2: Overview of command line options for the *i*RAISE_release executable (continued)

Command line option	Use case	Explanation
- -screen-no <number>	S	Only for expert users for parallel screening: If partitions shall be screened in parallel, this option specifies, how many partitions are screened, started at the partition which was given with the - -screen-start option.

C.2.3 Structure of a screening project

Once a screening project has been created (see section C.3.1 for how to to this), a project folder with the following structure has been created:

- raise_resources_*givenname**
  - **idx** This folder contains the generated index of protein descriptors. It is needed for screening. The file `cataloging.log` in this folder contains log information about the indexing.
  - **data** This folder contains the proteins stored in a database in the file `complex.db`. After a screening run, here solution databases for the query compounds are also located. If a consolidation (with option `-E`) has been carried out, this folder also contains the complex solutions database file `SolutionsComplete.db`
  - **solutions** Here the solutions of *i*RAISE screenings are found. For each query molecule which has been used to screen the database, a text file with the scores is written to this folder. If no screening has been carried out, this folder is empty. If the same molecule is screened multiple times against the project, the results are appended to that file. The name of the solution file is a combination of the molecule name and the corresponding molecule file name. Also, if the poses of a ligand have been written, a subfolder named `poses` has been created here, which contains the sdf files of ligand poses for each target separately.

*The *givenname* can be supplied by the user with the `-n` option of the *i*RAISE tool.

C.3 Example use cases

Together with the *i*RAISE executable, example data and scripts are provided in the *i*RAISE package. This data shall demonstrate how the *i*RAISE tool shall be used. In the first script, a project is created from a list of proteins and reference ligand. Another script shows, how to create a project from proteins without reference ligands. Then, the created project is screened with a ligand. Finally, the results of this screen are consolidated and poses of the ligand are extracted which can then be further examined or used in other tools.
As example data, the 85 complexes of the Astex Diverse Set are used.

C.3.1 How to create a screening project with reference ligands

The script for an example of how to create a screening project can be found in the Scripts folder of the *i*RAISE package. The script is called `CreatingScreeningProject.sh`. It can be used to create a screening project with the example Astex Data. The script has to be called from the iRAISE folder since it contains relative paths:

```
 #!/bin/bash

# input variables
your_path=.
proteins=${your_path}/ExampleData/pdb.list
referenceligs=${your_path}/ExampleData/lig.list
executable_path=${your_path}/bin/
projectpath=${your_path}
projectname=AstexTest

#call of iRAISE tool with creating a project option
${executable_path}/iRAISE_release \
-C ${proteins}  \
--refs ${referenceligs} \
--name ${projectname} \
-I ${projectpath}
```

The pdb codes for the proteins as are the reference ligands are given by a list (For the lists see folder ExampleData). The project was named 'AstexTest' and because the project path was given as ".", the project with the name 'raise_resources_AstexTest' was stored in the

iRAISE folder from which the script was called. After calling this script, the project is ready for screening. The project contains now all 85 targets of the Astex Diverse Set. For creating own projects, the user only has to exchange the projectname, give the lists of the proteins and the reference ligands that shall be used in the screening project.

C.3.2 How to create a screening project without reference ligands

The script for an example of how to create a screening project with pdb files only without a reference ligand list can be found in the Scripts folder of the *i*RAISE package. The script is called `CreatingScreeningProject2.sh`. It also uses the AstexData, but this time only the pdb-files of the proteins. The script has to be called from the *i*RAISE folder since it contains relative paths:

```
#!/bin/bash

# input variables
your_path=.
proteins=${your_path}/ExampleData/pdb.list
executable_path=${your_path}/bin/
projectpath=${your_path}
projectname=AstexTest2

#call of \textit{i}RAISE tool with indexing option
${executable_path}/iRAISE_release \
-C ${proteins}  \
--name ${projectname} \
-I ${projectpath}
```

This time, only the list is given which contains the proteins. The *i*RAISE tool now on itself generates active sites, wherever it finds a ligand in the pdb file which is no buffer or cofactor or crystallization agent. Since a new project name 'AstexTest2' was given, a separate project will be created.

C.3.3 How to screen a project with a query molecule

For screening a project with a molecule, the prepared screening project is necessary as well as a file which contains the query molecule with 3D coordinates. A script which screens the previously created project with the name "AstexTest" with a ligand from the Astex Diverse Set, can be found in the Scripts folder with the name `Screening.sh`. The script needs the pdb code of the ligand that shall be screened in lowercase as a command line argument. One example call of that script would be
`./Scripts/Screening.sh 1g9v`
The script also has to be called from the *i*RAISE folder since it contains relative paths:

```
#!/bin/bash

# input variables
your_path=.
library=${your_path}/ExampleData/cryst_ligs/
executable_path=${your_path}/bin/
projectpath=${your_path}
projectname=AstexTest

#call of \textit{i}RAISE tool with screening option
${executable_path}/\textit{i}RAISE_release \
 -S ${library}/$1_crysth.mol2 \
--name ${projectname} \
-I ${projectpath} \
-c 200
```

In this script, the number of conformations that shall be generated are limited to 200 by the option `-c` which creates conformations. The solutions of the screening can be found in text form in the solutions-folder of the screening project. See the next paragraph for how to extract the ligand poses from the solutions.

C.3.4 How to extract ligand poses from a screening project

The script `ConsolidateAndWritePoses.sh` shows, how poses can be exported to sdf files for further use, e.g. binding mode evaluation. The script first of all gathers all so far available screening results in a database. The number of solutions that shall be exported can be given with the option -e. The given number decides, for how many targets for each ligand poses are written. For example, if the option -e 5 is given, the poses for the 5 best scored targets are exported. The poses are then written to the solution folder of the project. For one target, all poses (maximally 10) are written to a separate multi-sdf file. The file is named after the following scheme:

lignd-name_target-rank_pocket-ID_protein-name.sdf

Example: The file name `1g9v_1_2h_1HVY.sdf` indicates, that the file contains poses of the molecule with the name 1g9v for the first ranked target with the name 1HVY in pocket ID 2 (only of relevance if proteins with several pockets are contained in the screening project).

```
#!/bin/bash

# input variables
your_path=.
executable_path=${your_path}/bin/
projectpath=${your_path}
projectname=AstexTest

#call of iRAISE tool for exporting poses
${executables}/iRAISE_release \
  -E \
  --name ${projectname} \
  -I ${projectpath} \
  -e 5
```

The `-E` and `-e` options can be combined together with the -S option, then the consolidation and export is done immediately after screening. However, if several screening partitions are screened in parallel, the `-E` option should be called only after all screening runs are finished.

C.4 Limitations

Note the following limtations of the *i*RAISE software

- Covalently bound metals: Ligands with covalently bound metals can currently not be handled by the tool.
- The *i*RAISE concept cannot handle query molecules which contain no hydrogen acceptors or donors at all.
- Huge amounts of proteins are stored in the data partitions of the project . One data partition contains about 100 proteins, depending on the size of the active site. For parallel screening, thus it is recommended to screen several of these data partitions in parallel (see command line options). The results however, are stored in separate solution text files for each partition, meaning that the user has to merge the solutions manually e.g. by joining the text files containing the scores.
- Since the project is still in development status, no recommendations on computing capabilities are made, however, for each partition that is screened in parallel, it is assumed that 8GB Ram are available.

D

ComplexViewer Userguide

D.1 About ComplexViewer

The ComplexViewer is a viewer for the protein database content of an *i*RAISE project and the solutions of an *i*RAISE screening. It is a basic graphic user interface in development status and developed for usage in combination with an *i*RAISE project only. Its functionality is therefore limited to this application.
This user guide contains instructions how to use the ComplexViewer. First, it is explained how to browse the proteins contained in an *i*RAISE project. Subsequently, it is shown, how solutions of an *i*RAISE screening can be viewed by browsing targets with the predicted binding modes.

D.2 Starting ComplexViewer

The ComplexViewer is currently available for Linux-operated systems only. For starting the GUI, type

```
./bin/ComplexViewer_release
```

into the command line.

D.3 Browsing *i*RAISE proteins

The first use case of the ComplexViewer is the browsing of the protein content of an *i*RAISE project. Firstly, the database file has to be loaded. Choose the `Load Database` option from the Menu. In the now popping up dialog, choose the `complex.db` file of your *i*RAISE project located in the `data` folder of the project.

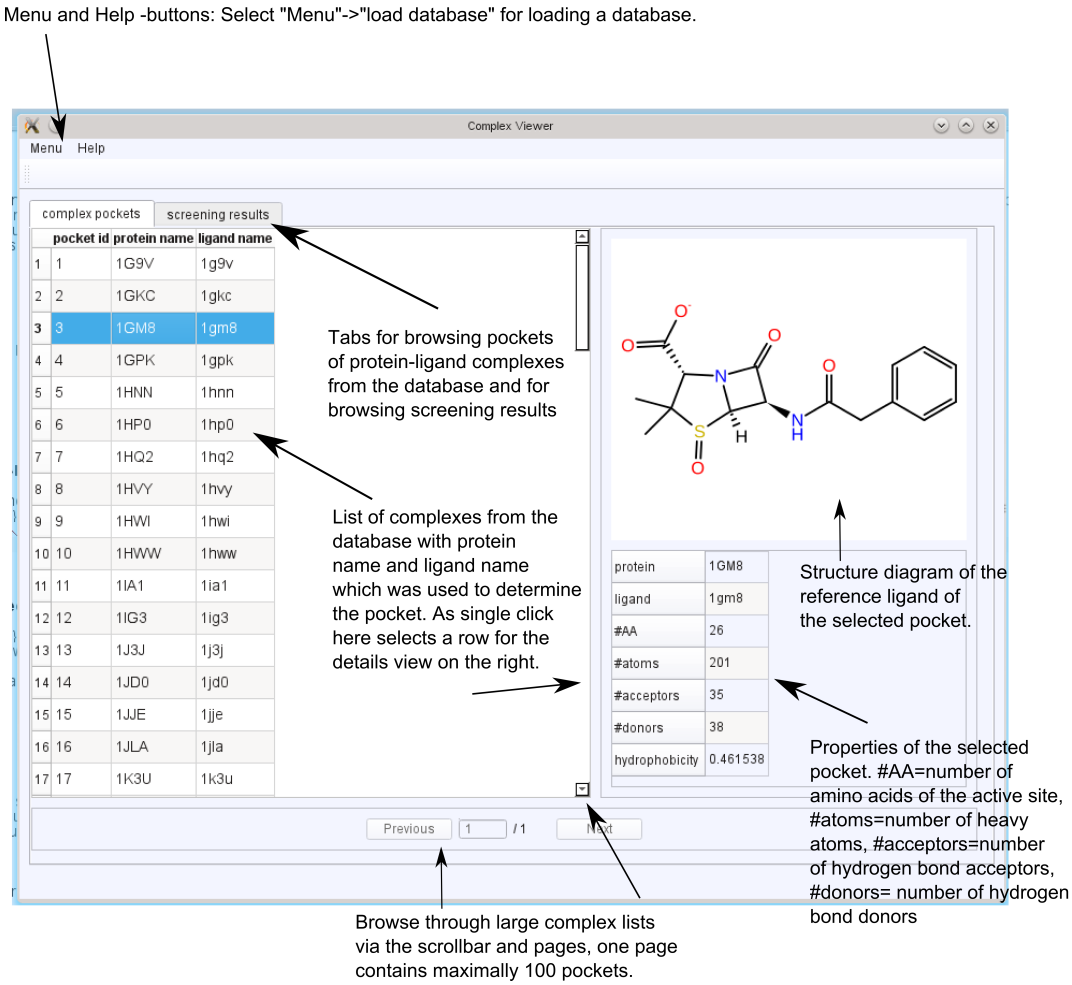


Figure D.1: Screenshot of the ComplexViewer with explanatory text.

In Figure D.1, the ComplexViewer is shown after a database has been loaded. In the `complex`

`pockets` tab, the content of the protein pocket database of the current *i*RAISE project is shown in form of a table. It lists a pocket id (in case one protein is registered with several pockets), the protein name and the name of the reference ligand which was used to determine the active site. A single click on a row of this table updates the details section (see Figure D.1), where the reference ligand is depicted in form of a 2D structure diagram as well as some properties of the active site are listed.
The proteins contained in the loaded database are organized on pages, with 100 proteins per page. `Next` and `previous` buttons and a page number input allows to browse through all proteins.

D.3.1 3D viewer of pockets

A double click on a row of the protein pocket table opens a 3D viewer in a separate window. This viewer depicts the amino acids of the active site and the reference ligand in 3D (see Figure D.2). The content of the 3D viewer is updated every time a new pocket is selected by a double click on the respective row. The following options are available for the 3D display of a pocket:

- Showing water molecules
- Displaying the protein backbone as chain trace
- Showing the ligand in a sphere representation of the van-der-Waals radii of its atoms

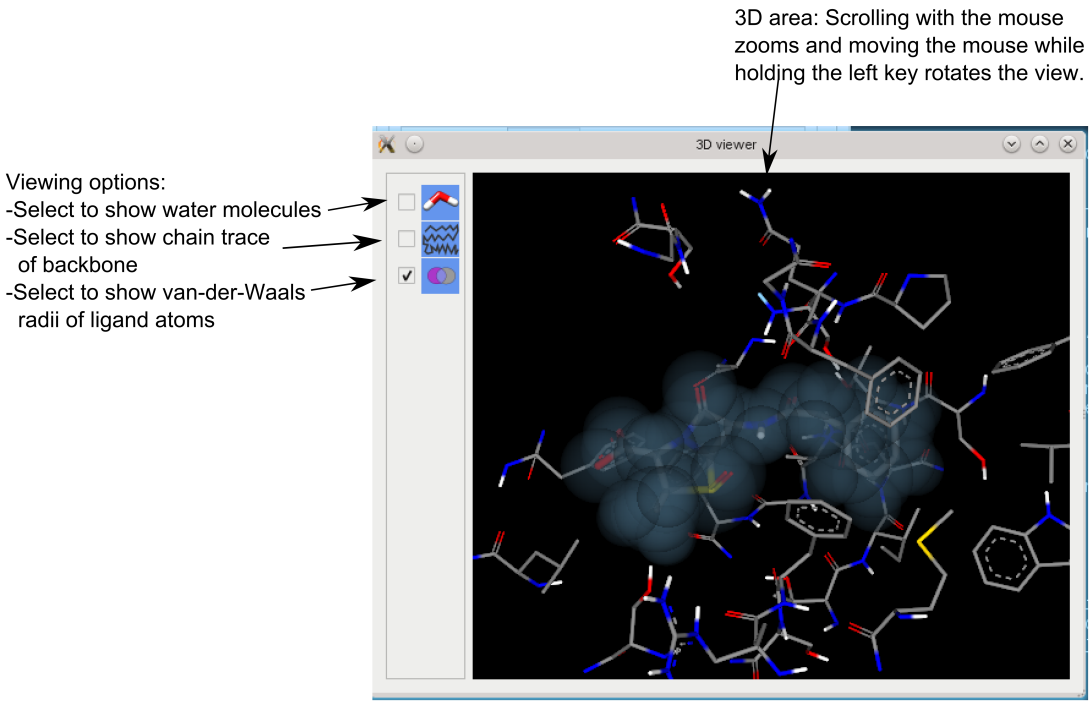


Figure D.2: Screenshot of the 3D viewer with explanatory text.

D.4 Inspecting *i*RAISE screening solutions

The ComplexViewer can further be used for browsing the solutions of an *i*RAISE screening. In Figure D.3, the `screening results` tab of the ComplexViewer is shown. This tab is only filled with data if the *i*RAISE project from which the database has been loaded was already screened with one or several ligands. A table on the left side of the view contains the names of these ligands and the number of target matches which were found in the screening. Details of a screening run are displayed in the target solution table after the respective row has been selected in the left table by a double click. The target solution table is located at the right side of the `screening results` tab (see Figure D.3). It contains the name of the query ligand, the name of the protein, the ID of the pocket (for identification if one protein was registered with several pockets) and the number of ligand poses, of which maximally 10 are shown, sorted by score.

