Ligand-based Virtual Screening Utilizing Partial Shape Constraints

Dissertation with the aim of achieving a doctoral degree
at the Faculty of Mathematics, Informatics and Natural Sciences

Department of Informatics
of Universität Hamburg

submitted by Mathias Michael von Behren (5910653)
30.11.2016 in Hamburg

Vorsitzender der Prüfungskommission:
Prof. Dr. Menzel

Gutachter:
Prof. Dr. Matthias Rarey
Prof. Dr. Johannes Kirchmair

Tag der Disputation: 21.03.2017

Abstract

In drug discovery, the identification of lead structures as basis for the development of new drugs is of vital importance. Computational methods present an efficient way to search for promising structures without the costs of exhaustive experimental evaluations in the first stages of the drug discovery process. Ligand-based virtual screening is a well established method to search for new lead structures based on known active ligands for a certain target of interest. In this thesis, the development of a new method for highly efficient ligand-based virtual screening is presented. The mRAISE method furthermore addresses open challenges of the field, utilizing new approaches for partial-shape matching.

mRAISE uses special triangle descriptors, originally developed for structure-based virtual screening, to initially compare molecules on a coarse level and subsequently uses matching descriptors to calculate molecular alignments. To enable rapid screening, the descriptors of a compound library are preprocessed and the descriptors of all compounds are stored in a special bitmap index. Molecular alignments are scored using atom-centered Gaussian functions with weights representing the similarity of physicochemical properties of the respective atoms. Based on the local shape description of the utilized descriptor, partial shape constraints are incorporated into the screening procedure. These constraints can either be automatically derived from protein-ligand complexes or be manually defined via an atom selection of the user.

The method has been evaluated on multiple datasets in terms of active enrichment as well as alignment accuracy. In comparison to other methods, mRAISE is always among the top ranks regarding the screening performance as well as the runtime.

For the interactive features of mRAISE, a graphical user interface has been developed providing the complete functionality of the method combined with further options to visualize query descriptors and screening results. Furthermore, the graphical user interface is necessary for the manual definition of partial shape constraints by the user.

Kurzfassung

Die Identifikation von Leitstrukturen als Grundlage für die Entwicklung neuer Medikamente ist von entscheidender Bedeutung für den Wirkstoffentwurf. Computergestützte Methoden stellen hierbei eine effiziente Möglichkeit dar nach vielversprechenden Strukturen zu suchen, ohne die Kosten für erschöpfende Experimente in den ersten Phasen des Wirkstoffentwurfs. Ligandbasiertes virtuelles Screening ist eine etablierte Methode für die Suche nach neuen Leitstrukturen, basierend auf bekannten, aktiven Liganden für ein bestimmtes Zielprotein. In dieser Arbeit wurde eine neue Methode für effizientes ligandbasiertes virtuelles Screening entwickelt. Die mRAISE Methode widmet sich außerdem ungelösten Herausforderungen durch die Nutzung von neuen Ansätzen für einen partiellen Formvergleich.
mRAISE nutzt spezielle Dreiecksdeskriptoren, die ursprünglich für strukturbasiertes virtuelles Screening entwickelt wurden, um Moleküle zunächst grob zu vergleichen und anschließend auf Basis von übereinstimmende Deskriptoren molekulare Überlagerungen zu berechnen. Um einen schnellen Vergleich zu ermöglichen, werden die Deskriptoren einer Molekülbibliothek vorberechnet und in einem speziellen Bitmap Index gespeichert. Molekulare Überlagerungen werden bewertet mithilfe von Atom-zentrierten Gauß Funktionen mit Gewichtungen zur Berücksichtigung der physikochemischen Eigenschaften der jeweiligen Atome. Basierend auf der Beschreibung der lokalen Form des Deskriptors werden Einschränkungen für den partiellen Formvergleich in das Screening integriert. Diese Einschränkungen können entweder automatisch von Protein-Ligand Komplexen abgeleitet oder manuell durch einen Nutzer mittels der Selektion von Atomen definiert werden.
Die Methode wurde auf mehreren Datensätzen evaluiert, sowohl mit Hinblick auf die Anreicherung von aktiven Liganden als auch auf die Genauigkeit der Überlagerungen. Im Vergleich mit anderen Methoden ist mRAISE immer auf den besten Plätzen bezüglich der Screening Leistung sowie der Laufzeit.
Für die interaktiven Funktionalitäten von mRAISE wurde eine grafische Benutzeroberfläche entwickelt, welche die volle Funktionalität der Methode zusammen mit weiteren Optionen zur Visualisierung von Anfragedeskriptoren und Screeningresultaten zur Verfügung stellt. Außerdem ist die graphische Benutzeroberfläche erforderlich für die manuelle Definition von Beschränkungen für den partiellen Formvergleich durch den Nutzer.

Danksagung

An dieser Stelle möchte ich zumindest einigen der Menschen danken, ohne die diese Arbeit auf die eine oder andere Art nicht möglich gewesen wäre.
Als erstes möchte ich mich bei Prof. Dr. Matthias Rarey bedanken, der es mir ermöglichte meine Promotion mit einem sehr spannenden und herausfordernden Projekt durchzuführen und meine Ergebnisse auch international zu präsentieren. Außerdem danke ich ihm für die hervorragende Betreuung und die außergewöhnlich tolle Atmosphäre in seiner Arbeitsgruppe, die wahrlich ihresgleichen sucht.
Die großen Fortschritte und guten Ergebnisse, die ich in meiner Arbeit erzielen konnte, wären ohne die Vor- und Mitarbeit einiger ehemaliger und aktueller Kollegen nicht möglich gewesen, daher danke ich vor allem Karen Schomburg und Angela Henzler und auch allen weiteren Kollegen, die an der Entwicklung unserer Dreiecksdeskriptoren und deren Anwendung beteiligt waren für die tolle Zusammenarbeit. Ein besonderer Dank gilt auch Andrea Volkamer, die mich durch die Betreuung meiner Masterarbeit und auch späterer Zusammenarbeit in meiner Entscheidung zu diesem Weg bestärkt und große Teile dieser Arbeit einem prüfenden Blick unterzogen hat.
Auch allen anderen Kollegen am ZBH möchte ich danken für die schöne Zeit und die vielen anregenden Gespräche und Diskussionen, allen voran meinen Büronachbarn Thomas Otto und Eva Nittinger. Besonderer Dank gilt auch Dirk Willrodt für anregende technische Diskussionen und wohltuende Zerstreuung in der Mittagspause.
Meinen Eltern danke ich dafür, dass sie mir mein Studium und damit auch diesen weiteren Weg ermöglicht haben und dafür das sie immer hinter mir standen. Auch meiner kleinen Schwester und dem Rest meiner Familie danke ich für stete Ermutigung und Rückhalt. Meinen guten Freunden danke ich für den regelmäßigen Ausgleich und die Freude mit der sie mein Leben bereichern, egal ob sie direkt in Hamburg, im näheren Umland oder ganz im Sauerland zu Hause sind. Last but not least danke ich Viola für ihre Unterstützung und ihr Verständnis besonders in den letzten stressigen Phasen dieser Arbeit.

Abreviations

1D	One-Dimensional
2D	Two-Dimensional
3D	Three-Dimensional
Å	Ångström
ADMET	Pharmacokinetics: Absorption, Distribution, Metabolism, Excretion and TOXicity
AUC	Area Under the Curve
EF	Enrichment Factor
ER	Enrichment
HR	Hitrate
HTS	High Throughput Screening
IUPAC	International Union of Pure and Applied Chemistry
LBVS	Ligand-based Virtual Screening
NMR	Nuclear Magnetic Resonance
PDB	Protein Data Bank
RAISE	RApid Index-based Screening Engine
RMSD	Root Mean Square Deviation
ROC	Receiver Operator Characteristic
SBVS	Structure-based Virtual Screening
SQL	Structured Query Language
TP	True Positive
VLS	Virtual Ligand Screening
VS	Virtual Screening

Contents

Appendices

1 Chapter 1. Introduction

1.1. Motivation

Rational drug discovery aims at identifying small molecules, which interact in a beneficial way with a protein associated with a disease. Hereby, large amounts of available compounds have to be investigated systematically. The development of a new Drug is, therefore, a time consuming and expensive process. Starting from the definition of a therapeutic target until the marketing of a new drug is estimated to take between 12 and 14 years and costs approximately 1 billion dollars. [1,2]
Therefore, computational methods used in this process aim on reducing the costs and duration as well as on identifying the most promising candidates for a new drug. This process of identifying compounds, which are potentially able to interact with a certain target and influence its activity, is realized by screening large compound libraries and is called "hit" identification or lead structure search. The screening procedure can either be realized *in vitro* using high throughput screening (HTS) or *in silico* by using virtual ligand screening (VLS). In todays drug discovery processes, it is well established to use combinations of both approaches [3].
The advantage of *in silico* methods is that they are relatively fast, depending on the computational capacities of the research group, and require only little financial investment [4]. Therefore, these methods are generally used as a first filtering step on large compound libraries, and only the most promising predicted candidates are further considered for experimental evaluation [5].
The computational approaches for VLS can be divided into two broad categories, structure-based and ligand-based virtual screening. Methods for ligand-based virtual screening (LBVS) require known bioactive ligands and search for potential lead structures usually based on ligand-similarity. For structure-based virtual screening (SBVS), the three-dimensional structure of the target protein needs to be available

to which the binding affinity of compounds is predicted based on shape complementarity as well as possible intermolecular interactions.
Both categories have their own strengths and weaknesses. Therefore, the decision on which method should be used in a screening project not only depends on the available information, i.e. if only an active ligand or also the structure of the protein is known, but also on its individual capabilities. In general, LBVS allows a variety of queries based on different molecule descriptors and representations. Due to the comparably low complexity of these queries, it is a lot more efficient for the screening of large compound libraries than SBVS. However, SBVS has a higher selectivity since it also implies the restrictions of the protein structure. The high complexity of these constraints ideally allows more descriptive solutions but also restrict the practical use on large libraries. In recent years, methods for SBVS and LBVS are often combined to exploit all available chemical as well as structural information and, therefore, benefit of the strengths of both approaches. [6] Therefore, for the development of new methods, it should be considered that all available information can be used to enhance the screening performance. This presents new opportunities to address open challenges in virtual screening. For example, new methods for LBVS could combine the effectiveness of the ligand-based approach with the selectivity of constraints derived from the protein structure. Since often only certain parts of a ligand interact with a protein and are therefore important for binding, only partial constraints need to be applied to a query ligand. Another promising source for such constraints even if no protein structure is available, is the knowledge of an experienced user about the most important regions of a molecule. Ideally, a new method should provide integrated functionality to efficiently incorporate different partial shape constraints into the LBVS approach.

1.2. Overview of Content

In the following, the content of this thesis is shortly summarized.
Firstly, the current status of lead structure identification in drug design is discussed in detail in Chapter 2. Methods for this purpose can basically be divided into experimental and computational methods. The chapter will mainly focus on the computational side and discusses the advantage and disadvantage of different approaches. A consecutive analysis of the open challenges and unsolved problems in LBVS highlights the issued, which shall be addressed in the research project of this thesis.
Following this description, the aims and objectives of the thesis are listed and discussed, and the preconditions of the work are described in Chapter 3.
In Chapter 4, the used and developed methods of the new LBVS method mRAISE are described.

Then, in Chapter 5, the datasets for evaluation and comparison studies used in this thesis are described. This includes datasets from the literature as well as a newly designed dataset for the evaluation of the accuracy of molecular alignments. Furthermore, in this chapter, the performed experiments using the respective datasets are described.
Following the general description of the experiments, in Chapter 6 the used evaluation strategies as well as the used performance metrics are explained and discussed.
In Chapter 7, the results of the previously introduced experiments using the respective evaluation metrics are shown. Individual interesting cases are discussed and mRAISE is compared to other methods of the field. Furthermore, the influence of the different partial shape constrain approaches realized are analyzed.
Finally, in Chapter 8, the results and conclusions of the performed experiments are summarized and remaining problems and challenges are discussed. Further possible developments to increase the performance of mRAISE and to address remaining challenges are discussed in an outlook.

2 Chapter 2. State of the Art

In this chapter, the state of the art in rational drug discovery and virtual screening is described. The main focus hereby lies on LBVS, since this is the focus of this research project.
Section 2.1 explains the different phases of the drug discovery process and highlights the application areas of virtual screening methods. Subsequently, the basic concepts of different virtual screening approaches are described.
Finally, in Section 2.2, the state of the art in LBVS is discussed in detail. This not only includes existing methods and different approaches but also common evaluation strategies, benchmark datasets and open challenges.

2.1. Rational Drug Discovery

At the end of the 19th century, Ehrlich developed the concept of drugs binding selectively to certain receptors [7], based on the principle of lock and key introduced by Fisher [8]. This was followed by further considerable work of Langley [9] who introduced the idea of receptors as molecular switches that can be turned on and off, as well as further work of Ehrlich [10] introducing the idea of exhaustively testing variations of ligands to find new active substances. Today, this can be seen as the origin of modern drug discovery.
In the following decades, drug discovery evolved from a field depending on serendipities and individual imagination to focused research projects run by interdisciplinary teams. Further developments the 20th century of technologies like x-ray crystallography [11] and nuclear magnetic resonance spectroscopy [12, 13] allowed the scientists to investigate molecular structures on an atomic level and to this day the amount of available structural information is increasing rapidly.

Due to the growing number of solved protein structures, the drug discovery process shifted to more rational approaches for identification of active compounds. With the introduction of experimental HTS and combinatorial chemistry in the 1980s, the capabilities to explore the chemical space for potential drug candidates increased massively. Nevertheless, the high expectations and hopes on finding high amounts of new active compounds were not met [14–17] and the amount of considerable information is growing to a scale, which is hardly manageable using only experimental methods. As a consequence, nowadays computer-aided approaches play an essential role in the drug discovery process.

The modern drug discovery pipeline is a complex system divided in multiple stages and involving a variety of different technologies. Computer-based methods are a well-established part of this process. The pipeline is divided into the following stages:

- **Target-identification:** Drugs target special proteins, which are associated with the disease that should be cured, to regulate their activity. Therefore the first step in a drug discovery project is the identification of the target protein. This is mostly done by experimental evaluation. An identified protein has to be validated, to ensure its modulation has an influence on the state of the disease. Furthermore, activity assays have to be established to later test the activity of possible candidates on the respective target. If the structure of the target of interest is available or it is possible to obtain it by X-ray crystallography, NMR or homology models, this information is acquired as well.

- **Lead-identification:** In the next stage, lead structures are searched for the identified target. A lead structure is a ligand, which already has a high affinity i.e. binds strongly to the target. These can be either known, natural ligands, other reported binders from the literature or completely new compounds. Based on one or multiple known active ligands, LBVS can be used to screen compound libraries for new lead structures which are easier to synthesize, directly obtainable by vendors or already show higher affinities than known ligands. If the structure of the protein is known, SBVS can also be applied for this purpose. If no active ligand is known for the target, HTS can be used to screen compound libraries in search for potential lead structures.
  Hits identified by VS methods still have to be validated experimentally. However, the use of these methods allows the evaluation of much larger compound libraries and experimental methods only have to focus on small subsets of the libraries which are considered as promising hits by the computational methods.

- **Lead-optimization:** The optimization of lead structures in terms of activity, selectivity and the very important ADMET properties (absorption, metabolism, excretion and toxicology) is a very important part of the drug discovery process. Also in this stage VS methods can be used, for example to predict ADMET properties or to predict the affinity as consequence of certain substitutions (QSAR).
- **Clinical trials:** The last phase of the drug discovery process are pre-clinical and clinical studies. The first test drugs in animal models and the second perform actual clinical studies with human patients. If this phase is successful, a new drug can be introduced to the market.

As can be seen, the use of computational methods highly depends on the available information, which determines what methods are applicable at a certain stage. In general, computational methods can assist the drug discovery pipeline in multiple ways: binding mode elucidation, active cite prediction and analysis, binding site comparison, lead identification, lead optimization, and de-novo design.
VS hereby mainly assists in the lead identification, providing efficient methods for a rapid screening of large compound libraries without the need of a high resolution protein structure. This drastically decreases the costs for experimental evaluation and synthesis of inactive compounds.

In the following, VS technologies in general and subsequently, the special field of LBVS will be described.

2.1.1. Virtual Screening

Virtual screening has been defined by the International Union of Pure and Applied Chemistry (IUPAC), as a process which selects compounds based on the rating scheme of an underlying computational model. [18] Methods for VS can be divided into multiple categories depending on criteria like the required input information and the requirements for hits in the screening library. The most common classification of VS methods divides the field into LBVS methods, searching for compounds similar to a query ligand and SBVS methods, searching for compounds fitting into the targets binding site. The first require an active ligand as input and the second requires the structure of the target protein. Both approaches can be applied for lead structure identification without any other preconditions or restrictions and are discussed in the following in more detail.
In recent years, multiple attempts have been started to combine LBSV and SBVS methods to exploit all available information and benefit from the advantages of both approaches [19–25]. These approaches use methods from broth areas either in an hierarchical oder parallel manner. In hierarchical concepts, the faster method is

used as first filter and the computational more expensive method is used on the most promising results of the previous method. In parallel approaches, the methods are executed on the same data and hits are determined either complementary (taking the best hits from each methods) or consensual (taking the best hits found by both methods). Although the results of this concept are mixed [26], the idea of exploiting all available information holds significant potential [6].

Ligand-based Virtual Screening

Ligand-based approaches focus on the identification of lead structures based on their similarity to a known active ligand for the target of interest. Therefore, these methods work without the need of a solved protein structure.
The focus of this work lies on LBVS methods working with three-dimensional ligand structures: In the first step, the actual molecular alignment of a query molecule and compounds of the screening library are calculated. This is a complex problem and multiple different approaches to achieve meaningful alignments have been evaluated. This calculation includes many degrees of freedom like translation, rotation and the flexibility of the ligands. Based on the calculated alignments, the similarity of the ligands is quantified using so called scoring functions. Since mRAISE is a method for LBVS, different methodologies for this approach, as well as common evaluation strategies and datasets are presented in more detail in Section 2.2.

Structure-based Virtual Screening

In SBVS, new lead structures are searched based on their complementarity to the binding site of the target protein. This process therefore requires high quality protein structures either generated by experimental methods or homology models. SBVS methods need to address two problems: First, the ligands of the screening library have to be docked into the active site with respect to sterical fit as well as the creation of favorable protein-ligand interactions. Secondly, the calculated binding poses have to be scored using a mathematical algorithm, i.e. scoring function to predict the binding affinity of the ligand. Based on the ranking of all compounds with respect to their calculated scores, top ranked hits are then selected for further experimental evaluation.

For docking methods, it is best to separately analyze the part of the methods generating the ligand poses in the binding site and the scoring functions for affinity prediction, since both parts show a variety of different approaches.

Search Strategies

- **Methods using multiple ligand conformations:** These algorithms address the molecular flexibility of the compounds by previously generating multiple conformations for the molecules of the screening library. During the actual docking procedure, both the ligand and the binding site are then handled as being rigid, which eliminates multiple degreed of freedom during the calculation of ligand poses. Methods following this approach, like DOCK 3.0 [27], FLOG [28], FRED [29], and TrixX-BMI [30] calculate transformations placing the compounds into the binding site based on a simplified representation of the ligands as well as the descriptor. These representations are for example graphs, atom-centered Gaussian functions, pharmacophoric features or surface descriptions. Complex algorithms are used to detect matches between the respective descriptors and transformations are calculated accordingly.

- **Fragment-based methods:** Fragment-based methods create possible ligand conformations iteratively within the protein binding site and therefore try to avoid the generation of clashing conformations. Therefore, all molecules of the compound library are initially fragmented by cutting the molecules at each rotatable bond. Algorithms then either the place one initial fragment into the binding site and then connect the remaining parts incrementally while evaluating multiple different torsion angles (e.g. FlexX [31], HAMMERHEAD [32], and DOCK 4.0 [33]), or try to fit all fragments into the binding site and connect them afterwards (e.g. SURFLEX [34], EHITS [35]).

- **Stochastic methods:** Stochastic methods use optimization algorithms to ideally find an optimal solution for the placement of a ligand into the binding site at minimal energy costs. Based on the target function, this can allow different degrees of freedom up to full flexibility of all structures. The algorithm then tries to vary parameters at random to find the best possible solution (global minimum). A variety of optimization algorithms has been used for this purpose in the past including Monte-Carlo methods (ICM [36], QXP [37]), genetic algorithms (AUTODOCK [38], GOLD [39]) and ant colony optimization algorithms (PLANTS [40]).

- **Simulation methods:** The simulation of molecular dynamics with algorithms like MD-simulations are extremely complex and time consuming and therefore usually can not be used for real VS experiments. They can however be used for post optimization of individual poses.

- **Hierarchical methods:** Hierarchical methods use combinations of the above described approaches. An example of a method combining systematic with stochastic methods is GLIDE [41].

Scoring Functions

- **Empirical scoring:** Empirical scoring functions try to estimate the Gibbs free energy based on a simple sum of uncorrelated energy terms. These terms for example estimate the contributions of hydrogen bonds, ionic interactions and lipophilic contacts. An empiric scoring function is for example used in FlexX [31].
- **Knowledge-based scoring:** Knowledge-based scoring functions try to estimate the affinity based on observations in studied protein-ligand complexes. Frequently observed contacts between pairs of ligand and protein atoms are considered as favorable for the binding energy and based on the frequency distribution of observed contacts, distance dependent pairwise atom potentials are calculated. For new ligand pose by a docking algorithm, a score is calculated as the sum of all pairwise potentials between protein and ligand atoms. An example for such a scoring function is the potential of mean force[151].
- **Forcefield-based scoring:** Forcefield-based scoring functions are used for example in GOLD [39], GLIDE [41] and AUTODOCK [42] and usually quantify the sum of the protein-ligand interaction energy and the internal ligand energy by classic potentials of interactions between individual atoms.
- **Consensus Scoring:** Consensus scoring again combines different scoring functions to overcome weaknesses of the individual scoring functions and re-ranks the hits accordingly.

2.2. Ligand-based Virtual Screening

Like VS in general, LBVS can be further divided into different approaches based on the required data required in order to use the respective methods. For example, Sheridan and Kearsley, actually divided VS into 4 categories, namely docking (SBVS), similarity searching, QSAR methods, and substructure search [43]. While the first category has already been discussed, similarity searching will be the actual focus of this section, since the research project of this dissertation focused on the development of a method for similarity searching based on one known, active ligand. Furthermore, similarity searching is of special interest, since it generally is the most widely used approach in VS [6].

QSAR and substructure search can to a certain degree also be applied to lead structure identification, but require a lot more previous knowledge of known actives with measured affinity values or of special substructure patterns that are mandatory for the ligands activity. Therefore, both approaches are only briefly

described in the following, since they are not in the focus of this work. Another recent trend in LBVS, which requires high amounts of data of active ligands are machine learning approaches. These trends will also be briefly described in the following:

- **QSAR:** The abbreviation QSAR stands for Quantitative Structure-Activity Relationship. QSAR methods try to quantify the relation between the structure of physicochemical properties of a ligand with its bioactivity [44]. The underlying models have to be trained on a preferably large set of known actives with measured affinity values in order to subsequently predict the affinity of other compounds. Since the required preconditions for QSAR methods are quite high, they are best suited for already well studied targets or later stages of the drug discovery process like the lead optimization.

- **Substructure search:** Methods searching for special substructures in a compound library can be applied if a special scaffold is known, which is directly related to the affinity of a ligand. Depending on the complexity of the substructure, this will most likely results in close analogues of the query structure and the found hits are likely to be active to the same target as well. [45] Since such information is not necessarily available at the first stages of a drug discovery process, these methods can not be generally applied for lead structure identification. Furthermore. and the aim of virtual screening initially is the identification of structurally diverse compounds as lead structures [46].

- **Machine learning:** In recent years, the amount of publicly available bioactivity data has increased significantly, this led to a rising interest in data mining and machine learning algorithms to make active use of this data [46]. The general hope is to be able to train models based on known active and inactive compounds using different molecule descriptors and to subsequently used these models to predict the likelihood of other molecules to be active as well. The most common methodologies applied in this field are support vector machines, Bayesian methods and decision trees. The quality of the trained models depends on factors like training set diversity and the ability of parameters to cover the active and inactive chemical space [47]. Nevertheless, machine learning models capable of screening large compound libraries can be developed if datasets of sufficient size and diversity are available [48–50].

2.2.1. Similarity Searching

As mentioned earlier, similarity searching is the main focus of LBVS and the most applied VS approach in general. The key advantage of this approach is that it does not require a structure of the target protein and is applicability using only one known, active ligand as input. In general, similarity searching follows the assumption that globally similar compounds are most likely to show the same bioactivity [51].

In the following, different approaches for ligand-based similarity searching are presented. The methods are hereby divided into alignment-free descriptor-based methods and alignment-based methods.

Descriptor-based Similarity

Alignment-free descriptor-based approaches usually encode multiple different molecular properties in fingerprints, i.e. binary feature vectors. The similarity of two molecules based on fingerprint representations can then be quantified according to matching and mismatching features using different metrics like the Tanimoto similarity or the Euclidean distance. Based on the utilized information, these methods can be further divided as follows:

1D Methods: One-dimensional properties can be derived directly simple countable features, which can be directly derived from the structural formula of by summing up properties of the individual atoms. Examples for one-dimensional properties are the number of heavy atoms, the number of rotational bonds, the number of potential hydrogen-bond forming groups, the molecular weight or the logP value of a molecule. These properties are usually used for basic filtering of compound libraries, for example to retrieve only leadlike compounds following the rules of Lipinski [52] or Oprea [53].

2D Methods: Two-dimensional methods describe the topological connectivity of a molecule. This can for example be used in form of canonic SMILES descriptions and SMARTS pattern matching, to identify and filter unwanted structures or substructures in a screening library. Other approaches compare molecules using fingerprints registering the presence or absence of certain previously defined molecular fragments like MACCS and BCI descriptors [54]. More complex topological fingerprints like the Daylight fingerprint [55] enumerate all substructures up to a certain length present in the molecule. The ECFP (Extended Connectivity Fingerprint) [56], encodes the atom types, charges and connectivity information of

the circular surroundings of individual atoms. ECFPs can achieve high hit rates, but due to using topological descriptors, the results lack structural diversity [57,58]. To increase the diversity of the results, FCFPs (Function Class Fingerprint), a variant of the ECFP, abstract the atoms to functional features like hydrogen bond acceptors and donors, aromatic and halogens. Finally, the CATS descriptor [59] annotates all atoms of a molecule with one of five functional features and measures pairwise distances between atoms with the same features. The resulting distance histograms can also be encoded in a topological fingerprint. The advantage of the actual structure to more functional representations has the advantage to be less strict concerning the topology and therefore allows the discovery of new scaffolds during VS.

3D Methods: Three-dimensional methods depend on the conformation of the individual compound because they exploit the information of atomic coordinates. Therefore, the problem of molecular flexibility has to be addressed in these methods. This can either be done by enumerating possible conformations of a molecule or by flexible alignments procedures, which adept the atomic coordinates as needed. Three-dimensional descriptors can for example be numeric like the van der Waals volume, the molecular-, and polar surface area or pairwise distances between atoms. An interesting recent approach for alignment-free similarity-based LBVS is **LisiCa** (ligand similarity using clique algorithm [60], which can use either 2D or 3D representations of the molecules as input. In general, the method represents molecules as graphs and calculates the similarity of the molecules based on the size of the maximum common subgraph (MCS) determined using maximum clique detection on a calculated product graph. While for 2D representations, the graphs represent atoms as vertices and bonds as edges as usual, in 3D representations, edges are placed between all pairs of vertices and annotated with the pairwise distance of the respective atoms. Therefore, this algorithm allows the detection atom mappings between the molecules in 3D space.

Molecular Alignments

While 1D methods only rate the similarity of molecules based on simple physicochemical properties and 2D methods can only incorporate the topology of the molecules, 3D methods have the opportunity to compare the actual structural similarity in 3D space. However, the structural similarity can only be evaluated based on molecular alignments of the query ligand and the compounds of the screening library. The task to find optimal alignments of two different molecules is a challenging problem and a variety of methods has been developed to address this problem by different ways of molecular representations and scoring schemes. To incorporate the molecular flexibility, which is a crucial part in order to find

meaningful alignments, either alternative conformations have to be generated for the compounds followed by a rigid alignments process, or the alignments algorithm has to somehow allow flexible alignments. In the following, an overview of multiple different methods and alignment approaches is given. The focus of the list lies on recently developed methods as well as the methods used for the comparison of mRAISE. The presented methods are divided into rigid and flexible approaches:

Rigid alignment methods:

ROCS (Rapid Overlay of Chemical Structures) [61, 62], abstracts molecules to a continuous description of the molecular shape using atom-centered Gaussian functions for all heavy atoms of the molecule. Initially four differently oriented starting alignments are created by superimposing the "shape centroids" of the molecules. Starting from these orientations, the alignments are optimized by rotation operations around different axis with respect to a scoring function using the Gaussian-based shape description to maximize the volume overlap of the molecules. During this process, similarity and dissimilarity of user-defined physicochemical properties of the underlying atoms can be incorporated. This includes charges, hydrogen-bond donors and acceptors, hydrophobicity as well as ring membership.

SimG [63], is a similar approach to the previously introduced ROCS. Additional to the basic concepts of ROCS, SimG is also capable of structure-based virtual screening by deriving a shape representation from protein binding sites and matching ligands into this shape. Furthermore, the scoring function used for the alignment optimization differs, besides a term describing the volume overlap of the molecules, a second scoring term has been introduced representing the sum of aligned similar pharmacophoric features of the underlying atoms with respect to the total number of features.

SHAEP [64] combines the strengths of shape-based approaches with that of field-based approaches. For the calculation of the initial alignment, two molecules are represented as graphs encoding the electrostatic potential as well as the local shape at points near the molecular surfaces. The maximal common subgraph between both graphs is then used as basis for the alignment. A subsequent optimization is then again performed to optimize the volume overlap of the molecules with respect to a Gaussian based shape description.

Other methods following the general principals of Gaussian-based shape representation followed by rigid volume overlap optimization are for example **Align-It** [65] and **MolShaCS** [66].

In **LigMatch** [67] a geometric hashing algorithm is used to calculate possible alignments. Herein molecules are represented by a set of triangle descriptors with

atoms as corners and interatomic distances as side lengths. All matching triangles with the same atom types as corners and similar distances are then used to align the molecules. An alignment is scored with respect to the number of coincident atoms in an alignment.

Flexible alignment methods:

LIGSIFT [68], is a very recent approach for flexible alignments using a Gaussian-based shape descriptions of the molecules. In this case, the initial alignments are generated by aligning the molecules with respect to their principal axes of the moment of inertia tensors and by aligning enumerated triplets of atoms with similar chemical nature. Based on the sub-optimal initial alignments, a flexible optimization is performed using a Metropolis Monte-Carlo simulation.

In **Screen3D** [69] an initial alignment is calculated by calculating the largest possible mapping of ligand features based on atom-atom distance histograms. To obtain the best possible starting alignment, the molecules are aligned using an RMSD minimization algorithm on the respective atom mappings. Starting from this initial alignment, an optimization algorithm is used to again maximize the volume overlap of the aligned molecules. The method can perform this optimization either flexible, allowing changes in the rotatable binds of the molecule during the optimization of completely rigid. In the second case, initially a certain number of random conformations of one ligand are enumerated.

FlexS [70] is a LBVS approach based on the flexible SBVS method FlexX. Like in the previously described docking algorithm, the compounds of the screening library are divided into fragments by cutting them at rotational bonds. Based on an initially aligned fragment, the remaining fragments of the compound are iteratively reassembled during the search process guided by a similarity-based scoring function. Each assembly step hereby tries different torsion angles to account for molecular flexibility.

Surflex-sim [71] is another adaptation of an SBVS method for the purpose of LBVS. it uses a modified version of the fragmentation and reconstruction algorithm used in HAMMERHEAD for a flexible alignment of rigid fragments onto an also rigid query structure. The key feature of this method is the representation of the query molecule as surface points, to which the fragments are aligned during the reconstruction phase and scored according to the achieved accuracy of the alignment.

pharmACOphore [72] is an LBVS method based on the previously introduced SBVS method PLANTS. As in the docking approach, ant colony optimization is used to calculate the most favorable alignment according to the scoring function by

allowing translational and rotational movements within a defined sphere around the query structure. The algorithm furthermore allows flexible superimposition onto a rigid query structure by adapting the rotational bonds of the target structure during the optimization. For the scoring function, each atom is assigned with a pharmacophoric feature and all correlating pharmacophoric features contribute to the score based on their pairwise distance.

ICMsim [73] is the LBVS version of the ICM [74] approach also calculates fully flexible alignments using a Monte-Carlo simulation for the optimization of an initial alignment.

Pharmacophore Matching

The search for lead structures based on pharmacophore concepts can actually either be based on ligand or protein-ligand complex structures. Nevertheless, it is usually associated more with LBVS than SBVS [6], because it does not necessarily need a protein structure as input, although it can benefit from the additional information. A pharmacophore is an abstract model representing the essential features of a ligand to bind to a certain target. It is defined by IUPAC as a set of electronic as well as steric features, which is required to trigger or block the biological response of a compound [18].
The abstract definition of pharmacophoric features is independent from the actual molecular structure, since it encodes features and not special functional groups. Represented features can for example be hydrogen-bond interactions or lipophilic areas. A pharmacophore can either be derived based on common features of aligned active ligands or based on complementary features between the ligand and the target protein. A once derived pharmacophore can then be encoded using descriptors and is then used as input to a similarity search on the screening library. Prominent methods for VS using pharmacophore matching are for example Phase [75], Pharmer [76] and LigandScout [77].
Like all methods based on 3D structures, pharmacophore matching is conformation dependent and relies on strategies to handle molecular flexibility.

2.2.2. Evaluation and Benchmarking

To ensure the predictive power of VS methods and to allow a user to make a reasonable decision on which method to choose for a drug discovery project, evaluation studies have to be performed. Ideally, these studies should be performed with a consecutive experimental verification of the binding affinity of predicted hits. [78] However, the necessary resources for such evaluation studies are often not available. [79]

The alternative for the evaluation of VS methods is the execution of retrospective experiments allowing the assessment of two important criteria, the enrichment of active compounds in the respective data as well as the accuracy of the predicted binding modes. [80]

Another important aspect for the introduction of a new methods to the field of LBVS is besides the general evaluation of its capabilities in those aspects, an extensive comparison to other state of the art methods in order to proof its value for the field. Critical for such comparison studies are not only commonly used datasets, but also the identical preparation and handling of the respective data as well as an accurate documentation of the experiments.

2.2.3. Evaluation Data

In general, benchmarking datasets for virtual screening methods include two types of compounds for one or multiple targets. Firstly, a dataset needs to include active compounds with a known, documented activity to the respective target and secondly, a large set of inactive compounds. In general, the inactive compounds are only assumed not to bind to the respective targets (decoys), since validated inactive compounds are seldom reported in the literature and therefore generally not available in the necessary quantities [80]. An overview of available benchmark datasets for VS enrichment studies can be seen in Table 2.1.
To evaluate the quality of molecular alignments calculated by 3D LBVS methods, special datasets of prealigned ligand ensembles are required. Such ensembles consist of different ligands binding to the same protein. The reference alignment poses of those ligands are obtained indirectly by superimposing the identical binding sites along with the bound ligands. Ideally, LBVS methods should be able reproduce the reference alignments when comparing ligands of the same ensemble. This evaluates their capability on aligning the most important similar regions of similar ligands and to reproduce the conserved binding mode of the members of the ensemble needed to bind to their common target.

In contrast to the datasets composed for enrichment evaluation, no recent, commonly used benchmarking datasets existed for the purpose of alignment quality evaluation at the start of this project. As a matter of fact, the evaluation of this aspect of LBVS was rarely performed in recent publications introducing new methods and if it was addressed, most of the time only small datasets were used, which were not publicly available afterwards.

Table 2.1.: Average AUC values for all DUD targets.

dataset	publication year	targets	actives	decoys
Bissantz *et al.* [81]	2000	2	20	1980
McGovern *et al.* [82]	2003	10	2200	95579
Diller *et al.* [83]	2003	6	958	32000
Lorber *et al.* [84]	2005	7	2201	98500
Irwin *et al.* [85]	2005	5	862	95579
Miteva *et al.* [86]	2005	4	49	65611
Pham and Jain [87]	2006	29	226	1861
DUD [88]	2006	40	2950	95316
DUD-E [89]	2012	102	66695	1420433
GLL - GDD [90]	2012	147	25145	980655
DEKOIS 2.0 [91]	2013	81	3240	97200
NRLiSt BDB [92]	2014	54	9905	458981
MUDB-HDACs [93]	2015	14	631	24609

Data taken from [94]

2.3. Open Challenges

Despite the large amount of available methods introduced in recent years using a variety of different approaches, the field of LBVS still remains interesting for future developments and there are still open challenges, which can be addressed.

Firstly, the alignment problem is not yet solved in a completely satisfying manner [95]. Algorithms following flexible approaches are usually slow in comparison and therefore not as practical as other methods for lead-structure identification in large compound libraries. Efficient ways have to be developed achieving highly accurate alignments without exhaustive flexible optimization algorithms, for example by somehow discretizing the space of possible alignments by meaningful descriptors with a high likelihood to produce nearly optimal alignments.

A topic related to the quality of the calculated alignments, is the general handling of molecular flexibility. Molecular alignments can only be as good as the available conformations of the compared molecules. Solutions for this issue might be better methods for the enumeration of conformations, faster flexible optimizations or general concepts matching molecules without the need of global shape similarity.

Another remaining challenge in the field of LBVS is the complex topic of partial shape matching [96], which could also address other problems like the dependency on similar molecular conformations for screening compounds in order to find meaningful alignments. The focus only on the most important parts of a ligand,

responsible for its bioactivity could result in more diverse but still active hits and allow the user new ways to influence the screening procedure as desired.

Despite the fact that the number of targets for which both, an active ligand as well as a 3D protein structure assist, combinations of LBVS and SBVS approaches are still quite rare [6]. New methods could introduce new ways to make use of all available information for VS.

It is also important to note that the problem of correct similarity-based compound ranking is also far from being solved and there is still a need for better scoring functions for the ranking of screened compounds.

Generally, new methods have to face the classic challenges of providing high quality results in reasonable amounts of time in order to be applicable for real VS projects It is therefore of great importance that this is proofed in comparison to other methods of the field.

3 Chapter 3. Research Aims and Preconditions

Based on the overview of the state of the art of LBVS in the preceding chapter, the following chapter will define the main aims and objectives of this dissertation. A special focus hereby lies on the obstacles and open challenges of the field, which shall be addressed in this project. Furthermore, the preconditions at the beginning of the project will be shown.

3.1. Aims and Objectives

The main aim of this dissertation project is the development of a new computational method for LBVS that provides the ability to define special partial shape constraints. Furthermore, the method should be able to compete with state of the art LBVS methods both in terms of screening result quality as well as computing time requirements.

The analysis of the literature has shown how competitive the field of LBVS is but at the same time pointed out that there are still open challenges in the field. Therefore, primarily a new method that is introduced to the field has to compete with existing methods both in terms of screening capabilities as well as computing time. Furthermore, for the purpose of this project, the challenging topic of meaningful partial shape matching will be addressed.
Thus, for this research project, the following obstacles are defined:

- **Efficient data handling:** For the performance as well as the reliability of a LBVS method, it is important to handle the compounds of a screening

library consistently and efficiently, including the calculation and storage of conformations.

- **Meaningful abstraction of ligand information:** An efficient to handle representation of the most important features of a ligand is needed to enable a rapid comparison of complex three-dimensional structures. Furthermore, in a pre-processing step, this representation allows the elimination of obviously dissimilar candidates in the compound library.

- **Knowledge-based partial shape constraints:** The integration of meaningful partial shape matching in LBVS is a difficult task, since the constraints have to be created based on accurate knowledge of the most important regions of a query ligand.

- **High quality three-dimensional alignments:** Accurate molecular alignments are an important part of 3D LBVS methods. Only good alignments allow an accurate similarity rating by superimposing the most important common features of two ligands.

- **Scoring of molecular similarity:** A detailed scoring function needs to capture the most important similarities as well as the dissimilarities of two aligned ligands. The similarity score should provide an accurate ranking of compound library hits.

- **Evaluation data:** New methods need to be evaluated on reliable datasets. Herby, highlighting the strengths and weaknesses of a method is just as important as the comparison to other state of the art methods. However, there is no diverse dataset of statistically sufficient size available in the literature that could be used as standard for the evaluation of alignment quality.

- **Usability:** In addition to the basic development of a new method, also the usability for potential users is of great importance. Therefore, the software should be easy to use and assist the user in all important steps of its application. While a good, easy to use command line interface might be sufficient for general screening runs, a graphical user interface becomes necessary as soon as constraints can be derived manually. Furthermore, a graphical user interface allows the visualization of results for evaluation purposes as well as of query ligands and the respective descriptors to get a better understanding of the method.

3.2. Preconditions and Course of the Project

The work at hand was prepared at the Center for Bioinformatics (ZBH) at the University of Hamburg in the research group for Computational Molecular Design from February 2012 to November 2016.

Herein 'mRAISE' is introduced, a new software for LBVS. Fundamentally this work is based on algorithms developed for indexing and screening in Trixx BMI by Dr. Jochen Schlosser in his dissertation [30]. These algorithms were reimplemented by K. Schomburg, S. Urbaczek and A. Henzler at the beginning of this dissertation as the 'Trixx' and the 'FastBitIndex' libraries. This includes the functionality for interaction point calculation and descriptor generation. Furthermore, this work makes use of the MoleculeDB as developed by M. Hilbig for MONA [97] and the 'NAOMI' library developed at the ZBH and the BioSolveIT (www.biosolveit.de). As external libraries, mRAISE uses Qt (http://qt-project.org) and FastBit [98]. Qt was used for the development of a graphical user interface for mRAISE, which uses the 3D visualization library developed by the BioSolveIT also utilizing Qt.

The mRAISE method has been published in the *Journal of Computer Aided Molecular Design* in one publication that is already published as of writing this thesis and a second, which is still in reviewing status. Furthermore, the results have been presented at an international conference in form of a poster. A list of all publications and posters can be seen in Appendix F.

Chapter 4.

4 Methods

In the following chapter, the different methods used in mRAISE are described. A special focus lies hereby on the aspects that have been either adapted or newly developed for the ligand-based setup in mRAISE.
The basis of mRAISE is formed by modules that were already available in libraries developed at the Center for Bioinformatics Hamburg (see Section 4.1). Utilizing those libraries made it possible to focus the development mainly on the special challenges of LBVS as well as on the new concepts for partial shape constraints. The libraries provide a wide range of functionality including, among others, functions to read and write molecule and protein files, a chemically accurate digital representation of molecular structures, a generator for alternative conformations of small molecules, databases for the efficient storage of molecules alongside associated information, and a basic framework for three-dimensional visualization.
Following the basic components, a general description of the workflow of mRAISE is given (see Section 4.2.
The key technologies used in mRAISE are a triangle descriptor representation of the molecules specially developed for virtual screening and a bit encoded index to efficiently store and access those descriptors. Therefore, the descriptor as well as the index are described in more detail in Section 4.3. This is followed by the required adaptations to the descriptor for the purpose of LBVS (Section 4.4), and new developments for partial shape constraints (Section 4.5). Afterwards, another important part for a LBVS method, the scoring function for the evaluation of molecular similarity based on structural alignments, is introduced (Section 4.6). Finally, the output formats of mRAISE are explained (Section 4.7) and a newly developed graphical user interface (GUI) for the visualization of screening results as well as the manual definition of query constraints is shown (Section 4.8). A user guide for the command line version (Appendix C) as well as for the GUI version (Appendix D) of mRAISE can be found in the Appendices.

4.1. Basic Libraries

This section introduces the existing basic concepts and modules that have been used in mRAISE.

4.1.1. NAOMI

The NAOMI library developed at the Center for Bioinformatics Hamburg provides the basic functionalities needed to work with chemical structures in a chemoinformatical context. This includes the initialization of molecule and protein structures from various available file formats as well as the functionality to write the initialized structures back into such formats. During the initialization, the molecules and proteins are converted into complex data structures following a very strict chemical model. As a result, files containing chemically invalid information are discarded to maintain the integrity of the model. A further limitation of the NAOMI library is that it can not initialize ligands with covalently bound metals. [99,100]

4.1.2. Conformation Generation

To take into account the flexibility of molecular structures during virtual screening, the most common approach is the generation of multiple conformational representations of the molecules in the compound library. In mRAISE, conformations are generated automatically using the knowledge-based CONFECT [101] algorithm. This algorithm enumerates conformations for a given molecule by assigning new torsion angles to rotatable bonds based on a torsion library which has been derived from crystallographic data [102]. This way, the accessible conformational space is explored using only the most likely and, therefore, low-energy torsion angles. For the sampling of ring conformations, CONFECT uses a set of precalculated forcefield-optimized templates. After the enumeration of conformations, a conclusive clustering reduces the number of generated structures down to a desired ensemble size while trying to maintain the diversity of the enumerated ensemble. In the latest iteration of CONFECT, as introduced in UNICON [103], the torsion library has been refined [104] and a new RMSD-based clustering algorithm has been integrated to further increase the diversity of the generated ensemble. This version, furthermore, introduces three quality levels for the conformation generation and the most accurate (quality level 3) is used as default in mRAISE.

4.1.3. MoleculeDB

During the development of the molecular filtering tool MONA [97,105], an SQLite database for the efficient storage and retrieval of molecules has been developed. For the purpose of filtering chemical libraries in MONA, additionally physicochemical properties of the molecules are calculated and stored as associated data to the molecule entries. However, the implementation of the database also allows to add further information or to plainly store the molecules.
This so called MoleculeDB has consecutively been used in other software developments working with large molecule libraries like cRAISE [106] and now mRAISE [107]. The primary representation of a molecule in the database is the so called MolString, a canonized string-representation of all atoms and bonds of the molecule which is suited as an identifier for molecules with the same topology. As a consequence, the database differentiates between unique molecules and additional instances with the same topology as an already registered molecule. If an entry with the same MolString exists in the database, the new molecule is simply inserted as an additional instance. Therefore, only the alternative coordinates of the atoms are stored and a special ID refers to the existing molecule entry to which this instance belongs. The MolString representation together with a BLOB (Binary Large Object) encoding the Cartesian coordinates of all atoms and some additional data like the molecule's name allows a complete and fast reinitialization of the stored molecules. Besides the usage as efficient storage for the molecules of the screening library, the MoleculeDB has also been used to optionally store screening results for later reexamination and visualization.

4.1.4. Three-dimensional Visualization

To visualize screening results as well as to explore and define query ligands and partial shape constraints, mRAISE does not only provide a command-line interface but also a GUI. Herein, query molecules, descriptors, structural alignments, and molecule surrounding residues derived from a complex structure can be visualized in three-dimensional space.
This visualization has been realized utilizing a component available at the ZBH and developed by the BioSolveIT, designed for user interfaces created with the cross-platform GUI development framework Qt(http://qt-project.org). This component is a special implementation of the Qt class QWidget and allows to easily create areas in a GUI to display and interact with three-dimensional structures. Together with this QWidget a multitude of utilities are provided for example to create visualizations of molecules and proteins directly from the internally used data structures as well as to visualize a variety of geometric primitives like additional spheres and lines.

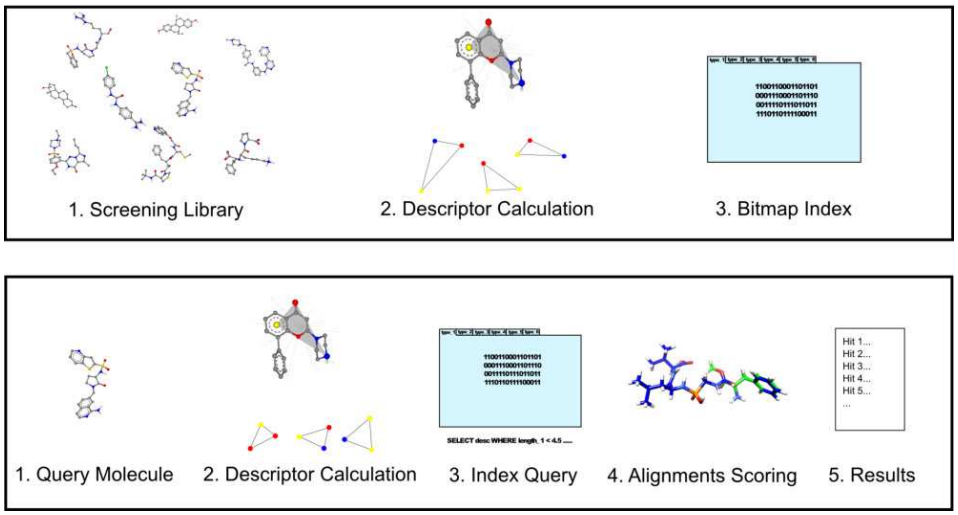


Figure 4.1.: Overview of the mRAISE workflow divided in registration (above) and screening (below). Reprinted from [107] with permission of Springer.

4.2. mRAISE Workflow

The general workflow of mRAISE and all other applications using the TrixX technology as developed in TrixX-BMI [30] and now RAISE, can be divided into two basic steps, a registration phase and a screening phase. An overview of the mRAISE workflow can be seen in Figure 4.1.

4.2.1. Registration

In the registration phase, molecules of a compound library are initialized and the descriptor index is created as described in Section 4.3.3 for efficient screening. In detail this includes the following steps:

1. Screening Library
    - Molecule initialization from an input file
    - Generation of additional conformations for each input molecule (if desired)
    - Storage of all molecules and conformations in a MoleculeDB
2. Calculation of descriptors for each structure
3. Creation of a descriptor index

4.2.2. Screening

During the screening phase descriptors are calculated for a query molecule and then matched against a previously prepared descriptor index as described in Section 4.3.4. Matching molecules are aligned to the query molecule and the similarity of the molecules is scored. In mRAISE the phase can be divided into the following steps.

1. Initialization of the query molecule from an input file
2. Calculation of query descriptors
3. Matching of each query descriptor to the entries of the descriptor index
4. Alignment of the query molecule and compounds based on matching descriptors
5. Calculation of the similarity score for each alignment and storing the best score per conformation

4.3. TrixX Methodology

The fundamental basis of mRAISE is the so called TrixX triangle descriptor which has initially been developed for the purpose of structure-based virtual screening by Schellhammer and Rarey [108]. Later, the descriptor has been extended by a representation of the local surrounding shape and a bitmap-based index representation to efficiently store and compare descriptors of preprocessed compound libraries [30]. In its latest iteration for structure-based virtual screening this technology has been used in a tool called cRAISE [106]. This version is now based on the NAOMI framework (see Section 4.1.1), and provides new possibilities to efficiently guide the screening process. Besides this, the TrixX technology has also been successfully applied to other areas of virtual screening like protein binding site comparison in TrixP [109] and inverse protein-ligand screening in iRAISE [110]. Despite the use of this technology in other areas of virtual screening, the name TrixX is mainly associated with the original structure-based application. As a consequence, new developments refer to the descriptor index-technology as RApid Index-based Screening Engine (RAISE) and the respective tools are named referring to this abbreviation: cRAISE [106], iRAISE [110] and now mRAISE [107].
In the following, the descriptor and its application for virtual screening will be explained in detail, since mRAISE only operates with ligands, the descriptor generation for proteins will not be explained here.

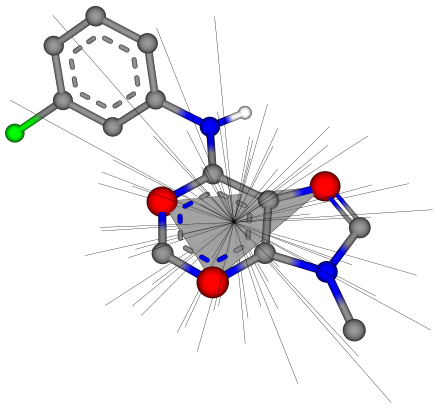


Figure 4.2.: Visualization of a TrixX descriptor derived from the cdk2 kinase query ligand of the DUD dataset. The displayed descriptor is based on three hydrogen bond acceptors highlighted as red spheres.

4.3.1. Descriptor

For the purpose of virtual screening, the abstraction of highly complex structures using descriptors encoding only important features reduces the complexity of the comparison problem and therefore enables rapid screening of large libraries. Comparing descriptors alone of course is only a very coarse measure of molecular similarity, but on the other hand it already excludes obviously dissimilar combinations. Therefore, the comparison of descriptors is usually done as a first step and is then followed by a more complex and time-consuming evaluation procedure.
The TrixX descriptor is based on a triangle descriptor, i.e., three-point pharmacophore, annotated with additional information for an efficient but also meaningful abstraction of molecular structures. An exemplary TrixX descriptor is depicted in Figure 4.2. In general, the descriptor includes the following information:

- **Triangle corners:** A descriptor is defined by three so called interaction points. These points indicate spots where interactions between a ligand and a protein might occur and can be of type hydrogen bond donor, hydrogen bond acceptor or hydrophobic. The determination of the interaction points for a molecule is described in Section 4.3.2.

- **Coordinates:** For the matching of two descriptors the types of the interaction points are sufficient, however, for the superposition of the respective molecules, the coordinates of the interaction points are needed as well. Like in iRAISE, these coordinates are stored with the descriptor. This increases the space requirement of a stored descriptor but at the same time avoids recalculation of the descriptors to receive the information needed for the superposition.

- **Directions:** For polar interaction points, the descriptor also stores the potential interaction directions. For details on the determination of these directions and their representation in the descriptor see Section 4.3.2.
- **Side lengths:** The side lengths of the triangle descriptor correspond to the distances between the interaction points and are also stored in the descriptor.
- **Shape representation:** The shape of a molecule is represented by the lengths of 80 canonized rays radiating from the center of the triangle. This part of the descriptor plays an important role for partial shape matching during virtual screening with mRAISE and is described in detail in Section 4.3.2.
- **IDs:** Since all molecules of the compound library and the calculated conformations are stored in a MoleculeDB (see Section 4.1.3), the identifiers needed to reinitialize the respective entry from the database are included in the descriptor.

To ensure that the descriptor generation is deterministic and all properties of the descriptors are directly comparable, canonization is a crucial part during the descriptor generation. For details on the canonization process see Section 4.3.2.
Each final descriptor has an associated type expressed as a number between zero and eight which is based on its combination of interaction points. A triangle type of ”0” for example indicates a descriptor with three hydrogen bond donors. A triangle of type ”1” indicates a descriptor with two hydrogen bond donors and one hydrogen bond acceptor. Triangles with only hydrophobic interaction points are discarded, this results in a total of nine possible triangle type values. The triangle type plays an important role during the creation of the descriptor index (see Section 4.3.3) and during the screening process (see Section 4.3.4).

4.3.2. Descriptor Generation

For the purpose of LBVS in mRAISE descriptors only need to be generated for small molecules. Therefore, in the following, only the ligand-based descriptor generation will be explained. An overview of the different steps of the generation process can be seen in Figure 4.3.
Descriptors are generated for all possible combinations of three interaction points of a molecule. Each resulting triangle must fulfill the following constraints not to be discarded:

- No more than two hydrophobic interaction points
- Two interaction points can not correspond to the same atom. The involved atoms have to be separated by at least two bonds.

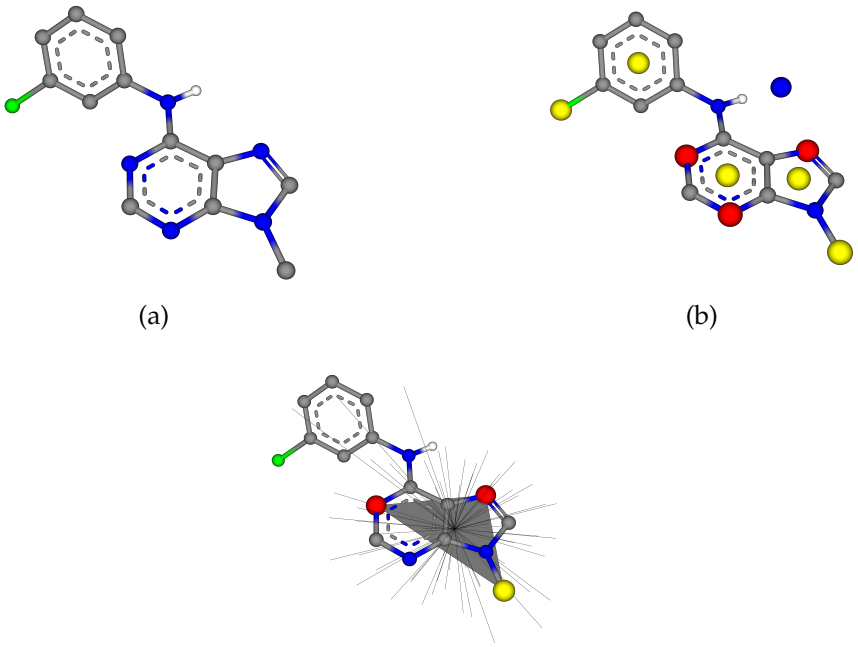


Figure 4.3.: Visualization of the descriptor generation using the cdk2 query molecule of the DUD dataset. a) Initialized ligand. b) Ligand with interaction points (hydrogen bond donor = blue, hydrogen bond acceptor = red, hydrophobic = yellow) c) Example of a resulting descriptor.

- The side lengths of the triangle must all be between 1.0Å and 9.6Å.
- The internal angles of the triangle must exceed 0.15rad

The remaining triangles are canonized based on their interaction point types and their side lengths (see Canonization). Finally, the interaction directions (see Interaction Points) and the shape descriptor (see Shape Descriptor) are also annotated to the descriptor alongside the MoleculeDB IDs of the respective ligand.

Interaction Points

The interaction points used as basis for the descriptor generation are calculated for each ligand based on the cRAISE [106] interaction model and follow two different approaches, one for polar interactions, i.e., potential hydrogen bond donors and acceptors, and another one for potential hydrophobic interaction spots. For the polar interactions, this also includes the calculation and storage of possible interaction directions.

Polar interactions can either be of type *Donor* or *Acceptor*. As can be seen in Figure 4.3b, Acceptor interaction points, indicated by red spheres, lie on the associated heavy atom of a hydrogen bond acceptor. Donor interaction points on the other hand, indicated by blue spheres, lie on a hypothetical heavy atom of an optimally placed interacting acceptor. This is estimated at 2.8Å distance from the heavy atom of the donor following the direction of the hydrogen atom.
Possible interaction directions are calculated based on the respective chemical group of the interaction point. For hydrogen bond acceptors, these are the directions of all present lone pairs. It is, therefore, possible that an Acceptor interaction point has more than one interaction direction. In case of hydrogen bond donors, the interaction direction is already given following the direction of the hydrogen atom. This information is nevertheless saved for the interaction point as well. As a result, Donor interaction points can only have one possible interaction direction.
Also included in the interaction point generation is the flexibility of rotational groups. For a hydroxyl group, e.g., not only the direction represented in the molecule structure is used. In addition, interaction points are sampled following steps of 72° around the rotatable atom. All resulting interaction points and their respective interaction directions are kept as long as they point in accessible area and not into the ligands own volume.
For the representation of interaction directions in the descriptor an icosahedron is centered on the hydrophilic interaction points and canonically oriented with respect to the triangle (see Section 4.3.2). A face of the icosahedron is marked, if an interaction direction points through that special face. This way the information of interaction directions can be reduced to a bit vector of size 20 where each bit corresponds to one face of the icosahedron and the bit is set if the respective face is marked.