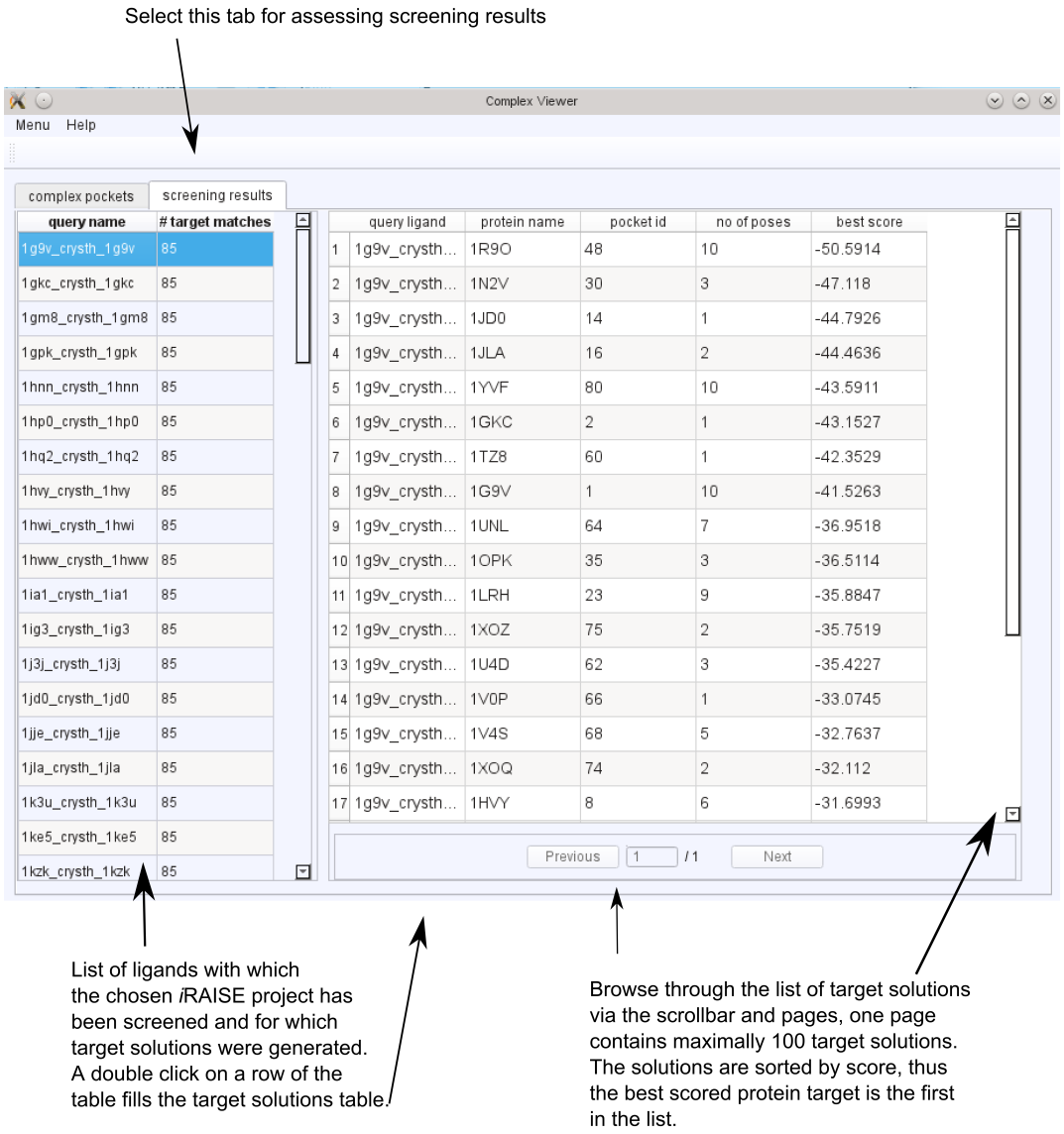


Figure D.3: Screenshot of the `screening results` tab of the ComplexViewer with explanatory text.

D.4.1 3D viewer of solutions

A double click on a row of the target solutions table opens the 3D viewer in a separate window. In Figure D.4, a screenshot is shown. The top panel of the window shows the name of the currently shown protein pocket, the score of the current pose and the number of the current pose. The generated poses can be browsed by the scrollbar at the bottom, which changes the pose of the ligand and updates the score and the number. A checkbox in the top panel allows to include the reference ligand into the viewer. A button labeled `PROTOSS` can be used to align the hydrogen atoms of the currently shown complex after the best hydrogen bond network calculated with Protoss. This step has to be initiated manually with the button and is not re-calculated automatically with each pose.

Next to the same viewing options as contained in the pocket 3D viewer, the list of available options is expanded in the solution viewer with the following options:

- Showing the reference ligand of the current pocket.
- Showing all atoms which are accessible to the ligand, i. e., located near a ligand atom. This feature helps in studying with which atoms the ligand may interact.
- Highlighting of not-covered ligand atoms with pink spheres.
- Highlighting of not-covered pocket atoms with pink spheres (i.e. atoms which have no pocket or ligand atom in a radius of 4.5 Å in their neighborhood)
- Showing a grid representation of the active site.

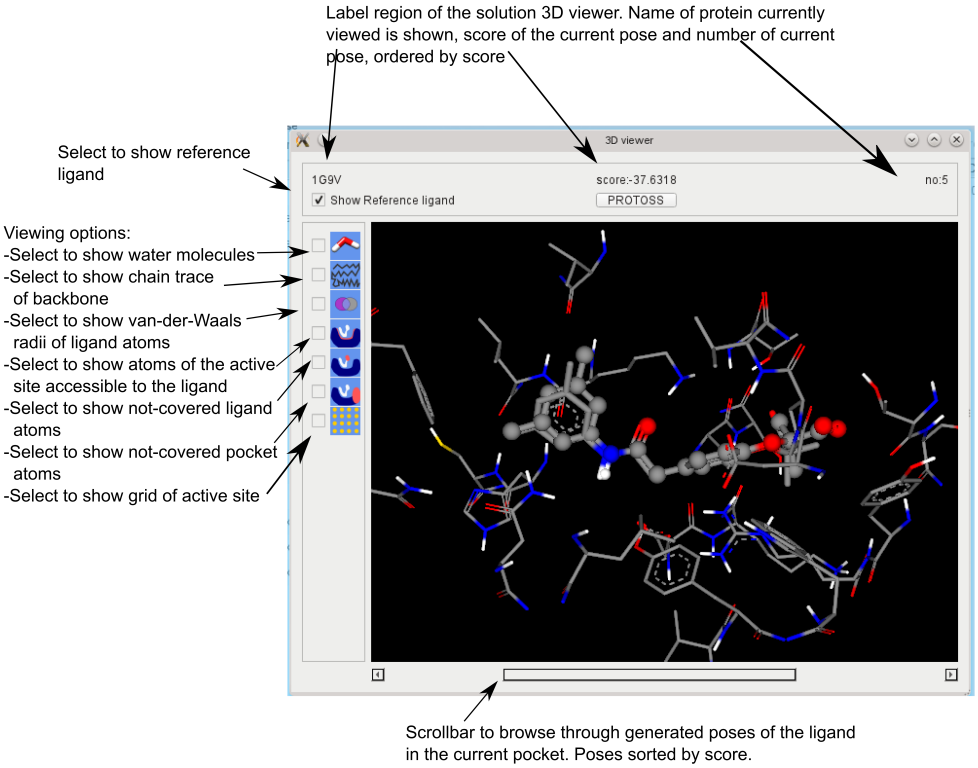


Figure D.4: Screenshot of the 3D viewer of a target solution with explanatory text.

E

Drugs/sc-PDB data set

E.1 Discarded structures from the sc-PDB

1adb, 1a46, 1aqx, 1e5o, 1esm, 1etr, 1jk3, 1k4y, 1kgi, 1m2p, 1m2q, 1mf4, 1o2w, 1od1, 1osh, 1pxo, 1zrb, 2e14, 2j3q, 2jji, 2nsd, 2qnn, 2r3y, 2uwu, 2wu3, 2x6o, 2x9d, 2xb5, 2xd6, 2xh1, 2xpw, 2xui, 2xx2, 2xx4, 2xx5, 2y6q, 2znn, 3a7d, 3ebo, 3eq0, 3fk7, 3fl5, 3fp0, 3gh8, 3h6l, 3hat, 3inw, 3inx, 3ktj, 3mhn, 3mz6, 3n3v, 3nsq, 3nzi, 3oe4, 3oe5, 3osi, 3osw, 3ozs, 3ozt, 3pba, 3pwd, 3q07, 3q9y, 3ql8, 3r2a, 3rtu, 3rzq, 3sfd, 3sin, 3sqp, 3sz1, 3tfy, 3tiy, 3tvq, 3tvs, 3ut5, 3v8p, 3vgn, 3zz2, 4aux, 4dgn, 4ef7, 4eha, 4enx, 4frs

E.2 True positive structures for the 72 ligands

1 Dorzolamide (DB00869)
1a42, 1bn1, 1bn3, 1bn4, 1bnn, 1bnq, 1bnt, 1bnu, 1bnv, 1bnw, 1cil, 1cim, 1cin, 1cnw, 1cnx, 1eou, 1g45, 1g48, 1g4o, 1g52, 1i8z, 1i90, 1i91, 1i9l, 1i9m, 1i9n, 1i9o, 1i9p, 1i9q, 1if7, 1if8, 1if9, 1kwq, 1okl, 1okm, 1okn, 1oq5, 1ttm, 1xpz, 1xq0, 1ze8, 2h15, 2hd6, 2hkk, 2hoc, 2nn7, 2nnv, 2pou, 2pow, 2q1q, 2qo8, 2wd2, 2weh, 2wej, 2x7s, 2x7t, 2x7u, 3b4f, 3bet, 3bl1, 3c7p, 3d8w, 3d9z, 3da2, 3dbu, 3dcc, 3dd8, 3eft, 3f4x, 3f7b, 3f8e, 3hku, 3hlj, 3hs4, 3iai, 3ibl, 3ibn, 3igp, 3k2f, 3k34, 3kig, 3kne, 3l14, 3lxe, 3m04, 3m14, 3m2n, 3m2x, 3m3x, 3m40, 3m5e,

3m67, 3m96, 3m98, 3mdz, 3mhc, 3mhi, 3mhl, 3mhm, 3mho, 3ml5, 3myq, 3nb5, 3ni5, 3oik, 3oil, 3oim, 3oku, 3okv, 3oy0, 3oyq, 3oys, 3p25, 3p29, 3p4v, 3p58, 3p5a, 3p5l, 3po6, 3qyk, 3r16, 3r17, 3ryj, 3ryv, 3ryx, 3ryy, 3ryz, 3rz0, 3rz1, 3rz5, 3rz7, 3rz8, 3s71, 3s72, 3s73, 3s74, 3s8x, 3s9t, 3sap, 3sax, 3sbh, 3sbi, 3t5z, 3t82, 3t84, 3t85, 3ucj, 3v5g, 3v7x, 3vbd, 4e3d, 4e3f

2 Lovastatin (DB00227)
1hw8, 1hw9, 1hwi, 1hwk, 1hwl, 1rd4, 1xdd, 1xdg, 1xuo, 2ica, 2o7n, 2q6c, 3bgl, 3bqm, 3bqn, 3cct, 3ccw, 3cd0, 3cd5, 3cd7, 3cda, 3cdb, 3e2m, 3m6f, 3mz4, 3rqd, 3sff, 3sfh

3 Estradiol (DB00783)
1d2s, 1e6w, 1fds, 1l2i, 1l2j, 1lhn, 1lho, 1lhu, 1lhv, 1lhw, 1n6a, 1qkm, 1r5k, 1sj0, 1tw4, 1u3r, 1u3s, 1uom, 1x76, 1x78, 1x7b, 1x7e, 1x7r, 1xp1, 1xp6, 1xp9, 1xpc, 1xqc, 1yim, 1yin, 1yy4, 1yye, 1zaf, 1zky, 2ayr, 2i0g, 2i0j, 2iog, 2iok, 2j7x, 2jfa, 2jj3, 2nv7, 2ouz, 2p15, 2pog, 2q70, 2qab, 2qe4, 2qgt, 2qgw, 2qr9, 2qtu, 2r6w, 2r6y, 2yat, 2yjd, 2z4b, 3akm, 3cbp, 3cv3, 3dt3, 3erd, 3ert, 3hlv, 3l03, 3m58, 3oll, 3omo, 3omp, 3omq, 3os5, 3os9, 3osa, 3uu7, 3uua, 3uuc, 3uud, 4dma, 4e47

4 Efavirenz (DB00625)
1a30, 1ajv, 1ajx, 1c0t, 1c0u, 1c1b, 1c1c, 1d4h, 1d4i, 1dmp, 1ebw, 1eby, 1ebz, 1ec0, 1ec1, 1ec2, 1eet, 1ep4, 1fk9, 1hbv, 1hih, 1hpz, 1hvh, 1hwr, 1hxb, 1ikx, 1iky, 1jla, 1jlc, 1jlq, 1klm, 1npa, 1npw, 1odw, 1odx, 1qbr, 1qbs, 1qbu, 1rev, 1rt1, 1rt3, 1rt5, 1rt6, 1rt7, 1rth, 1rti, 1s1t, 1sbg, 1t7k, 1tkt, 1tkx, 1tkz, 1tl1, 1tl3, 1tv6, 1vrt, 1vru, 1w5v, 1w5w, 1w5x, 1w5y, 2b6a, 2fde, 2ic3, 2ops, 2rf2, 2rki, 2uxz, 2uy0, 2vg5, 2vg7, 2won, 2ykm, 2zd1, 3di6, 3dle, 3dlg, 3dmj, 3dok, 3dol, 3dox, 3drp, 3dya, 3e01, 3ffi, 3gga, 3i0r, 3i0s, 3irx, 3is9, 3k4v, 3lak, 3lal, 3lam, 3lan, 3lp1, 3m8q, 3mec, 3med, 3mee, 3meg, 3n3i, 3nbp, 3nu3, 3ok9, 3psu, 3qaa, 3t1a, 3tam, 3tl9, 3tlh, 3tof, 3toh, 3v81

5 Delavirdine (DB00705)
1a30, 1ajv, 1ajx, 1c0t, 1c0u, 1c1b, 1c1c, 1d4h, 1d4i, 1dmp, 1ebw, 1eby, 1ebz, 1ec0, 1ec1, 1ec2, 1eet, 1ep4, 1fk9, 1hbv, 1hih, 1hpz, 1hvh, 1hwr, 1hxb, 1ikx, 1iky, 1jla, 1jlc, 1jlq, 1klm, 1npa, 1npw, 1odw, 1odx, 1qbr, 1qbs, 1qbu, 1rev, 1rt1, 1rt3, 1rt5, 1rt6, 1rt7, 1rth, 1rti, 1s1t, 1sbg, 1t7k, 1tkt, 1tkx, 1tkz, 1tl1, 1tl3, 1tv6, 1vrt, 1vru, 1w5v, 1w5w, 1w5x, 1w5y, 2b6a, 2fde, 2ic3, 2ops, 2rf2, 2rki, 2uxz, 2uy0, 2vg5, 2vg7, 2won, 2ykm, 2zd1, 3di6, 3dle, 3dlg, 3dmj, 3dok, 3dol, 3dox, 3drp, 3dya, 3e01, 3ffi, 3gga, 3i0r, 3i0s, 3irx, 3is9, 3k4v, 3lak, 3lal, 3lam, 3lan, 3lp1, 3m8q, 3mec, 3med, 3mee, 3meg, 3n3i, 3nbp, 3nu3, 3ok9, 3psu, 3qaa, 3t1a, 3tam, 3tl9, 3tlh, 3tof, 3toh, 3v81

6 Niflumic acid (DB04552)
1db4, 1fdk, 1ht5, 1fv0, 1jq8, 1kqu, 1kvo, 1oxl, 1pxx, 1q7a, 1skg, 1td7, 1tg4, 1tj9, 1tk4, 2arm, 2b04, 2b17, 2gns, 2oye, 2oyu, 2pws, 2qhw, 2que, 2qvd, 2wm3, 2wq5, 3cv3, 3fo7, 3g8f, 3kk6, 3l30, 3ln0, 3ln1, 3mqe, 3n8w, 3n8x, 3n8y, 3n8z, 3nju, 3nt1, 3ntb, 3ntg, 3osh, 3q7d, 3qmo, 3rr3, 4cox, 4dbk, 4fga, 4fm5, 4gld, 6cox

7 Nevirapine(DB00238)
1a30, 1ajv, 1ajx, 1c0t, 1c0u, 1c1b, 1c1c, 1d4h, 1d4i, 1dmp, 1ebw, 1eby, 1ebz, 1ec0, 1ec1, 1ec2, 1eet, 1ep4, 1fk9, 1hbv, 1hih, 1hpz, 1hvh, 1hwr, 1hxb, 1ikx, 1iky, 1jla, 1jlc, 1jlq, 1klm, 1npa, 1npw, 1odw, 1odx, 1qbr, 1qbs, 1qbu, 1rev, 1rt1, 1rt3, 1rt5, 1rt6, 1rt7, 1rth, 1rti, 1s1t, 1sbg, 1t7k, 1tkt, 1tkx, 1tkz, 1tl1, 1tl3, 1tv6, 1vrt, 1vru, 1w5v, 1w5w, 1w5x, 1w5y, 2b6a, 2fde, 2ic3, 2ops, 2rf2, 2rki, 2uxz, 2uy0, 2vg5, 2vg7, 2won, 2ykm, 2zd1, 3di6, 3dle, 3dlg, 3dmj, 3dok, 3dol, 3dox, 3drp, 3dya, 3e01, 3ffi, 3gga, 3i0r, 3i0s, 3irx, 3is9, 3k4v, 3lak, 3lal, 3lam, 3lan, 3lp1, 3m8q, 3mec, 3med, 3mee, 3meg, 3n3i, 3nbp, 3nu3, 3ok9, 3psu, 3qaa, 3t1a, 3tam, 3tl9, 3tlh, 3tof, 3toh, 3v81

8 Galantamine (DB00674)
1dx4, 1e3q, 1e66, 1eve, 1gpn, 1h22, 1h23, 1j07, 1n5r, 1odc, 1q83, 1q84, 1qon, 1u65, 1ut6, 1w4l, 1w6r, 1zgb, 1zgc, 2cmf, 2gyu, 2gyw, 2ha6, 2jez, 2ph9, 2w6c, 2wu4, 2xi4, 2xud, 2xuf, 3i6m, 3i6z, 3uon, 3zv7, 4a16, 4a23, 4b0o

9 Sitagliptin (DB01261)
1rwq, 1wcy, 1x70, 2aj8, 2bgr, 2bub, 2buc, 2hha, 2i3z, 2i78, 2iit, 2iiv, 2jid, 2oae, 2oag, 2ogz, 2ole, 2onc, 2oph, 2oqi, 2qjr, 2qoe, 2qt9, 2qtb, 2rgu, 2rip, 3ccc, 3d4l, 3eio, 3f8s, 3g0b, 3g0c, 3g0d, 3g0g, 3h0c, 3hab, 3hac, 3kwf, 3kwj, 3o95, 3o9v, 3oc0, 3opm, 3qbj, 3sww, 3sx4, 3vjm, 4a5s

10 Tadalafil (DB00820)
1ptw, 1ro6, 1ro9, 1ror, 1so2, 1t9s, 1tbb, 1tbf, 1uho, 1xlx, 1xlz, 1xm4, 1xm6, 1xmu, 1xom, 1xoq, 1xor, 1xos, 1xot, 1xoz, 1xp0, 1y2e, 1y2h, 1y2k, 2h42, 2h44, 2o8h, 2oun, 2ouq, 2our, 2ovv, 2ovy, 2pw3, 2qyn, 2y0j, 3bjc, 3d3p, 3dba, 3dy8, 3dyl, 3frg, 3g3n, 3g45, 3g4i, 3g4k, 3g58, 3gwt, 3hmv, 3hqw, 3hqy, 3hqz, 3hr1, 3i8v, 3iad, 3iak, 3ib8, 3itu, 3jsi, 3jsw, 3jwr, 3k3e, 3k3h, 3k4s, 3kkt, 3lxg, 3o56, 3o57, 3qi4, 3qpn, 3qpo, 3qpp, 3shy, 3shz, 3sie, 3sl4, 3sl5, 3sl6, 3sl8, 3sn7, 3sni, 3snl, 3tge, 3tgg, 3tvx, 3ui7, 3uuo, 3v94, 3v9b, 4ael, 4ddl, 4dff

11 Imatinib (DB00619)
1agw, 1fgi, 1fpu, 1ht5, 1m52, 1pkg, 1pxx, 1xbb, 2e2b, 2f4j, 2fgi, 2g2f, 2hen, 2hiw, 2hyy,

2hz4, 2hzi, 2hzn, 2i0v, 2i1m, 2itp, 2itx, 2ivs, 2ivt, 2ivv, 2oh4, 2oye, 2oyu, 2p0c, 2p2h, 2p2i, 2pvf, 2qoh, 2qu5, 2qu6, 2r4b, 2rfn, 2rfs, 2rgp, 2rl5, 2v7a, 2vwu, 2vwv, 2vww, 2vwx, 2vwy, 2vwz, 2vx0, 2vx1, 2wd1, 2wgj, 2wkm, 2wqb, 2x2k, 2x2l, 2x2m, 2x9f, 2xb7, 2xba, 2xir, 2xvd, 2xyu, 2y6o, 2yfx, 2zm3, 3a4p, 3aox, 3b2t, 3b8q, 3b8r, 3be2, 3bea, 3bel, 3bpr, 3brb, 3bu5, 3ce3, 3cjf, 3cjg, 3cp9, 3cpb, 3cpc, 3cth, 3ctj, 3d94, 3dk3, 3dko, 3dpk, 3dtw, 3dzq, 3efj, 3efk, 3efl, 3ekk, 3ekn, 3eta, 3ewh, 3f5p, 3f66, 3f82, 3fxx, 3g0e, 3gqi, 3gql, 3hng, 3i5n, 3js2, 3kf4, 3kfa, 3kk6, 3l8p, 3l8v, 3lcd, 3lcs, 3lct, 3lmg, 3ln0, 3ln1, 3lq8, 3lvp, 3lw0, 3lzb, 3mqe, 3ms9, 3n8w, 3n8x, 3n8y, 3n8z, 3nt1, 3ntb, 3ntg, 3nw5, 3nw6, 3o23, 3oxz, 3oy3, 3pls, 3poz, 3pp0, 3pyy, 3q6w, 3q7d, 3qmo, 3qqu, 3qrj, 3qrk, 3qti, 3qup, 3r7o, 3rgz, 3rhx, 3ri1, 3rr3, 3tcp, 3tt0, 3u6h, 3u6i, 3u6j, 3ue4, 3ug2, 3uim, 3v5q, 3vhe, 3vid, 3vjn, 3vnt, 3zxz, 3zze, 4aoj, 4at3, 4at4, 4at5, 4aw5, 4cox, 4dce, 4deg, 4deh, 4dei, 4f63, 4f64, 4f65, 4fm5, 4fny, 4fnz, 4fob, 4foc, 4fod, 6cox

12 Succhinylcholine (DB00202)
2ha6, 3uon, 4b0o

13 Apixaban (DB06605)
1ezq, 1f0r, 1f0s, 1fjs, 1g2l, 1ioe, 1iqg, 1iqh, 1iqi, 1iqj, 1iql, 1iqm, 1iqn, 1ksn, 1lpg, 1lqd, 1mq5, 1mq6, 1nfu, 1nfw, 1nfx, 1nfy, 1v3x, 1xka, 1z6e, 2bmg, 2boh, 2bok, 2bqw, 2ei6, 2ei7, 2ei8, 2j2u, 2j34, 2j38, 2j4i, 2j94, 2j95, 2jkh, 2p16, 2p3t, 2p3u, 2p93, 2p94, 2p95, 2phb, 2pr3, 2q1j, 2ra0, 2uwl, 2uwo, 2uwp, 2vh0, 2vh6, 2w3i, 2w3k, 2wyg, 2wyj, 2xbv, 2xbw, 2xbx, 2xc0, 2xc4, 2xc5, 2y5f, 2y5g, 2y5h, 2y7x, 2y7z, 2y80, 2y81, 2y82, 3cen, 3cs7, 3ffg, 3kl6, 3kqb, 3kqc, 3kqd, 3kqe, 3m36, 3m37, 3q3k, 3sw2, 3tk5, 3tk6, 4a7i

14 Cocaine (DB00907)
2pgz, 3uon

15 Dichlorphenamide (DB01144)
1a42, 1bn1, 1bn3, 1bn4, 1bnn, 1bnq, 1bnt, 1bnu, 1bnv, 1bnw, 1cil, 1cim, 1cin, 1cnw, 1cnx, 1eou, 1g45, 1g48, 1g4o, 1g52, 1i8z, 1i90, 1i91, 1i9l, 1i9m, 1i9n, 1i9o, 1i9p, 1i9q, 1if7, 1if8, 1if9, 1kwq, 1okl, 1okm, 1okn, 1oq5, 1ttm, 1xpz, 1xq0, 1ze8, 2h15, 2hd6, 2hkk, 2hoc, 2nn7, 2nnv, 2pou, 2pow, 2q1q, 2qo8, 2wd2, 2weh, 2wej, 2x7s, 2x7t, 2x7u, 3b4f, 3bet, 3bl1, 3c7p, 3d8w, 3d9z, 3da2, 3dbu, 3dcc, 3dd8, 3eft, 3f4x, 3f7b, 3f8e, 3hku, 3hlj, 3hs4, 3iai, 3ibl, 3ibn, 3igp, 3k2f, 3k34, 3kig, 3kne, 3l14, 3lxe, 3m04, 3m14, 3m2n, 3m2x, 3m3x, 3m40, 3m5e, 3m67, 3m96, 3m98, 3mdz, 3mhc, 3mhi, 3mhl, 3mhm, 3mho, 3ml5, 3myq, 3nb5, 3ni5, 3oik, 3oil, 3oim, 3oku, 3okv, 3oy0, 3oyq, 3oys, 3p25, 3p29, 3p4v, 3p58, 3p5a, 3p5l, 3po6, 3qyk, 3r16, 3r17, 3ryj, 3ryv, 3ryx, 3ryy, 3ryz, 3rz0, 3rz1, 3rz5, 3rz7, 3rz8, 3s71, 3s72, 3s73, 3s74,

3s8x, 3s9t, 3sap, 3sax, 3sbh, 3sbi, 3t5u, 3t5z, 3t82, 3t84, 3t85, 3ucj, 3v5g, 3v7x, 3vbd, 4e3d, 4e3f

16 Proflavine (DB01123)
1a2c, 1a4w, 1a5g, 1a61, 1ae8, 1afe, 1b5g, 1bhx, 1c4u, 1c4v, 1c5n, 1ca8, 1d3d, 1d3p, 1d3t, 1d4p, 1d6w, 1d9i, 1eb1, 1ets, 1ett, 1fpc, 1g30, 1g32, 1ghv, 1ghx, 1gj4, 1gj5, 1h8d, 1jwt, 1k21, 1k22, 1kts, 1ktt, 1mu6, 1mu8, 1mue, 1nm6, 1nrs, 1nt1, 1ny2, 1nzq, 1o0d, 1o2g, 1o5g, 1oyt, 1riw, 1rkw, 1rpw, 1sb1, 1sl3, 1t4u, 1t4v, 1ta2, 1ta6, 1tom, 1uvt, 1vzq, 1way, 1ype, 1ypg, 1ypj, 1ypk, 1ypl, 1ypm, 1z71, 1zgi, 1zgv, 2fes, 2g0e, 2gby, 2hgt, 2jh0, 2jh5, 2jh6, 2r2m, 2uuj, 2uuk, 2v3h, 2v3o, 2v57, 2zc9, 2zda, 2zdv, 2zf0, 2zff, 2zfp, 2zfq, 2zfr, 2zg0, 2zgb, 2zgx, 2zhe, 2zhf, 2zhw, 2zi2, 2ziq, 2znk, 2zo3, 3biu, 3bqz, 3br2, 3bti, 3bv9, 3c27, 3da9, 3dhk, 3egk, 3f68, 3hth, 3ldx, 3p17, 3p70, 3pm1, 3po1, 3qdz, 3qto, 3qtv, 3qwc, 3qx5, 3rlw, 3rly, 3rm0, 3rm2, 3rml, 3rmm, 3rmn, 3rmo, 3sha, 3shc, 3si3, 3si4, 3sv2, 3t5f, 3tu7, 3utu, 4ax9, 4ayv, 4ayy, 4az2, 4e7r

17 Phenylbutazone (DB00812)
1ht5, 1pxx, 2oye, 2oyu, 3b99, 3kk6, 3ln0, 3ln1, 3mqe, 3n8w, 3n8x, 3n8y, 3n8z, 3nt1, 3ntb, 3ntg, 3q7d, 3qmo, 3rr3, 4cox, 4fm5, 6cox

18 Minocycline (DB01017)
1gkc, 1gkd, 2ovx, 2ovz, 2ow0, 2ow1, 2ow2

19 Indomethacine (DB00328)
1bh5, 1db4, 1fdk, 1fro, 1hk1, 1hk3, 1ht5, 1i7i, 1jq8, 1k7l, 1kkq, 1knu, 1kqu, 1kvo, 1nyx, 1oxl, 1pxx, 1q7a, 1skg, 1td7, 1tg4, 1tj9, 1tk4, 1v3t, 1zeo, 1zgy, 2arm, 2b04, 2b17, 2bx8, 2bxf, 2g0g, 2gns, 2h7c, 2hwq, 2hwr, 2i4j, 2i4p, 2jez, 2npa, 2om9, 2oye, 2oyu, 2p4y, 2p54, 2pob, 2pws, 2q5p, 2q61, 2q6s, 2q8s, 2qhw, 2que, 2qvd, 2rew, 2vna, 2vue, 2w4q, 2w98, 2wq5, 2xvq, 2xvu, 2y05, 2yfe, 2za0, 2zb3, 2zb4, 2zb7, 2zb8, 2zno, 2zvt, 3a73, 3adt, 3adv, 3adw, 3adx, 3an3, 3an4, 3b0q, 3b1m, 3b3k, 3b9l, 3cdp, 3cds, 3cs8, 3cv3, 3cwd, 3d6d, 3et1, 3et3, 3fei, 3fej, 3fo7, 3fur, 3g8f, 3g8i, 3g9e, 3gbk, 3h0a, 3ho0, 3hod, 3ia6, 3k8s, 3kdt, 3kdu, 3kk6, 3kmg, 3l30, 3lmp, 3ln0, 3ln1, 3lu7, 3lu8, 3mqe, 3n8w, 3n8x, 3n8y, 3n8z, 3nju, 3noa, 3nt1, 3ntb, 3ntg, 3ogw, 3osh, 3q7d, 3qmo, 3qt0, 3r5n, 3r8a, 3r8i, 3rr3, 3s9s, 3sp6, 3t03, 3tdl, 3ty0, 3v9t, 3v9v, 3v9y, 3vi8, 3vjh, 3vji, 3vn2, 4cox, 4dbk, 4e99, 4f9m, 4fga, 4fm5, 4gld, 4prg, 6cox

20 Pentoxifylline (DB00806)
1ptw, 1q91, 1ro6, 1ro9, 1ror, 1so2, 1t9s, 1tbb, 1tbf, 1uho, 1xlx, 1xlz, 1xm4, 1xm6, 1xmu,

1xom, 1xoq, 1xor, 1xos, 1xot, 1xoz, 1xp0, 1y2e, 1y2h, 1y2k, 2h42, 2h44, 2o8h, 2oun, 2ouq, 2our, 2ovv, 2ovy, 2pw3, 2qyn, 2y0j, 2ydo, 2ydv, 3arr, 3bjc, 3d3p, 3dba, 3dy8, 3dyl, 3eml, 3frg, 3g3n, 3g45, 3g4i, 3g4k, 3g58, 3gwt, 3hmv, 3hqw, 3hqy, 3hqz, 3hr1, 3i8v, 3iad, 3iak, 3ib8, 3itu, 3jsi, 3jsw, 3jwr, 3k3e, 3k3h, 3k4s, 3kkt, 3lxg, 3o56, 3o57, 3qak, 3qi4, 3qpn, 3qpo, 3qpp, 3shy, 3shz, 3sie, 3sl4, 3sl5, 3sl6, 3sl8, 3sn7, 3sni, 3snl, 3tge, 3tgg, 3tvx, 3ui7, 3uuo, 3v94, 3v9b, 4ael, 4ddl, 4dff, 4fe3

21 Chlormaphenicol (DB00446)
1c1y, 1csn, 1di8, 1di9, 1dm2, 1eh4, 1g5s, 1grq, 1grr, 1gua, 1h1r, 1h1s, 1jnk, 1kv1, 1kv2, 1m7q, 1mru, 1p2a, 1p5e, 1pkd, 1pme, 1pmn, 1pmu, 1pxl, 1pxm, 1pxn, 1pye, 1q23, 1q97, 1q99, 1qca, 1r78, 1unl, 1v0o, 1xjd, 1zrz, 2a19, 2a2d, 2cch, 2csn, 2p0e, 2qt1, 2uxp, 3cla, 3cr3, 3erk, 3gbu, 3ih0, 3lij, 3u9f, 4erk

22 Finasteride (DB01216)
3bur, 3buv, 3caq, 3cas, 3cav, 3g1r, 3uzw, 3uzx, 3uzy

23 Topiramate (DB00273)
1a42, 1bn1, 1bn3, 1bn4, 1bnn, 1bnq, 1bnt, 1bnu, 1bnv, 1bnw, 1cil, 1cim, 1cin, 1cnw, 1cnx, 1eou, 1g45, 1g48, 1g4o, 1g52, 1i8z, 1i90, 1i91, 1i9l, 1i9m, 1i9n, 1i9o, 1i9p, 1i9q, 1if7, 1if8, 1if9, 1kwq, 1okl, 1okm, 1okn, 1oq5, 1ttm, 1xpz, 1xq0, 1ze8, 2h15, 2hd6, 2hkk, 2hoc, 2nn7, 2nnv, 2pou, 2pow, 2q1q, 2qo8, 2wd2, 2weh, 2wej, 2x7s, 2x7t, 2x7u, 3b4f, 3bet, 3bl1, 3c7p, 3d8w, 3d9z, 3da2, 3dbu, 3dcc, 3dd8, 3eft, 3f4x, 3f7b, 3f8e, 3hku, 3hlj, 3hs4, 3iai, 3ibl, 3ibn, 3igp, 3k2f, 3k34, 3kig, 3kne, 3l14, 3lxe, 3m04, 3m14, 3m2n, 3m2x, 3m3x, 3m40, 3m5e, 3m67, 3m96, 3m98, 3mdz, 3mhc, 3mhi, 3mhl, 3mhm, 3mho, 3ml5, 3myq, 3nb5, 3ni5, 3oik, 3oil, 3oim, 3oku, 3okv, 3oy0, 3oyq, 3oys, 3p25, 3p29, 3p4v, 3p58, 3p5a, 3p5l, 3po6, 3qyk, 3r16, 3r17, 3ryj, 3ryv, 3ryx, 3ryy, 3ryz, 3rz0, 3rz1, 3rz5, 3rz7, 3rz8, 3s71, 3s72, 3s73, 3s74, 3s8x, 3s9t, 3sap, 3sax, 3sbh, 3sbi, 3t5z, 3t82, 3t84, 3t85, 3ucj, 3v5g, 3v7x, 3vbd, 4e3d, 4e3f