Apolar interactions are placed based on an initial selection of hydrophobic candidate atoms. These candidates are:

- Carbon atoms with four single bonds, three or more of them have to bind carbon, hydrogen or halogen atoms.
- Carbon atoms with two single bonds and one double bond and only carbon or hydrogen neighbors.
- Carbons with one single and one triple bond.
- Sulfur and halogen atoms.

A following placement procedure decides based on these candidates where hydrophobic interaction points will be placed. The procedure is divided into three hierarchical parts, each part places interaction points and removes candidate markings:

1. All candidates that are members of the same ring, which is either aromatic or has a maximum of nine atoms, are removed and an interaction point is placed in the center of the ring.
2. For each bond of the molecule, the algorithm checks if the connected atoms are both candidates. If that is the case, the markings on both atoms are removed and an interaction point is placed at the middle of the bond.
3. All remaining isolated candidates are then used as interaction points

Shape Descriptor

Each triangle descriptor includes a description of the local surrounding shape of the molecule. Internally, this is represented by the lengths of 80 rays radiating from the center of the triangle and ending at the surrounding molecular surface (see Figure 4.2). To ensure an equal distribution, the rays are sent through the center of the faces of an icosahedron, which has been further refined into 80 faces. This refinement is done by placing new corners at the middle of each icosahedron edge and connecting them so that each original face is divided into four equal faces. For each ray, the length from its origin to the point where it exists the molecular surface of the ligand is measured and stored. Hereby, a minimum length of 1.0Å and a maximum length of 7.1Å has to be adhered. The center of a descriptor can in some cases lie outside of the molecular surface. In this case, the minimal length is applied to rays that never reach the molecular surface since they point away from the ligand (see Figure 4.4). The maximal length on the other hand is set for rays that would exit the molecular surface but in a distance of more than specified 7.1Å. To be able to compare the shape descriptors of two triangles, it is necessary to canonize the descriptor relative to a local coordinate system defined by the triangle. This way, it is ensured that after superimposing descriptors with matching triangle corners onto each other, the 80 numbered rays of the descriptor lie perfectly on top of their respective counterpart in the other descriptor. As a result, the 80 distance values can be directly compared during screening (see Section 4.3.4). The details of the canonization procedure is described in the following section.

Canonization

To allow the direct comparison of TrixX descriptors, the most crucial part of its generation is the canonization of some of its features. Without this part, it would not be possible to compare descriptors derived from different structures and to obtain meaningful matches. This includes the canonization of the triangle itself by sorting its corners as well as the deterministic orientation of the icosahedron used

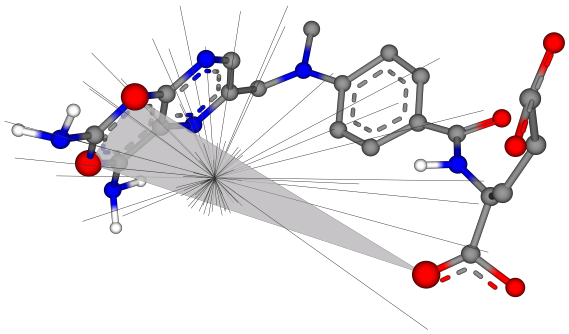


Figure 4.4.: Visualization of a descriptor with rays of minimal length.

for the representation of possible interaction direction and for the orientation of the shape descriptor.
For the basic triangle generation based on interaction point triplets, this also means, that not all possible combinations of the same interaction point triplet have to be enumerated into individual triangles. For all molecules of the screening library, it is sufficient to create one combination of each interaction point triplet if it is deterministically canonized into the same final triangle.
In the following, we define the type of a descriptor as its combination of ordered interaction points and represent them by three uppercase letters, e.g. 'XYZ'. The usage of the same letter indicates same types on the respective interaction type. We further define the edges of the triangle with e_1 being the edge between the first and the second e_2 as the edge between the second and the third interaction point and e_3 as the edge between the third and the first interaction point. Descriptor triangles are canonized as follows:

- First, triangle corners are sorted ascending by type with Donor = 0, Acceptor = 1 and Hydrophobic = 2
- Second, descriptors are sorted by side lengths while preserving the order of interaction point types
  - Descriptors of type XXX are rearranged to $e_1 \leq e_2 \leq e_3$
  - Descriptors of type XYY are rearranged to $e_1 \leq e_3$, holding e_2 in place
  - Descriptors of type XXY are rearranged to $e_2 \leq e_3$, holding e_1 in place

This procedure has more than one correct solution for isosceles triangles. Therefore, it is important to enumerate all options to guarantee that every possible match is found. Nevertheless, it is sufficient to do this only for the descriptors of the query ligand.

- For isosceles triangles of type XXX all five other permutations of the triangle corners are generated.
- For isosceles triangles of type XYY or XXY the one different ordering obtainable by swapping the identical interactions is generated.

The key to canonize the representation of the interaction directions as well as the orientation of the bulk rays is the icosahedron on which both implementations are based. Whenever an icosahedron is used, it is deterministically oriented with respect to the respective already canonized triangle it is used in. This is achieved by the following procedure:

- The ray pointing through the first face of the icosahedron is rotated onto the axis from the center of the triangle to the first corner
- The second ray is rotated onto the first edge (e_1) of the triangle

After this, the whole icosahedron is conclusively oriented based on a local coordinate system defined by the triangle and the result will always be the same for identical triangles.

4.3.3. Descriptor Index

For the rapid screening of large compound libraries, Schlosser and Rarey [30] introduced a bitmap index for the efficient storage and comparison of precalculated TrixX descriptors. The FastBit index developed by Kesheng Wu [98] for the efficient handling of data from experimental physics was found to be best suited for the application in virtual screening. FastBit is optimized for read-only access of consistent multidimensional data. Therefore, a key technology used in FastBit is the Word-Aligned Hybrid compression (WAH) for the bitmaps, reducing the space requirement while still allowing to access the data with logical operations without the need of decompression. Here, sequences of equal bits are encoded by a representation consisting of the bit value and the length of the respective bit sequence. Other parts of the bitmap remain uncompressed and the bitmap is then grouped with respect to the CPU word size. To further increase the efficiency of queries, the bitmaps can be specially encoded to either optimize for the usage of equality or range comparisons.
To efficiently handle all properties of the TrixX descriptor in the index, some of them need to be binned. For the continuous properties like lengths of triangle

sides and the 80 bulk rays, specific binning schemes are used. For the side lengths, bins with a range of 0.1Å are used resulting in a representation using 85 bits and for the bulk ray lengths bins of size 0.4Å are used resulting in a total of 15 bits. For discrete values, no binning is necessary. The same holds for interaction directions which already are represented as a sequence of 20 bits. On basis of this information, FastBit creates the index structure and compressed bitmaps for each of its dimensions. Regarding the encoding of the bitmaps for efficient comparison, the bitmaps for the side lengths and the bulk rays are encoded especially for range queries.
Another strategy to speed up the screening process based on this index is a special partitioning of the descriptor data. During the creation of the FastBit index, descriptors are stored in individual partitions based on their triangle type (see Section 4.3.1), a partition in this case is an independent subindex. The size of a partition is hereby limited to never exceed 2 GB so that one partition can always be held in memory completely. Thus, if necessary there might be multiple partitions of the same triangle type in one index.
As a result of using this technology, compound libraries have to be preprocessed for screening and a once created index can be screened as often as desired.

4.3.4. Descriptor Matching

During the screening procedure, descriptors are generated for a query molecule and then used to formulate queries to the descriptor index. For this purpose, the descriptors are sorted by their triangle type and only corresponding partitions of the index are screened (see Section 4.3.3). This way, automatically only descriptors with matching interaction points are compared. To minimize the reading operations on the hard drive, a partition is kept in memory as long as necessary. This means that in the case of multiple partitions per type, all query descriptors of that type are screened against one partition before the next is loaded.
The descriptor index is addressed using SQL-like database queries comparing all properties of the descriptor combined with a logical AND. For some properties tolerance values are applied for a less strict comparison. Matching descriptors, therefore, have to fulfill all of the following criteria:

- All side lengths have to match with a tolerance of ± 1.0Å
- Bulk rays can be matched in two different ways depending on the area of application:
  - For protein-ligand scenarios, all rays of the ligand have to be shorter or equal to the protein rays ($+0.5$Å tolerance), to fit into the binding site.

  - For protein-protein or ligand-ligand scenarios, all rays need to be of equal lengths ($\pm$0.5Å tolerance).

- The interaction directions of polar interaction points have to match. This is the case, if a bitwise AND operation on the bit strings representing the directions results in at least one set bit.

4.3.5. Alignments Calculation

For each matching descriptor, the triangle corner coordinates as well as the IDs, needed to retrieve the molecule from the MoleculeDB in its respective conformation, are returned from the descriptor index. Using the coordinates of the triangle corners, the transformation matrix to superimpose the matching target triangle onto the query triangle is calculated and stored alongside the IDs in a special data structure. These data structures are then grouped based on the conformation IDs of the matching compounds to ensure that a molecule has to be reinitialized only once even if it has multiple different matches.
In the final step, each matching molecule conformation is reinitialized once and all corresponding transformations are applied to calculate the the similarity of the molecules based on the respective structural alignment (see Section 4.6).

4.3.6. Partial Shape Approach

Some of the biggest challenges in virtual screening are the handling of molecular flexibility and the discovery of unapparent new active ligands. Regarding these challenges, it appears to be too restrictive if the matching criteria do not allow tolerances and flexibility. To a certain degree, tolerance is already incorporated in the descriptor matching in the FastBit index. Nevertheless, due to the original design for SBVS, all 80 rays of the shape descriptor need to match at once (see Section 4.3.4). For a protein-ligand scenario this makes perfect sense, since this only eliminates ligand poses which would definitely clash with the protein binding site. However, for the comparison of binding sites or ligands only, more flexibility has is needed, especially in the shape comparison in order to not only find close derivatives of the query structure. Since high degrees of unspecific tolerance might lead to a lot of false positive hits, meaningful partial shape constraints, enforcing high similarity in certain areas of a molecule while allowing flexibility in others remains the real challenge.
A first general concept for partial shape matching using the bulk descriptor has been developed by Christin Schaerfer in her diploma thesis [111]. This concept only requires a certain percentage of coherent rays to match at the same time and has

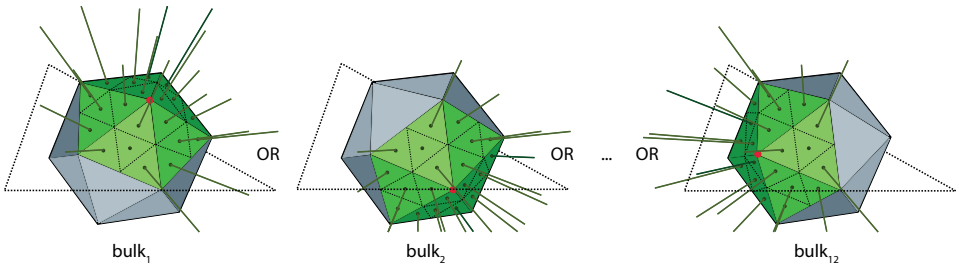


Figure 4.5.: Visualization of the partial bulk concept with 25% shape similarity requirement. Reprinted with permission from [109]. Copyright 2013 American Chemical Society

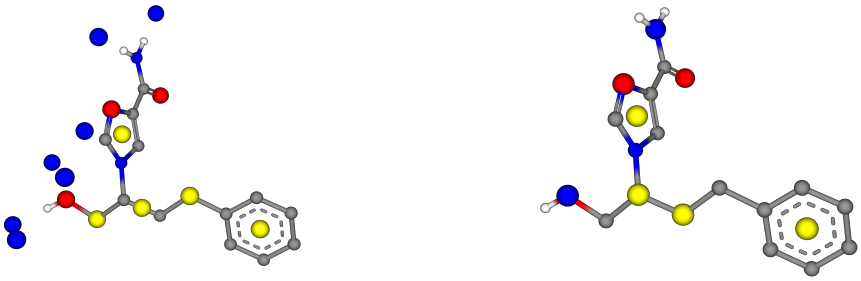
since then successfully been applied to binding site comparison in TrixP [109] and is now also incorporated in mRAISE. The algorithm utilizes the properties of the icosahedron, used within the creation of the bulk rays to define multiple subsets of neighboring rays representing a certain percentage of shape similarity. These subsets are then individually inserted into the FastBit query and combined with a logical OR. This way it is sufficient for a match if only one subset of rays matches a descriptor in the index.
Different subset sizes and, therefore, different percentages of shape similarity requirement can be generated as follows:

- 25% shape matching is acquired by only selecting rays going through triangles surrounding the same icosahedron vertex. Since a vertex is surrounded by five triangles this leads to a selection of 20 rays. Using all vertices the same way, this leads to 12 possible subsets (see Figure 4.5).

- 40% shape matching can be achieved by selecting all rays going through triangles surrounding the same icosahedron edge. This leads to 30 subsets of 32 selected rays.

- 50% shape matching requires the selection of all rays going through triangles surrounding the icosahedron triangle. In total, this results in 20 subsets of 40 selected rays.

Figure 4.5 shows an example for the calculation of a subset of rays surrounding the same icosahedron vertex for 25% shape similarity. A partial shape requirement of 25% and 50% is available as options in mRAISE.



(a) (b)

Figure 4.6.: Comparison of interaction point placements before (a) and after (b) the changes used in mRAISE. Blue spheres = hydrogen bond donors, red spheres = hydrogen bond acceptors, yellow spheres = hydrophobic.

4.4. mRAISE Adaptations

In the following, the changes to the interaction points for LBVS as well as the new concepts for knowledge-based partial shape constraints used in mRAISE are explained.

4.4.1. Interaction Points

During the development of mRAISE an alternative placement of interaction points representing hydrogen bond donors was found to be more efficient for the purpose of LBVS. Furthermore, while comparing ligand descriptors, a weakness of the original algorithm to place hydrophobic interaction points has been detected and a new placement algorithm has been introduced. Figure 4.6 shows the influence of the changes compared to the original version of the interaction points.

Polar Interaction Points

The original approach to generate polar interaction points as described in Section 4.3.2 was designed for SBVS and therefore incorporated protein-ligand complementarity. This is especially the case for the placement of the hydrogen bond donor interaction points, which are not placed on the respective ligand atoms but on coordinates where potential interacting protein atoms might occur (see Figure 4.6a). For LBVS applications, this is not essentially a bad thing, since matching descriptors would not necessarily superimpose the hydrogen donor groups onto each other, but they would superimpose the ligands in a way that both are likely to form an

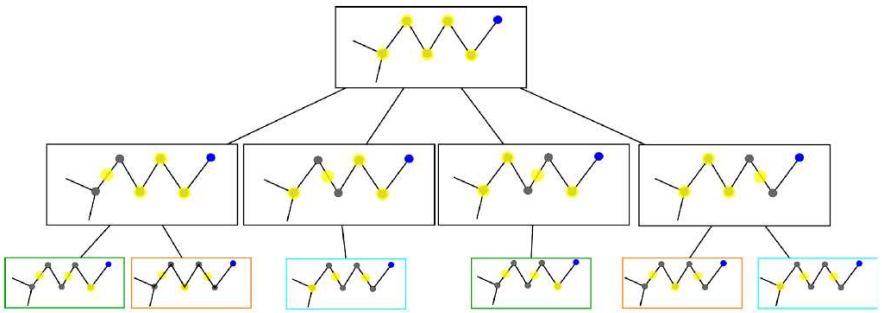


Figure 4.7.: Example of different placements of hydrophobic interaction points depending on the order of processes bonds.

interaction to the same hydrogen bond acceptor group of a protein.
However, during the development of mRAISE, it turned out to be beneficial to place the hydrogen bond donor interaction points on the respective heavy atom of the ligand like it is done for the hydrogen bond acceptors. This not only guarantees the superimposition of the respective atoms of the ligands upon matching, it also reduces the number of interaction points and consequently the number of descriptors significantly. This is because hydrogen bond donors with more than one interaction direction and especially rotatable ones no longer result in separate interaction points for all possible interaction direction but in one interaction point annotated with all possible interaction directions.
As a rough estimation of the influence of this change, a small statistic on all compounds of the DUD dataset has been calculated. On average it reduced the number of descriptors per conformation by 18.7% while producing comparable or even better results.

Apolar Interaction Points

The placement of hydrophobic interaction points on the bonds of the ligand as described before had an algorithmic problem. It turned out, that the placement of the interaction points depended on the order of the atoms and bonds in the respective data structures and was therefore not deterministic. In cRAISE this had no influence because all ligands were stored in a MoleculeDB before descriptor calculation and therefore the order of atoms and bonds was canonized for these structures, in iRAISE this also was not significant since only one ligand at the time is used as query structure. In mRAISE, however, library molecules are canonized like in cRAISE but the query ligand is not. In some cases, this resulted in different placements of hydrophobic interaction points for otherwise identical ligands. The problem occurred during the second step of the placement of hydrophobic

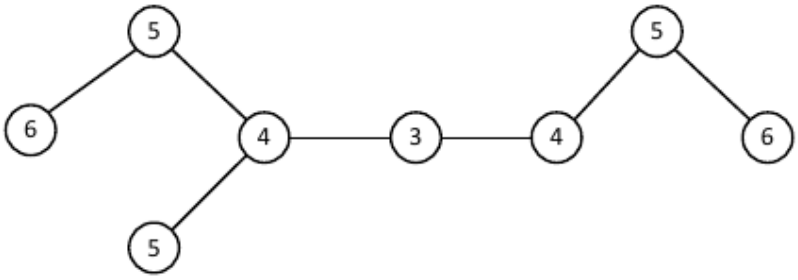


Figure 4.8.: Example of a molecule graph where each atom is annotated with the maximum shortest path to any other atom in the molecule.

interaction points for chains of carbon atoms that were all candidates for hydrophobic points. An example for this problem can be seen in Figure 4.7 for five atoms connected by four bonds. Starting from the shown initial situation, the algorithm would process the first bond available in the data structure which could be either one of the four and places a new interaction point on that bond while removing the candidate markings on the connected carbons. As shown in the picture, depending on the selected bond, this can result in four different situations with one placed interaction point and three remaining marked atoms. Depending on the distribution of those marked atoms either one or two bonds connecting two marked atoms remain, which can be selected next to place another hydrophobic interaction point on the respective bond. At the end of the second step, two of the four bonds have an apolar interaction point on them and one marker remains on one of the carbon atoms. As a result, the candidate marker would become an own interaction point in the third step. As can be seen, depending on the order of the processed bonds, there are three solutions for the placement of hydrophobic interaction points in this example both with two interaction points on bonds and another on the remaining atom which is not attached to that bond. For mRAISE, it is of great importance that the same molecules always result in the same set of descriptors. Therefore, a new algorithm for the second step of the placement procedure has been introduced. The new algorithm makes use of the fact that molecules are internally represented as graphs with bonds as edges and atoms as nodes in NAOMI (see Section 4.1.1). In a first step, the Floyd–Warshall algorithm is used to compute the shortest paths between all pairs of atoms in the molecule and to store them in a distance matrix M of size $n \times n$, where n is the number of atoms in the respective molecule. Each edge is hereby handled as if it had a weight of 1. After this calculation, each row M_i in the matrix M stores the shortest paths to all other atoms in the molecule starting from the atom i. The maximum among this values is therefore the longest direct path between this atom and another atom of the molecule. Next, each atom of the molecule is annotated with that maximum value of its respective row. An

example for such an annotation can be seen in Figure 4.8. Finally, like in the original implementation of this placement step, the algorithm identifies all atoms which are candidates for an hydrophobic interaction point and have at least one neighbor which is also a candidate. An interaction point is placed directly onto this atom, if the path annotation on that atom is an odd number.
After this step, candidate markings are removed from all atoms that became hydrophobic interaction points as well as from their direct neighbors. This way, the third step can consecutively process all remaining candidates.

4.5. Knowledge-based Partial Shape Constraints

In mRAISE a new approach for meaningful partial shape constraints to guide the screening process is introduced. These constraints make use of additional information derived from protein-ligand complexes or the experience of the user who can define shape constraints manually.

4.5.1. Complex-based Partial Shape Constraints

Recent applications of virtual screening methods in application studies often involved combinations of structure-based as well as ligand-based approaches to make use of all available information and achieve the best possible results. This of course is a reasonable strategy if the structure of the protein-ligand complex is available. Therefore, during the creation of triangle descriptors mRAISE can also make use of this information to derive partial shape constraints with respect to the binding site of the protein to which the query ligand is bound.
Two different modes for complex-based partial shape queries are available in mRAISE. Based on the local shape description available in each descriptor, one uses the information to only match descriptors that would fit into the binding site (inclusion queries) and the other tries to maintain close contacts between the ligand and the protein on matching (contact queries).

Inclusion Queries

Without any information about the binding site to which a query ligand binds, even promising highly similar hits from a LBVS campaign might not show any activity to the same targets. One reason for this might simply be steric incompatibility of the structures. In those cases, steric constraints of the binding site would be of more interest for virtual screening than the shape of only one ligand.

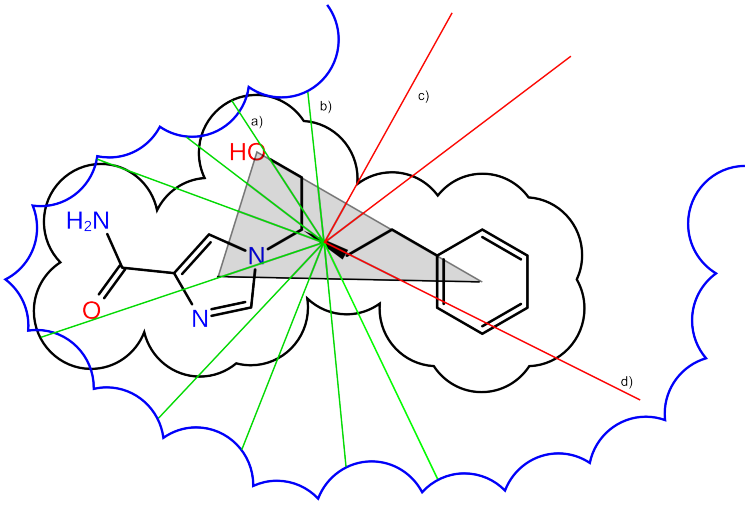


Figure 4.9.: Simplified depiction of the selection of bulk rays for inclusion queries. Rays are adapted and selected depending on the distance to the protein surface (blue): Rays are selected if they reach the protein surface. The length of the ray is set either to the distance of the ligand surface (a) or of the protein surface (b), depending on which one is further (green rays). Not selected are rays which never reach the binding site (c) or are at maximum length before doing so (d) (red rays). Reprinted from [112].

Inclusion queries therefore match descriptors that would fit into the same area of the binding site instead of those that roughly feature the same shape as the query ligand. During the generation of the shape description of each query descriptor, each ray is extended to not end at the point of exiting the molecular surface of the ligand but at the point where it enters the molecular surface of the protein binding site while still respecting the maximal possible length as described in Section 4.3.2. An exception is made for situations where the van der Waals radii of protein and ligand atoms overlap and the distance for a ray to enter the protein binding site molecular surface would actually be shorter than the usually calculated distance. In this case, the length of the ray is still set to the point where the ray exits the molecular surface of the ligand. If a ray does not reach the protein molecular surface, it is not considered during descriptor comparison. A depiction of this process is shown in Figure 4.9.
During the descriptor comparison using the FastBit index, the query for the bulk ray lengths is changed to only match descriptors with equal or shorter lengths than the ones of the query descriptor. Since not all of the 80 rays are used, if they do not interact with the binding site, some descriptors might become relatively unspecific and would therefore lead to insignificant matches. To avoid this, all descriptors with less than 40 remaining rays are discarded.

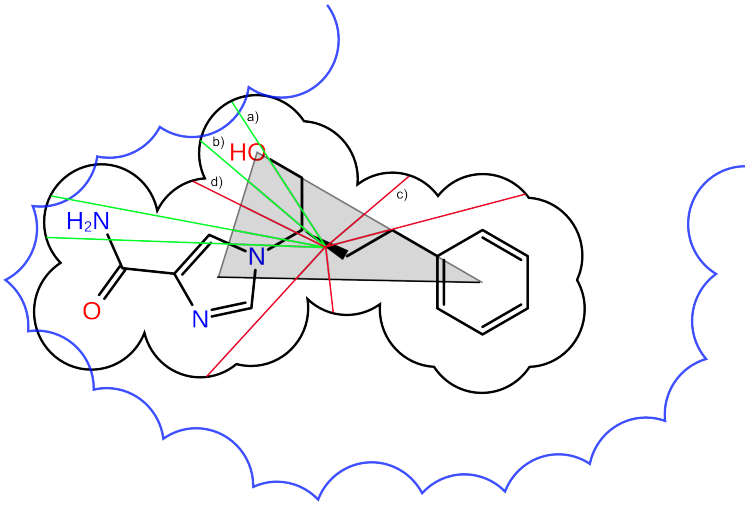


Figure 4.10.: Simplified depiction of the selection of bulk rays for contact queries. Rays are selected by the distance between the ligand surface (black) and the protein surface (blue): Rays are selected if the surfaces overlap (a) or are in close contact (b) (green rays). Not selected are rays pointing towards bulk (c) and rays where ligand and protein surface are too far away from each other (d) (red rays). Reprinted from [112].

Contact Queries

The contact queries are another approach to derive important information from the protein-ligand complex. During binding, the protein and the ligand form interaction which require geometrically close contacts. These contacts are therefore very important for the activity and highlight the most important features of the ligand structure. The information of close contacts can be incorporated in mRAISE again using partial shape constraints. During the generation of the shape descriptor, only those rays are used which would intersect with the protein molecular surface 0.5Å distant from the point where they left the ligands molecular surface. This process is illustrated in Figure 4.10.
Like in the case of the inclusion queries, some descriptors will only have a few used rays remaining following this approach, especially since such close contacts only occur at a few special areas of a ligand. To prevent completely insignificant matches, only descriptors with at least five remaining rays are used during screening.

4.5.2. Manual Partial Shape Constraints

It remains a major benefit of LBVS that it can be applied in situations where no crystal structure of the protein is available. In such situations, the only available

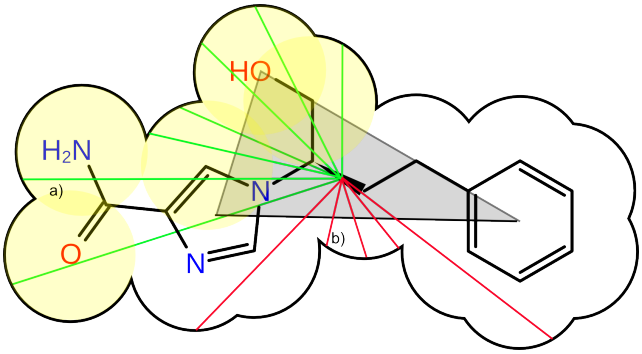


Figure 4.11.: Simplified depiction of the selection of bulk rays for manual selection queries. Rays are selected if they end within a selected atom sphere (a) (green rays). Not selected are all other rays (b) (red rays). Reprinted from [112].

resource for meaningful partial shape constraints is the expert knowledge of the user. To enable the user to incorporate his experience-based ideas about important and unimportant regions of a molecule into the screening procedure, mRAISE introduces a new concept for the manual definition of partial shape constraints.
A user can interact with a query molecule using a specially developed graphical user interface (see Section 4.8) by freely selecting atoms of the ligand and, thereby, defining regions considered as important for the activity of the ligand. As a result of this selection, descriptors are generated and the bulk rays are used to represent these special constraints, by only selecting rays which pierce through the molecular surface belonging to these atoms. Depending on the number of selected atoms, the number of used rays can again be quite low. To discard insignificant descriptors, like in the contact queries, only descriptors with at least 5 remaining rays are used for screening. In detail this corresponds to all rays ending within the van der Waals radius of a selected atom and are not of maximal length. Figure 4.11 shows a simplified depiction of this process.
An important feature of the graphical user interface is furthermore that the manually defined queries can be saved to an annotated SDF file. This way the query can also be loaded by the command line version of mRAISE for exhaustive screening runs (see Section 4.8.1).

4.6. Scoring

The similarity of two aligned molecules is scored in mRAISE using basic atom-centered Gaussian functions. This general concept is commonly used to estimate the volume overlap of molecules. The Gaussian function basically returns a value

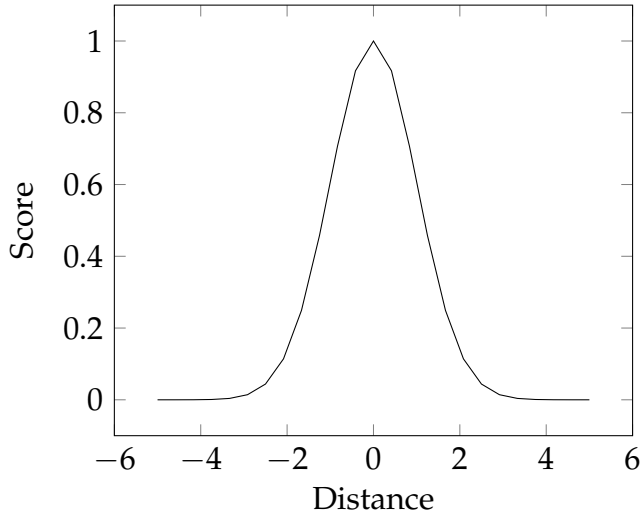


Figure 4.12.: Plot of $e^{-0.5*(x^2)}$ showing the score behavior based on the distance (x) of two atoms.

between 0 and 1 based on the distance between two atoms and is usually parameterized with respect to the van der Waals radii of certain atoms so that a score of 0.5 is achieved at the distance where the atom spheres would half overlap. Physicochemical properties can also be included as weights for these functions. The similarity between a query molecule Q and a target molecule T is calculated in mRAISE as follows:

$$s(Q,T) = \sum_{q \in Q} \sum_{t \in T} w_{ch}(q,t) w_{ri}(q,t) w_{ia}(q,t) \alpha e^{\beta(qp-tp)^2} \tag{4.1}$$

Here q, t are the atoms of the query and target molecule and gp, tp are their respective coordinates. Furthermore, the weights w_{ch}, w_{ri} and w_{ia} incorporate matching or mismatching features of the atoms into the score. These features are charges (ch) (see Table A.1), ring membership (ri) (see Table A.2) and the ability to be a hydrogen bond donor or acceptor (ia) (see Table A.3). With α set to 1.0 and β set to -0.5, the parameters of the Gaussian function are chosen in mRAISE with respect to the van der Waals radius of a carbon atom and used the same way for each calculation. The function yields a score of 0.5 at an atomic distance of 1.18Å, which is the distance at which the intersecting volume of two carbon atoms reaches half of the atomic sphere volume (see Figure 4.12).