24 Papaverine (DB01113)
1ptw, 1ro6, 1ro9, 1ror, 1so2, 1t9s, 1tbb, 1tbf, 1uho, 1xlx, 1xlz, 1xm4, 1xm6, 1xmu, 1xom, 1xoq, 1xor, 1xos, 1xot, 1xoz, 1xp0, 1y2e, 1y2h, 1y2k, 2h42, 2h44, 2o8h, 2oun, 2ouq, 2our, 2ovv, 2ovy, 2pw3, 2qyn, 2y0j, 3bjc, 3d3p, 3dba, 3dy8, 3dyl, 3frg, 3g3n, 3g45, 3g4i, 3g4k, 3g58, 3gwt, 3hmv, 3hqw, 3hqy, 3hqz, 3hr1, 3i8v, 3iad, 3iak, 3ib8, 3itu, 3jsi, 3jsw, 3jwr, 3k3e, 3k3h, 3k4s, 3kkt, 3lxg, 3o56, 3o57, 3qi4, 3qpn, 3qpo, 3qpp, 3shy, 3shz, 3sie, 3sl4, 3sl5, 3sl6, 3sl8, 3sn7, 3sni, 3snl, 3tge, 3tgg, 3tvx, 3ui7, 3uuo, 3v94, 3v9b, 4ael, 4ddl, 4dff

25 Balsalazide (DB01014)
1ht5, 1i7i, 1k7l, 1kkq, 1knu, 1nyx, 1pxx, 1zeo, 1zgy, 2g0g, 2hwq, 2hwr, 2i4j, 2i4p, 2npa, 2om9, 2oye, 2oyu, 2p4y, 2p54, 2pob, 2q5p, 2q61, 2q6s, 2q8s, 2rew, 2yfe, 2z98, 2zno, 2zvt, 3adt, 3adv, 3adw, 3adx, 3an3, 3an4, 3b0q, 3b1m, 3b3k, 3cdp, 3cds, 3cs8, 3cwd, 3d6d, 3et1, 3et3, 3fei, 3fej, 3fur, 3g8i, 3g9e, 3gbk, 3h0a, 3ho0, 3hod, 3ia6, 3k8s, 3kdt, 3kdu, 3kk6, 3kmg, 3lmp, 3ln0, 3ln1, 3lt5, 3mqe, 3n8w, 3n8x, 3n8y, 3n8z, 3noa, 3nt1, 3ntb, 3ntg, 3q7d, 3qmo, 3qt0, 3r5n, 3r8a, 3r8i, 3rr3, 3s9s, 3sp6, 3t03, 3ty0, 3v9t, 3v9v, 3v9y, 3vi8, 3vjh, 3vji, 3vn2, 4cox, 4f9m, 4fm5, 4prg, 6cox

26 Ethoxzolamide (DB00311)
1a42, 1bn1, 1bn3, 1bn4, 1bnn, 1bnq, 1bnt, 1bnu, 1bnv, 1bnw, 1cil, 1cim, 1cin, 1cnw, 1cnx, 1eou, 1g45, 1g48, 1g4o, 1g52, 1i8z, 1i90, 1i91, 1i9l, 1i9m, 1i9n, 1i9o, 1i9p, 1i9q, 1if7, 1if8, 1if9, 1kwq, 1okl, 1okm, 1okn, 1oq5, 1ttm, 1xpz, 1xq0, 1ze8, 2h15, 2hd6, 2hkk, 2hoc, 2nn7, 2nnv, 2pou, 2pow, 2q1q, 2qo8, 2wd2, 2weh, 2wej, 2x7s, 2x7t, 2x7u, 3b4f, 3bet, 3bl1, 3c7p, 3d8w, 3d9z, 3da2, 3dbu, 3dcc, 3dd8, 3eft, 3f4x, 3f7b, 3f8e, 3hku, 3hlj, 3hs4, 3iai, 3ibl, 3ibn, 3igp, 3k2f, 3k34, 3kig, 3kne, 3l14, 3lxe, 3m04, 3m14, 3m2n, 3m2x, 3m3x, 3m40, 3m5e, 3m67, 3m96, 3m98, 3mdz, 3mhc, 3mhi, 3mhl, 3mhm, 3mho, 3ml5, 3myq, 3nb5, 3ni5, 3oik, 3oil, 3oim, 3oku, 3okv, 3oy0, 3oyq, 3oys, 3p25, 3p29, 3p4v, 3p58, 3p5a, 3p5l, 3po6, 3qyk, 3r16, 3r17, 3ryj, 3ryv, 3ryx, 3ryy, 3ryz, 3rz0, 3rz1, 3rz5, 3rz7, 3rz8, 3s71, 3s72, 3s73, 3s74, 3s8x, 3s9t, 3sap, 3sax, 3sbh, 3sbi, 3t5z, 3t82, 3t84, 3t85, 3ucj, 3v5g, 3v7x, 3vbd, 4e3d, 4e3f

27 Gefitinib (DB00317)
1xkk, 2hen, 2i0v, 2i1m, 2itp, 2itx, 2ivs, 2ivt, 2ivv, 2oh4, 2p0c, 2p2h, 2p2i, 2pvf, 2qu5, 2qu6, 2r4b, 2rfn, 2rfs, 2rgp, 2rl5, 2vwu, 2vwv, 2vww, 2vwx, 2vwy, 2vwz, 2vx0, 2vx1, 2wd1, 2wgj, 2wkm, 2wqb, 2x2k, 2x2l, 2x2m, 2x9f, 2xb7, 2xba, 2xir, 2xvd, 2xyu, 2y6o, 2yfx, 2zm3, 3a4p, 3aox, 3b2t, 3b8q, 3b8r, 3be2, 3bea, 3bel, 3bpr, 3brb, 3bu5, 3ce3, 3cjf, 3cjg, 3cp9, 3cpb, 3cpc, 3cth, 3ctj, 3d94, 3dko, 3dpk, 3dtw, 3dzq, 3efj, 3efk, 3efl, 3ekk, 3ekn, 3eta, 3ewh, 3f5p, 3f66, 3f82, 3fxx, 3g0e, 3gql, 3hng, 3i5n, 3js2, 3l8p, 3l8v, 3lcd, 3lcs, 3lct, 3lmg, 3lq8, 3lvp, 3lw0, 3lzb, 3nw5, 3nw6, 3o23, 3pls, 3poz, 3pp0, 3q6w, 3qqu, 3qti, 3qup, 3r7o, 3rgz, 3rhx, 3ri1, 3tcp, 3tt0, 3u6h, 3u6i, 3u6j, 3ug2, 3uim, 3v5q, 3vhe, 3vid, 3vjn, 3vnt, 3zxz, 3zze, 4aoj, 4at3, 4at4, 4at5, 4aw5, 4dce, 4deg, 4deh, 4dei, 4f63, 4f64, 4f65, 4fny, 4fnz, 4fob, 4foc, 4fod

28 Prazosin (DB00457)
2rh1, 3d4s, 3ny8, 3ny9, 3owx

29 Pyrimethamine (DB00205)
1dr2, 1dr7, 1dyh, 1dyj, 1e26, 1hfr, 1jom, 1ly3, 1ly4, 1mvt, 1ohk, 1s3v, 1s3w, 1s3y, 1u70, 1u71, 2ano, 2bla, 2cd2, 2dhf, 2oip, 2qk8, 2w3a, 2w3b, 2w3m, 2w3v, 2w3w, 2w9g, 2w9s, 2zza, 3cd2, 3clb, 3cse, 3d80, 3d84, 3dau, 3dga, 3e0b, 3eej, 3eek, 3eel, 3eem, 3eig, 3f0b, 3f0q, 3f0s, 3f0u, 3f0v, 3f0x, 3fl8, 3fl9, 3fq0, 3fqc, 3fqf, 3fqo, 3fqv, 3fqz, 3fra, 3frb, 3frd, 3frf, 3fy8, 3fy9, 3fyv, 3fyw, 3ghv, 3ghw, 3gi2, 3hbb, 3i8a, 3ia4, 3inv, 3irm, 3iro, 3ix9, 3jsu, 3jvx, 3jwf, 3jwk, 3k45, 3k47, 3kfy, 3kjs, 3m09, 3ntz, 3nu0, 3nxo, 3nxt, 3nxv, 3nxx, 3oaf, 3qfx, 3qg2, 3qgt, 3ql0, 3ql3, 3qlx, 3qly, 3qlz, 3r33, 3rg9, 3ro9, 3roa, 3s3v, 3s7a, 3s9u, 3sa1, 3sa2, 3sai, 3sgy, 3sh2, 3sqy, 3sr5, 3srr, 3srs, 3sru, 3srw, 3td8, 3tq8, 3tq9, 3tqb, 3um5, 3um6, 3um8, 4dfr

30 Tolmetin (DB00500)
1ht5, 1pxx, 2oye, 2oyu, 3e08, 3kk6, 3ln0, 3ln1, 3mqe, 3n8w, 3n8x, 3n8y, 3n8z, 3nt1, 3ntb, 3ntg, 3ogw, 3q7d, 3ql6, 3qmo, 3rr3, 3s3g, 4cox, 4fm5, 6cox

31 Thiabendazole (DB00730)
3sfe

32 Ethacrynic acid (DB00903)
1hk1, 1hk3, 2bx8, 2bxf, 2vct, 2vue, 2xvq, 2xvu, 3a73, 3b9l, 3lu7, 3lu8, 3n9j, 3tdl, 4e99

33 Abacavir (DB01048)
1a30, 1adb, 1ajv, 1ajx, 1b14, 1b15, 1b16, 1c0t, 1c0u, 1c1b, 1c1c, 1cdo, 1d1t, 1d4h, 1d4i, 1dmp, 1e3e, 1ebw, 1eby, 1ebz, 1ec0, 1ec1, 1ec2, 1eet, 1ep4, 1fk9, 1hbv, 1hdz, 1hih, 1hpz, 1hso, 1hsz, 1hvh, 1hwr, 1hxb, 1ikx, 1iky, 1jla, 1jlc, 1jlq, 1klm, 1mc5, 1mg5, 1mp0, 1npa, 1npw, 1odw, 1odx, 1qbr, 1qbs, 1qbu, 1rev, 1rt1, 1rt3, 1rt5, 1rt6, 1rt7, 1rth, 1rti, 1s1t, 1sbg, 1t7k, 1tkt, 1tkx, 1tkz, 1tl1, 1tl3, 1tv6, 1u3w, 1vrt, 1vru, 1w5v, 1w5w, 1w5x, 1w5y, 2b6a, 2eer, 2fde, 2ic3, 2jhf, 2ops, 2rf2, 2rki, 2uxz, 2uy0, 2vg5, 2vg7, 2won, 2xaa, 2ykm, 2zd1, 3cv3, 3di6, 3dle, 3dlg, 3dmj, 3dok, 3dol, 3dox, 3drp, 3dya, 3e01, 3ffi, 3gga, 3i0r, 3i0s, 3irx, 3is9, 3k4v, 3lak, 3lal, 3lam, 3lan, 3lp1, 3m8q, 3mec, 3med, 3mee, 3meg, 3n3i, 3nbp, 3nu3, 3ok9, 3oq6, 3ox4, 3psu, 3qaa, 3qj5, 3rj5, 3t1a, 3tam, 3tl9, 3tlh, 3tof, 3toh, 3v81, 3vrj

34 Varenicline (DB01273)
4afg

35 Captopril (DB01197)
1gkc, 1gkd, 1xuc, 1xud, 1xur, 1you, 2oc2, 2ovx, 2ovz, 2ow0, 2ow1, 2ow2, 2ozr, 2xhm, 2yig, 3bkk, 3bkl, 3elm, 3i7g, 3i7i, 3kec, 3kej, 3kek, 3kry, 3lus, 3o2x, 3tvc, 456c, 4a7b, 4exs, 830c

36 Celecoxib (DB00482)
1h1w, 1hk1, 1hk3, 1ht5, 1oky, 1oq5, 1pxx, 1uu3, 1uu7, 1uu8, 1uu9, 1z5m, 2bx8, 2bxf, 2oye, 2oyu, 2pe1, 2pe2, 2r7b, 2vue, 2xch, 2xck, 2xvq, 2xvu, 3a73, 3b9l, 3h9o, 3hrc, 3ion, 3iop, 3kk6, 3ln0, 3ln1, 3lu7, 3lu8, 3mqe, 3n8w, 3n8x, 3n8y, 3n8z, 3nt1, 3ntb, 3ntg, 3nun, 3nuy, 3orz, 3q7d, 3qc4, 3qcq, 3qcs, 3qcx, 3qcy, 3qd0, 3qd3, 3qd4, 3qmo, 3rr3, 3rwp, 3rwq, 3sc1, 3tdl, 4cox, 4e99, 4fm5, 6cox

37 Levonorgestrel (DB00367)
1a28, 1e3k, 1gs4, 1i38, 1l2i, 1l2j, 1lhv, 1n6a, 1qkm, 1r5k, 1sj0, 1sqn, 1sr7, 1t65, 1u3r, 1u3s, 1uom, 1x76, 1x78, 1x7b, 1x7e, 1x7r, 1xnn, 1xow, 1xp1, 1xp6, 1xp9, 1xpc, 1xqc, 1yim, 1yin, 1yy4, 1yye, 1z95, 1zaf, 1zky, 2ax6, 2ax8, 2ax9, 2ayr, 2hvc, 2i0g, 2i0j, 2ihq, 2iog, 2iok, 2jfa, 2jj3, 2nv7, 2nw4, 2ouz, 2ovh, 2oz7, 2p15, 2pnu, 2pog, 2q70, 2q7i, 2q7k, 2qab, 2qe4, 2qgt, 2qgw, 2qr9, 2qtu, 2r6w, 2r6y, 2w8y, 2yat, 2yjd, 2z4b, 3b5r, 3b65, 3b66, 3b67, 3b68, 3cbp, 3d90, 3dt3, 3erd, 3ert, 3g0w, 3g8o, 3hlv, 3kba, 3l03, 3l3x, 3m58, 3oll, 3omo, 3omp, 3omq, 3os5, 3os9, 3osa, 3rlj, 3rll, 3uu7, 3uua, 3uuc, 3uud, 3zr7, 3zra, 4a2j, 4apu, 4dma, 4e47

38 Zonisamide (DB00909)
1a42, 1bn1, 1bn3, 1bn4, 1bnn, 1bnq, 1bnt, 1bnu, 1bnv, 1bnw, 1cil, 1cim, 1cin, 1cnw, 1cnx, 1eou, 1g45, 1g48, 1g4o, 1g52, 1i8z, 1i90, 1i91, 1i9l, 1i9m, 1i9n, 1i9o, 1i9p, 1i9q, 1if7, 1if8, 1if9, 1kwq, 1okl, 1okm, 1okn, 1oq5, 1ttm, 1xpz, 1xq0, 1ze8, 2bxr, 2h15, 2hd6, 2hkk, 2hoc, 2nn7, 2nnv, 2pou, 2pow, 2q1q, 2qo8, 2v5z, 2v60, 2v61, 2vvl, 2wd2, 2weh, 2wej, 2x7s, 2x7t, 2x7u, 2xfn, 3b4f, 3bet, 3bl1, 3c7p, 3d8w, 3d9z, 3da2, 3dbu, 3dcc, 3dd8, 3eft, 3f4x, 3f7b, 3f8e, 3hku, 3hlj, 3hs4, 3iai, 3ibl, 3ibn, 3igp, 3k2f, 3k34, 3kig, 3kne, 3l14, 3lxe, 3m04, 3m14, 3m2n, 3m2x, 3m3x, 3m40, 3m5e, 3m67, 3m96, 3m98, 3mdz, 3mhc, 3mhi, 3mhl, 3mhm, 3mho, 3ml5, 3myq, 3nb5, 3ni5, 3oik, 3oil, 3oim, 3oku, 3okv, 3oy0, 3oyq, 3oys, 3p25, 3p29, 3p4v, 3p58, 3p5a, 3p5l, 3po6, 3po7, 3qyk, 3r16, 3r17, 3ryj, 3ryv, 3ryx, 3ryy, 3ryz, 3rz0, 3rz1, 3rz5, 3rz7, 3rz8, 3s71, 3s72, 3s73, 3s74, 3s8x, 3s9t, 3sap, 3sax, 3sbh, 3sbi, 3t5z, 3t82, 3t84, 3t85, 3ucj, 3v5g, 3v7x, 3vbd, 4a79, 4a7a, 4e3d, 4e3f

39 Ampicillin (DB00415)
1hk1, 1hk3, 1ikg, 1nx9, 1pw1, 2bx8, 2bxf, 2vue, 2xvq, 2xvu, 3a73, 3b9l, 3lu7, 3lu8, 3ndv, 3tdl, 4e99

40 Penicillin V (DB00417)
1a8t, 1dd6, 1gm7, 1gm9, 1ikg, 1jje, 1jjt, 1kr3, 1l2s, 1pw1, 1pzo, 1pzp, 1xgi, 2aio, 2doo, 2r9x, 2wzz, 2z71, 3g2z, 3gqz, 3iof, 3iog, 3ly4, 3m8t, 3pag, 3q6x, 3sh7, 3sh8, 4ddy, 4de0, 4de1, 4de3, 4exs, 4ey2, 4eyb, 4eyf

41 Meticillin (DB01603)
1ikg, 1pw1, 3kp4,

42 Penicillin G (DB01053)
1gm7, 1ikg, 1pw1

43 Progesterone (DB00396)
1a28, 1e3k, 1l2i, 1l2j, 1mrq, 1n6a, 1qkm, 1r5k, 1sj0, 1sqn, 1sr7, 1u3r, 1u3s, 1uom, 1x76, 1x78, 1x7b, 1x7e, 1x7r, 1xp1, 1xp6, 1xp9, 1xpc, 1xqc, 1y9r, 1yim, 1yin, 1yy4, 1yye, 1zaf, 1zky, 2a3i, 2aa6, 2aax, 2ab2, 2aba, 2ayr, 2hzq, 2i0g, 2i0j, 2iog, 2iok, 2jfa, 2jj3, 2nv7, 2ouz, 2ovh, 2p15, 2pog, 2q70, 2qab, 2qe4, 2qgt, 2qgw, 2qr9, 2qtu, 2r6w, 2r6y, 2w8y, 2yat, 2yjd, 2z4b, 3cbp, 3d90, 3dt3, 3erd, 3ert, 3g8o, 3hlv, 3kba, 3l03, 3m58, 3oll, 3omo, 3omp, 3omq, 3os5, 3os9, 3osa, 3uu7, 3uua, 3uuc, 3uud, 3vhv, 3zr7, 3zra, 4a2j, 4apu, 4dma, 4e47

44 Testosterone (DB00624)
1afs, 1gs4, 1hk1, 1hk3, 1i38, 1j96, 1jtv, 1q13, 1t65, 1xnn, 1xow, 1z95, 2ax6, 2ax8, 2ax9, 2bx8, 2bxf, 2bxr, 2hvc, 2ihq, 2ipf, 2ipj, 2nw4, 2oz7, 2pnu, 2q7i, 2q7k, 2v5z, 2v60, 2v61, 2vue, 2vvl, 2xfn, 2xvq, 2xvu, 3a73, 3b5r, 3b65, 3b66, 3b67, 3b68, 3b9l, 3bur, 3g0w, 3l3x, 3lu7, 3lu8, 3po7, 3rlj, 3rll, 3tdl, 4a79, 4a7a, 4e99

45 Spironolactone (DB00421)
1gs4, 1i38, 1t65, 1xnn, 1xow, 1y9r, 1z95, 2a3i, 2aa6, 2aax, 2ab2, 2ax6, 2ax8, 2ax9, 2hvc, 2ihq, 2nw4, 2oz7, 2pnu, 2q7i, 2q7k, 3b5r, 3b65, 3b66, 3b67, 3b68, 3g0w, 3l3x, 3rlj, 3rll, 3vhv

46 Diazepam (DB00829)
1hk1, 1hk3, 2bx8, 2bxf, 2vue, 2xvq, 2xvu, 3a73, 3b9l, 3lu7, 3lu8, 3tdl, 4e99

47 Midazolam (DB00683)
3cv3, 3u5k

48 Diclofenac (DB00586)
1db4, 1fdk, 1hk1, 1hk3, 1ht5, 1ict, 1ie4, 1jq8, 1kgj, 1kqu, 1kvo, 1oxl, 1pxx, 1q7a, 1skg, 1sn0, 1td7, 1tg4, 1tj9, 1tk4, 1tz8, 2arm, 2b04, 2b17, 2bx8, 2bxf, 2gns, 2oye, 2oyu, 2pws, 2qhw, 2que, 2qvd, 2vue, 2wek, 2wq5, 2xvq, 2xvu, 3a73, 3b9l, 3cv3, 3fo7, 3g8f, 3ib0, 3kk6, 3l30, 3ln0, 3ln1, 3lu7, 3lu8, 3mqe, 3n8w, 3n8x, 3n8y, 3n8z, 3nju, 3nt1, 3ntb, 3ntg, 3osh, 3q7d, 3qmo, 3rr3, 3tdl, 4cox, 4dbk, 4e99, 4fga, 4fm5, 4gld, 6cox

49 Mefenamic acid (DB00784)
1ht5, 1pxx, 2oye, 2oyu, 2xn3, 3kk6, 3ln0, 3ln1, 3mqe, 3n8w, 3n8x, 3n8y, 3n8z, 3nt1, 3ntb, 3ntg, 3q7d, 3qmo, 3r43, 3rr3, 4cox, 4fm5, 6cox

50 Meclofenamic acid (DB00939)
1ht5, 1pxx, 2oye, 2oyu, 2xn3, 3kk6, 3ln0, 3ln1, 3mqe, 3n8w, 3n8x, 3n8y, 3n8z, 3nt1, 3ntb, 3ntg, 3q7d, 3qmo, 3rr3, 4cox, 4fm5, 6cox

51 Methotrexate (DB00563)
1an5, 1axw, 1dr2, 1dr7, 1dyh, 1dyj, 1e26, 1f4e, 1f4f, 1f4g, 1hfr, 1jg0, 1jom, 1jtq, 1juj, 1jut, 1lce, 1ly3, 1ly4, 1mvt, 1ohk, 1p33, 1s3v, 1s3w, 1s3y, 1syn, 1tsd, 1tsn, 1u70, 1u71, 1vzc, 2ano, 2bla, 2cd2, 2dhf, 2oip, 2qk8, 2tsr, 2vf0, 2w3a, 2w3b, 2w3m, 2w3v, 2w3w, 2w9g, 2w9s, 2zza, 3apy, 3b5b, 3b9h, 3bgx, 3bhl, 3bnz, 3byx, 3c06, 3cd2, 3clb, 3cse, 3d80, 3d84, 3dau, 3dga, 3e0b, 3eej, 3eek, 3eel, 3eem, 3eig, 3f0b, 3f0q, 3f0s, 3f0u, 3f0v, 3f0x, 3fl8, 3fl9, 3fq0, 3fqc, 3fqf, 3fqo, 3fqv, 3fqz, 3fra, 3frb, 3frd, 3frf, 3fsu, 3fy8, 3fy9, 3fyv, 3fyw, 3ghv, 3ghw, 3gi2, 3hbb, 3i8a, 3ia4, 3ijz, 3ik0, 3ik1, 3inv, 3irm, 3iro, 3ix9, 3jsu, 3jvx, 3jwf, 3jwk, 3k2h, 3k45, 3k47, 3kfy, 3kjs, 3m09, 3n2a, 3nrr, 3ntz, 3nu0, 3nxo, 3nxt, 3nxv, 3nxx, 3oaf, 3ob7, 3pyz, 3qfx, 3qg2, 3qgt, 3qj7, 3ql0, 3ql3, 3qlx, 3qly, 3qlz, 3r33, 3rg9, 3ro9, 3roa, 3s3v, 3s7a, 3s9u, 3sa1, 3sa2, 3sai, 3sgy, 3sh2, 3sqy, 3sr5, 3srr, 3srs, 3sru, 3srw, 3td8, 3tq8, 3tq9, 3tqb, 3um5, 3um6, 3um8, 4dfr, 4dq1, 4e5o, 4eb4, 4f2v, 4gev

52 Raltitrexed (DB00293)
1an5, 1axw, 1f4e, 1f4f, 1f4g, 1jg0, 1jtq, 1juj, 1jut, 1lce, 1syn, 1tsd, 1tsn, 1vzc, 2oip, 2tsr, 2vf0, 3b5b, 3b9h, 3bgx, 3bhl, 3bnz, 3byx, 3c06, 3clb, 3dga, 3hbb, 3ijz, 3ik0, 3ik1, 3inv, 3irm, 3iro, 3jsu, 3k2h, 3kjs, 3n2a, 3nrr, 3ob7, 3pyz, 3qg2, 3qgt, 3qj7, 3um5, 3um6, 3um8, 4dq1, 4e5o, 4eb4, 4f2v, 4gev

53 Clomipramine (DB01242)
18gs, 1aqv, 1hk1, 1hk3, 2bx8, 2bxf, 2pgt, 2q6h, 2vue, 2xvq, 2xvu, 3a73, 3b9l, 3csj, 3gss, 3gus, 3ie3, 3lu7, 3lu8, 3n9j, 3tdl, 4e99, 4pgt

54 Imipramine (DB00458)
1gs4, 1i38, 1t65, 1xnn, 1xow, 1z95, 2ax6, 2ax8, 2ax9, 2hvc, 2ihq, 2nw4, 2oz7, 2pnu, 2q72, 2q7i, 2q7k, 2rh1, 3b5r, 3b65, 3b66, 3b67, 3b68, 3d4s, 3g0w, 3l3x, 3ny8, 3ny9, 3rlj, 3rll, 3uon

55 Desipramine (DB01151)
1gs4, 1i38, 1t65, 1xnn, 1xow, 1z95, 2ax6, 2ax8, 2ax9, 2hvc, 2ihq, 2nw4, 2oz7, 2pnu, 2q7i, 2q7k, 2qb4, 2rh1, 3b5r, 3b65, 3b66, 3b67, 3b68, 3d4s, 3g0w, 3l3x, 3ny8, 3ny9, 3rlj, 3rll, 3uon

56 Trimethoprim (DB00440)
1an5, 1axw, 1dr2, 1dr7, 1dyh, 1dyj, 1e26, 1f4e, 1f4f, 1f4g, 1hfr, 1jg0, 1jom, 1jtq, 1juj, 1jut, 1lce, 1ly3, 1ly4, 1mvt, 1ohk, 1s3v, 1s3w, 1s3y, 1syn, 1tsd, 1tsn, 1u70, 1u71, 1vzc, 2ano, 2bla, 2cd2, 2dhf, 2oip, 2qk8, 2tsr, 2vf0, 2w3a, 2w3b, 2w3m, 2w3v, 2w3w, 2w9g, 2w9s, 2zza, 3b5b, 3b9h, 3bgx, 3bhl, 3bnz, 3byx, 3c06, 3cd2, 3clb, 3cse, 3d80, 3d84, 3dau, 3dga, 3e0b, 3eej, 3eek, 3eel, 3eem, 3eig, 3f0b, 3f0q, 3f0s, 3f0u, 3f0v, 3f0x, 3fl8, 3fl9, 3fq0, 3fqc, 3fqf, 3fqo, 3fqv, 3fqz, 3fra, 3frb, 3frd, 3frf, 3fy8, 3fy9, 3fyv, 3fyw, 3ghv, 3ghw, 3gi2, 3hbb, 3i8a, 3ia4, 3ijz, 3ik0, 3ik1, 3inv, 3irm, 3iro, 3ix9, 3jsu, 3jvx, 3jwf, 3jwk, 3k2h, 3k45, 3k47, 3kfy, 3kjs, 3m09, 3nrr, 3ntz, 3nu0, 3nxo, 3nxt, 3nxv, 3nxx, 3oaf, 3ob7, 3qfx, 3qg2, 3qgt, 3qj7, 3ql0, 3ql3, 3qlx, 3qly, 3qlz, 3r33, 3rg9, 3ro9, 3roa, 3s3v, 3s7a, 3s9u, 3sa1, 3sa2, 3sai, 3sgy, 3sh2, 3sqy, 3sr5, 3srr, 3srs, 3sru, 3srw, 3td8, 3tq8, 3tq9, 3tqb, 3um5, 3um6, 3um8, 4dfr, 4dq1, 4e5o, 4eb4, 4f2v, 4gev

57 Trimetrexate (DB01157)
1dr2, 1dr7, 1dyh, 1dyj, 1e26, 1hfr, 1jom, 1ly3, 1ly4, 1mvt, 1ohk, 1s3v, 1s3w, 1s3y, 1u70, 1u71, 2ano, 2bla, 2cd2, 2dhf, 2oip, 2qk8, 2w3a, 2w3b, 2w3m, 2w3v, 2w3w, 2w9g, 2w9s, 2zza, 3cd2, 3clb, 3cse, 3d80, 3d84, 3dau, 3dga, 3e0b, 3eej, 3eek, 3eel, 3eem, 3eig, 3f0b, 3f0q, 3f0s, 3f0u, 3f0v, 3f0x, 3fl8, 3fl9, 3fq0, 3fqc, 3fqf, 3fqo, 3fqv, 3fqz, 3fra, 3frb, 3frd, 3frf, 3fy8, 3fy9, 3fyv, 3fyw, 3ghv, 3ghw, 3gi2, 3hbb, 3i8a, 3ia4, 3inv, 3irm, 3iro, 3ix9, 3jsu, 3jvx, 3jwf, 3jwk, 3k45, 3k47, 3kfy, 3kjs, 3m09, 3ntz, 3nu0, 3nxo, 3nxt, 3nxv, 3nxx, 3oaf, 3qfx, 3qg2, 3qgt, 3ql0, 3ql3, 3qlx, 3qly, 3qlz, 3r33, 3rg9, 3ro9, 3roa, 3s3v, 3s7a, 3s9u, 3sa1, 3sa2, 3sai, 3sgy, 3sh2, 3sqy, 3sr5, 3srr, 3srs, 3sru, 3srw, 3td8, 3tq8, 3tq9, 3tqb, 3um5, 3um6, 3um8, 4dfr

58 Dexamethasome (DB01234)
1m2z, 1nhz, 3bqd, 3cld, 3e7c, 3gn8, 3k22, 3k23, 3mnp, 3mzs, 3n9y, 3na0, 3na1

59 Hydrocortisone (DB00741)
1m2z, 1nhz, 2vdy, 3bqd, 3cld, 3e7c, 3k22, 3k23

60 Fludrocortisone (DB00687)
1gs4, 1i38, 1m2z, 1nhz, 1t65, 1xnn, 1xow, 1y9r, 1z95, 2a3i, 2aa6, 2aax, 2ab2, 2ax6, 2ax8, 2ax9, 2hvc, 2ihq, 2nw4, 2oz7, 2pnu, 2q7i, 2q7k, 3b5r, 3b65, 3b66, 3b67, 3b68, 3bqd, 3cld, 3e7c, 3g0w, 3k22, 3k23, 3l3x, 3rlj, 3rll, 3vhv

61 Chlorpromazine (DB00477)
2rh1, 3apx, 3d4s, 3ny8, 3ny9, 4b0o

62 Trifluoperazine (DB00831)
1ctr, 1gs4, 1i38, 1lin, 1t65, 1wrk, 1xnn, 1xow, 1z95, 2ax6, 2ax8, 2ax9, 2e1q, 2e3t, 2hvc, 2ihq, 2nw4, 2oz7, 2pnu, 2q7i, 2q7k, 2rh1, 2vn9, 2wel, 3ax9, 3b5r, 3b65, 3b66, 3b67, 3b68, 3bdj, 3cv3, 3d4s, 3g0w, 3ko0, 3l3x, 3nvw, 3ny8, 3ny9, 3rlj, 3rll, 3rv5

63 Alogliptin (DB06203)
1rwq, 1wcy, 1x70, 2aj8, 2bgr, 2bub, 2buc, 2hha, 2i3z, 2i78, 2iit, 2iiv, 2jid, 2oae, 2oag, 2ogz, 2ole, 2onc, 2oph, 2oqi, 2qjr, 2qoe, 2qt9, 2qtb, 2rgu, 2rip, 3ccc, 3d4l, 3eio, 3f8s, 3g0b, 3g0c, 3g0d, 3g0g, 3h0c, 3hab, 3hac, 3kwf, 3kwj, 3o95, 3o9v, 3oc0, 3opm, 3qbj, 3sww, 3sx4, 3vjm, 4a5s

64 Linagliptin (DB08882)
1rwq, 1wcy, 1x70, 2aj8, 2bgr, 2bub, 2buc, 2hha, 2i3z, 2i78, 2iit, 2iiv, 2jid, 2oae, 2oag, 2ogz, 2ole, 2onc, 2oph, 2oqi, 2qjr, 2qoe, 2qt9, 2qtb, 2rgu, 2rip, 3ccc, 3d4l, 3eio, 3f8s, 3g0b, 3g0c, 3g0d, 3g0g, 3h0c, 3hab, 3hac, 3kwf, 3kwj, 3o95, 3o9v, 3oc0, 3opm, 3qbj, 3sww, 3sx4, 3vjm, 4a5s

65 Naproxene (DB00788)
1ht5, 1pxx, 2oye, 2oyu, 3cv3, 3kk6, 3ln0, 3ln1, 3mqe, 3n8w, 3n8x, 3n8y, 3n8z, 3nt1, 3ntb, 3ntg, 3q7d, 3qmo, 3r58, 3rr3, 3ufy, 4cox, 4fjp, 4fm5, 6cox