To obtain a normalized final score for the similarity of two molecules Q and T, the Hodgkin Similarity is used:

$$hs(Q,T) = \frac{2s(Q,T)}{s(Q,Q) + s(T,T)} \tag{4.2}$$

Based on all descriptor matches between the query ligand and a matching ligand conformation of the screening library, each possible alignment is scored using this function and only the best score for each conformation is written to an output file and eventually stored into a SolutionDB as explained in the following.

4.7. Results

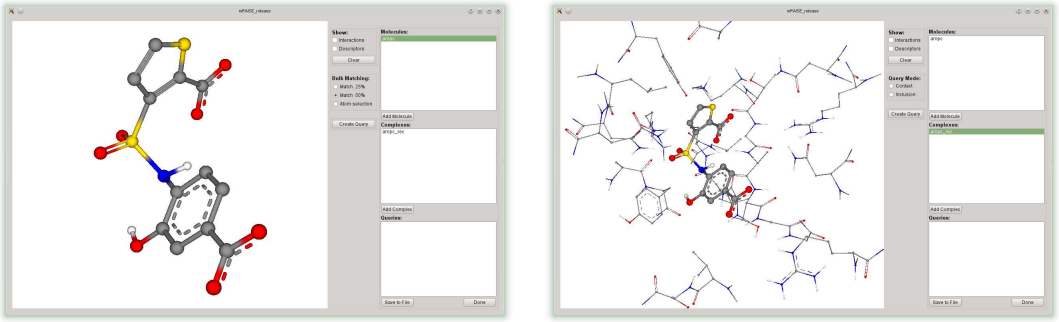
The results of a LBVS run can be written to a file in order to be evaluated. Using the command line version of mRAISE, each run produces a list of matching molecule conformations with their respective score. An example of the content of such a file is shown below:

ZINC03327557,4305,294447,0.376295
ZINC03327557,4305,294611,0.394193
ZINC00618696,2448,163516,0.357897
ZINC00618696,2448,163517,0.357079

Each line holds four entries, the first is the molecule name, the second and the third entry are the IDs of the molecule and the respective conformation from the MoleculeDB and the fourth entry is the calculated similarity score. A second optional way to store the screening results is in the form of a new MoleculeDB which stores all matching molecule conformations together with the respective score. This option needs a lot more space on the hard drive, but it allows the visual inspection of results using the GUI version of mRAISE (see Section 4.8). Furthermore, the database enables more evaluation options like writing the best scored conformations to SDF files. For the purpose of evaluation mRAISE furthermore is able to combine multiple solution databases if the screening procedure has been split into multiple parallel runs in order to safe time.
Using the GUI version (see Section 4.8) of mRAISE, screening results always have to be stored in a MoleculeDB, since the preparation and visualization of the results in the GUI is based information only available using the database.

4.8. GUI

mRAISE is available in two different versions, one is a tool with just a command line interface and the other is a tool with a graphical user interface. Both version are sufficient to create or load descriptor indices, to load molecules or complexes to create queries, and to perform complete screening runs (see Appendix C and



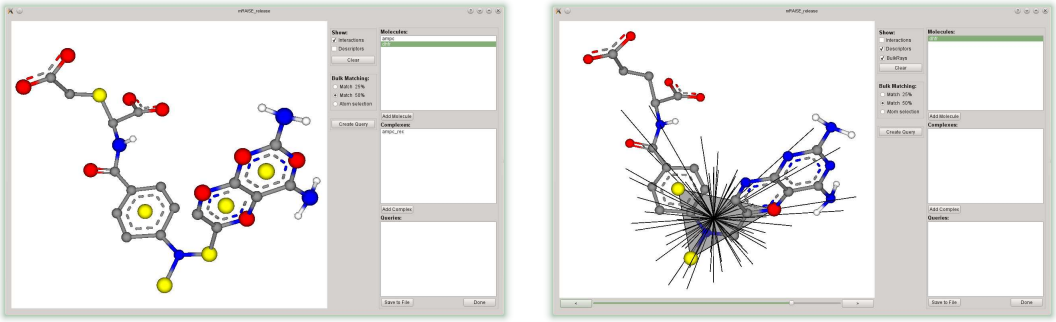
(a) (b)

Figure 4.13.: Visualization of query ligands. a) Only a ligand loaded. b) Ligand loaded together with a protein structure.

Appendix D for the respective user guides).
While the command line version obviously is better suited for exhaustive screening runs for example on computer clusters, the GUI version provides some unique interactive features. The most important of which is definitely the interface to create and save queries with manual selected shape constraints (see Section 4.5.2). Another important feature of this version is the visualization of screening results, not only as sorted lists but also by visualizing the respective molecular alignments of selected hits.

4.8.1. Query Preparation

Figure 4.13 shows the query preparation window of mRAISE, here, molecules can be loaded from various file formats as well as protein-ligand complexes from a PDB file in combination with an extra file containing the ligand of interest. The structures can be inspected and the calculated triangle descriptors as well as the interaction points can be visualized (see Figure 4.14). Queries can be created with all available options like 25% and 50% shape matching as well as complex derived constraints. Furthermore, the atoms of a displayed ligand can be freely selected and a query can be created based on the current atom selection as well (see Figure 4.15). Defined queries are hold in storage as long as the program is not terminated. Created queries can be inspected, used, and saved at all time. Ligand-based queries (25% shape, 50% shape and atom selections) can furthermore be saved to annotated SDF files which can be directly processed by the command line version of mRAISE or be reloaded into the mRAISE GUI. Associated data can be annotated in SDF files directly following the molecule information (indicated by 'M END') and before the entry separator ('$$$$'). Multiple entries can be defined using a header in the



(a) (b)

Figure 4.14.: Visualization of query features. a) Interaction points. (yellow = hydrophobic, blue = hydrogen bond donor, red = hydrogen bond acceptor b) Example of a displayed descriptor.

form of '> <name>', where the name can be freely chosen followed by newline separated data entries. The header for mRAISE query information is

> <mRAISE_matching>

followed by the matching mode

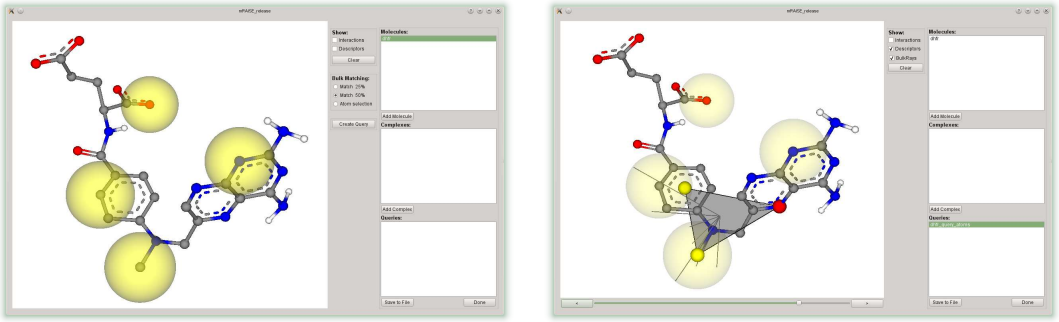
- mode 0 = 25% shape matching
- mode 1 = 50% shape matching
- mode 2 = atom selection

In case of 'mode 2' the entry also stores the IDs of the selected atoms with respect to the SDF file in separated by newlines. A random example of a saved atom selection annotation can be seen below:

```
> <mRAISE_matching>
mode  2
8
6
7
5
```

4.8.2. Screening

In the screening tab (see Figure 4.16) of the mRAISE GUI, the query preparation window (described in the previous section) can be opened and a list of already



(a) (b)

Figure 4.15.: Visualization of the manual selection of partial shape constraints a) Selection of four molecules indicated by yellow spheres. b) One of the remaining descriptors based on the atom selection with two hydrophobic and one hydrogen bond acceptor corners.

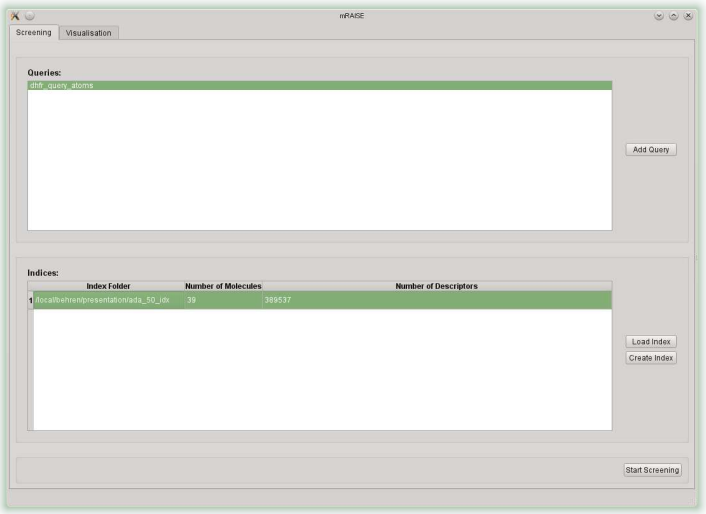
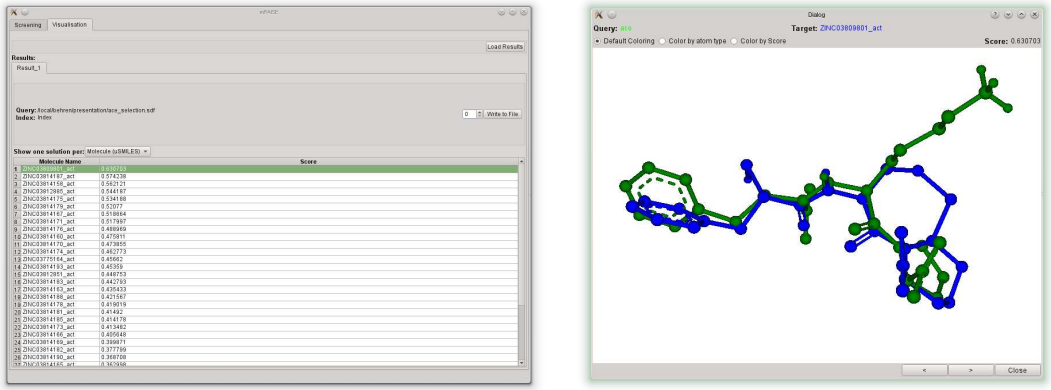


Figure 4.16.: Picture of the screening tab in the mRAISE GUI.

prepared queries is shown. Furthermore, new indices can be created or existing indices can be loaded for screening. All indices ready for screening are also displayed in a separate list. Once a prepared query and an initialized index are selected, a screening run can be started.



(a) (b)

Figure 4.17.: Visualization of screening results using the mRAISE GUI. a) Sorted list of hits with scores. b) Alignment of two ligands with the query ligand in green and the target ligand in green.

4.8.3. Result Visualization

Figure 4.17a shows the general visualization of screening results, here a sorted list of hits is presented with the respective molecule names and their scores. The data shown in the list can be adapted by choosing either one displayed entry per Molecule, Conformation or Name. Each entry of the hit list can be selected and a new window will open showing the respective alignments of the entry and the query ligand (see Figure 4.17b). The visualization can be changed to either highlight the different structures by individual colors, by just coloring both ligands based on their atom types or by coloring each atom of the query ligand based on the individual score it achieves. The color scale hereby goes from green (good score) to gray (score of zero). At last, the visualization tab also allows to load results directly from database files created by mRAISE during screening runs and to write a specified number of best scored hit list entries to SDF files.

5 Chapter 5. Datasets

For the development of computational models in the field of chemoinformatics reliable experimental data is indispensable. For the development of methods dedicated to LBVS, ligands known to have similar activity to the same targets as well as an ideally even larger set of ligands, which definitely have no activity to the respective targets, are required. For the purpose of alignment validation even the 3D structures of protein-ligand complexes representing different ligands bound to the same target are required. In the best case these complex structures are experimentally determined by X-ray crystallography and available in high resolution. Structures determined by Nuclear Magnet Resonance (NMR) are also a potential source, but those structures are limited to solvable proteins and are still of debatable quality [113, 114].
Different datasets are required at different stages of the development process for a new virtual screening method. They range from small, diverse datasets for the purpose of the initial method development and parametrization to larger datasets for the validation of the method and for the comparison to other methods in the field. While the first category of datasets are important for guiding the development process, the second are inalienable for the introduction of a new method into a field as contested as LBVS. Therefore, the following sections will focus on the datasets used for this purpose.

The datasets used in this project were chosen by their potential to compare mRAISE to a preferable high amount of well-known and recent methods as well as by their quality to serve as an objective validation foundation. For this purpose, the datasets used in comparison studies as well as in introductions of new methods have to be publicly available and the data preparation has to be well documented. In the following sections, first difficulties on choosing appropriate datasets and general problems of validation studies in the literature are discussed. Following this, the

used datasets for the validation of mRAISE and the respective preparation steps are presented.

5.1. The Directory of Useful Decoys

The Directory of Useful Decoys (DUD) was composed by Huang *et al.* in 2006 [88] as a benchmarking set for molecular docking. It consists of 2950 ligands for 40 different targets and 36 topologically distinct but physically similar selected decoys for each of those ligands.
All targets were selected based on the availability of annotated ligands and crystal structures as well as their eventual usage in previous docking studies. Initially, the 2950 annotated ligands available for those targets were combined with 3.5 million molecules from the ZINC database [115], which followed the Lipinski rules for druglikeness [52]. The topological dissimilarity between the ligands and their respective decoys was ensured using type 2 substructure keys of CACTVS [116] as well as the standard Daylight fingerprint [55]. Only considering molecules with a Tanimoto coefficient below a certain threshold to any annotated ligand using these fingerprints already reduced the number of ZINC molecules to 1.5 million compounds considered as topologically dissimilar. Out of these remaining molecules the 36 physically most similar compounds were selected for each annotated ligand using QikProp(Schrödinger, LLC, New York, NY) for the calculation of physicochemical properties and QikSim(Schrödinger, LLC, New York, NY) for the priorization of similar compounds. Weights were used to emphasize properties important for druglikeness (molecular weight, number of hydrogen bond acceptors and donors, number of rotational bonds, and logP), followed by the numbers of functional groups (amine, amide, amidine, and carboxylic acids) with a lower weight. All other properties were ignored using a weight of zero.
It has to be noted that by following this procedure the same molecule can be used multiple times as decoy for different ligands and the total number of decoys therefore does not equal the amount of annotated ligands times 36. Furthermore, Huang *et al.* only assume that the topologically different decoys are actually true negatives, which is not necessarily true in each case [117]. Another important aspect of the DUD dataset that has to be taken into account is that it includes duplicates of the same molecules like protomers and tautomers. The consistent handling of those compounds during a VS experiment is crucial in order to enable any conclusions about the performance of a method in comparison to other methods.

Besides the criticism of the DUD (see Section 5.2), the dataset has been chosen for validation studies with mRAISE, since one setup enables an objective comparison to a variety of different methods. The data preparation is explained below and

enables a direct comparison to reported performances of: LIGSIFT, Align-It, ROCS, ShaEP, MolShaCS, Surflex-sim, FlexS, and ICMsim.

5.2. The Directory of Useful Decoys Enhanced

The Directory of Useful Decoys Enhanced (DUD-E) was composed by Mysinger *et al.* in 2012 [89] as an extended and more challenging version of the original DUD. The dataset contains 22866 clustered ligands for 102 targets, with 50 physicochemical similar decoys for each ligand.
The frequent use of the original DUD dataset for virtual screening benchmarks [118–125] revealed weaknesses in the ligands as well as in the decoys. Good and Oprea noted, that some of the ligand sets are dominated by only a few different chemotypes, which leads to high enrichments with only one scaffold achieving top ranks [126]. A clustering of the ligand scaffolds would reduce the size of the dataset to only 13 targets with more than 15 remaining ligands. Furthermore, for the decoys, multiple studies observed an imbalance in the net formal charge [127–129], which made it easier to discriminate between actives and decoys based on this property. It also turned out that some decoys are actually false negatives and bind to their respective targets despite the 2D dissimilarity criteria [130]. This, combined with the low target diversity for example concerning membrane domain proteins, emphasized the need for more targets with more ligands and better decoys.
In the DUD-E, the number of targets was extended from 40 to 102, focusing on targets with many ligands and multiple available structures in the RCSB PDB [131]. The final set includes 38 of the original targets used in DUD and covers a variety of diverse protein categories with 26 kinases, 15 proteases, 11 nuclear receptors, five GPCRs, two ion channels, two P450s, 36 other enzymes, and five miscellaneous proteins. All ligands included in the dataset have been drawn from the ChEMBL09 [132] database and had to feature measured affinities reported in the literature. To increase scaffold diversity and at the same time reduce the ligand set sizes, all ligands of one target were clustered using the Bemis-Murcko atomic frameworks [133], which still led to an average of 224 ligands per target. If more than 100 clusters were created, only one representative with the highest affinity was chosen from each cluster. For targets with less than 100 clusters, more than one representative was drawn from each cluster until more than 100 ligands were selected. Lastly, if more than 600 clusters were present, the affinity threshold was reduced until fewer than 600 frameworks remained. As in the original DUD, property matched decoy sets were generated for each ligand (see Section 5.1). In addition to the previously used properties, the net charge was added during this procedure to address the noticed deficiencies. Furthermore, the problem of false decoys has been addressed by a more strict filtering with regard to

Table 5.1.: Overview of the mRAISE dataset.

Protein Class	Unique Ligands
Trypsin	29
Thrombin	11
Alpha-Mannosidase II	15
Matrix metalloproteinase-12 (MMP-12)	10
CDK2 Kinase	8
Carbonic Anhydrase II	41
Thermolysin	9
CYP121	8
HIV Protease	24 (10)
Bromodomain-containing protein 4 (BRD4)	9
Isopenicillin N Synthase	16

Since the HIV Protease ensemble has more than one conformation for its ligands, the number of unique ligands is given in brackets.

topological dissimilarity and whenever possible, experimentally validated decoys were included. In the new approach, the topological dissimilarity is ensured using the ECFP4 fingerprint to remove 75% of the most similar decoys.

The motivation of using the DUD-E is not the ability to compare the performance of mRAISE to other methods, but the intent to use a second, more challenging dataset without the flaws recognized in the DUD.

5.3. The mRAISE Dataset

The mRAISE dataset was introduced in 2016 [107] for the validation of the quality of calculated molecular alignments in mRAISE due to a lack of freely available prepared datasets of sufficient size and quality in the literature. It consists of 180 prealigned ligands for 11 diverse targets, which were aligned based on their binding to identical binding sites (see Table 5.1). As basis of the dataset, a previously published subset of the PDB consisting only of high-resolution structures [134] has been used. This dataset has initially been compiled for a validation study on water positions, but the high-resolution criteria (resolution $\leq$ 1.5Å) also ensures good overall structural quality. Another important feature of the dataset is that it is not filtered for unique proteins and therefore includes identical binding sites with different bound ligands. The search for such structures and the calculation of a superimposition of identical binding sites has been done using the protein

ensemble assembly tool SIENA [135].
In a first step, unwanted ligands are excluded from the dataset using the list of unwanted HET codes published with iRAISE [110]. This list includes co-factors, solution buffer agents, crystallization agents, ions, and ligands with covalently bound metals that can not be initialized using NAOMI (see Section 4.1.1). It was created joining three previously published lists [136–138] and has been extended by additional codes. All remaining ligands were used to define query binding sites in SIENA by applying a binding site radius of 6.5 Å around each bound ligand. The resulting queries were used to search for identical binding sites in the high-resolution dataset and SIENA created an ensemble for each query consisting of all found structures exhibiting a backbone RMSD of less 0.5Å or less to the query binding site. A consecutive clustering is performed to reduce the redundancy in the resulting ensembles and to generate a diverse final set. During this process, all ensembles with at least one common structure were joined into the same cluster. Afterwards, one representative ensemble is chosen out of each cluster with respect to the highest number of included topologically unique structures. If there was more than one valid choice for one cluster, the ensembles were ordered based on the PDB codes of their query binding sites as well as the occurrence of reference ligands in the PDB file. In a final step the ligands were drawn from all remaining ensembles and were then filtered with MONA and further scripts to ensure the following physicochemical properties and criteria:

- Each ligand consists of at least ten heavy atoms (for more pharmaceutically relevant ligands).
- No ligand has more than 15 rotatable bonds and no macrocycles with more than eight atoms (excluding highly flexible ligands).
- Each ensemble has to consist of at least eight unique molecules according to the most strict initialization criteria of MONA (for statistical relevance).
- Each ligand shares more than five overlapping atoms with each other member of the ensemble (to ensure a shared binding mode). This is described in further detail in the following.

The overlap criteria was introduced to guarantee a shared volume of all ligands within the binding site. Hereby, a pair of atoms was defined as overlapping if the atom centers were less than 1Å distant from each other. After calculating the atom overlap between all pairs of molecules within an ensemble, the molecule with the highest count of insufficient overlaps to other molecules of the ensemble was removed. This step has been repeated until all remaining ligands were sufficiently overlapping. To guarantee deterministic choices during this phase, whenever there was more than one possible decision, the ligand with smaller sum of overlapping atoms among all insufficient cases has been removed. If this criteria still produced

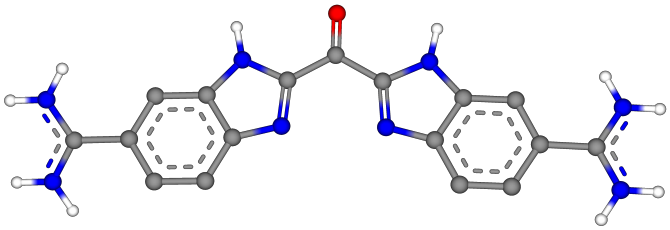


Figure 5.1.: Picture of the symmetric trypsin ligand BAK.

more than one solution, the ligand with that also had the lowest sum of overlapping atoms in all sufficient cases has been removed.
Two of the ensembles remaining after these steps had to be handled separately. Firstly, an ensemble of human transthyretin binding sites was excluded from the dataset because the ligands showed no unique binding mode due to a pseudo symmetry within this ligand family. Secondly, one ligand had to be removed from the trypsin ensemble since it was completely planar and symmetric (HET code: BAK, see Figure 5.1). In such a case, LBVS methods would not be able to distinguish between the two symmetric alignment solutions possible for this ligand. Lastly, a special feature of ensembles created with SIENA has to be considered using the mRAISE dataset. In case of the HIV protease ensemble more than one alignment solution might occur for the same PDB entry, this is due to the fact that the ligands of the HIV protease bind to a symmetric interface of a homodimer. Two alignments occur if an active site might be aligned to both units within the matching parameters and in some cases even four solutions occur if additionally an alternative ligand conformation is available. Of these multiple alignments and their respective binding poses none has been discarded since they all are equally valid. Therefore, while evaluating alignment solutions of a LBVS method, all available poses of the same ligand are considered and the minimum RMSD between the calculated poses and the available poses in the dataset is used.
A list of all included PDB codes together with the HET codes of the respective ligands can be found in Table E.1.

5.3.1. Data Preparation

For each target of the **DUD** as well as the **DUD-E** dataset a descriptor index has been created with up to 250 conformations for each ligand. As query ligand for each created index, the respective ligand from the query complex provided by the datasets is used. The query ligand for aa2ar in the DUD-E was not used as provided in the dataset but downloaded from the PDB for technical reasons. During the evaluation, the included duplicates of the same molecule (e.g. protomers and

tautomers) are excluded from the result files. This way only the highest ranked occurrence of a molecule is kept in the results.
For each individual ligand in each ensemble of the **mRAISE** dataset an index is generated with also up to 250 conformations. During the evaluation, the input structure in the index is skipped and only generated conformations are considered.

6 Chapter 6. Evaluation

In the following chapter, the evaluation and comparison strategy is embarked during the introduction and development of mRAISE are described.
Within the last decades, multiple different approaches have been developed to address the problems of LBVS. With that multitude of available options, newly introduced methods should optimally perform exhaustive evaluation experiments and be compared to the latest and commonly used tools of the field if possible. In general LBVS methods are evaluated in retrospective experiments, which means that they are tested to retrieve molecules known to be bioactive to the same target or target class from a background of at least assumed inactive molecules. An important criteria is therefore the choice of well suited datasets as well as the utilization of established and well suited performance metrics for this purpose. The datasets used for the evaluation of mRAISE have already been described in the previous chapter (see Chapter 5) and the used performance metrics will be described in the following (see Section 6.1).
The performed experiments basically evaluate two important features of a ligand-based screening method: the enrichment, meaning the ability to rank true positive hits of a dataset correctly, and the quality of the calculated molecular alignments. While the first feature is evaluated commonly and a variety of methods can directly be compared due to well commented reproducible experiments on freely available datasets, the quality of the molecular alignments is often neglected during the evaluation of new methods and no recent datasets of statistically sufficient size existed during the development of mRAISE. For this reason, experiments have been performed using the newly introduced mRAISE dataset for the validation of alignments calculated by LBVS (see Section 5.3).
Since mRAISE offers a variety of different approaches to influence the virtual screening performance by partial shape constraints, the experiments have not only been used to compare mRAISE to other methods if possible but also to compare

the influence of these different modes on virtual screening with mRAISE.
The different modes of mRAISE will be referred to as follows:

- **mRAISE_classic** uses 50% shape requirement (see Section 4.3.6).
- **mRAISE_inclusion** uses complex derived inclusion queries (see Section 4.5.1).
- **mRAISE_contact** uses complex derived contact queries (see Section 4.5.1).
- **mRAISE_manual** uses manually selected shape constraints (see Section 4.5.2).

The following experiments have been performed using the different modes of mRAISE:

1. Enrichment study on the DUD
2. Enrichment study on a DUD subset
3. Enrichment study on the DUD-E
4. The influence of manual partial shape constraints
5. Alignments quality evaluation

The experiments 1, 2, 3 and 5 evaluate the performance of mRAISE_classic as well as the influence of mRAISE_inclusion and mRAISE_contact. Furthermore, experiments 1 and 2 also compare the performance of mRAISE to other state of the art LBVS methods. Experiment 4 analyzes the possible impact of mRAISE_manual.

In the following sections, first, the used performance metrics and evaluation criteria are discussed and second, the performed experiments are described. For the results of the experiments as well as their discussion see Chapter 7.

6.1. Criteria and Metrics

The evaluation criteria and measures described in the following are commonly used and mostly chosen because they allow the direct comparison of mRAISE to other methods. The requirement of the used measures is hereby to objectively highlight the strengths as well as the limitations of the method and to advise a user in which scenario mRAISE and its individual modes are beneficial. In order to be of any value for future developments in the area of LBVS, it is important to note that all experiments always have to be reproducible. This includes a detailed documentation of used datasets and their preparations (see Chapter 5).
For a ligand-based screening method two basic capabilities have to be shown by experiments. Firstly, the most important feature of an LBVS method is its ability

to identify other active molecules among large compound libraries. In doing so, the active molecules also have to be ranked high in the hit list because only a small set of top-ranked hits is usually considered for subsequent experimental evaluation. Secondly, the reliability of the calculated molecular alignments has to be validated. Such an evaluation proofs the ability of the method to correctly quantify the similarity of two molecules based on their most important biochemical features, which are required for binding to the same target.
A descriptor-based approach like mRAISE does not necessarily provide a score for each screened compound, since there might be cases with not even one matching descriptor. In those cases, a score of 0 is assumed for the respective compound during evaluation.

Enrichment Evaluation The evaluation of the enrichment power, especially with respect to the early recognition of active molecules is an important task and several metrics exist for this purpose. Here, multiple different metrics have been chosen mostly regarding reproducible evaluation experiments of other LBVS methods. This includes the most common and established metrics like the AUC (Area und the ROC curve), which is an easily interpretable metric for the overall performance of a virtual screening method, and the Enrichment Factor at different percentages of considered ranked hits, which allows the comparison of the early enrichment of different methods on the same data.
An important question during the calculation of these metrics is how to handle compounds which produced no matches and are therefore unranked and have the same score (0). In cases where there are still actives remaining in the unranked data, an even distribution of actives among these compounds is assumed for evaluation of used metrics.
The used metrics are defined as follows:

- **Area under the ROC curve** (AUC)
  A ROC (Receiver operating characteristic) curve is a measure of the overall screening performance. Therefore, the plot visualizes the true positive rate versus the false positive rate. The area under the curve can be calculated as follows: [139]

$$AUC = \frac{1}{N_D} \sum_{N_D}^{i=1} TPR_i \tag{6.1}$$

  Here, N_D is the total number of decoys and TPR_i is the true positive rate at decoy i (number of true positives rated higher than the decoy i divided by the total number of actives). The value of the AUC is bound by 0 and 1, with 1 representing a perfect enrichment and 0.5 representing a completely random performance.

- **Enrichment Factor** (EF)

$$EF_{X\%} = \frac{TP_{x\%}/Hits_{x\%}}{N_{Actives}/N_{Total}} \tag{6.2}$$

  Here, $TP_{x\%}$ is the number of true positives present within the first x% of ranked hits, $Hits_{x\%}$ is the total number of hits at x% and $N_{Actives}/N_{Total}$ is the ratio of actives in the whole dataset.
  The Enrichment Factor is the standard measure for the early enrichment of actives among a fraction of ranked hits. This fraction (X%) is usually chosen between 1% and 10% to resemble the amount of hits considered for experimental evaluation after a LBVS campaign.
  Since the EF is calculated with respect to the ratio of actives present in a dataset and the size of the dataset, the EF is not comparable among different experiments and the actual values are not interpretable as a good or bad enrichment. During this evaluation, the EF is calculated at 1%, 5% and 10%.

- **Enrichment** (ER)
  A simpler but easier to interpret metric, simply referred to as the Enrichment, has been used by Giganti *et al.* [140] to compare the performance of different methods on 11 targets of the DUD dataset. The Enrichment is the number of actives found at a specific fraction of the dataset divided by the total number of actives.

$$ER_{x\%} = \frac{TP_{x\%}}{N_{Actives}} \tag{6.3}$$

  For the comparison to the published data, the ER has been calculated at 1% and 10%.

- **Hit Rate** (HR) The Hit Rate is an attempt to make the EF easier to interpret by putting it in relation to the best obtainable EF on the respective dataset at the specific fraction of data.

$$HR_{X\%} = \frac{actualEF_{x\%}}{idealEF_{x\%}} \tag{6.4}$$

  Therefore, this metric does no longer depend on the ratio of actives and decoys in a dataset. The $idealEF_{x\%}$ is calculated with the formula shown above while assuming that as many actives as possible are found within x% of the dataset. The HR has also been calculated at 1%, 5% and 10%.

Alignment Evaluation The most common approach to access the quality of calculated molecular alignments, is the calculation of the Root Mean Square Deviation (RMSD) between the two poses p_{pred} and p_{ref} of a molecule M. Hereby, p_{pred} is calculated by screening the molecule M versus a molecule N and p_{ref} is derived from aligning the protein-ligand complexes of M and N.

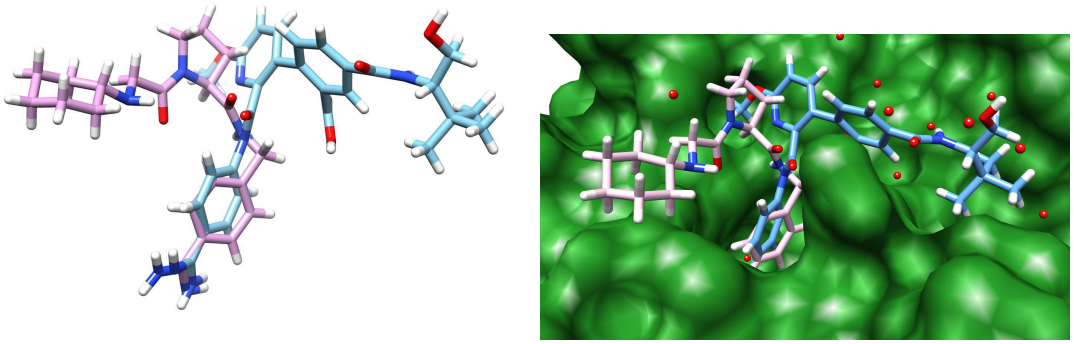


Figure 6.1.: Example of a reference alignment with only partially overlapping ligands. Displayed are the trypsin ligands from 3LJO (pink) and 2AYW (blue). Reprinted from [107] with permission of Springer.

- **Root Mean Square Deviation** (RMSD)
  The RMSD compares the Cartesian coordinates p of a generated pose with the respective coordinates c in the reference alignment.

$$RSMD = \sqrt{\frac{\sum_{n}^{i=1}(c_i - p_i)^2}{n}} \tag{6.5}$$

  This calculation is only done with the n heavy atoms of a molecule, since the coordinates of hydrogen atoms are usually not resolved in a crystallographic structure. Furthermore, for molecules with symmetric groups, no unique assignments of atoms is possible, therefore the RMSD is calculated based on the best mapping of topological identical atoms.

- **RMSD-O**
  A slight modification to the RMSD calculation has been performed to restrict the calculation of the RMSD to only those atoms of a molecule which are actually part of a conserved binding mode represented by the reference alignment. The RMSD-O is restricted to all atoms a of M for which at least one atom b exists in the reference alignment of M and N that is within a distance of 2.0Å to a. Figure 6.1 shows an example of a reference alignment of two trypsin ligands from the mRAISE dataset. It can be seen that the ligands overlap only partially, which makes only a part of this alignment recoverable using LBVS. The restricted RMSD, representing the actual overlay region of the reference alignments, is called RMSD-O and is used in all evaluations of alignments calculated with mRAISE.

For experiments evaluating a calculated pose of a molecule with respect to a reference derived from a crystal structure, an RMSD of less than 2.0Å is considered

a good result. For the evaluation, average and median RMSD-O values of the best scored pose as well as the lowest RMSD-O among the ten best scored poses for each ensemble of the dataset are reported.

6.2. Experiments

In the following, the performed experiments are described in detail. For the preparation of the used dataset see Chapter 5.

6.2.1. Enrichment Study on the DUD

The DUD dataset (see Section 5.1) offers the opportunity to compare the enrichment performance of mRAISE to a variety of different methods using multiple evaluation metrics. Each of the 40 targets is screened using the respective crystallographic query structure provided by the dataset.
The resulting ranked hit lists were used to calculate the AUC as well as the ER and HR at 1%, 5% and 10% of ranked hits. The experiment allows a direct and detailed comparison of the performance on each target to the methods LIGSIFT, Align-It and ROCS. Furthermore, the average AUC values are used to compare to the performance of ShaEP and MolShaCS, as well as to the 2D fingerprint method ECFP4. In addition, the individual performances of different mRAISE modes are compared and the influence of the automatically derived partial shape constraints is analyzed.

6.2.2. Enrichment Study on a DUD Subset

A small experiment on 11 targets of the DUD dataset is performed to compare the performance of mRAISE to the additional methods Surflex-sim, FlexS, ICMsim, and again ROCS, using the ER metric.
The screening procedure is equal to that used in the complete DUD experiment and the ER is calculated for 1% and 10% of ranked hits.

6.2.3. Enrichment Study on the DUD-E

Besides its frequent use, some issues of the DUD dataset motivated the creation of the more challenging and more diverse DUD-E dataset (see Section 5.2). Reproducible evaluation studies on this dataset allowing direct comparison are quite rare

in the literature. Nevertheless, to further highlight the capabilities of mRAISE and to have a broader basis for the statistical analysis of the influence of partial shape constraints, the whole DUD-E dataset has been screened.
As in the classical DUD experiments, all 102 targets of the DUD-E have been screened using the provided ligand of the query structure. All modes with automatically derived partial shape constraints in mRAISE were compared using the AUC as well as the ER and HR at 1%, 5% and 10% of ranked hits.

6.2.4. The Influence of Manual Partial Shape Constraints

The ability to define manual partial shape constraints as provided by mRAISE (see Section 4.5.2) is an extremely complex and variable tool. Therefore, it is extraordinary difficult to design experiments for an objective evaluation. The definition of meaningful constraints requires an experienced user with a good biochemical background and preferably expert knowledge about the target or compound class of interest. Five targets, which were considered difficult in terms of ligand flexibility and size as well as showing insufficiently high AUC values (less than 0.7) using mRAISE_classic were chosen from the DUD dataset to nevertheless show the possible impact of manually defined shape constraints. For these five targets, manual shape constraints are defined and used for screening of the respective descriptor indices. As evaluation metrics, again, the AUC as well as the ER and HR at 1%, 5% and 10% of ranked hits have been used.

6.2.5. Alignment Quality Evaluation

To evaluate the quality of molecular alignments calculated with the automated modes of mRAISE, each ligand of each ensemble of the mRAISE dataset has been screened against each other member of the same ensemble. As described before, average and median RMSD-O values are calculated based on the best scored poses for each conformation of the target molecule and the respective pose from the reference alignment. For the actual screening process, the query conformation is taken directly from the input file whereas only generated conformations are used for the screened target molecule.
Average and median RMSD-O values are reported with respect to the best scored pose for each comparison as well as with respect to the best RMSD-O within the ten best scored conformations.

7 Chapter 7. Results and Discussion

The following chapter presents and discusses the results of the experiments introduced in Section 6.2 to evaluate and compare the performance of mRAISE as well as to analyze the influence of the different partial shape concepts.
In the first part, the ranking capabilities and the enrichment of the method are discussed based on experiments performed using the DUD as well as the DUD-E dataset(see Section 7.1). Herein the DUD is primarily used to compare to other LBVS methods, while the more challenging DUD-E is used to reinforce the performance results of mRAISE. The different automated modes for the generation of partial shape constrains are also analyzed based on their performance on both datasets and the potential of manually selected partial shape constraints is highlighted on a special selection of DUD targets (see Section 7.1.4). In the second part of the chapter, the results of the alignments experiments on the mRAISE dataset are shown and discussed. Again, the influence of the automatically generated partial shape constraints is analyzed (see Section 7.2). Finally, in the last part of the chapter, the computing time of mRAISE is compared to other methods provided that this information is available (see Section 7.3).
For the differentiation between the different possible setups of mRAISE, including the partial shape approaches, the terminology introduced in Chapter 6 is used.

7.1. Enrichment Experiments

Enrichment studies analyze the ability of a LBVS method to separate compounds active to the same targets as a query ligand from inactive compounds. For this purpose, retrospective experiments are performed on datasets with annotated actives and decoys to specific targets (see Chapter 5). Based on the used evaluation

Table 7.1.: Average AUC values for all DUD targets.

	avg. AUC	median AUC
LIGSIFT	0.79 ± 0.20	0.82
mRAISE	0.76 ± 0.19	0.84
Align-It	0.75 ± 0.23	0.79
ROCS	0.73 ± 0.20	0.78
SHAEP	0.64 ± 0.17	NA
MolShaCS	0.63 ± 0.08	NA

Values with standard deviation. Reprinted from [107] with permission of Springer.

metrics, the overall performance as well as the specific enrichment among top-ranked hits can be validated (see Section 6.1).

7.1.1. Enrichment Study on DUD

The DUD dataset has been available for a decade and since its release it has been used in multiple evaluation and comparison studies. In the field of LBVS it is crucial to compare the performance of a new method with the best and most recent available methods as well as to highlight the individual strengths of a method. This made it almost inevitable to embrace the opportunity to use the DUD dataset as basis for a broad comparison study using standard evaluation metrics.

Overall Enrichment

On average, mRAISE_classic achieves an AUC of 0.76 ± 0.19 on the DUD dataset (see Table 7.1). In comparison to the average performances of five other methods, this performance is second best with only LIGSIFT showing an higher average AUC of 0.79 ± 0.20. However, looking at the less outlier-dependent median value of AUC on all DUD targets, mRAISE achieves the highest value with 0.84 compared to LIGSIFT as second best with a median AUC of 0.82.

The complete overall screening results of mRAISE_classic on all 40 targets of the DUD in comparison to the performances of ROCS, Align-It and LIGSIFT can be seen in Table A.4. Looking at the individual results for each target, mRAISE_classic achieves a performances better than random selection (AUC >0.5) in 36 of the 40 cases. Comparing the individual performances with the other methods, mRAISE has the best or equals the best performance compared to the other three methods

Table 7.2.: Average EF at 1%, 5% and 10 % for all targets of the DUD.

	$EF_{1\%}$	$EF_{5\%}$	$EF_{10\%}$
LIGSIFT	20.8 ± 12.6	9.3 ± 6.0	5.4 ± 3.0
mRAISE	20.2 ± 12.1	9.4 ± 6.0	5.4 ± 3.0
ROCS	19.4 ± 12.9	8.4 ± 6.0	5.2 ± 3.0
Align-it	16.9 ± 12.5	8.1 ± 5.9	4.9 ± 3.2

Values with standard deviation. Reprinted from [107] with permission of Springer.

Table 7.3.: Average HR at 1%, 5% and 10 % for all targets of the DUD.

	$HR_{1\%}$	$HR_{5\%}$	$HR_{10\%}$
LIGSIFT	59.0 ± 35.6	46.6 ± 30.2	54.4 ± 29.9
mRAISE	55.5 ± 33.3	46.7± 30.0	53.9 ± 30.6
ROCS	54.6 ± 36.3	38.3 ± 30.0	46.0 ± 30.4
Align-it	48.0 ± 35.5	40.6 ± 29.6	49.4 ± 31.8

Values with standard deviation. Reprinted from [107] with permission of Springer.

in terms of the AUC on 13 targets.

Early Enrichment

Besides the overall performance, the ability to rank true actives to especially the top ranks of a score-ordered list can be of even more interest, since usually only a fraction of the top scored hits is taken into consideration for further studies. Table 7.2 shows the average EF and Table 7.3 shows the average HR at 1%, 5% and 10% of ranked hits for mRAISE, LIGSIFT, ROCS and Align-It.

Looking only at the first percent of the hit list, mRAISE achieves the second best EF with 20.2 ± 12.1, again only just exceeded by LIGSIFT with 20.8 ± 12.6. According to the HR, mRAISE on average retrieves 55.5% ± 33.3% of the actives which possibly could be found only looking at 1% of the respective dataset which is about 1% more than ROCS and 7.5% more than Align-It, but 3.5% less than LIGSIFT.

Considering the top 5% of ranked hits however, mRAISE exceeds the other methods slightly with an EF of 9.4 ± 6.0 and an HR of 46.7± 30.0.

Finally, looking at the first 10% of ranked hits, mRAISE and LIGSIFT both have the best EF of 5.4 ± 3.0, while the HR shows, that LIGSIFT is slightly closer to the optimal enrichment with 54.4 ± 29.9 compared to 53.9 ± 30.6 of mRAISE.

It has to be noted that while a comparison of the different methods is possible based on the DUD experiments for early as well as overall enrichment capabilities, it only allows a rough ranking of the used methods. Since the observed standard deviations are very high, no statistically significant differences can be observed comparing the methods. Furthermore even looking into the results per target, there is no method superior to the other methods in all cases. Another important point for the comparison of 3D virtual screening methods, is the fact that the performance also depends on the quality of the used conformations. The methods presented in this experiment used different algorithms for this purpose. For example ROCS and LIGSIFT used conformations generated with OMEGA while the conformations used in mRAISE were calculated using the latest version of CONFECT (see Section 4.1.2). Nevertheless, the DUD experiment shows that mRAISE provides a good overall as well as early enrichment of screening libraries, which is comparable and in individual cases superior to other state of the art methods.

Class-specific Enrichment

Looking at protein families instead of single targets, can provide the insight if a method is especially beneficial for the screening of special families. The protein families represented in the DUD dataset are Nuclear hormone receptors, Kinases, Serine proteases, Metalloenzymes, Folate enzymes. Figure 7.1 shows the average AUC of mRAISE, LIGSIFT, Align-It and ROCS for the six protein families represented in the DUD dataset. All four methods show similar trends on all families including the best performances on the Folate enzymes and the worst performance on the Kinases. mRAISE achieves an AUC of 0.8 and better for four of the six families and has its worst performance on the Kinases with an average AUC of 0.6.

As can be seen, the performance of mRAISE for the different protein classes follows the same trend as the other LBVS methods and while it does not significantly exceed the performance of the other methods in any case, it is also not inferior to the other methods. However, it should be noted that some of the families only have very few members, e.g. four Metalloenzymes, three Serine proteases and two Folate enzymes. It is therefore difficult to draw statistically relevant conclusions based on this information.

Influence of Partial Shape Constraints

histogram The most important features integrated in mRAISE are the new concepts to derive partial shape constraints for virtual screening. Following the general performance analysis of mRAISE_classic on the DUD, now the influence of mRAISE_inclusion and mRAISE_contact will be shown and discussed.

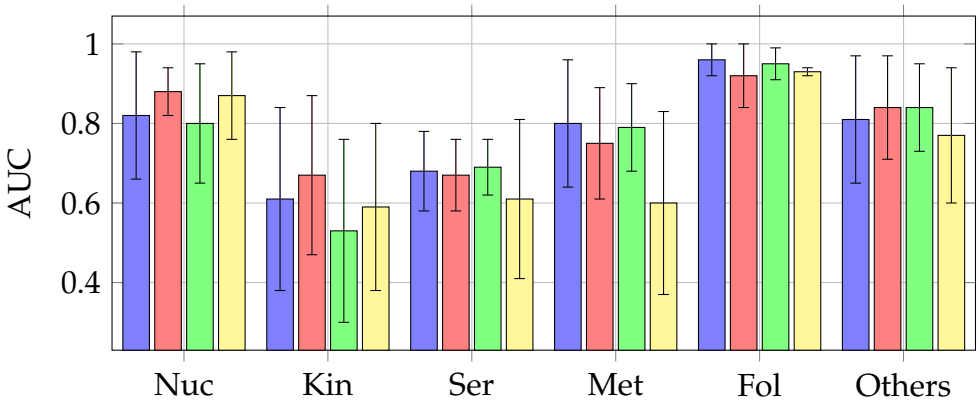


Figure 7.1.: Average AUC on the protein families present in DUD using mRAISE (, LIGSIFT ( ), Align-It( ) and ROCS( ) . Nuc = Nuclear hormone receptors, Kin = Kinases, Ser = Serine proteases, Met = Metalloenzymes, Fol = Folate enzymes)

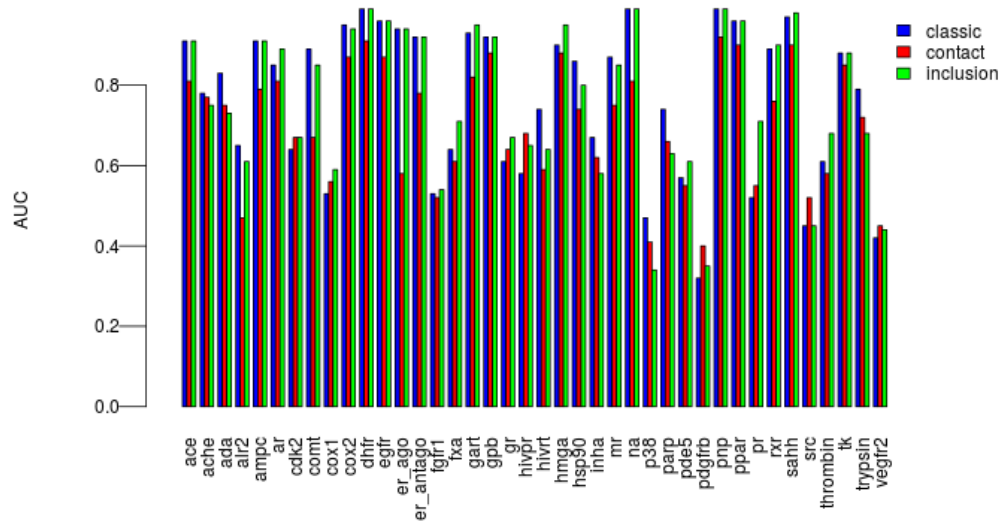


Figure 7.2.: Overview of the results on the DUD dataset for mRAISE_classic, mRAISE_inclusion and mRAISE_contact. Reprinted from [112].

Table 7.4.: Average and median AUC values of the ROC curves on the DUD dataset.

Mode	avg. AUC	median AUC
mRAISE_classic	0.76 ± 0.19	0.84
mRAISE_contact	0.70 ± 0.15	0.73
mRAISE_inclusion	0.76 ± 0.19	0.78

Values with standard deviation. Reprinted from [112].

Table 7.4 shows the average and median AUC values for all 40 targets of the DUD using the three different modes of mRAISE. Concerning the overall performance the different modes perform quite similar, mRAISE_inclusion has the same AUC as mRAISE_classic with 0.76 ± 0.19. Looking at the median value shows a slightly better performance of mRAISE_classic with a value of 0.84 in comparison to 0.78 using mRAISE_inclusion. mRAISE_comparison on the other hand shows a weaker overall performance with an average AUC of 0.70 ± and a median of 0.73. The overall similar performance and the trend for mRAISE_classic and mRAISE_inclusion to perform equally and at the same time slightly better than mRAISE_contact can also be seen in Figure 7.2.
Comparing the individual performances per target shows that mRAISE_classic has the best or is equal to the best performance for 23 out of the 40 targets. However, for some targets, the partial shape constraints cause notable improvements on the AUC. For mRAISE_inclusion, the most apparent improvements can be seen on the Progesterone receptor (+0.19), Factor Xa (+0.07), Thrombin (+0.07), Cyclooxygenase 1 (+0.06), Glucocorticoid receptor (+0.05), and Hydroxymethylglutaryl-CoA reductase (+0.05). On the other hand, mRAISE_contact also shows improved screening performance on some individual targets, e.g. on the HIV protease (+0.1), Platelet derived growth factor receptor kinase (+0.08), and Tyrosine kinase SRC (+0.07).
Looking at the early enrichment, mRAISE_classic has the highest $EF_{1\%}$ with 20.2 ± 12.1 compared to 19.3 ± 12.1 of mRAISE_contact and 19.9 ± 12.3 of mRAISE_inclusion. However looking at 5% and 10% of ranked hits, mRAISE_inclusion performs slightly better than mRAISE_classic with 9.5 ± 6.0 and 5.5 ± 3.1 compared to 9.4 ± 6.0 and 5.5 ± 3.1. Again, mRAISE_contact shows a slightly worse performance with 8.5 ± 5.4 and 4.8 ± 2.7 respectively.
The average HR as shown in Table 7.6 further illustrates the observations of the EF. At 1% of ranked data, mRAISE_classic shows a slightly better performance than mRAISE_inclusion, while the performance of mRAISE_inclusion is best at 5% and 10% of considered hits and mRAISE_contact shows the worst performance on all three percentages. Nevertheless, individual performances on certain DUD targets again show a significant improvement using the complex-derived constraints. Looking at the detailed results of the enrichment at the first percentage of ranked hits (see Table A.8), which is usually of the highest interest, ten out of the 40 targets

Table 7.5.: Average enrichment factor on the DUD dataset at one five and ten percent of ranked hits.

Mode	$EF_{1\%}$	$EF_{5\%}$	$EF_{10\%}$
mRAISE_classic	20.2 ± 12.1	9.4 ± 6.0	5.4 ± 3.0
mRAISE_contact	19.3 ± 12.1	8.5 ± 5.4	4.8 ± 2.7
mRAISE_inclusion	19.9 ± 12.3	9.5 ± 6.0	5.5 ± 3.1

Values with standard deviation. Reprinted from [112].

Table 7.6.: Average hitrate on the DUD dataset at one five and ten percent of ranked hits.

Mode	$HR_{1\%}$	$HR_{5\%}$	$HR_{10\%}$
mRAISE_classic	55.5 ± 33.3	46.7 ± 30.0	53.9 ± 30.6
mRAISE_contact	53.0 ± 33.2	42.1 ± 26.5	48.0 ± 27.2
mRAISE_inclusion	54.6 ± 33.9	47.3 ± 30.0	54.6 ± 30.5

Values with standard deviation. Reprinted from [112].

show an improvement compared to mRAISE_classic using either mRAISE_contact or mRAISE_inclusion. In five of these cases, the HR increases by more than 10, meaning that 10% more actives are found at that point. Figure 7.3 shows the average performance of the three mRAISE modes with respect to the six protein families represented in the DUD dataset. For all six classes, the results correspond to the other comparisons based on the overall and early enrichment. As can be seen, mRAISE_classic and mRAISE_inclusion perform similar while mRAISE_contact is generally slightly inferior to the other to modes.

The performed experiment showed that the complex derived constraints do not generally increase or decrease the performance of mRAISE_classic on average but still can have a significant impact on individual targets, e.g. on the HIV protease and the Progesterone receptor. A general rule for cases where the new partial shape constraints are especially beneficial could not be derived from the 40 targets of the DUD.

Comparison to a 2D Fingerprint

Besides the comparison to other 3D LBVS methods, the comparison to a state of the art 2D fingerprint method is also of high interest, since such methods usually only need a fraction of the time a 3D methods needs for a virtual screening run and are also not dependent on the conformation of a molecule. Table 7.7 shows a comparison of the average values for AUC, EF and HR between mRAISE and the

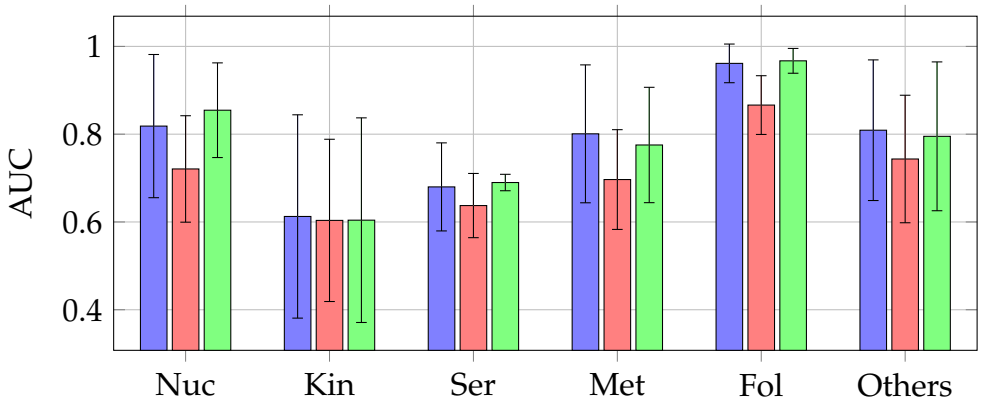


Figure 7.3.: Average AUC on the protein families present in DUD using mRAISE_classic (, mRAISE_contact ( ) and mRAISE_inclusion( ). Nuc = Nuclear hormone receptors, Kin = Kinases, Ser = Serine proteases, Met = Metalloenzymes, Fol = Folate enzymes)

Table 7.7.: Average performance using the $ECFP_4$ fingerprint on the DUD in comparison to mRAISE.

Method	AUC	$EF_{1\%}$	$EF_{5\%}$	$EF_{10\%}$
$ECFP_4$	0.74 ± 0.21	18.8 ± 11.8	8.8 ± 5.5	5.4 ± 3.0
mRAISE	0.76 ± 0.19	20.2 ± 12.1	9.4 ± 6.0	5.4 ± 3.0
	$HR_{1\%}$	$HR_{5\%}$	$HR_{10\%}$	
$ECFP_4$	51.5 ± 32.4	43 ± 27.6	53.84 ± 29.6	
mRAISE	55.5 ± 33.3	46.7 ± 30.0	53.9 ± 30.6	

Reprinted from [107] with permission of Springer.

$ECFP_4$ fingerprint.
While the overall performance between the 2D and the 3D method are very similar, especially the early enrichment is worse for the $ECFP_4$ with an $HR_{1\%}$ of 51.5 ± 32.4 compared to 55.5 ± 33.3.

The good results of the $ECFP_4$ fingerprint on the DUD is no surprise, since the dataset is especially designed to have topological dissimilarity between actives and decoys in order to prevent false negatives in the sets of decoys, which of course is an ideal premise for a topological fingerprint. Nevertheless, mRAISE_classic performs equally in terms of overall enrichment and slightly better in terms of the early enrichment with on average 4% more found actives at 1% of ranked hits.

Table 7.8.: Percentage of found actives at 1% and 10% ranked results.

	mRAISE		Surflex-sim		ROCS		FlexS		ICMsim	
Target	1%	10%	1%	10%	1%	10%	1%	10%	1%	10%
ada	**17.39**	**39.13**	5.13	23.08	7.69	33.33	15.38	30.77	10.26	28.21
cdk2	16.00	36.00	2.78	8.33	**22.22**	**38.89**	5.56	11.11	9.72	30.56
dhfr	**33.33**	**98.51**	4.88	25.61	19.02	61.95	8.05	27.80	20.24	80.73
er_ant	**17.95**	**89.74**	10.26	82.05	10.26	**89.74**	15.38	71.79	**17.95**	76.92
fxa	4.93	23.24	1.37	2.74	4.11	4.11	5.48	19.86	**11.64**	**46.58**
hivrt	20.00	40.00	9.30	18.60	20.93	25.58	**27.91**	**58.14**	18.60	39.53
na	32.65	**97.96**	16.33	55.10	**34.69**	89.80	20.41	69.39	16.33	73.47
p38	10.94	16.02	9.03	21.15	8.81	15.42	**12.56**	**26.21**	11.23	17.84
thrombin	**3.08**	6.15	1.39	15.28	0.75	15.28	2.78	12.5	2.78	**76.39**
tk	**31.81**	**63.64**	22.73	50.00	22.73	50.00	9.09	54.55	22.73	**63.64**
trypsin	2.27	11.36	0.00	59.18	4.08	34.69	**18.37**	18.37	10.20	**95.92**
Average	**17.31**	47.43	7.56	32.83	14.12	41.71	12.82	36.41	13.79	**57.25**

Highest values for each target highlighted in bold. Reprinted from [107] with permission of Springer.

7.1.2. Enrichment Study on a DUD Subset

In a comparison study by Giganti *et al.* [141], the early enrichment of four methods has been compared using only a subset of the DUD but otherwise under the same conditions as the previous DUD experiments. This enables a direct comparison to these tools even if it only is based on 11 of the 40 available targets, which have been selected by the authors based on their use in the literature for benchmarking studies and their diversity in terms of binding site properties. The used metric for this experiment is the ER, the plain percentage of found actives at a certain percentage of considered hits. The complete results can be seen in Table 7.8. As can be seen, no method is superior to the other methods in all of the cases. However, mRAISE_classic shows the highest percentage of found actives at 1% as well as at 10% of ranked hits for five of the 11 targets.

This study allowed an even further comparison between mRAISE and other methods of the field. Overall, for the 11 used targets of the DUD, mRAISE_classic shows the best amount of found actives at the first 1% of ranked hits and is also the second best at 10% following ICMsim.