66 Nabumetone (DB00461)
1ht5, 1pxx, 2oye, 2oyu, 3kk6, 3ln0, 3ln1, 3mqe, 3n8w, 3n8x, 3n8y, 3n8z, 3nt1, 3ntb, 3ntg, 3ogw, 3q7d, 3ql6, 3qmo, 3rr3, 3taj, 4cox, 4fm5, 6cox

67 Hydrochlorothiazide (DB00999)
1a42, 1bn1, 1bn3, 1bn4, 1bnn, 1bnq, 1bnt, 1bnu, 1bnv, 1bnw, 1cil, 1cim, 1cin, 1cnw, 1cnx, 1eou, 1g45, 1g48, 1g4o, 1g52, 1i8z, 1i90, 1i91, 1i9l, 1i9m, 1i9n, 1i9o, 1i9p, 1i9q, 1if7, 1if8, 1if9, 1kwq, 1okl, 1okm, 1okn, 1oq5, 1ttm, 1xpz, 1xq0, 1ze8, 2h15, 2hd6, 2hkk, 2hoc, 2nn7, 2nnv, 2pou, 2pow, 2q1q, 2qo8, 2wd2, 2weh, 2wej, 2x7s, 2x7t, 2x7u, 3b4f, 3bet, 3bl1, 3c7p, 3d8w, 3d9z, 3da2, 3dbu, 3dcc, 3dd8, 3eft, 3f4x, 3f7b, 3f8e, 3hku, 3hlj, 3hs4, 3iai, 3ibl, 3ibn, 3igp, 3ik6, 3k2f, 3k34, 3kig, 3kne, 3l14, 3lxe, 3m04, 3m14, 3m2n, 3m2x, 3m3x, 3m40, 3m5e, 3m67, 3m96, 3m98, 3mdz, 3mhc, 3mhi, 3mhl, 3mhm, 3mho, 3ml5, 3myq, 3nb5, 3ni5, 3oik, 3oil, 3oim, 3oku, 3okv, 3oy0, 3oyq, 3oys, 3p25, 3p29, 3p4v, 3p58, 3p5a, 3p5l, 3po6, 3qyk, 3r16, 3r17, 3ryj, 3ryv, 3ryx, 3ryy, 3ryz, 3rz0, 3rz1, 3rz5, 3rz7, 3rz8, 3s71, 3s72, 3s73, 3s74, 3s8x, 3s9t, 3sap, 3sax, 3sbh, 3sbi, 3t5z, 3t82, 3t84, 3t85, 3ucj, 3v5g, 3v7x, 3vbd, 4e3d, 4e3f

68 Hydroflumethiazide (DB00774)
1a42, 1bn1, 1bn3, 1bn4, 1bnn, 1bnq, 1bnt, 1bnu, 1bnv, 1bnw, 1cil, 1cim, 1cin, 1cnw, 1cnx, 1eou, 1g45, 1g48, 1g4o, 1g52, 1i8z, 1i90, 1i91, 1i9l, 1i9m, 1i9n, 1i9o, 1i9p, 1i9q, 1if7, 1if8, 1if9, 1kwq, 1okl, 1okm, 1okn, 1oq5, 1ttm, 1xpz, 1xq0, 1ze8, 2h15, 2hd6, 2hkk, 2hoc, 2nn7, 2nnv, 2pou, 2pow, 2q1q, 2qo8, 2wd2, 2weh, 2wej, 2x7s, 2x7t, 2x7u, 3b4f, 3bet, 3bl1, 3c7p, 3d8w, 3d9z, 3da2, 3dbu, 3dcc, 3dd8, 3eft, 3f4x, 3f7b, 3f8e, 3hku, 3hlj, 3hs4, 3iai, 3ibl, 3ibn, 3igp, 3ik6, 3k2f, 3k34, 3kig, 3kne, 3l14, 3lxe, 3m04, 3m14, 3m2n, 3m2x, 3m3x, 3m40, 3m5e, 3m67, 3m96, 3m98, 3mdz, 3mhc, 3mhi, 3mhl, 3mhm, 3mho, 3ml5, 3myq, 3nb5, 3ni5, 3oik, 3oil, 3oim, 3oku, 3okv, 3oy0, 3oyq, 3oys, 3p25, 3p29, 3p4v, 3p58, 3p5a, 3p5l, 3po6, 3qyk, 3r16, 3r17, 3ryj, 3ryv, 3ryx, 3ryy, 3ryz, 3rz0, 3rz1, 3rz5, 3rz7, 3rz8, 3s71, 3s72, 3s73, 3s74, 3s8x, 3s9t, 3sap, 3sax, 3sbh, 3sbi, 3t5z, 3t82, 3t84, 3t85, 3ucj, 3v5g, 3v7x, 3vbd, 4e3d, 4e3f

69 Sildenafil (DB00203)
1t9s, 1tbf, 1uho, 1xoz, 1xp0, 2h42, 2h44, 3bjc, 3dba, 3dy8, 3dyl, 3hqw, 3hqy, 3hqz, 3hr1, 3jsi, 3jsw, 3jwr, 3k3e, 3k3h, 3lxg, 3qi4, 3qpn, 3qpo, 3qpp, 3shy, 3shz, 3sie, 3sn7, 3sni, 3snl, 3tge, 3tgg, 3ui7, 3uuo, 4ael, 4ddl, 4dff

70 Vardenafil (DB00862)
1t9s, 1tbf, 1uho, 1xot, 1xoz, 1xp0, 2h42, 2h44, 3bjc, 3dba, 3dy8, 3dyl, 3hqw, 3hqy, 3hqz, 3hr1, 3jsi, 3jsw, 3jwr, 3k3e, 3k3h, 3lxg, 3qi4, 3qpn, 3qpo, 3qpp, 3shy, 3shz, 3sie, 3sn7, 3sni, 3snl, 3tge, 3tgg, 3ui7, 3uuo, 4ael, 4ddl, 4dff

71 Ursodeoxycholic acid (DB01586)
1ihi, 1j96, 1lwi, 1s1p, 1s2c, 2hdj, 2ipj, 3h7r, 3h7u, 3o02, 3r43, 3r58, 3r6i, 3r7m, 3r8g, 3r8h, 3r94, 3ufy, 3ugr, 3uwe, 4dbs, 4dbu

72 Chenodeoxycholic acid (DB06777)
1ihi, 3bej, 3dct, 3dcu, 3fxv, 3hc5, 3l1b, 3o02, 3okh, 3oki, 3olf, 3omk, 3omm, 3oof, 3ook, 3p88, 3p89, 3rut, 3ruu

F

Unwanted HET codes

The following HET codes are not used as reference ligands if the active site is determined automatically by *i*RAISE. These molecules are cofactors or solution, buffer or crystallization agents. Further, ion HET codes and HET codes of molecules which cannot be initialized by NAOMI (due to covalently bound metals or elements not supported by NAOMI) are contained in the 'unwanted HET codes' set.

Ligand HET codes: 0HH, 10A, 12H, 12P, 13S, 140, 144, 15P, 16A, 16D, 192, 1AB, 1AC, 1AN, 1BO, 1BP, 1CB, 1CM, 1GN, 1LU, 1MA, 1MC, 1MR, 1MZ, 1PE, 1PG, 202, 20S, 217, 233, 24T, 25T, 26D, 2AC, 2AF, 2AP, 2BM, 2BR, 2CH, 2CM, 2EZ, 2FU, 2HA, 2HP, 2IB, 2IM, 2KT, 2ME, 2MG, 2MP, 2MZ, 2NO, 2OS, 2PA, 2PC, 2PE, 2PN, 2PO, 34A, 3AP, 3BB, 3BR, 3CH, 3CL, 3CN, 3EP, 3FA, 3GR, 3HL, 3MC, 3MF, 3MO, 3MP, 3MT, 3NP, 3OH, 3OL, 3PH, 3PO, 3PP, 3PY, 3TR, 4AP, 4AX, 4CB, 4HA, 4IP, 4MV, 4MZ, 4PA, 5AN, 5BR, 5IP, 5MC, 5MP, 5MU, 6NA, 6PC, 749, 7MG, 9CS, A2G, A3B, A48, A5P, A6P, AAB, AAC, AAE, ABA, ABH, ABN, ABU, AC0, AC5, ACA, ACD, ACE, ACM, ACN, ACR, ACT, ACY, ADA, ADE, ADM, ADP, AE3, AEM, AFB, AG2, AGA, AGU, AHG, AHI, AHR, AI2, AIB, AIO, AJ3, AKB, AKR, ALA, ALG, ALQ, AMB, AMC, AML, AMT, AMV, ANL, AOA, APB, ARE, ARF, ARS, ART, AS, AS2, ASN, ASP, AST, ATJ, ATO, ATQ, AZI, B2A, B2F, B2I, B2V, BAL, BAM, BBU, BCD, BCT, BDB, BDD, BDP, BEM, BEN, BEO, BEQ,

BET, BEZ, BGC, BGL, BGX, BH1, BHH, BHL, BIB, BJH, BJI, BJP, BLA, BLE, BLV, BLY, BMA, BME, BME, BML, BMM, BMT, BNG, BNO, BNS, BNZ, BO3, BO4, BOC, BOG, BOM, BOR, BPH, BR5, BRB, BRC, BRJ, BRP, BTB, BTL, BTN, BU1, BU2, BU3, BUA, BUB, BUQ, BVC, BVF, BVG, BZB, BZF, BZI, BZP, C21, C2A, C2B, C2N, C8E, CA1, CAB, CAC, CAD, CAM, CAN, CAP, CAQ, CAS, CAT, CCM, CCN, CCP, CDL, CE1, CE8, CE9, CEF, CEJ, CEP, CEQ, CFA, CFQ, CFT, CGU, CH2, CHM, CHT, CHX, CIG, CIR, CIT, CKP, CLB, CLD, CLL, CLX, CM6, CME, CMP, CMS, CMT, CNH, CNN, CO2, CO3, COI, COM, CP, CP4, CPS, CRD, CRN, CRS, CRT, CSD, CSO, CSS, CSW, CTB, CTR, CTT, CVB, CXB, CXE, CXF, CXL, CXM, CYC, CYH, CYI, CYS, CYT, D12, D1D, D2P, DA1, DAL, DAO, DAS, DAV, DBP, DCE, DCY, DDQ, DEM, DEN, DEP, DER, DFP, DFX, DG6, DGA, DGG, DGL, DGY, DHA, DHD, DHK, DHM, DHS, DIA, DIB, DIO, DLE, DM1, DMF, DMG, DMN, DMR, DMS, DOD, DP3, DP4, DPE, DPF, DPJ, DPN, DPO, DPR, DRN, DSG, DSN, DSS, DTD, DTH, DTI, DTL, DTO, DTT, DTU, DTV, DUC, DVA, DXE, DXX, DZZ, EAP, EDO, EEE, EFS, EGC, EGD, ELA, EMM, ENC, EOH, EPE, EPO, ESA, ETA, ETF, ETI, ETM, ETN, ETP, ETX, ETY, F09, F6P, FA1, FA6, FAC, FAG, FBA, FCA, FCB, FCL, FCN, FFP, FLA, FLC, FLM, FMN, FMS, FMT, FNG, FOA, FOP, FOR, FPI, FPN, FPO, FPR, FPY, FRU, FU2, FUC, FUM, FUX, FX3, G16, G1P, G3P, G4D, G4S, G6P, GAG, GAI, GAL, GAQ, GAU, GB, GBD, GBL, GC4, GCO, GCU, GDM, GEG, GER, GG6, GLA, GLC, GLL, GLN, GLO, GLR, GLS, GLU, GLV, GLY, GM1, GOA, GOL, GPM, GSC, GSH, GUA, GVE, GVH, GYP, GZZ, H01, H02, H2O, H2S, H2U, H4B, HAE, HAI, HAV, HBA, HBR, HBS, HCA, HCS, HDA, HDS, HE2, HE4, HED, HEQ, HEX, HEZ, HGU, HHO, HIO, HIU, HLE, HLT, HMC, HMF, HMN, HOA, HOH, HOZ, HP6, HPA, HPH, HPN, HPY, HSE, HSL, HSM, HSW, HT, HTO, HTS, HXA, HY1, HYA, HYF, HYP, I, I3P, I4P, IAP, IAS, IBO, IBS, IBZ, ICN, ICP, ICT, IDH, IDM, IDR, IDS, IHG, IHP, ILE, IMD, IMR, IND, IOB, IOL, IOM, IP5, IPA, IPH, IPM, IPU, ISP, ISU, ITU, IUR, IVA, IZC, JEF, KCX, KDF, KIV, KMT, KOS, KPH, L1P, L2P, L3P, L4P, LAC, LAF, LAR, LAT, LBT, LCP, LDA, LDM, LDY, LEA, LEN, LEU, LG3, LG4, LG5, LG6, LGV, LI1, LIO, LIS, LLP, LMT, LMU, LNK, LNL, LPC, LPG, LTL, LVG, LXP, LYS, LYT, M1N, M1P, M2G, M2M, M6D, M6P, MAE, MAH, MAK, MAL, MAN, MAS, MAV, MAW, MBD, MBN, MBR, MBT, MBV, MCR, MCT, MD2, MDD, ME2, MEC, MED, MEE, MES, MET, MEV, MEZ, MFU, MG8, MGO, MGX, MHN, MHO, MIC, MLA, MLE, MLI, MLM, MLP, MLR, MLT, MMA, MMP, MMQ, MMZ, MNA, MNC, MOH, MOR, MPA, MPC, MPD, MPG, MPI, MPJ, MPO, MR3, MRC, MRD, MRY, MSE, MSF, MSM, MTG, MTL, MUR, MVA, MVL, MXE, MYR, MYS, N2O, N2P, N8E, NAG, NAK, NBE, NBF, NBN, NBT, NBU, NBZ, NCA, NCM, ND4, NDG, NEH, NEN, NEQ, NET, NGA, NGS, NH2, NH3, NH4, NHE, NHV, NHY, NIO, NIS, NLE, NME, NMH, PGE, , NO, NO2, NO3, NOE, NOY, NPB, NPN, NPY, NS1, NS5, NSM, NT, NTA, NTB, NTC,

NTJ, NTN, NVA, NVI, NXA, O, OAA, OC9, OCA, OCT, OCY, ODS, OMC, OMG, OPE, ORN, OSM, OXD, OXE, OXL, OXM, OXN, OXP, OXQ, P1R, P2O, P33, P4C, P4G, P6G, PAE, PAH, PAM, PBA, PBC, PBR, PCA, PCR, PCT, PCZ, PDO, PDT, PE3, PE4, PE5, PE6, PE7, PE9, PEA, PED, PEG, PEL, PEO, PEU, PG0, PG3, PG4, PG5, PG6, PGA, PGE, PGH, PGO, PGR, PH1, PHB, PHN, PHO, PHS, PHZ, PI, PID, PIH, PIM, PIN, PIP, PIS, PLD, PLM, PLP, PMP, PNZ, PO2, PO4, POA, POL, PON, POP, PPB, PPF, PPI, PPK, PPV, PRI, PRO, PS5, PSE, PSL, PSU, PTD, PTL, PTR, PUB, PUT, PXY, PY7, PYC, PYD, PYE, PYF, PYG, PYL, PYM, PYQ, PYR, PYS, PYT, PYZ, PZO, QPS, QV4, R5P, RAF, RCL, RCO, REA, RET, RGI, RIB, RIP, RNS, RNT, RPD, RPL, RUB, S0H, SAR, SAT, SB1, SBD, SBE, SBL, SBO, SBT, SCC, SCH, SCN, SCS, SDS, SE, SE4, SEP, SER, SES, SFO, SGL, SGM, SGN, SHF, SHO, SHV, SIA, SIF, SIN, SLE, SLF, SM2, SM3, SM4, SMB, SMC, SO2, SO3, SO4, SO4, SOA, SOR, SPA, SPH, SPM, SPN, SPO, SQU, SRB, SRD, SRT, SS1, SS2, STA, SUC, SUF, SUM, T1A, T42, TAM, TAR, TAS, TAU, TBU, TC4, TCB, TCZ, TDA, TDR, TEO, TF4, TFA, TFB, THE, THJ, THP, THR, TLA, TMA, TME, TMT, TMZ, TOU, TP5, TPO, TRA, TRC, TRD, TRE, TRI, TRS, TRT, TSM, TSZ, TTN, TWT, TYI, TYS, TZC, TZE, TZL, TZZ, U10, UNA, UND, UNK, UPL, URA, URE, URF, URP, UVW, V35, V36, VA1, VAL, VIG, VSO, VX, VXA, WBU, WZ1, WZ2, XAP, XDL, XDN, XIF, XLS, XPE, XUL, XYD, XYH, XYL, XYP, XYS, YAN, YG, YRR, ASG, DCL, DKA, MBG, MH6, PC1, PHQ, RAM, T32, 2LU, 3MG, 4SC, 6CT, 9MR, ACI, AGL, B1F, B4G, BCG, BRM, BZO, CBI, CBM, CBX, CBZ, CEC, CH3, CM5, CMO, CRY, CYA, DIS, DOM, DOX, DUM, EHN, EOX, ETD, ETO, FLO, FUB, FX1, G2I, GAC, GCM, GCS, GLB, GLD, GS1, GTE, H1M, HP3, HPG, HS2, HYD, ILG, IOH, IPS, KDA, KDB, KDD, KDE, KDR, KFG, KO2, KO4, MAB, MCB, MCE, MTO, MTT, NMO, NYT, OBD, OHE, OMB, OME, OTE, OXA, OXO, OXY, OXZ, PAR, PER, PPM, PSS, PVL, QPU, SBU, SEO, SFN, SGC, SOH, SOM, SUL, TMN, TPH, UNF, UNL, X4S, X5S, YT3.

Ion HET codes: 3CO, 4MO, 6MO, AG, AL, AR, AU, AU3, BA, BR, BR, CA, CD, CE, CL, CMO, CO, CR, CS, CU, CU1, CYN, EU, EU3, F, FE, FE2, GA, GD, GD3, HDZ, HG, HO, HO3, IOD, IR3, K, KR, LA, LI, LU, MG, MN, NA, NI, OS4, OXY, PB, PD, PER, PO3, PR, PT, RB, SM, SR, SX, TB, TL, U1, VO4, W, XE, YB, YT3, ZN.

Non-NAOMI-conform HET codes: 6HE, 6WO, 7HE, A71, A72, AAS, AF3, ALF, AMS, APW, AUC, B12, B69, B70, BCL, BF2, BPT, C1O, C2O, CFM, CFN, CFO, CLF, CLN, CLO, CLP, CN1, CNB, CNF, CON, CUA, CUB, CUM, CUN, CUO, CUZ, DAZ, DEF, DHE, DOZ, DRU, DTZ, DW2, DWC, EMC, EMT, F3S, F3S, FCO, FEA, FEL, FEO, FES, FNE, FS1, FS2, FSO, HC0, HC1, HDD, HE6, HEC, HEG, HEM, HEV, HF3, HF5, HG2,

HGB, HGI, HNI, IME, ISP, IUM, JM1, KEG, KYS, LCO, LPT, MAC, MAP, MBO, MF4, MGF, MH2, MM4, MMC, MN3, MNH, MNR, MO, MO7, MOO, MOS, MP1, MTD, NCO, NCP, NFE, NFO, NFR, NFS, OEC, OFO, OMO, OS, PBM, PC4, PCL, PEJ, PHF, PHG, PMB, POR, R4A, R6A, RBU, REO, REP, RHX, RTA, RTB, RTC, SB, SEK, SF3, SF4, SF4, SMO, SRM, TCN, U10, UNX, WCC, WO4, WO5, XCC, Y1, YBT.

G

ROC curves for Drugs/sc-PDB enrichment experiment

ROC curves for all 72 ligands of the Drugs/sc-PDB data set. The plots are shown separately for the categories 'excellent', 'good', 'medium' and 'bad' enrichment.
Subsequently, the ROC curves of the clustered drugs of the Drugs/sc-PDB data set are shown. For these, two ROC curves are shown in one diagram for each drug: One curve if only the targets annotated in the DrugBank are considered as true targets for a drug and another curve representing the case if the targets of one cluster are combined and counted for all members of one cluster as true targets.

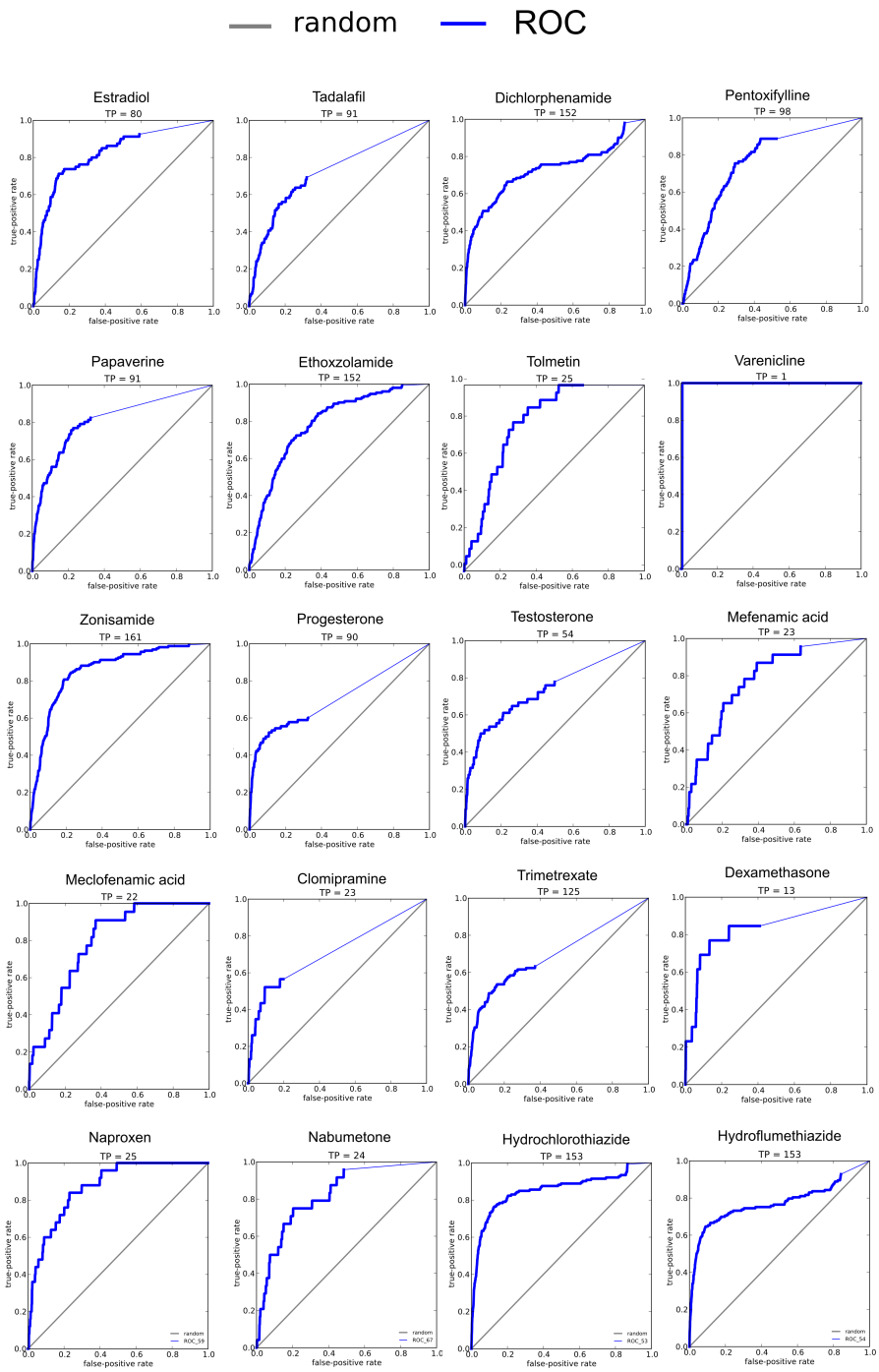


Figure G.1: ROC plots for 20 ligands categorized as 'excellent'. Thick blue lines show the true positives found, if not all true positives were identified, a thin line is drawn from the last found positive on assuming random distribution from there on. TP=Number of true positives.

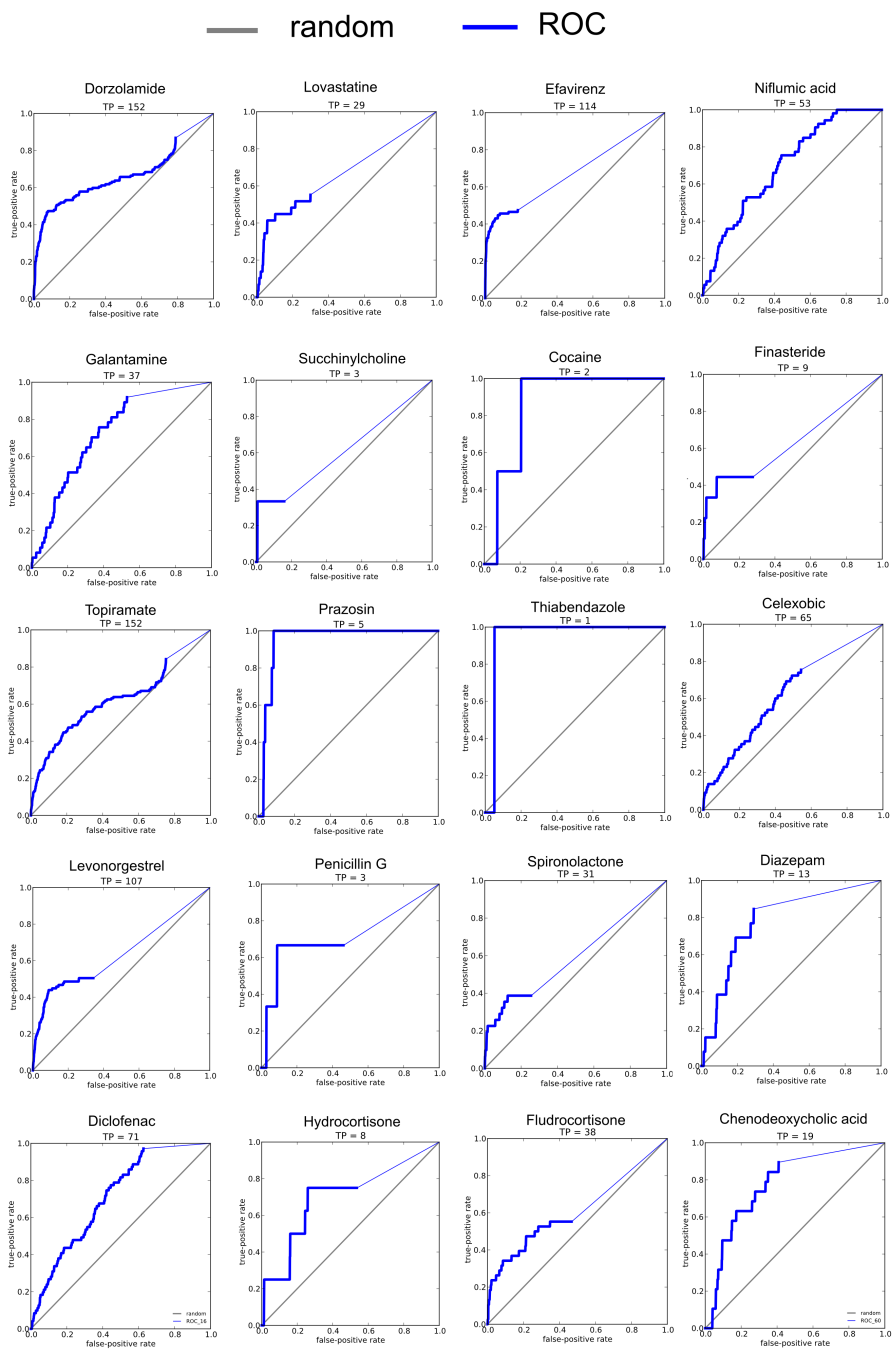


Figure G.2: ROC plots for 20 ligands categorized as 'good'. Thick blue lines show the true positives found, if not all true positives were identified, a thin line is drawn from the last found positive on assuming random distribution from there on. TP=Number of true positives.

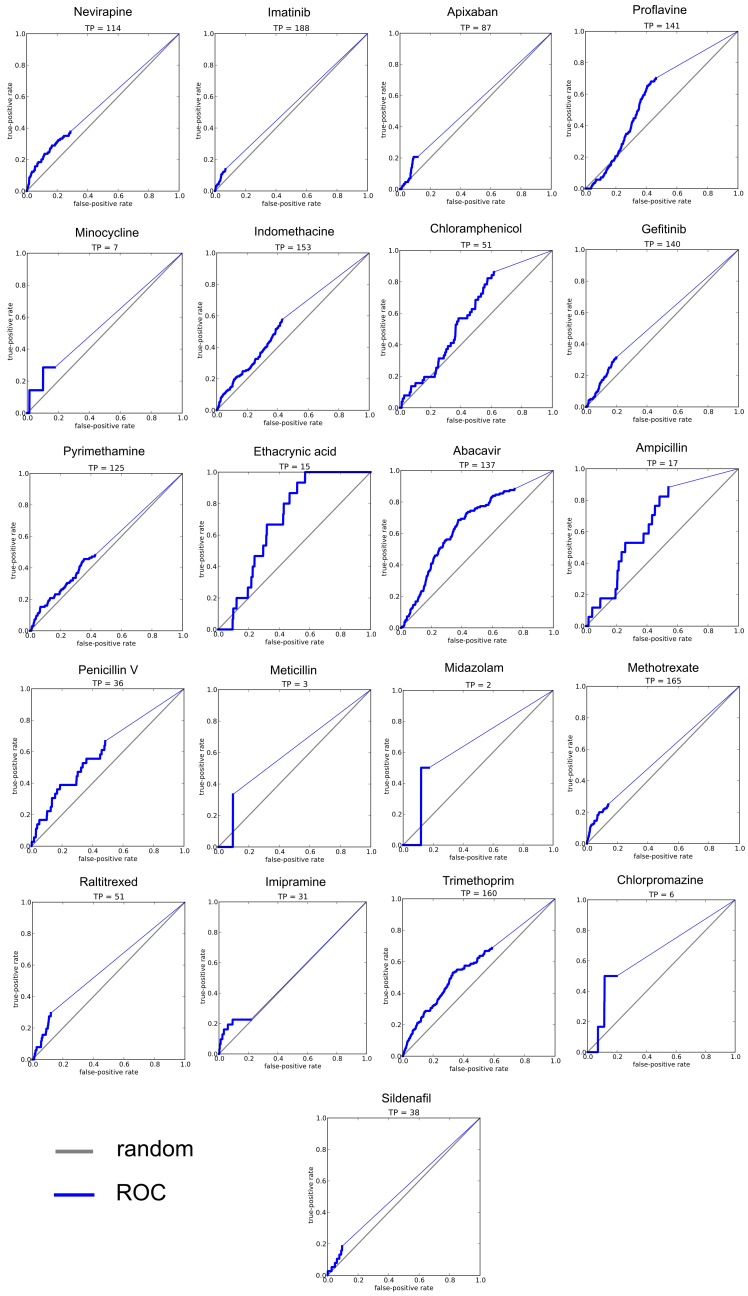


Figure G.3: ROC plots for 21 ligands categorized as 'medium'. Thick blue lines show the true positives found, if not all true positives were identified, a thin line is drawn from the last found positive on assuming random distribution from there on. TP=Number of true positives.

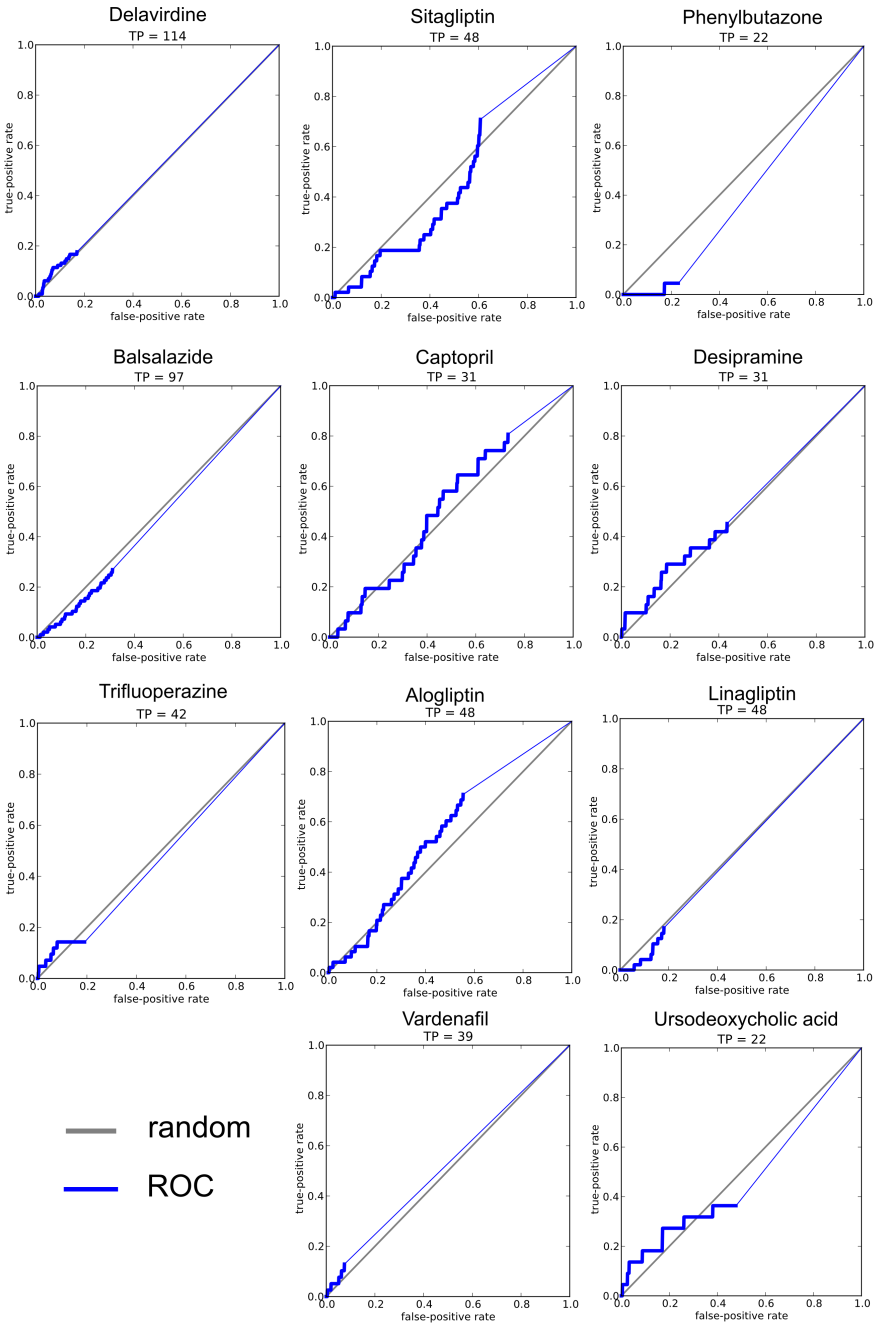


Figure G.4: ROC plots for 11 ligands categorized as 'bad'. Thick blue lines show the true positives found, if not all true positives were identified, a thin line is drawn from the last found positive on assuming random distribution from there on. TP=Number of true positives.