7.1.3. Enrichment Study on DUD-E

To underline the results of mRAISE on the DUD on a more challenging and diverse dataset, as well as to further analyze the influence of partial shape constraints, an experiment on the DUD-E has been performed. The detailed results on all 102

Table 7.9.: Average and median AUC values of the ROC curves on the DUD-E dataset.

Mode	avg. AUC	median AUC
mRAISE_classic	0.74 ± 0.15	0.73
mRAISE_contact	0.72 ± 0.16	0.76
mRAISE_inclusion	0.72 ± 0.16	0.75

Values with standard deviation. Reprinted from [112].

targets of the DUD-E using mRAISE_classic can be seen in Table A.5 and Table A.6, this includes the AUC as well as the EF and HR at 1%, 5% and 10%. The same information can also be found for mRAISE_contact in Table A.9, Table A.10, and Table A.11 and for mRAISE_inclusion in Table A.12, Table A.13, and Table A.14.

Overall Enrichment

A comparison of the different mRAISE modes using the average results can be found in Table 7.9 On average mRAISE_classic achieves an AUC of 0.74 ± 0.15 on the DUD-E with only 8 of 102 targets showing an AUC of 0.5 of less. Comparing this performance to mRAISE_inclusion and mRAISE_contact shows that all three modes almost perform equally on average with both other modes achieving an average AUC of 0.72 ± 0.16. Looking at the median value, mRAISE_contact and mRAISE_inclusion are even slightly better than mRAISE_classic with 0.76 and 0.75 respectively compared to 0.73. Regarding the number of cases with worse than or equal to random performance however, mRAISE_contact has an AUC of 0.5 or less for 14 targets and mRAISE_inclusion for 10.

In Table 7.10 all targets showing an improved overall performance using complex-derived partial shape constraints by an AUC increase of 0.05 or more. If both modes show an increased AUC on the same target, only the best compared performance is listed. A special case is the HIV protease, since the AUC increase is the same for both methods.

Comparing the detailed results furthermore highlights, that the most apparent improvements of +0.1 and more occurred almost exclusively on targets with highly flexible actives. Apart from the Mineralocorticoid receptor (mcr), which shows an improved AUC by +0.11 using mRAISE_contact, all other cases occurred on targets with an average of eight or more rotatable bonds within the active compounds. This can also be seen in Table 7.11, which shows the percentage of targets showing an improved or equal overall performance compared to mRAISE_classic with at least eight, nine, or ten average rotatable bonds within the active compounds.

Table 7.10.: List of targets of the DUD-E which show an increased AUC using mRAISE_contact or mRAISE_equality in comparison to mRAISE_classic.

inclusion		contact	
Target	Increased AUC	Target	Increased AUC
bace1	+0.16	pa2ga	+0.16
ppard	+0.12	fkb1a	+0.15
xiap	+0.10	mcr	+0.11
fa10	+0.09	fak1	+0.10
pyrd	+0.08	fnta	+0.09
mapk2	+0.07	jak2	+0.08
mk10	+0.06	gcr	+0.06
dhi1	+0.05	pde5a	+0.05
dpp4	+0.05	pgh1	+0.05
hdac2	+0.05		
cdk2	+0.05		
hivpr +0.13			

The increased AUC is calculated with respect to the results of mRAISE_classic. The AUC of hivpr increases by the same amount for both modes. Reprinted from [112].

Table 7.11.: Percentage of DUD-E targets with equal or improved performance compared to mRAISE_classic and a certain number of rotatable bonds.

Average Number of Rotational Bonds	inclusion	contact
$\geq$8	52.4	52.4
$\geq$9	72.7	63.6
$\geq$10	100.0	100.0

Average numbers calculated using the actives for each target present in the DUD dataset. Reprinted from [112].

Table 7.12.: Average enrichment factor on the DUD-E dataset at one, five, and ten percent of ranked hits.

Mode	$EF_{1\%}$	$EF_{5\%}$	$EF_{10\%}$
mRAISE_classic	23.45 ± 17.00	7.78 ± 4.92	4.69 ± 2.50
mRAISE_contact	22.67 ± 17.10	7.37 ± 4.96	4.46 ± 2.53
mRAISE_inclusion	22.76 ± 17.04	7.45 ± 4.99	4.45 ± 2.51

Values with standard deviation. Reprinted from [112].

Table 7.13.: Average hitrate on the DUD-E dataset at one, five, and ten percent of ranked hits.

Mode	$HR_{1\%}$	$HR_{5\%}$	$HR_{10\%}$
mRAISE_classic	37.95 ± 26.36	38.94 ± 24.44	46.98 ± 24.78
mRAISE_contact	36.79 ± 27.04	36.96 ± 24.60	44.66 ± 25.09
mRAISE_inclusion	37.12 ± 27.14	37.37 ± 24.73	44.60 ± 24.93

Values with standard deviation. Reprinted from [112].

Early Enrichment

For the early enrichment, the average results of mRAISE_classic, mRAISE_contact and mRAISE _inclusion can be found in Table 7.12 for the EF and Table 7.13 for the HR. As can be seen, the early enrichment shows the same trend as the overall performance on the 102 DUD-E targets. While mRAISE_classic shows a slightly better performance at 1%, both complex-based modes exceed the performance of mRAISE_classic at 5% and 10%. However, individual cases highlight the strengths of each mode and the benefit of the derived constraints for virtual screening.

As can be seen, the experiments executed on the DUD-E further highlight the conclusions drawn from the DUD dataset. Overall, the three automated modes of mRAISE show a good but also very similar performance. However, looking into the individual cases where the complex derived constraints improved the screening performance allows new conclusions based on the larger DUD-E dataset. It has been shown, that the complex-derived constraints especially improve the performance on targets with large and highly flexible ligands and therefore reduce the dependency on the quality of the generated conformations. Since such cases are difficult to handle for 3D virtual screening methods in general, the complex-derived constraints could be a promising tool to address this problem in future studies.

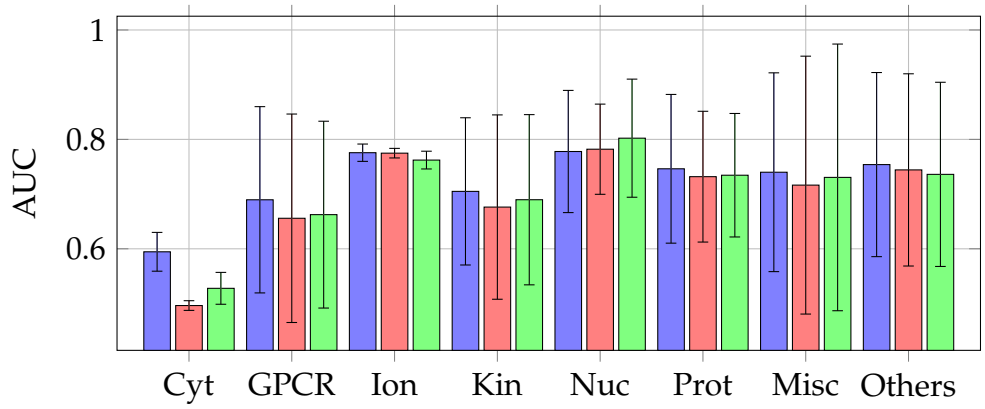


Figure 7.4.: Average AUC on the protein families present in DUD using mRAISE_classic (, mRAISE_contact ( ) and mRAISE_inclusion( ). Cyt = Cytochrome P450s, GPCR = GPCRs, Ion = Ion channels, Kin = Kinases, Nuc = Nuclear receptors, Prot = Proteases, Misc = Miscellaneous, Others = Other enzymes.)

Class-specific Enrichment

Since the DUD-E represents an extended number of targets compared to the DUD, it is even better suited for the evaluation of the screening performance based on the included protein classes. The protein classes represented in the DUD-E are Kinases, Proteases, Cytochrome P450s, Ion Channels, GPCRs, Nuclear Receptors and two unspecific groups containing ”other enzymes” and ”miscellaneous” proteins.
Figure 7.4 shows the average AUC values of the different mRAISE modes for each of the respective protein classes represented in the DUD-E. Like in the previous experiments on the DUD, the performance of all modes is very similar. Noticeable are the Cytochrome P450s and the GPCRs which show a better performance using mRAISE_classic, and the Nuclear receptors which show a slightly better performance for mRAISE_contact as well as mRAISE_inclusion.

Despite the small performance differences, the overall performance per protein class are comparable for almost all cases and the same trends are shown for each mode. Even looking at the classes showing a slightly improved average performance using one mode compared to the others is difficult, since the number of members in the specific classes is quite low and does not allow statistical significant conclusions. Especially the Cytochrome P450s only include two and the GPCRs only five targets. While the Nuclear receptors at least include 11 targets, the actual difference between the average AUC values is only 0.02 with a standard deviation between $\pm$ 0.08 and $\pm$ 0.11.

Table 7.14.: Overview of the five selected DUD targets for the mRAISE_selection experiment.

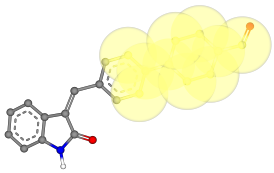
	Avg number of rotatable bonds	heavy atoms	AUC (mRAISE_classic)
hivpr	9.3	37.6	0.58
thrombin	7.1	32.7	0.61
fxa	6.8	32.5	0.64
pde5	5.7	31.2	0.57
fgfr1	5.4	29.7	0.53

Average numbers calculated using the actives for each target present in the DUD dataset. Reprinted from [112].

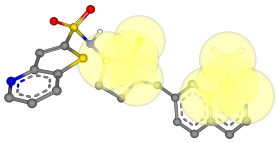
7.1.4. Manual Partial Shape Constraints

User-defined partial shape constraints via a manual selection of atoms is a very promising tool for virtual screening but it highly depends on the expert knowledge of the user. Therefore, an objective validation of this method is not a trivial task. Nonetheless, in order to highlight the possible impact of manually defined constraints, an experiment has been designed focusing only on a small number of particularly challenging targets selected from the DUD dataset.
In total, five targets from the DUD were chosen which were considered as difficult and also showed a weak performance (AUC <0.7) using mRAISE_classic. Difficulty is hereby defined by two criteria, the first is the average number of rotational bonds and the second is the average number of heavy atoms among the sets of actives of the target. These criteria follow the assumption that the larger and more flexible ligands are, the more challenging it is to find optimal solutions using a LBVS method. Both criteria in combination with the performance threshold lead to the same five targets which can be seen in Table 7.14.
Each query ligand for the respective targets has been loaded into the GUI version of mRAISE and a manual selection of atoms has been done only based on information drawn from the respective PDB entries as well as visual inspections of the protein-ligand complexes (see Figure 7.5).
The performance of these queries can be seen in Table 7.15 in comparison to the results of mRAISE_classic. For all five targets the AUC increased significantly with differences between 0.09 and 0.22.

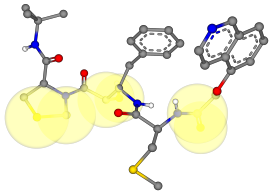
The shown results clearly highlight the potential of the manually selected partial shape constraints for LBVS with mRAISE. All five cases represented difficult targets and three of them only achieved performances just above random selection using mRAISE_classic. The actual selection of atoms will obviously always depend on the knowledge of the user, but for experienced users the impact can be drastic. Even in



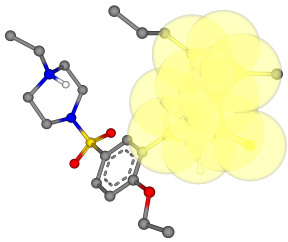
(a) fgfr1



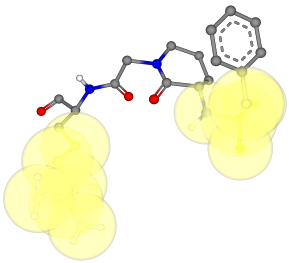
(b) fxa



(c) hivpr



(d) pde5



(e) thrombin

Figure 7.5.: Manual selection of atoms for the query ligands of five targets of the DUD dataset. The van der Waals radius of selected atoms is highlighted in yellow. Reprinted from [112].

Table 7.15.: AUC values of mRAISE with and without aid of manual selection.

	mRAISE_classic	mRAISE_selection
hivpr	0.58	0.69
thrombin	0.61	0.70
fxa	0.64	0.73
pde5	0.57	0.68
fgfr1	0.53	0.75

Reprinted from [112].

the shown cases an even better suited selection of atoms might further increase the displayed results.

7.2. Alignment Experiments

An important but often neglected part during the evaluation of 3D LBVS methods is the quality of the calculated three-dimensional alignments. However, the alignments are one of the important benefits of a three-dimensional method in comparison to two-dimensional methods which are usually faster and not necessarily far worse regarding the overall enrichment capabilities (see Section 7.1.1). Furthermore, the evaluation of the alignment quality shows if a method is able to correctly reproduce the binding mode of ligands binding to the same target and to thus superimpose the most important biochemical features of the molecules. For an experienced user, the alignment of top-ranked hits can also provide insight into the features which were prioritized by the respective method and especially in case of LBVS under manually selected constraints this can lead to a refinement of the applied selections. Lastly, the visual inspection of the alignments might encourage the user during an individual selection of promising hits for further experiments.

7.2.1. Alignment Quality Evaluation

The experiments to access the quality of molecular alignments calculated with mRAISE has been performed as described in Section 6.2.5. The summarized results for mRAISE_classic in comparison to the performances of mRAISE_contact as well as mRAISE_inclusion can be seen in Table 7.16 and Table 7.17 for the average and median RMSD-O values of the top-ranked conformation and for the best RMSD-O within the ten top-ranked conformations respectively. As mentioned in the Section 6.1, for the assessment of the alignment quality based on the RMSD of atom coordinates, an RMSD (in case of this evaluation an RMSD-O) of less than 2.0Å is considered as a successful recreation of the real binding mode.
Looking only at the top-ranked conformation, mRAISE_classic achieves an average RMSD-O of less than 2.0Å for four of the 11 ensembles while mRAISE_contact and mRAISE_inclusion achieve this for only three ensembles. However, for the ensemble showing an average RMSD-O of less than 2.0Å only for mRAISE_classic, the other modes also achieve values only slightly above 2.0Å. Looking at the less outlier dependent median RMSD-O values, mRAISE_classic succeeds only in three cases, while mRAISE_contact and mRAISE_inclusion achieve median RMSD-O values of less than 2.0Å for seven ensembles.
The results regarding the best RMSD-O values within the ten top-ranked hits

Table 7.16.: Comparison of the different methods on the mRAISE dataset. The shown results are the average (left) and median (right) RMSD-O value considering the best ranked conformations.

	classic		contact		inclusion	
Trypsin	**1.64**	**1.06**	**1.80**	**1.06**	**1.63**	**1.06**
Thrombin	2.95	2.43	2.93	**1.99**	3.20	2.20
ALPHA-MANNOSIDASE II	**1.96**	**1.77**	2.08	**1.40**	2.06	**1.45**
Matrix metalloproteinase-12 (MMP-12)	3.17	2.12	3.04	2.06	2.88	**1.89**
CDK 2 Kinase	2.83	**1.98**	2.50	**1.81**	2.44	**1.93**
Carbonic Anhydrase II	**1.70**	**1.53**	**1.71**	**1.54**	**1.69**	**1.53**
Thermolysin	3.19	2.16	2.18	**1.57**	2.07	**1.47**
CYP121	3.94	4.87	3.51	4.38	3.55	4.27
HIV Protease	2.93	2.56	2.26	2.16	2.51	2.44
Bromodomain-containing protein 4	3.62	4.98	4.50	5.81	3.54	3.24
Isopenicillin N Synthase	**1.74**	**1.63**	**1.74**	**1.63**	**1.57**	**1.47**

RMSD values smaller than 2.0Å highlighted in bold. Reprinted from [112].

Table 7.17.: Comparison of the different methods on the mRAISE dataset. The shown results are the average (left) and median (right) RMSD-O value considering the best value of the ten top-ranked conformations only.

	classic		contact		inclusion	
Trypsin	**1.40**	**0.95**	**1.55**	**0.94**	**1.42**	**0.95**
Thrombin	2.28	**1.72**	2.14	**1.57**	2.26	**1.69**
ALPHA-MANNOSIDASE II	**1.38**	**0.90**	**1.46**	**0.82**	**1.52**	**0.88**
Matrix metalloproteinase-12 (MMP-12)	2.74	**1.95**	2.52	**1.73**	2.34	**1.55**
CDK 2 Kinase	2.26	**1.28**	**1.87**	**1.26**	**1.91**	**1.45**
Carbonic Anhydrase II	**1.23**	**1.19**	**1.29**	**1.21**	**1.27**	**1.19**
Thermolysin	2.67	**1.74**	**1.83**	**1.52**	**1.76**	**1.42**
CYP121	2.85	3.35	2.77	3.30	2.83	3.69
HIV Protease	2.70	2.33	**1.93**	**1.78**	2.09	**1.89**
Bromodomain-containing protein 4 (BRD4)	3.16	4.55	3.85	4.84	2.75	2.40
Isopenicillin N Synthase	**1.53**	**1.51**	**1.49**	**1.46**	**1.39**	**1.34**

RMSD values smaller than 2.0Å highlighted in bold. Reprinted from [112].

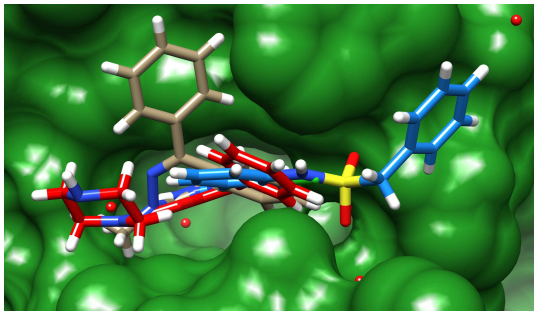


Figure 7.6.: Binding site of 4HXS and its bound ligand (blue) together with two additional members of the BRD 4 ensemble (3U5L in brown and 4CFL in red). Reprinted from [107] with permission of Springer.

further highlight an improved overall alignment quality using the complex-derived partial shape constraints. mRAISE_contact achieves average RMSD-O values of less than 2.0Åfor seven and mRAISE_inclusion for six of the 11 ensembles, while mRAISE_classic achieves this still only for the same four ensembles. Looking at the median RMSD-O values shows the same trend with mRAISE_contact and mRAISE_inclusion succeeding in nine cases while mRAISE_classic succeeds only in eight cases.
A detailed analysis of the performances on the different ensembles highlights particularly difficult cases and also shows limitations of the different methods and LBVS in general.

In the following, eight of the eleven ensembles are discussed in further detail. The ensembles of Serine protease, ALPHA-MANNOSIDASE II and Carbonic Anhydrase II are omitted, since they show a good performance for all modes already considering only the top-ranked conformation. Of special interest are the CYP121 and the BRD4 ensemble, since all mRAISE modes fail to achieve good average and median RMSD-O values for those ensembles.

The **BRD4** ensemble seems to be one of the most challenging ensembles of the mRAISE dataset. This is due to the fact that the binding site of BRD4 is a rather narrow and all ligand poses of the ensemble are only loosely fixed by one hydrogen bond in the center of the pocket. Apart from that one directed interaction, substantial parts of the ligands fold into different directions on the surrounding surface of the binding site (see Figure 7.6). As a consequence, the ligands only overlap partially in the are of the hydrogen bond, but besides that have no compulsion to further lie upon each other. While neither the classic version nor the complex-derived constraints succeed in achieving an average or median RMSD-O value of less than 2.0Å, especially mRAISE_inclusion significantly decreases the median RMSD-O value by 2.15 compared to mRAISE_classic. This effect can be explained by the fact that mRAISE_classic as well as mRAISE_contact are failing to find even one

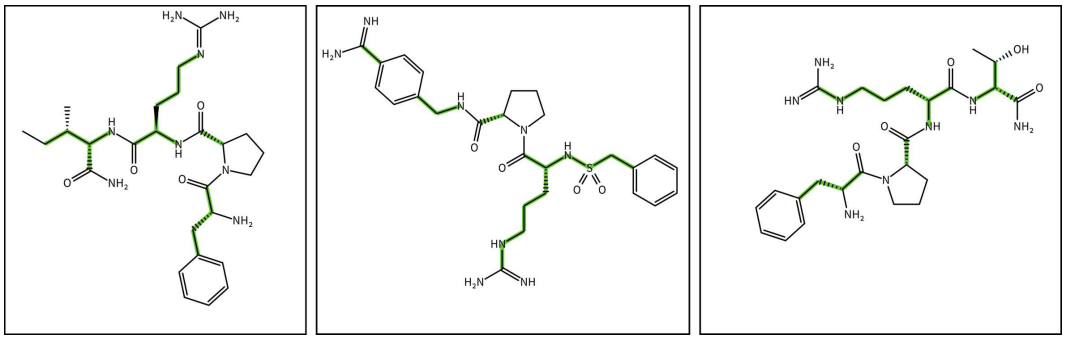


Figure 7.7.: Ligands of 3U8O, 3U8R and 3U98 with rotatable bonds highlighted in green. Reprinted from [107] with permission of Springer.

matching descriptor for 16 and 17 pairwise comparisons respectively. However, using mRAISE_inclusion not only lowers the average RMSD-O of all pairwise comparisons, it also finds matches for 14 of the problematic cases, explaining the significant difference in the median value. As can be seen, using constraints based on the binding site surface improved the performance on this structurally diverse ensemble.

CYP121 also seems to be a difficult target for ligand-based screening, since even looking at the ten best ranked conformations the best performance is achieved by mRAISE_inclusion with an average RMSD-O of 2.77 and a median RMSD-O of 3.30. Looking into the ensemble shows that the included ligands do not seem to have a conserved binding mode fixed by common hydrogen bonds, the only observed interaction present in each protein-ligand complex is an aromatic pi-pi interaction between hydrophobic rings of different sizes in the ligands and the residue Phe168A in the binding site. This is aggravated by the fact that all ligands of the ensemble have two to five functional groups that could function as a hydrogen bond donor or acceptor despite the fact that no common hydrogen bonds can be observed. Due to the fact that the scoring function used in mRAISE highly prioritizes the superimposition of potential candidates for hydrogen bonds (see Section 4.6), mRAISE struggles with the recreation of these binding poses mainly driven by an aromatic interaction.

For the **Thrombin** ensemble it is conspicuous that all three modes are able to achieve median RMSD-O values of less than 2.0Å at least within the ten top-ranked hits, but none achieves an average RMSD-O of less than 2.0Å. This can be attributed to three special members of the ensemble which show an overall worse RMSD-O than all other ligands of the ensemble. All three ligands (3U8O, 3U8R and 3U98) are conformationally very difficult with 14 rotatable bonds present in each structure as can be seen in Figure 7.7. The conformational difficulty of these ligands is highlighted by the fact that even the self-comparison of the respective

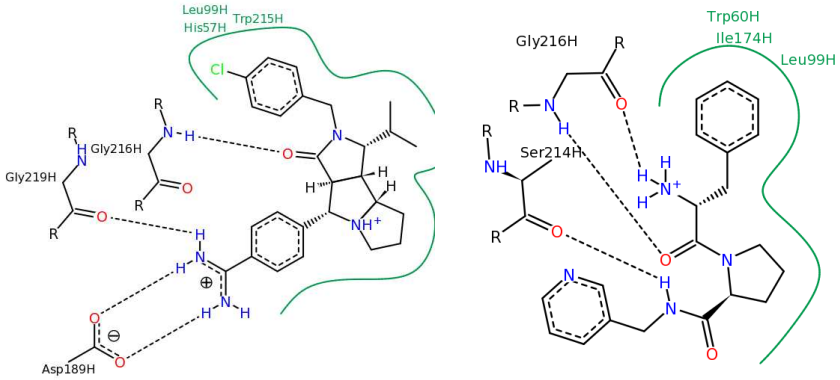
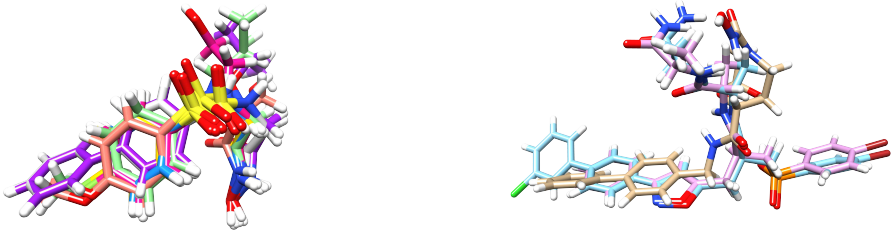


Figure 7.8.: Depiction of the different interaction modes of 2CF8 (top) and 3P17 (bottom). The top picture shows the interactions of the amidino group, which are missing only in 2ZFF and 3P17 (picture created with Poseview [142]). Reprinted from [107] with permission of Springer.

crystal structure against generated conformations is unable to achieve alignments with an RMSD-O of less than 2.0Å at first rank. This is further highlighted by the fact that most screening runs using one of the three crystal structures as query against any other member of the ensemble generally yields good RMSD-O values, while the screening of other structures against generated conformation of the three ligands usually results in very high RMSD-O values. Another interesting case are the ligands of 2ZFF and 3P17, which show a different binding mode in the protein-ligand complexes than the rest of the ensemble. While all other ligands form two characteristic hydrogen bonds Asp189 and Gly219 residues of the protein, they do not have such a group and form different interactions to the residues Gly216 and Ser214 (see Figure 7.8). For mRAISE_classic, this fact is especially problematic resulting in high RMSD-O values up to 7.46Å and in seven cases even no matching descriptors. However, mRAISE_contact is able to find matches for five of these cases and mRAISE_inclusion is actually able to align all seven of them.

The results for the **Trypsin** ligands are very good for all three modes. Nevertheless, the ligands in 4AB9, 4ABA, 4ABD, and 4ABE are a quite interesting case, because they are only small bound fragments. Small molecules are challenging for an approach like mRAISE, since they result in only a small amount of descriptors, making the matching more difficult. The best performing of the three fragments is the one bound to 4AB9, which achieves RMSD-O values of less than 2.0Å for a majority of the comparisons where matches are found. This is due to the fact that the other three ligands show a slightly different shale compared to the rest of the ensemble. This shapes leads to less prolate triangles, which are not present in

Figure 7.9.: Badly overlapping small fragments. Reprinted from [107] with permission from Springer.



(a) (b)

Figure 7.10.: Picture of the aligned active ligands of the Matrix metalloproteinase-12 with (a) and without (b) the common sulfonamide group. Reprinted from [112].

the other structures of the ensemble (see Figure 7.9). This different triangle base is also the reason why there is no significant difference for these cases when using complex-based shape constraints. Nevertheless, due to the size of the ensemble the overall result of the ensemble is still very good. Because of the amount of cases where no match is found while comparing other ligands to the four fragments, the median is the more reliable value to measure the performance on this ensemble.

The **MMP-12** ensemble is another case where all modes achieve a median RMSD-O of less than 2.0Å within the ten top-ranked hits, but none achieves this for the average RMSD-O. Again, three outliers can be identified consistently showing much higher RMSD-O values than the rest of the ensemble (2HU6, 4GR0, and 4GR8). Due to the relatively small size of the ensemble, the average RMSD-O value is strongly affected by these ligands. All three structures have in common, that they do not have a sulfonamide group which is present in all members of the ensemble, which is a good anchoring point for the alignment of those structures (see Figure 7.10). Comparing the three modes of mRAISE, mRAISE_inclusion handles these ligands the best which can also be seen in the reduction of the average RMSD-O by 0.4Å

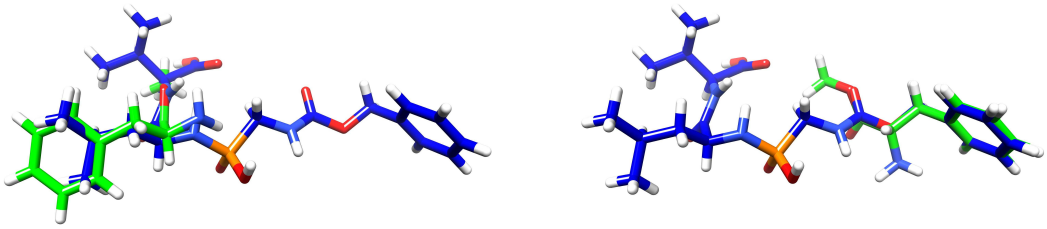


Figure 7.11.: Left ref alignment, right favorable alignments for volume overlap). Reprinted from [112] with permission of Springer.

for the ten-top ranked conformations.

The **Thermolysin** shows a significant improvement using the complex-derived partial shape constraints. While mRAISE_classic only achieves an average RMSD.O of 2.67Å even when considering the ten top-ranked conformations, mRAISE_contact as well as mRAISE_inclusion achieve average RMSD-O values of less than 2.0Å. The reason for this is the comparably very small ligand of 3QGO, which only consists of 13 heavy atoms. This is just half the number of the heavy atoms in the next smallest ligand in the ensemble. In Figure A.1 an overview of all contained molecules is shown. Besides the size of the ligand of 3QGO, the real problematic aspect is that the phenyl ring in this molecule can be ideally superimposed onto respective rings occurring in most of the other ensemble members. This superimposition is highly preferable in order to maximize the volume overlap of the molecules, but it does not represent the actual binding mode of the ligand and would not even place it in the binding site at all (see Figure 7.11). For this reason, mRAISE_classic is not able to recover the actual binding mode in this case resulting in RMSD-O values of 10Å and more at first rank in six out of nine alignments. As mentioned before, the complex-derived constraints significantly improve the alignments quality even in this special case. mRAISE_contact achieves an average RMSD-O of 1.83Å and mRAISE_inclusion an average RMSD-O of 1.76Å. Looking at the 5 comparisons where mRAISE_classic showed RMSD-O values of 10Å and more, mRAISE_inclusion calculates alignments with an RMSD-O of less than 2.71Å for all sic cases, while mRAISE_inclusion achieves this for five of the six cases.