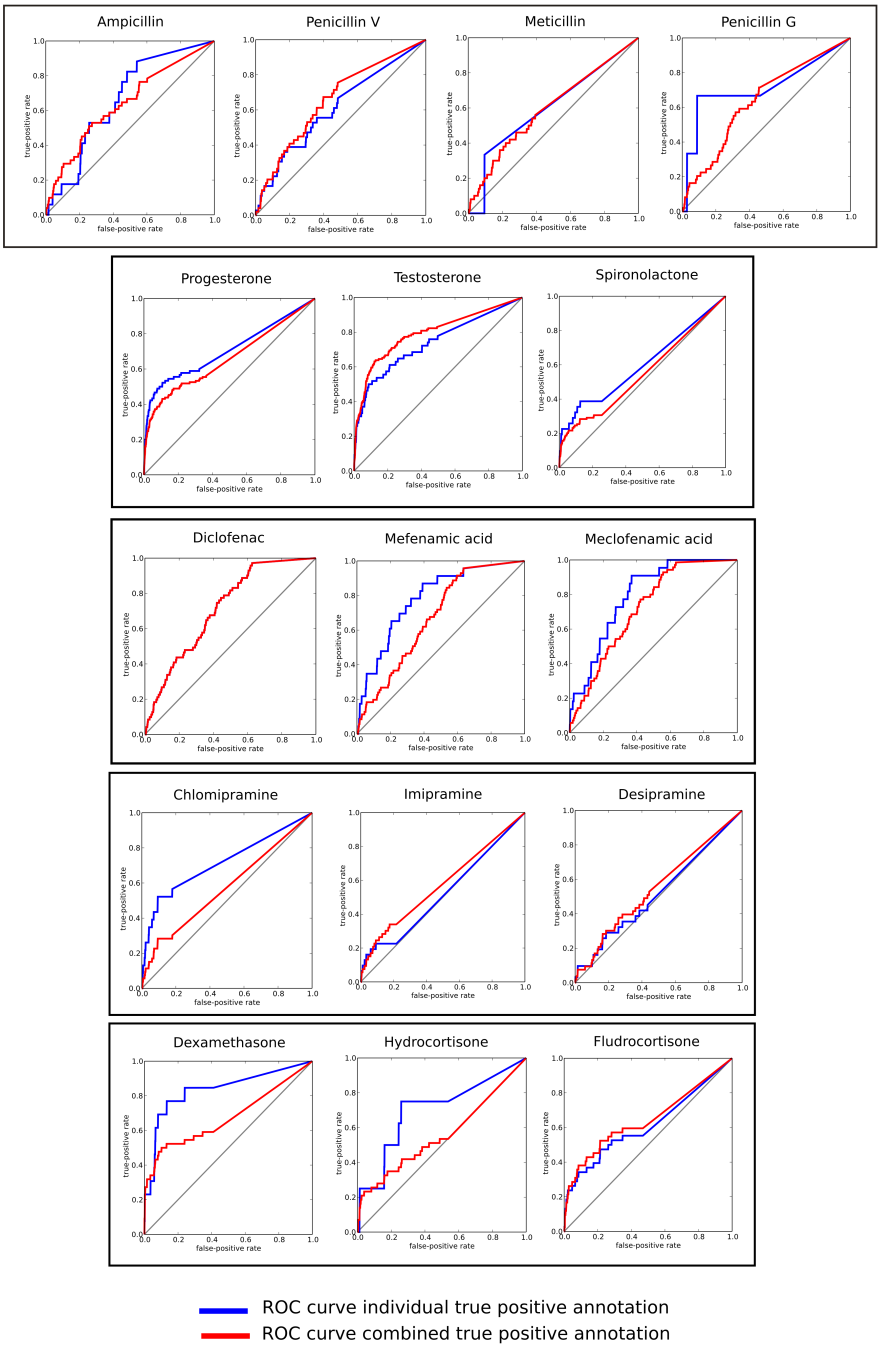


Figure G.5: ROC plots for drug clusters with three or more members. The blue line shows the ROC curve with individual target annotation, the red line shows the ROC curve if all targets of one clusters are combined as true targets. If only a red line is shown, then the individual and the combined true target annotaion was identical.

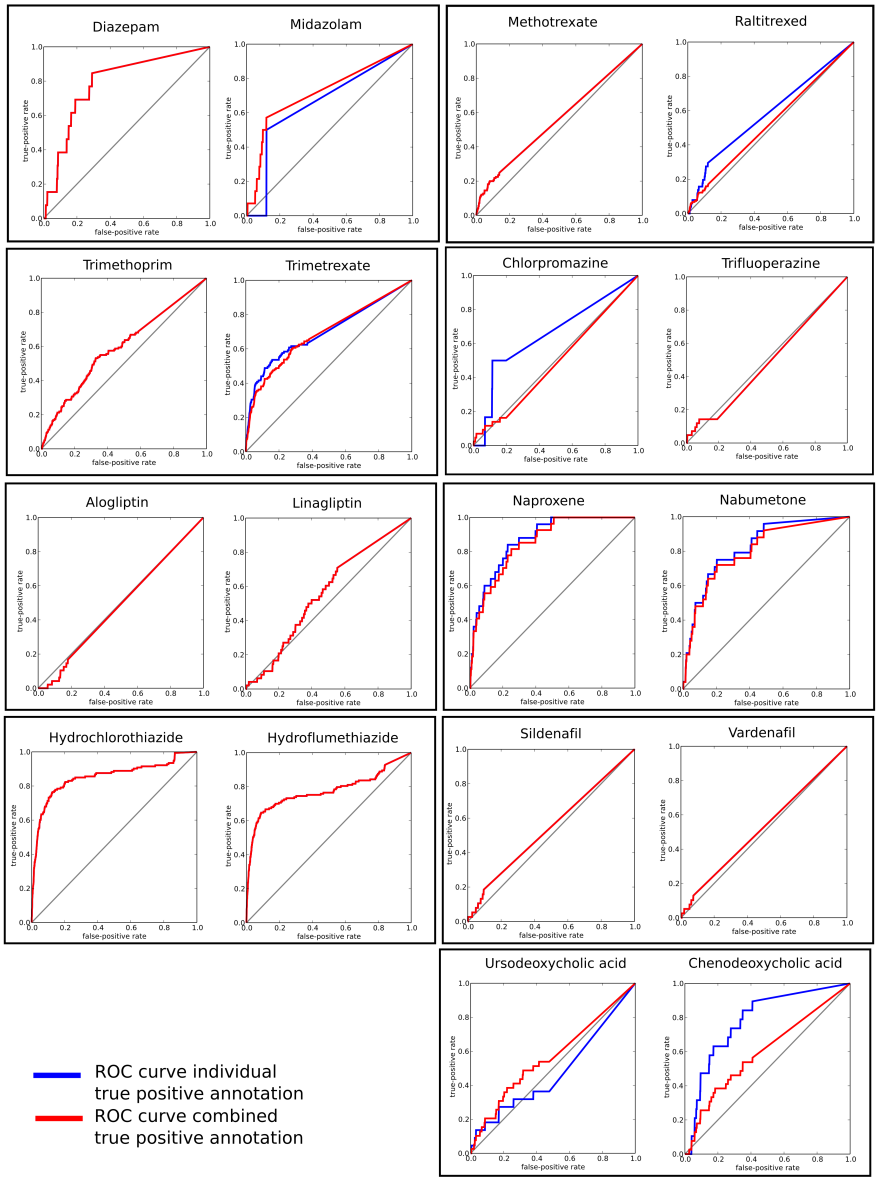


Figure G.6: ROC plots for drug clusters with two members. The blue line shows the ROC curve with individual target annotation, the red line shows the ROC curve if all targets of one clusters are combined as true targets. If only a red line is shown, then the individual and the combined true target annotaion was identical.

H

sc-PDB Diverse Set Results

List of ranks of the first true positive (TP) target for predictions of *i*RAISE on the 117 ligands of the sc-PDB Diverse Set.

Ligand HET code	*i*RAISE-flex	*i*RAISE-flex EC-TP	*i*RAISE-crystal	*i*RAISE-crystal EC-TP
61	79	8	6	6
115	287	287	1	1
215	80	33	2	2
356	299	299	1	1
501	111	111	1	1
669	258	258	1	1
760	76	76	75	75
783	2	2	1	1
792	7	7	1	1
839	2	2	1	1
905	6	6	1413	1413
961	4	2	1	1
984	55	55	2	2
20A	84	6	1	1

Ligand HET code	*i*RAISE-flex	*i*RAISE-flex EC-TP	*i*RAISE-crystal	*i*RAISE-crystal EC-TP
2IG	113	113	15	15
3B9	151	151	1	1
3CC	1	1	1	1
3LP	350	350	81	81
3QC	18	1	1	1
55V	2	2	1	1
5RM	4	4	12	4
6C3	25	25	10	8
87Y	2002	2002	1043	1043
A46	13	13	3	3
A80	2	1	1	1
AEE	218	1	2	2
AH1	1	1	1	1
AO5	108	108	1	1
AXX	39	4	1	1
AZZ	450	450	3	3
BAU	3	3	1	1
BFS	799	799	118	118
BHY	450	7	4	3
BIG	1090	19	1	1
BRZ	231	231	73	73
CBT	16	16	23	23
CEI	101	4	1	1
CEL	8	8	2	2
CIA	106	51	1	1
CMF	30	28	3	3
CRZ	5	5	6	2
CT5	9	4	57	4
D1L	130	130	1166	1166
DD2	2	2	1	1
DEO	1	1	1	1
DES	72	2	11	1
DEX	5	1	1	1
DZG	138	138	5	5
DZP	64	10	5	3
E4D	1	1	1	1
E89	1	1	1	1
EI1	32	6	11	2
EQI	34	34	27	27
ET	2	2	1	1
FR4	67	67	1	1
FSN	4	4	1	1
GB7	46	46	1	1
GNT	86	1	373	1
GRR	5	5	2	2
GVR	2	2	1	1
H11	8	8	64	64
H24	33	33	13	13
H7J	469	469	8	8
HA3	64	11	198	198
HEF	5	5	11	11
HM5	11	11	6	6

Ligand HET code	*i*RAISE flex	*i*RAISE flex EC-TP	*i*RAISE crystal	*i*RAISE crystal EC-TP
I84	257	237	5	5
IAD	10	10	1	1
IC1	69	2	51	2
IMN	59	10	3	3
IMQ	6	6	1	1
IXM	2	2	2	2
LG7	46	5	14	2
LI9	4	4	1	1
LQQ	120	120	4	4
LS1	6	6	5	5
MC9	8	1	2	1
MD7	28	28	7	7
MTI	353	324	2	2
NDR	8	1	5	1
NGH	2848	2848	2844	2844
OA1	10	10	1	1
OEF	2	2	1	1
P1S	1448	1448	828	828
P1Z	83	83	1	1
P34	2	2	10	7
P4A	1048	8	129	3
PBF	725	652	32	32
PFP	9	8	5	3
PVB	51	3	3	1
R6C	5	5	33	33
R78	1	1	1	1
R88	53	53	1	1
ROF	13	13	2	2
RRC	2	2	5	1
RXC	869	869	1573	1573
S22	8	8	2	2
SAG	15	15	18	18
SB8	311	311	1	1
SCT	84	84	1	1
SHM	6	6	1	1
SLX	33	33	1	1
STC	140	140	102	102
T74	1123	4	1	1
TCD	352	1	2024	1
TDZ	99	1	6	6
TIM	78	11	1	1
TNK	2	2	1	1
TPR	130	130	39	39
TSX	3	3	1	1
VDN	39	39	1	1
VGA	210	210	1292	1292
VGB	394	11	1	1
VGG	1	1	1	1
VIB	18	18	326	326
XM5	7	7	10	10
ZMA	31	3	3	3

I
Publications

In this Appendix, my scientific contributions are listed, sorted into publications in scientific journals, talks at conferences and poster presentation at conferences. The scientific work which is related to this thesis is highlighted in bold.

I.1 Publications in scientific journals

1. **Schomburg, K. T., Rarey, M. (2014). Benchmark Datasets for Structure-based Computational Target Prediction. *Journal of Chemical Information and Modeling*, DOI: 10.1021/ci500131x**

2. **Schomburg, K. T, Bietz, S., Briem, H., Henzler, A. M., Urbaczek, S., Rarey, M. (2014). Facing the Challenges of Structure-based Target Prediction by Inverse Virtual Screening. *Journal of Chemical Information and Modeling*, 54(6), 1676-1686**

3. von Behren, M., Volkamer, A., Henzler, A. M., Schomburg, K. T., Urbaczek, S., Rarey, M. (2013). Fast Protein Binding Site Comparison via an Index-Based Screening Technology. *Journal of Chemical Information and Modeling*, 53, 411-422

4. Schomburg, K.T, Wetzer, L., Rarey, M. (2013). Interactive design of generic chemical patterns. *Drug Discovery Today*. 18, 651- 658

5. **Schomburg, K. T., Ardao, I., Götz, K., Rieckenberg, F., Liese, A., Zeng, A. P., Rarey, M. (2012). Computational Biotechnology: Prediction of competitive substrate inhibition of enzymes by buffer compounds with protein-ligand docking. *Journal of Biotechnology*, 161, 391-401**

6. Schomburg, K., Ehrlich, H.-C., Stierand, K., Rarey, M. (2010). From Structure Diagrams to Visual Chemical Patterns, *Journal of Chemical Information and Modeling*, 50, 1529-1535

I.2 Talks

1. **Schomburg, K. T., Rarey M. *Facing the challenges of computational target prediction*, 9th German Conference on Chemoinformatics, 2013, Fulda, Germany**

2. **Schomburg, K. T., Ardao, I., Götz, K., Rieckenberg, F., Liese, A., Zeng, A.-P., Rarey, M. *Computational prediction of enzyme activity in different buffer solutions*. ProcessNet-Jahrestagung und Jahrestagung der Biotechnologen, 2012, Karlsruhe, Germany**

3. **Schomburg, K. T., Ardao, I., Götz, K., Rieckenberg, F., Liese, A., Zeng, A.-P., Rarey, M. *Computational prediction of enzymatic activity in different buffer solutions*, 243rd ACS National Meeting, 2012, California, USA**

4. Schomburg, K.T., Rarey, M. *How to design chemical patterns easily with an interactive editor*, 243rd ACS National Meeting, 2012, California, USA

5. Schomburg, K., Ehrlich, H.-C., Stierand, K., Rarey, M. *Chemical Pattern Visualization in 2D – The SMARTSviewer*, 6th German Conference on Cheminformatics, 2010, Goslar

I.3 Poster

1. **Schomburg, K. T., Rarey, M. *Challenges in the Evaluation of Inverse Virtual Screening*, Gordon Research Conference, 2013, West Dover, USA**

2. **Schomburg, K. T., Ardao, I., Götz, K., Rieckenberg, F., Liese, A., Zeng, A.-P., Rarey, M., *Using computational methods to predict a lowered enzyme activity due to competitive substrate inhibition by buffer compounds*, International Workshop on New and Synthetic Bioproduction Systems, 2012, Hamburg, Germany**

3. Schomburg, K. T., Rarey, M. *Simplifying the design and interpretation of chemical patterns: A visual approach*, 243rd ACS National Meeting, 2012, California, USA

4. **Schomburg, K. T., Ardao, I., Götz, K., Rieckenberg, F., Liese, A., Zeng, A.-P., Rarey, M. *Computational prediction of enzyme activity in different buffer solutions*, 7th Status Seminar Chemical Biology, 2011, Frankfurt am Main, Germany**

5. Schomburg, K., Ehrlich, H.-C., Stierand, K., Rarey, M. *Visual Chemical Patterns: From Automated Depiction to Interactive Design*, ICCS, 2011, Noordwijkerhood, Netherlands