Finally, the ligands of the **HIV protease** are another example for a beneficial influence of the complex-derived partial shape constraints. What makes this ensemble special are the highly flexible ligands, with nine out of the ten ligands having 12 or more rotatable bonds. The smaller average and median RMSD-O values of mRAISE_inclusion as well as mRAISE_contact in comparison to mRAISE_classic highlight the fact that the handling of highly flexible molecules is improved and

Table 7.18.: Percentage of pairs with an RMSD-O smaller than a certain threshold. Combined results of all ensembles.

Method	Percentage $\leq$ 2.5Å	Percentage $\leq$ 2.0Å	Percentage $\leq$ 1.5Å
mRAISE_classic	87.5	80.8	62.9
mRAISE_contact	86.8	80.2	62.6
mRAISE_inclusion	**87.8**	**81.1**	**65.7**

Highest values highlighted in bold. Reprinted from [112].

the dependency on the conformational quality is reduced. Especially the queries of mRAISE_contact demanding shape similarity only in certain areas of the molecule while allowing more flexibility in others were highly beneficial in this case. An improved handling of highly flexible ligands using the new concepts of partial shape constraints could already be shown for the HIV protease in the DUD and the the general DUD-E experiment (see Section 7.1.1 and Section 7.1.3) as well as the experiment evaluation the influence of manually defined constraints (see Section 7.1.4).

Apart from the ensemble-based evaluation of the alignment experiment, another interesting way to look at the results is the evaluating the overall performance as the percentage of pairwise comparisons achieving certain RMSD-O thresholds. Such an evaluation is shown in Table 7.18 for RMSD-O values of less than 2.5Å, 2.0Å and 1.5Å.

As can be seen, the performances of all three modes are relatively similar and mRAISE_inclusion only slightly exceeds the performance of the other two methods 87.8% achieving an RMSD-O of less than 2.5Å, 81.1% achieving an RMSD-O of less than 2.0Å and 65.7% achieving an RMSD-O of less than 1.5Å. This performance is closely followed by mRAISE_classic then by mRAISE_contact. Nevertheless, mRAISE_contact as well as mRAISE_inclusion showed an higher amount of ensembles with median and average RMSD-O values of less than 2.0Å (see Table 7.16 and Table 7.17). As a consequence, both modes have significantly less outliers with high RMSD-O values while otherwise showing equal RMSD-O values for the already good performing cases.

7.3. Computing Time

For the decision which LBVS method should be used for a screening project, two main aspects will be considered. The first is the performance of the methods as analyzed in the previous sections. The second aspect is the runtime of a method, which determines of how much practical use a method can be in an actual drug discovery

process. The time needed to separately screen the 40 targets of the DUD dataset has been measured for mRAISE_classic, mRAISE_contact, and mRAISE_inclusion as well as for the freely available method LIGSIFT. The DUD dataset provides a variety of different screening scenarios considering diverging library sizes as well as different query complexities. To directly compare the times, LIGSIFT has been used on the same conformations as mRAISE_classic. Furthermore, a general time performance of ROCS can be taken from the public documentation [143] for comparison.
Using mRAISE_classic the average time needed to screen one conformation is 6 milliseconds resulting in a screening performance of about 167 conformations per second. The complex-derived partial shape constraints available in mRAISE not only have an influence on the performance, but also on the screening time. While the time needed for the screening of the descriptor index is significantly reduced in both cases due to the less complex query of the shape descriptor rays, the number of matches and therefore the scoring time increased. For mRAISE_contact these changes balance out each other on average leading to a performance of 5.5 milliseconds per conformation. For mRAISE_inclusion this is not the case and the screening time increases to 12 milliseconds per conformation. Based on the documented information, ROCS is roughly a factor of four to five faster than mRAISE_classic, achieving a screening performance of 600-800 conformations per second. The computational expensive flexible handling of the ligands in LIGSIFT is expressed in the comparably higher screening of only one to two conformations per second, which is roughly a factor of 90 slower than mRAISE_classic.

It should however be noted, that the flexible approach of LIGSIFT also works without the need of additional generated conformations. This would reduce the screening time drastically to about the same order of magnitude of mRAISE, but at the same time also reduce the screening accuracy. All performances of LIGSIFT discussed in this chapter were calculated using the same number of conformations as in mRAISE.
While this study allows a general ranking of screening times for the presented methods, it must furthermore be noted that the runtimes varied significantly for the different target libraries of the dataset. Multiple different aspects can have a strong influence of the screening time. For mRAISE, first to mention is the number of descriptors of the query ligand. The fastest performance of mRAISE_classic on all DUD targets was for a query ligand with only 20 calculated descriptors. For this target the screening time was actually about a factor of three faster than the average performance of ROCS. The second aspect influencing the screening time of mRAISE is the diversity of the screening library, since the descriptor based approach could filter out highly dissimilar ligands already in early stages of the descriptor matching. The provided times of mRAISE were calculated on the assumption that descriptor indices were already created beforehand. Since a descriptor index would ideally be screened multiple times for different screening projects, the time needed for

the initial creation of the index should be considered separately. On average, the creation of a descriptor index with precalculated ligand conformations takes 30 milliseconds per conformation.

8 Chapter 8. Conclusion

In this thesis, a new method for LBVS has been introduced and integrated in the new software mRAISE. The main focus of the development was on the incorporation of knowledge-based partial shape constraints and accurate similarity scoring. In addition to the method development, another important part of the thesis was the evaluation of different aspects of the method and the comparison to other LBVS methods. This included the development of a new dataset for the purpose of alignment quality evaluation. For the manual definition of partial shape constraints as well as for the visual inspection of query descriptors and screening solutions, mRAISE was created in two different version. The first is a tool with a command line interface and the second is a tool with a GUI. In the following, the achievements of the newly developed method are discussed. Furthermore, remaining limitations and possible future improvements are shown.

8.1. Achievements

The discussion of the achievements is structured based on the objectives defined in Chapter 3. The evaluation of mRAISE showed that it can keep up with and in some cases even exceed the performance of the best methods in the field of LBVS. Hereby mRAISE achieves an excellent balance between the screening performance and the required computing time. Furthermore, the introduced concepts for partial shape matching are an innovation for the field and their benefits could be shown on multiple examples.

Efficient Data Handling

In mRAISE, molecules of a screening library can be initialized from different established file formats and are stored in a database using a memory-saving format and allowing rapid access during the actual screening procedure. Furthermore, the integrated CONFECT method allows to directly enumerate additional conformations for each molecule, which are stored as alternative instances of the respective original molecule entry in the database. Therefore, no additional software is needed to prepare a screening library for future studies. The usage of an SQLite database hereby allows easy portability and avoids the need of separate database servers.

Meaningful Abstraction of Ligand Information

A triangle descriptor representation for ligands, which has originally been developed for the field of SBVS is used in mRAISE for a rapid initial screening of compound libraries. After this phase, only molecules with matching descriptors need to be reinitialized and processed on the atomic level. The descriptor has been updated to better fit the requirements for ligand-based screening and new methods for partial matching of the included shape descriptor have been introduced. The usage of a bitmap-index allows the fast comparison and efficient storage of preprocessed descriptors for a whole compound library.
The information represented by the descriptor, e.g. interactions and a local shape description abstract molecules on a high level allowing the method to discard obviously dissimilar molecules at early stages and providing meaningful alignments as basis for the scoring procedure.

Knowledge-based Partial Shape Constraints

Three different modes for the creation of partial shape constraints based on different information sources have been developed. Additionally, a previously developed method matching molecules with a certain percentage of shape similarity is available as basic mode in mRAISE. Two of the newly developed modes utilize information derived from protein-ligand complexes. The first aims at creating queries that match molecules fitting into the same area of the binding site rather than being of the same shape as the query ligand. The second aims at finding matches, which are able to form close contacts to the protein in the same areas as the query ligand does and at the same time allows more flexibility in areas where no such contacts occur. Finally the third mode allows a user to manually define regions of the query ligand by selecting heavy atoms to define important regions of the ligand. The

screening procedure then matches molecules with similar shape in the respective regions and allows arbitrary shape in other regions.

Evaluation Data

For the enrichment evaluation of LBVS methods in retrospective studies, datasets providing known active ligands to certain targets together with a challenging set of decoys are needed. A multitude of datasets of various sizes and diversities are available in the literature. For the evaluation of mRAISE two of these datasets were chose. First, the DUD dataset is chosen because it allowed the direct comparison to a variety of different state of the art methods. Second, the more diverse as well as more challenging DUD-E dataset has been used to further confirm the results.
For the evaluation of the quality of molecular alignments, a new dataset has been introduced consisting of 180 prealigned ligands for 11 diverse targets.

Performance

Based on two datasets taken from the literature, enrichment experiments have been performed and multiple different evaluation metrics have been calculated to analyze the performance capabilities of mRAISE. Additionally, a newly developed dataset has been used for the often neglected evaluation of the quality of calculated molecular alignments.
The performance of mRAISE has been compared to multiple different 2D and 3D state of the art LBVS methods and the influence of complex-derived partial shape constraints on the results of all experiments have been analyzed. The experiments on the DUD dataset showed that mRAISE provides a good overall and early enrichment, which is comparable to and in individual cases even superior to the performance of other state of the art methods. Furthermore, the experiments on the DUD-E dataset confirmed these results using a more diverse and better designed dataset. While the complex-derived partial shape constraints were not able to increase the average performance of mRAISE, it could be seen that they especially benefit the screening performance in cases with highly flexible actives. Finally, a small experiment on five DUD targets using the manually selected partial shape constraints showed a significant improvement in the overall enrichment and highlighted the possible impact of this mode especially on highly flexible difficult targets. Concerning the alignments quality, mRAISE showed good results in reproducing biologically relevant ligand poses and in aligning the important features of ligands with respect to a conserved binding mode. The usage of the complex-derived partial shape constraints showed to be beneficial in creating more accurate molecular alignments.

Screening of compound libraries with mRAISE on average takes between 5.5 milliseconds (mRAISE_contact) and 12 (mRAISE_inclusion) milliseconds per conformation depending on the partial shape matching mode. This allows the rapid screening of large compound libraries in a reasonable amount of time. In comparison to two other methods, mRAISE achieves a good balance between computing time and screening quality with one other method being faster but also showing lower enrichments and another method showing slightly better enrichment results but at the same time being much slower.

Usability

The experiments have shown that mRAISE is capable of identifying ligands active to the same targets within a background of decoys and to align them accurately with respect to a shared binding mode.
Another important aspect of the development of a new software is the usability by unexperienced users without a background in computer science. For a LBVS method the user group it should be designed for consist mostly of medicinal chemists.
Therefore, mRAISE is not only designed as a command line tool, but also as a GUI version providing the same functionality with additional options for 3D visualization. While the command line version is important for exhaustive automated screening runs especially on computer clusters, the command line version additionally allows the user to browse query ligands and their respective descriptors as well as to define partial shape constraints by the manual selection of heavy atoms. Furthermore, the GUI version allows the visual inspection of the results aligned to the query molecule, which allows a medicinal chemist to evaluate the usefulness of top ranked hits for further investigations based on their expert knowledge.

8.2. Limitations

The research goals set for this project have mostly been addressed successfully and the developed method showed a good performances regarding enrichment as well as alignment quality studies. Nevertheless, the evaluation of the method showed some limitations to the method, which are discussed in the following:

Ligand flexibility The incorporation of the structural flexibility of ligands into virtual screening is still an ongoing challenge. For LBVS, can either be addressed by increasing the size of the screening library by additional conformations of the

included molecules or by computationally expensive algorithms simulating molecular flexibility during the screening procedure. Both approaches are difficult in their own way, but the generation of additional conformations is the more common approach, since rapid screening methods generally tend to be faster and not worse in the overall screening quality than recent approaches of flexible ligand alignments. In mRAISE conformations generated by the CONFECT algorithm are used to address this problem and the new methods for partial shape concepts developed in this project also showed a better handling of highly flexible compounds. Nevertheless, generating a fixed number of conformations can never really reproduce the complete conformational space of individual structures and mRAISE ist therefore always dependent on the quality of the calculated conformations.

Hydrophobic ligands During the evaluation of mRAISE, individual problematic molecules occurred, which are mostly hydrophobic. The descriptor generation for such molecules can be very difficult especially if the are few polar interactions present in the structure, since no descriptors with only hydrophobic triangle corners are generated. As a consequence the chance of finding matching descriptors decreases and parts of the ligand might not be represented by descriptors at all. mRAISE therefore struggles especially concerning the recreation of hydrophobic driven binding modes as could be seen during the alignment quality evaluation for the CYP121 ensemble of the mRAISE dataset.

Memory usage In the last phases of the evaluation of mRAISE, especially the complex-derived partial shape queries showed a significant increase in the amount of descriptor matches. For now, matches are hold in memory and are not processed until all matches are collected. This is due to the fact that the current version of the descriptor index does no longer follow the concept of descriptor partitioning in terms of certain numbers of ligands per partition, but in terms of certain numbers of descriptors per partition. As a consequence, it is no longer guaranteed that all matches corresponding to one special ligand are found after screening one partition. Therefore, to prevent that a ligand has to be reinitialized more than once, all partitions need to be screened before the scoring procedure can start.
During the development of mRAISE this was no problem, since the information stored in a descriptor match is quite small and the amount of matches in LBVS is not comparable to the numbers reached in SBVS. However, for complex-derived queries especially on large datasets the memory requirement is increased.

8.3. Outlook

Based on the previously discussed limitations as well as additional ideas to further improve the performance of mRAISE, possible future steps are presented:

- Alignment post-optimization. Individual inspection of molecular alignments showed that in some cases small changes on just one rotational bond of a ligand could already improve the overlap of the structures and therefore increase the similarity score significantly. A post-optimization of the conformations of matching compounds could improve the results and better cover the actual flexibility of the molecules than just a fix number of generated conformations.
- Partial shape constraints based on ligand ensembles. In case of multiple known ligands for the target of interest, partial shape constraints could be derived based on the noticeable similarities between aligned ligands. As a start, an easy way to use this information is to automatically derive atom selections and use them as if selected by a user for manual partial shape constraints.
- Incorporate partial shape constraints in the scoring function. Another interesting idea would be to use special weights during the scoring of atoms which are part of the selected partial shape constraints to value these regions of the ligands higher than the rest of the molecule.
- Hydrophobic directions. Assigning directions to hydrophobic interaction points would make them more meaningful during matching. Such a direction could for example represent the orientation of a respective ring or the position of outgoing bonds of a hydrophobic atom.
- Solution database. The increasing amount of matches observed for complex-derived queries could be faced using a database to store the match results on the hard drive as used in cRAISE, even if this might slow down the process.

Finally, it would be very interesting to use mRAISE in some real life lead identification processes with subsequent experimental validation of the results and to see how the method would perform in such a scenario.

Bibliography

[1] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg, and A. L. Schacht, "How to improve RD productivity: the pharmaceutical industry's grand challenge," *Nat Rev Drug Discov*, vol. 9, pp. 203–214, Mar 2010.

[2] S. Morgan, P. Grootendorst, J. Lexchin, C. Cunningham, and D. Greyson, "The cost of drug development: a systematic review," *Health Policy*, vol. 100, pp. 4–17, Apr 2011.

[3] J. Bajorath, "Integration of virtual and high-throughput screening," *Nat Rev Drug Discov*, vol. 1, pp. 882–894, Nov 2002.

[4] N. Moitessier, P. Englebienne, D. Lee, J. Lawandi, and C. R. Corbeil, "Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go," *Br. J. Pharmacol.*, vol. 153 Suppl 1, pp. 7–26, Mar 2008.

[5] Y. Tanrikulu, B. Kruger, and E. Proschak, "The holistic integration of virtual screening in drug discovery," *Drug Discov. Today*, vol. 18, pp. 358–364, Apr 2013.

[6] A. Lavecchia and C. Di Giovanni, "Virtual screening strategies in drug discovery: a critical review," *Curr. Med. Chem.*, vol. 20, no. 23, pp. 2839–2860, 2013.

[7] P. Ehrlich, "Ueber den zusammenhang von chemischer constitution und wirkung," *Münchener medizinische Wochenschrift*, pp. 1654–1655, 1898.

[8] E. Fischer, "Einfluss der configuration auf die wirkung der enzyme," *Berichte der deutschen chemischen Gesellschaft*, vol. 27, no. 3, pp. 2985–2993, 1894.

[9] J. N. Langley, "On the reaction of cells and of nerve-endings to certain poisons, chiefly as regards the reaction of striated muscle to nicotine and to curari," *The Journal of physiology*, vol. 33, no. 4-5, p. 374, 1905.

[10] P. Ehrlich, "Die grundlagen der experimentellen chemotherapie," *Angewandte Chemie*, vol. 23, no. 1, pp. 2–8, 1910.

[11] W. L. Bragg, "The specular reflection of x-rays.," *Nature*, vol. 90, p. 410, 1912.

[12] E. M. Purcell, H. Torrey, and R. V. Pound, "Resonance absorption by nuclear magnetic moments in a solid," *Physical review*, vol. 69, no. 1-2, p. 37, 1946.

[13] F. Bloch, "Nuclear induction," *Physical review*, vol. 70, no. 7-8, p. 460, 1946.

[14] P. Gribbon, R. Lyons, P. Laflin, J. Bradley, C. Chambers, B. S. Williams, W. Keighley, and A. Sewing, "Evaluating real-life high-throughput screening data," *Journal of biomolecular screening*, vol. 10, no. 2, pp. 99–107, 2005.

[15] A. R. Leach and M. M. Hann, "The in silico world of virtual libraries," *Drug Discovery Today*, vol. 5, no. 8, pp. 326–336, 2000.

[16] C. H. Reynolds, B. A. Tounge, and S. D. Bembenek, "Ligand binding efficiency: trends, physical basis, and implications," *Journal of medicinal chemistry*, vol. 51, no. 8, pp. 2432–2438, 2008.

[17] M. M. Hann, A. R. Leach, and G. Harper, "Molecular complexity and its impact on the probability of finding leads for drug discovery," *Journal of chemical information and computer sciences*, vol. 41, no. 3, pp. 856–864, 2001.

[18] C. Wermuth, C. Ganellin, P. Lindberg, and L. Mitscher, "Glossary of terms used in medicinal chemistry (iupac recommendations 1998)," *Pure and Applied Chemistry*, vol. 70, no. 5, pp. 1129–1143, 1998.

[19] D. B. Kitchen, H. Decornez, J. R. Furr, and J. Bajorath, "Docking and scoring in virtual screening for drug discovery: methods and applications," *Nat Rev Drug Discov*, vol. 3, pp. 935–949, Nov 2004.

[20] D. Q. Wei, R. Zhang, Q. S. Du, W. N. Gao, Y. Li, H. Gao, S. Q. Wang, X. Zhang, A. X. Li, S. Sirois, and K. C. Chou, "Anti-SARS drug screening by molecular docking," *Amino Acids*, vol. 31, pp. 73–80, Jul 2006.

[21] G. Barreiro, C. R. Guimaraes, I. Tubert-Brohman, T. M. Lyons, J. Tirado-Rives, and W. L. Jorgensen, "Search for non-nucleoside inhibitors of HIV-1 reverse transcriptase using chemical similarity, molecular docking, and MM-GB/SA scoring," *J Chem Inf Model*, vol. 47, no. 6, pp. 2416–2428, 2007.

[22] I. G. Tikhonova, C. S. Sum, S. Neumann, S. Engel, B. M. Raaka, S. Costanzi, and M. C. Gershengorn, "Discovery of novel agonists and antagonists of the free fatty acid receptor 1 (FFAR1) using virtual screening," *J. Med. Chem.*, vol. 51, pp. 625–633, Feb 2008.

[23] T. W. Lin, M. M. Melgar, D. Kurth, S. J. Swamidass, J. Purdon, T. Tseng, G. Gago, P. Baldi, H. Gramajo, and S. C. Tsai, "Structure-based inhibitor design of AccD5, an essential acyl-CoA carboxylase carboxyltransferase domain of Mycobacterium tuberculosis," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 103, pp. 3072–3077, Feb 2006.

[24] D. Vidal, M. Thormann, and M. Pons, "A novel search engine for virtual screening of very large databases," *J Chem Inf Model*, vol. 46, no. 2, pp. 836–843, 2006.

[25] J. Mestres and R. M. Knegtel, "Similarity versus docking in 3d virtual screening," *Perspectives in Drug Discovery and Design*, vol. 20, no. 1, pp. 191–207, 2000.

[26] C. Bissantz, C. Schalon, W. Guba, and M. Stahl, "Focused library design in gpcr projects on the example of 5-ht2c agonists: Comparison of structure-based virtual screening with ligand-based search methods," *Proteins: Structure, Function, and Bioinformatics*, vol. 61, no. 4, pp. 938–952, 2005.

[27] R. L. DesJarlais, R. P. Sheridan, G. L. Seibel, J. S. Dixon, I. D. Kuntz, and R. Venkataraghavan, "Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure," *Journal of medicinal chemistry*, vol. 31, no. 4, pp. 722–729, 1988.

[28] S. K. Kearsley, D. J. Underwood, R. P. Sheridan, and M. D. Miller, "Flexibases: a way to enhance the use of molecular docking methods," *Journal of computer-aided molecular design*, vol. 8, no. 5, pp. 565–582, 1994.

[29] M. R. McGann, H. R. Almond, A. Nicholls, J. A. Grant, and F. K. Brown, "Gaussian docking functions," *Biopolymers*, vol. 68, no. 1, pp. 76–90, 2003.

[30] J. Schlosser and M. Rarey, "Beyond the virtual screening paradigm: structure-based searching for new lead compounds," *J Chem Inf Model*, vol. 49, pp. 800–809, Apr 2009.

[31] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe, "A fast flexible docking method using an incremental construction algorithm," *Journal of molecular biology*, vol. 261, no. 3, pp. 470–489, 1996.

[32] W. Welch, J. Ruppert, and A. N. Jain, "Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites," *Chemistry & biology*, vol. 3, no. 6, pp. 449–462, 1996.

[33] T. J. Ewing, S. Makino, A. G. Skillman, and I. D. Kuntz, "Dock 4.0: search strategies for automated molecular docking of flexible molecule databases," *Journal of computer-aided molecular design*, vol. 15, no. 5, pp. 411–428, 2001.

[34] A. N. Jain, "Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine," *Journal of medicinal chemistry*, vol. 46, no. 4, pp. 499–511, 2003.

[35] Z. Zsoldos, D. Reid, A. Simon, S. B. Sadjad, and A. P. Johnson, "ehits: a new fast, exhaustive flexible ligand docking system," *Journal of Molecular Graphics and Modelling*, vol. 26, no. 1, pp. 198–212, 2007.

[36] R. Abagyan, M. Totrov, and D. Kuznetsov, "Icm—a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation," *Journal of computational chemistry*, vol. 15, no. 5, pp. 488–506, 1994.

[37] C. McMartin and R. S. Bohacek, "Qxp: powerful, rapid computer algorithms for structure-based drug design," *Journal of computer-aided molecular design*, vol. 11, no. 4, pp. 333–344, 1997.

[38] G. M. Morris, D. S. Goodsell, R. Huey, and A. J. Olson, "Distributed automated docking of flexible ligands to proteins: parallel applications of autodock 2.4," *Journal of computer-aided molecular design*, vol. 10, no. 4, pp. 293–304, 1996.

[39] G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor, "Development and validation of a genetic algorithm for flexible docking," *Journal of molecular biology*, vol. 267, no. 3, pp. 727–748, 1997.

[40] O. Korb, T. Stützle, and T. E. Exner, "Plants: application of ant colony optimization to structure-based drug design," in *International Workshop on Ant Colony Optimization and Swarm Intelligence*, pp. 247–258, Springer, 2006.

[41] R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, *et al.*, "Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy," *Journal of medicinal chemistry*, vol. 47, no. 7, pp. 1739–1749, 2004.

[42] R. Huey, G. M. Morris, A. J. Olson, and D. S. Goodsell, "A semiempirical free energy force field with charge-based desolvation," *Journal of computational chemistry*, vol. 28, no. 6, pp. 1145–1152, 2007.

[43] R. P. Sheridan and S. K. Kearsley, "Why do we need so many chemical similarity search methods?," *Drug discovery today*, vol. 7, no. 17, pp. 903–911, 2002.

[44] H. Kubinyi, G. Folkers, and Y. C. Martin, *3D QSAR in Drug Design: Volume 2: Ligand-Protein Interactions and Molecular Similarity*, vol. 2. Springer Science & Business Media, 1998.

[45] J. Klekota and F. P. Roth, "Chemical substructures that enrich for biological activity," *Bioinformatics*, vol. 24, no. 21, pp. 2518–2525, 2008.

[46] H. Geppert, M. Vogt, and J. Bajorath, "Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation," *Journal of chemical information and modeling*, vol. 50, no. 2, pp. 205–216, 2010.

[47] X. H. Ma, J. Jia, F. Zhu, Y. Xue, Z. R. Li, and Y. Z. Chen, "Comparative analysis of machine learning methods in ligand-based virtual screening of large compound libraries," *Combinatorial chemistry & high throughput screening*, vol. 12, no. 4, pp. 344–357, 2009.

[48] L. Han, X. Ma, H. Lin, J. Jia, F. Zhu, Y. Xue, Z. Li, Z. Cao, Z. Ji, and Y. Chen, "A support vector machines approach for virtual screening of active compounds of single and multiple mechanisms from large libraries at an improved hit-rate and enrichment factor," *Journal of Molecular Graphics and Modelling*, vol. 26, no. 8, pp. 1276–1286, 2008.

[49] X. Liu, X. H. Ma, C. Tan, Y. Jiang, M. Go, B. C. Low, and Y. Z. Chen, "Virtual screening of abl inhibitors from large compound libraries by support vector machines," *Journal of Chemical Information and Modeling*, vol. 49, no. 9, pp. 2101–2110, 2009.

[50] X. Ma, R. Wang, C. Tan, Y. Jiang, T. Lu, H. Rao, X. Li, M. Go, B. Low, and Y. Chen, "Virtual screening of selective multitarget kinase inhibitors by combinatorial support vector machines," *Molecular pharmaceutics*, vol. 7, no. 5, pp. 1545–1560, 2010.

[51] A. Bender, "How similar are those molecules after all? use two descriptors and you will have three different answers," *Expert opinion on drug discovery*, vol. 5, no. 12, pp. 1141–1151, 2010.

[52] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings," *Adv. Drug Deliv. Rev.*, vol. 46, pp. 3–26, Mar 2001.

[53] T. I. Oprea, A. M. Davis, S. J. Teague, and P. D. Leeson, "Is there a difference between leads and drugs? a historical perspective," *Journal of chemical information and computer sciences*, vol. 41, no. 5, pp. 1308–1315, 2001.

[54] J. M. Barnard and G. M. Downs, "Chemical fragment generation and clustering software §," *Journal of chemical information and computer sciences*, vol. 37, no. 1, pp. 141–142, 1997.

[55] C. A. James, D. Weininger, and J. Delany, "Daylight theory manual," 2016. [Accessed online on 8. November 2016].

[56] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *J Chem Inf Model*, vol. 50, pp. 742–754, May 2010.

[57] G. B. McGaughey, R. P. Sheridan, C. I. Bayly, J. C. Culberson, C. Kreatsoulas, S. Lindsley, V. Maiorov, J.-F. Truchon, and W. D. Cornell, "Comparison of topological, shape, and docking methods in virtual screening," *Journal of chemical information and modeling*, vol. 47, no. 4, pp. 1504–1519, 2007.

[58] D. M. Krüger and A. Evers, "Comparison of structure-and ligand-based virtual screening protocols considering hit list complementarity and enrichment factors," *ChemMedChem*, vol. 5, no. 1, pp. 148–158, 2010.

[59] G. Schneider, W. Neidhart, T. Giller, and G. Schmid, ""scaffold-hopping" by topological pharmacophore search: A contribution to virtual screening," *Angewandte Chemie International Edition*, vol. 38, no. 19, pp. 2894–2896, 1999.

[60] S. Lesnik, T. Stular, B. Brus, D. Knez, S. Gobec, D. Janezic, and J. Konc, "Lisica: A software for ligand-based virtual screening and its application for the discovery of butyrylcholinesterase inhibitors," *Journal of chemical information and modeling*, vol. 55, no. 8, pp. 1521–1528, 2015.

[61] J. A. Grant, M. A. Gallardo, and B. T. Pickup, "A fast method of molecular shape comparison: A simple application of a gaussian description of molecular shape," *Journal of Computational Chemistry*, vol. 17, no. 14, pp. 1653–1666, 1996.

[62] T. S. Rush, J. A. Grant, L. Mosyak, and A. Nicholls, "A shape-based 3-d scaffold hopping method and its application to a bacterial protein-protein interaction," *Journal of medicinal chemistry*, vol. 48, no. 5, pp. 1489–1495, 2005.

[63] C. Cai, J. Gong, X. Liu, D. Gao, and H. Li, "Simg: An alignment based method for evaluating the similarity of small molecules and binding sites," *Journal of chemical information and modeling*, vol. 53, no. 8, pp. 2103–2115, 2013.

[64] M. J. Vainio, J. S. Puranen, and M. S. Johnson, "ShaEP: molecular overlay based on shape and electrostatic potential," *J Chem Inf Model*, vol. 49, pp. 492–502, Feb 2009.

[65] J. Taminau, G. Thijs, and H. De Winter, "Pharao: pharmacophore alignment and optimization," *J. Mol. Graph. Model.*, vol. 27, pp. 161–169, Sep 2008.

[66] L. A. Vaz de Lima and A. S. Nascimento, "MolShaCS: a free and open source tool for ligand similarity identification based on Gaussian descriptors," *Eur J Med Chem*, vol. 59, pp. 296–303, Jan 2013.

[67] S. L. Kinnings and R. M. Jackson, "Ligmatch: a multiple structure-based ligand matching method for 3d virtual screening," *Journal of chemical information and modeling*, vol. 49, no. 9, pp. 2056–2066, 2009.

[68] A. Roy and J. Skolnick, "LIGSIFT: an open-source tool for ligand structural alignment and virtual screening," *Bioinformatics*, vol. 31, pp. 539–544, Feb 2015.

[69] A. Kalaszi, D. Szisz, G. Imre, and T. Polgár, "Screen3d: a novel fully flexible high-throughput shape-similarity search method," *Journal of chemical information and modeling*, vol. 54, no. 4, pp. 1036–1049, 2014.

[70] C. Lemmen, T. Lengauer, and G. Klebe, "FLEXS: a method for fast flexible ligand superposition," *J. Med. Chem.*, vol. 41, pp. 4502–4520, Nov 1998.

[71] A. N. Jain, "Ligand-based structural hypotheses for virtual screening," *J. Med. Chem.*, vol. 47, pp. 947–961, Feb 2004.

[72] O. Korb, P. Monecke, G. Hessler, T. Stutzle, and T. E. Exner, "pharmacophore: multiple flexible ligand alignment based on ant colony optimization," *Journal of chemical information and modeling*, vol. 50, no. 9, pp. 1669–1681, 2010.

[73] M. Totrov, "Atomic property fields: Generalized 3d pharmacophoric potential for automated ligand superposition, pharmacophore elucidation and 3d qsar," *Chemical Biology & Drug Design*, vol. 71, no. 1, pp. 15–27, 2008.

[74] R. Abagyan, M. Totrov, and D. Kuznetsov, "Icm—a new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation," *Journal of Computational Chemistry*, vol. 15, no. 5, pp. 488–506, 1994.

[75] S. L. Dixon, A. M. Smondyrev, E. H. Knoll, S. N. Rao, D. E. Shaw, and R. A. Friesner, "Phase: a new engine for pharmacophore perception, 3d qsar model development, and 3d database screening: 1. methodology and preliminary results," *Journal of computer-aided molecular design*, vol. 20, no. 10-11, pp. 647–671, 2006.

[76] D. R. Koes and C. J. Camacho, "Pharmer: efficient and exact pharmacophore search," *Journal of chemical information and modeling*, vol. 51, no. 6, pp. 1307–1314, 2011.

[77] G. Wolber and T. Langer, "Ligandscout: 3-d pharmacophores derived from protein-bound ligands and their use as virtual screening filters," *Journal of chemical information and modeling*, vol. 45, no. 1, pp. 160–169, 2005.

[78] J. J. Irwin, "Community benchmarks for virtual screening," *J. Comput. Aided Mol. Des.*, vol. 22, no. 3-4, pp. 193–199, 2008.

[79] A. R. Leach, B. K. Shoichet, and C. E. Peishoff, "Prediction of protein-ligand interactions. Docking and scoring: successes and gaps," *J. Med. Chem.*, vol. 49, pp. 5851–5855, Oct 2006.

[80] J. Kirchmair, P. Markt, S. Distinto, G. Wolber, and T. Langer, "Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection–what can we learn from earlier mistakes?," *J. Comput. Aided Mol. Des.*, vol. 22, no. 3-4, pp. 213–228, 2008.

[81] C. Bissantz, G. Folkers, and D. Rognan, "Protein-based virtual screening of chemical databases. 1. evaluation of different docking/scoring combinations," *Journal of medicinal chemistry*, vol. 43, no. 25, pp. 4759–4767, 2000.

[82] S. L. McGovern and B. K. Shoichet, "Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes," *Journal of medicinal chemistry*, vol. 46, no. 14, pp. 2895–2907, 2003.

[83] D. J. Diller and R. Li, "Kinases, homology models, and high throughput docking," *Journal of medicinal chemistry*, vol. 46, no. 22, pp. 4638–4647, 2003.

[84] D. M. Lorber and B. K. Shoichet, "Hierarchical docking of databases of multiple ligand conformations," *Current topics in medicinal chemistry*, vol. 5, no. 8, pp. 739–749, 2005.

[85] J. J. Irwin, F. M. Raushel, and B. K. Shoichet, "Virtual screening against metalloenzymes for inhibitors and substrates," *Biochemistry*, vol. 44, no. 37, pp. 12316–12328, 2005.

[86] M. A. Miteva, W. H. Lee, M. O. Montes, and B. O. Villoutreix, "Fast structure-based virtual ligand screening combining fred, dock, and surflex," *Journal of medicinal chemistry*, vol. 48, no. 19, pp. 6012–6022, 2005.

[87] T. A. Pham and A. N. Jain, "Parameter estimation for scoring protein-ligand interactions using negative training data," *Journal of medicinal chemistry*, vol. 49, no. 20, pp. 5856–5868, 2006.

[88] N. Huang, B. K. Shoichet, , and J. J. Irwin, "Benchmarking sets for molecular docking," *Journal of Medicinal Chemistry*, vol. 49, no. 23, pp. 6789–6801, 2006. PMID: 17154509.

[89] M. M. Mysinger, M. Carchia, J. J. Irwin, and B. K. Shoichet, "Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking," *J. Med. Chem.*, vol. 55, pp. 6582–6594, Jul 2012.

[90] E. A. Gatica and C. N. Cavasotto, "Ligand and decoy sets for docking to g protein-coupled receptors," *Journal of chemical information and modeling*, vol. 52, no. 1, pp. 1–6, 2011.

[91] M. R. Bauer, T. M. Ibrahim, S. M. Vogel, and F. M. Boeckler, "Evaluation and optimization of virtual screening workflows with dekois 2.0–a public library of challenging docking benchmark sets," *Journal of chemical information and modeling*, vol. 53, no. 6, pp. 1447–1462, 2013.

[92] N. Lagarde, N. Ben Nasr, A. Jeremie, H. Guillemain, V. Laville, T. Labib, J.-F. Zagury, and M. Montes, "Nrlist bdb, the manually curated nuclear receptors ligands and structures benchmarking database," *Journal of medicinal chemistry*, vol. 57, no. 7, pp. 3117–3125, 2014.

[93] J. Xia, E. L. Tilahun, E. H. Kebede, T.-E. Reid, L. Zhang, and X. S. Wang, "Comparative modeling and benchmarking data sets for human histone deacetylases and sirtuin families," *Journal of chemical information and modeling*, vol. 55, no. 2, pp. 374–388, 2015.

[94] N. Lagarde, J.-F. Zagury, and M. Montes, "Benchmarking data sets for the evaluation of virtual ligand screening methods: Review and perspectives," *Journal of chemical information and modeling*, vol. 55, no. 7, pp. 1297–1307, 2015.

[95] A. R. Leach, V. J. Gillet, R. A. Lewis, and R. Taylor, "Three-dimensional pharmacophore methods in drug discovery," *J. Med. Chem*, vol. 53, no. 2, pp. 539–558, 2010.

[96] P. W. Finn and G. M. Morris, "Shape-based similarity searching in chemical databases," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 3, no. 3, pp. 226–241, 2013.

[97] M. Hilbig, S. Urbaczek, I. Groth, S. Heuser, and M. Rarey, "MONA - Interactive manipulation of molecule collections," *J Cheminform*, vol. 5, p. 38, Aug 2013.

[98] K. Wu, "Fastbit: an efficient indexing technology for accelerating data-intensive science," *Journal of Physics: Conference Series*, vol. 16, no. 1, p. 556, 2005.

[99] S. Urbaczek, A. Kolodzik, J. R. Fischer, T. Lippert, S. Heuser, I. Groth, T. Schulz-Gasch, and M. Rarey, "NAOMI: on the almost trivial task of reading molecules from different file formats," *J Chem Inf Model*, vol. 51, pp. 3199–3207, Dec 2011.

[100] S. Urbaczek, A. Kolodzik, I. Groth, S. Heuser, and M. Rarey, "Reading PDB: perception of molecules from 3D atomic coordinates," *J Chem Inf Model*, vol. 53, pp. 76–87, Jan 2013.

[101] C. Scharfer, T. Schulz-Gasch, J. Hert, L. Heinzerling, B. Schulz, T. Inhester, M. Stahl, and M. Rarey, "CONFECT: conformations from an expert collection of torsion patterns," *ChemMedChem*, vol. 8, pp. 1690–1700, Oct 2013.

[102] C. Scharfer, T. Schulz-Gasch, H. C. Ehrlich, W. Guba, M. Rarey, and M. Stahl, "Torsion angle preferences in druglike chemical space: a comprehensive guide," *J. Med. Chem.*, vol. 56, pp. 2016–2028, Mar 2013.

[103] K. Sommer, N. O. Friedrich, S. Bietz, M. Hilbig, T. Inhester, and M. Rarey, "UNICON: A Powerful and Easy-to-Use Compound Library Converter," *J Chem Inf Model*, vol. 56, pp. 1105–1111, Jun 2016.

[104] W. Guba, A. Meyder, M. Rarey, and J. Hert, "Torsion Library Reloaded: A New Version of Expert-Derived SMARTS Rules for Assessing Conformations of Small Molecules," *J Chem Inf Model*, vol. 56, pp. 1–5, Jan 2016.

[105] M. Hilbig and M. Rarey, "MONA 2: A Light Cheminformatics Platform for Interactive Compound Library Processing," *J Chem Inf Model*, vol. 55, pp. 2071–2078, Oct 2015.

[106] A. M. Henzler, S. Urbaczek, M. Hilbig, and M. Rarey, "An integrated approach to knowledge-driven structure-based virtual screening," *J. Comput. Aided Mol. Des.*, vol. 28, pp. 927–939, Sep 2014.

[107] M. M. von Behren, S. Bietz, E. Nittinger, and M. Rarey, "mRAISE: an alternative algorithmic approach to ligand-based virtual screening," *J. Comput. Aided Mol. Des.*, vol. 30, pp. 583–594, Aug 2016.

[108] I. Schellhammer and M. Rarey, "TrixX: structure-based molecule indexing for large-scale virtual screening in sublinear time," *J. Comput. Aided Mol. Des.*, vol. 21, pp. 223–238, May 2007.

[109] M. M. von Behren, A. Volkamer, A. M. Henzler, K. T. Schomburg, S. Urbaczek, and M. Rarey, "Fast protein binding site comparison via an index-based screening technology.," *Journal of chemical information and modeling*, vol. 53(2), pp. 411–22, 2013.

[110] K. T. Schomburg, S. Bietz, H. Briem, A. M. Henzler, S. Urbaczek, and M. Rarey, "Facing the challenges of structure-based target prediction by inverse virtual screening," *J Chem Inf Model*, vol. 54, pp. 1676–1686, Jun 2014.

[111] C. Schärfer, "Sublinear ligand-based virtual screening using bitmap indices," Master's thesis, University of Hamburg, Center for Bioinformatics (ZBH), Bundesstrasse 43, 20146 Hamburg, 2008.

[112] M. M. von Behren and M. Rarey, "Ligand-based virtual screening under partial shape constraints," *J. Comput. Aided Mol. Des.*, Manuscript submitted for publication.

[113] C. A. Spronk, S. B. Nabuurs, A. M. Bonvin, E. Krieger, G. W. Vuister, and G. Vriend, "The precision of NMR structure ensembles revisited," *J. Biomol. NMR*, vol. 25, pp. 225–234, Mar 2003.

[114] S. B. Nabuurs, C. A. Spronk, E. Krieger, H. Maassen, G. Vriend, and G. W. Vuister, "Quantitative evaluation of experimental NMR restraints," *J. Am. Chem. Soc.*, vol. 125, pp. 12026–12034, Oct 2003.

[115] J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad, and R. G. Coleman, "ZINC: a free tool to discover chemistry for biology," *J Chem Inf Model*, vol. 52, pp. 1757–1768, Jul 2012.

[116] W. D. Ihlenfeldt, Y. Takahashi, H. Abe, and S.-i. Sasaki, "Computation and management of chemical properties in cactvs: An extensible networked approach toward modularity and compatibility," *Journal of chemical information and computer sciences*, vol. 34, no. 1, pp. 109–116, 1994.

[117] H. Matter, "Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors," *Journal of medicinal chemistry*, vol. 40, no. 8, pp. 1219–1229, 1997.

[118] H. Fan, J. J. Irwin, B. M. Webb, G. Klebe, B. K. Shoichet, and A. Sali, "Molecular docking screens using comparative models of proteins," *J Chem Inf Model*, vol. 49, pp. 2512–2527, Nov 2009.

[119] M. P. Repasky, R. B. Murphy, J. L. Banks, J. R. Greenwood, I. Tubert-Brohman, S. Bhat, and R. A. Friesner, "Docking performance of the glide program as evaluated on the Astex and DUD datasets: a complete set of glide SP results and selected results for a new scoring function integrating WaterMap and glide," *J. Comput. Aided Mol. Des.*, vol. 26, pp. 787–799, Jun 2012.

[120] S. R. Brozell, S. Mukherjee, T. E. Balius, D. R. Roe, D. A. Case, and R. C. Rizzo, "Evaluation of DOCK 6 as a pose generation and database enrichment tool," *J. Comput. Aided Mol. Des.*, vol. 26, pp. 749–773, Jun 2012.

[121] M. A. Neves, M. Totrov, and R. Abagyan, "Docking and scoring with ICM: the benchmarking results and strategies for improvement," *J. Comput. Aided Mol. Des.*, vol. 26, pp. 675–686, Jun 2012.

[122] R. Spitzer and A. N. Jain, "Surflex-Dock: Docking benchmarks and real-world application," *J. Comput. Aided Mol. Des.*, vol. 26, pp. 687–699, Jun 2012.

[123] N. Schneider, S. Hindle, G. Lange, R. Klein, J. Albrecht, H. Briem, K. Beyer, H. Claussen, M. Gastreich, C. Lemmen, and M. Rarey, "Substantial improvements in large-scale redocking and screening using the novel HYDE scoring function," *J. Comput. Aided Mol. Des.*, vol. 26, pp. 701–723, Jun 2012.

[124] J. W. Liebeschuetz, J. C. Cole, and O. Korb, "Pose prediction and virtual screening performance of GOLD scoring functions in a standardized test," *J. Comput. Aided Mol. Des.*, vol. 26, pp. 737–748, Jun 2012.

[125] F. N. Novikov, V. S. Stroylov, A. A. Zeifman, O. V. Stroganov, V. Kulkov, and G. G. Chilov, "Lead Finder docking and virtual screening evaluation with Astex and DUD test sets," *J. Comput. Aided Mol. Des.*, vol. 26, pp. 725–735, Jun 2012.

[126] A. C. Good and T. I. Oprea, "Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection?," *J. Comput. Aided Mol. Des.*, vol. 22, no. 3-4, pp. 169–178, 2008.

[127] P. C. Hawkins, G. L. Warren, A. G. Skillman, and A. Nicholls, "How to do an evaluation: pitfalls and traps," *Journal of computer-aided molecular design*, vol. 22, no. 3-4, pp. 179–190, 2008.

[128] J. J. Irwin, "Community benchmarks for virtual screening," *J. Comput. Aided Mol. Des.*, vol. 22, no. 3-4, pp. 193–199, 2008.

[129] M. M. Mysinger and B. K. Shoichet, "Rapid context-dependent ligand desolvation in molecular docking," *J Chem Inf Model*, vol. 50, pp. 1561–1573, Sep 2010.

[130] S. M. Vogel, M. R. Bauer, and F. M. Boeckler, "DEKOIS: demanding evaluation kits for objective in silico screening–a versatile tool for benchmarking docking programs and scoring functions," *J Chem Inf Model*, vol. 51, pp. 2650–2665, Oct 2011.

[131] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic acids research*, vol. 28, no. 1, pp. 235–242, 2000.

[132] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, and J. P. Overington, "ChEMBL: a large-scale bioactivity database for drug discovery," *Nucleic Acids Res.*, vol. 40, pp. D1100–1107, Jan 2012.

[133] G. W. Bemis and M. A. Murcko, "The properties of known drugs. 1. Molecular frameworks," *J. Med. Chem.*, vol. 39, pp. 2887–2893, Jul 1996.

[134] E. Nittinger, N. Schneider, G. Lange, and M. Rarey, "Evidence of water molecules–a statistical evaluation of water molecules based on electron density," *J Chem Inf Model*, vol. 55, pp. 771–783, Apr 2015.

[135] S. Bietz and M. Rarey, "SIENA: Efficient Compilation of Selective Protein Binding Site Ensembles," *J Chem Inf Model*, vol. 56, pp. 248–259, Jan 2016.

[136] H. Strombergsson and G. J. Kleywegt, "A chemogenomics view on protein-ligand spaces," *BMC Bioinformatics*, vol. 10 Suppl 6, p. S13, Jun 2009.

[137] J. Bostrom, A. Hogner, and S. Schmitt, "Do structurally similar ligands bind in a similar fashion?," *J. Med. Chem.*, vol. 49, pp. 6716–6725, Nov 2006.

[138] J. Meslamani, D. Rognan, and E. Kellenberger, "sc-PDB: a database for identifying variations and multiplicity of 'druggable' binding sites in proteins," *Bioinformatics*, vol. 27, pp. 1324–1326, May 2011.

[139] I. R. Craig, J. W. Essex, and K. Spiegel, "Ensemble docking into multiple crystallographically derived protein structures: an evaluation based on the statistical analysis of enrichments," *Journal of chemical information and modeling*, vol. 50, no. 4, pp. 511–524, 2010.

[140] D. Giganti, H. Guillemain, J. L. Spadoni, M. Nilges, J. F. Zagury, and M. Montes, "Comparative evaluation of 3D virtual ligand screening methods: impact of the molecular alignment on enrichment," *J Chem Inf Model*, vol. 50, pp. 992–1004, Jun 2010.

[141] D. Giganti, H. Guillemain, J. L. Spadoni, M. Nilges, J. F. Zagury, and M. Montes, "Comparative evaluation of 3D virtual ligand screening methods: impact of the molecular alignment on enrichment," *J Chem Inf Model*, vol. 50, pp. 992–1004, Jun 2010.

[142] K. Stierand, P. C. Maass, and M. Rarey, "Molecular complexes at a glance: automated generation of two-dimensional complex diagrams," *Bioinformatics*, vol. 22, pp. 1710–1716, Jul 2006.

[143] "Rocs method introduction," Feb. 2016. [Accessed online on February 2016].

[144] J. Kirchmair, S. Distinto, P. Markt, D. Schuster, G. M. Spitzer, K. R. Liedl, and G. Wolber, "How to optimize shape-based virtual screening: choosing the right query and including chemical information," *J Chem Inf Model*, vol. 49, pp. 678–692, Mar 2009.

A Appendix A. Detailed Results

Table A.1.: Calculation of the weight based on atom charges (w_{ch}).

	None	Negative	Positive
None	1	0.5	0.5
Negative	0.5	1	0
Positive	0.5	0	1

An atom is considered as negatively charged, if the formal charge of that atom split upon topological similar atoms (mesomeric structures) is less than zero. An atom is likewise considered positively charged, if that charge is greater than zero. Reprinted from [107] with permission of Springer.

Table A.2.: Calculation of the weight based on ring membership (w_{ri}).

	In Ring	Not in Ring
In Ring	1	0.5
Not in Ring	0.5	1

Reprinted from [107] with permission from Springer.

Table A.3.: Calculation of the weight based on potential interactions (w_{ia}).

	Neutral	Hydrophobic	Donor	Acceptor	Donor/Acceptor
Neutral	1	0.5	0.5	0.5	0.5
Hydrophobic	0.5	1	0.5	0.5	0.5
Donor	0.5	0.5	2	0	2
Acceptor	0.5	0.5	0	2	2
Donor/Acceptor	0.5	0.5	2	2	2

An atoms is labeled 'Neutral' if it is an hydrophobic atom with directly connected hydrophilic neighbors. Reprinted from [107] with permission of Springer.

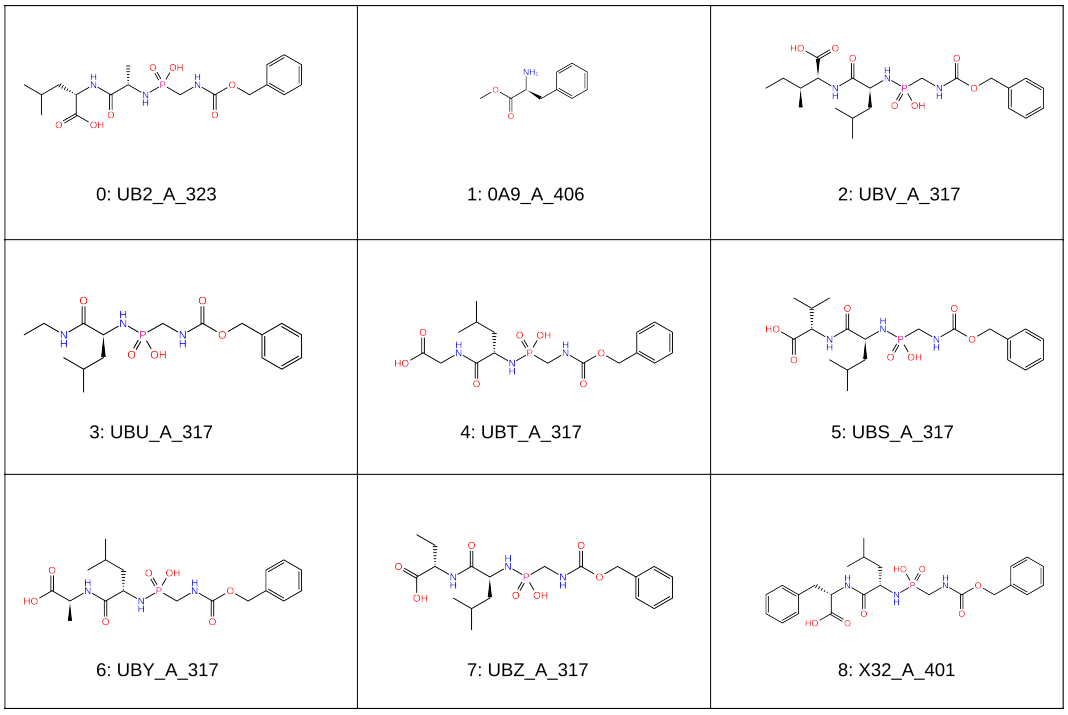


Figure A.1.: Overview of all ligands of the thermolysin example. Reprinted from [112].

Table A.4.: Detailed AUC values of the ROC curves for all DUD targets.

	LIGSIFT	mRAISE	Align-It	ROCS
ACE	0.79	0.91	0.86	0.7
ACHE	0.8	0.78	0.82	0.77
ADA	0.73	0.83	0.88	0.86
ALR2	0.69	0.65	0.71	0.57
AMPC	0.93	0.91	0.89	0.82
AR	0.83	0.85	0.79	0.79
CDK2	0.71	0.64	0.45	0.68
COMT	0.9	0.89	0.79	0.32
COX1	0.62	0.53	0.68	0.53
COX2	0.95	0.95	0.95	0.93
DHFR	0.97	0.99	0.97	0.92
EGFR	0.93	0.96	0.94	0.95
ERagonist	0.92	0.94	0.87	0.94
ERantagonist	0.9	0.92	0.94	0.98
FGFR1	0.62	0.53	0.59	0.49
FXA	0.77	0.64	0.62	0.39
GART	0.86	0.93	0.92	0.93
GPB	0.94	0.92	0.94	0.92
GR	0.87	0.61	0.56	0.79
HIVPR	0.79	0.58	0.78	0.56
HIVRT	0.78	0.74	0.63	0.66
HMGA	0.96	0.90	0.92	0.92
HSP90	0.87	0.86	0.65	0.66
INHA	0.72	0.67	0.77	0.72
MR	0.89	0.87	0.72	0.87
NA	0.96	0.99	0.88	0.97
P38	0.51	0.47	0.45	0.52
PARP	0.68	0.74	0.94	0.58
PDE5	0.57	0.57	0.64	0.53
PDGFRB	0.46	0.32	0.23	0.34
PNP	0.98	0.99	0.95	0.91
PPAR	0.85	0.96	0.91	0.92
PR	0.79	0.52	0.66	0.67
RXR	0.98	0.89	0.98	0.96
SAHH	0.97	0.97	0.96	0.97
SRC	0.38	0.45	0.38	0.38
THROMBIN	0.59	0.61	0.69	0.66
TK	0.92	0.88	0.78	0.86
TRYPSIN	0.64	0.79	0.75	0.78
VEGFR2	0.67	0.42	0.29	0.43
Average	0.79	0.76	0.75	0.73

Values for other methods taken from [144]. Reprinted from [107] with permission of Springer.

Table A.5.: Detailed results for all targets of the DUD-E dataset.

Target	AUC	$EF_{1\%}$	$EF_{5\%}$	$EF_{10\%}$	$HR_{1\%}$	$HR_{5\%}$	$HR_{10\%}$
aa2ar	0.89	31.21	11.12	6.79	47.02	55.60	67.84
abl1	0.71	20.38	5.94	4.12	33.94	29.67	41.21
ace	0.87	29.51	10.36	6.74	48.54	51.77	67.38
aces	0.64	3.10	3.80	3.80	5.26	18.98	37.97
ada17	0.79	21.09	8.50	5.55	30.85	42.48	55.45
ada	0.98	48.76	15.92	8.82	81.82	79.57	88.17
adrb1	0.72	24.43	7.21	4.41	37.50	36.03	44.13
adrb2	0.64	12.58	5.20	3.29	19.08	25.97	32.90
akt1	0.67	5.13	2.73	2.66	8.98	13.65	26.62
akt2	0.67	15.41	5.82	3.33	25.71	29.06	33.33
aldr	0.83	36.69	10.58	6.04	63.74	52.83	60.38
ampc	0.84	38.57	8.75	6.04	64.29	43.75	60.42
andr	0.74	19.35	7.14	4.50	35.62	35.69	44.98
aofb	0.48	0.82	1.31	1.07	1.43	6.56	10.66
bace1	0.47	2.48	0.92	0.78	3.83	4.59	7.77
braf	0.75	33.87	8.56	4.61	51.00	42.76	46.05
cah2	0.71	1.22	1.67	1.87	1.90	8.33	18.70
casp3	0.70	7.60	3.92	2.36	13.89	19.60	23.62
cdk2	0.69	14.98	5.40	3.74	25.09	27.00	37.34
comt	0.93	32.45	12.71	6.85	34.21	63.41	68.29
cp2c9	0.57	0.00	0.83	1.00	0.00	4.17	10.00
cp3a4	0.62	5.32	2.24	2.06	7.56	11.18	20.59
csf1r	0.72	26.53	6.63	3.80	35.77	33.13	37.95
cxcr4	0.76	40.54	10.52	6.01	47.06	52.50	60.00
def	0.87	33.91	11.80	7.17	59.65	58.82	71.57
dhi1	0.67	13.69	4.67	3.21	22.96	23.33	32.12
dpp4	0.69	12.96	4.92	3.17	16.67	24.58	31.71
drd3	0.43	3.75	1.00	0.67	5.22	5.00	6.67
dyr	0.92	32.91	10.82	6.84	43.68	54.11	68.40
egfr	0.87	43.08	12.85	7.40	65.63	64.21	73.99
esr1	0.91	46.58	13.74	7.84	84.76	68.67	78.33
esr2	0.83	27.86	12.16	7.36	49.76	60.76	73.57
fa10	0.62	14.38	6.07	3.48	37.56	30.35	34.82
fa7	0.92	25.68	12.32	7.12	46.03	61.40	71.05
fabp4	0.90	35.25	11.98	8.10	59.26	59.57	80.85
fak1	0.73	32.30	9.22	6.10	59.26	46.00	61.00
fgfr1	0.47	0.85	0.44	0.65	25.00	13.04	19.15
fkb1a	0.69	11.73	3.43	2.43	22.03	17.12	24.32
fnta	0.69	22.65	7.13	4.27	25.77	35.64	42.74
fpps	0.99	71.82	18.37	9.65	71.76	91.76	96.47
gcr	0.67	29.16	6.82	3.80	49.34	34.11	37.98
glcm	0.56	3.76	0.74	0.74	5.26	3.70	7.41
gria2	0.79	46.55	13.30	7.09	61.34	66.46	70.89
grik1	0.76	26.93	8.53	4.86	40.91	42.57	48.51
hdac2	0.53	7.08	3.03	1.95	12.50	15.14	19.46
hdac8	0.79	36.53	11.43	6.18	58.49	57.06	61.76
hivint	0.64	8.05	3.80	2.30	11.94	19.00	23.00
hivpr	0.70	7.28	3.58	3.10	10.77	17.91	30.97
hivrt	0.63	15.10	4.86	3.05	26.56	24.26	30.47

Performance on the DUD-E. Reprinted from [107] with permission of Springer.

Table A.6.: Continuation of the detailed results for all targets of the DUD-E dataset.

Target	AUC	$EF_{1\%}$	$EF_{5\%}$	$EF_{10\%}$	$HR_{1\%}$	$HR_{5\%}$	$HR_{10\%}$
hmdh	0.91	40.65	13.31	7.24	77.53	66.47	72.35
hs90a	0.77	32.05	10.26	6.26	57.14	51.14	62.50
hxk4	0.98	35.43	15.24	9.25	68.09	76.09	92.39
igf1r	0.79	19.68	7.04	4.46	30.85	35.14	44.59
inha	0.79	21.32	6.52	4.89	39.13	32.56	48.84
ital	0.44	18.17	3.63	2.03	29.07	18.12	20.29
jak2	0.79	26.18	8.04	4.86	42.42	40.19	48.60
kif11	0.73	37.41	10.87	5.52	62.32	54.31	55.17
kith	0.58	24.13	5.79	3.19	37.74	28.92	31.93
kit	0.48	8.79	3.87	2.29	17.24	19.30	22.81
kpcb	0.85	51.27	12.90	6.75	78.41	64.44	67.41
lck	0.64	5.48	2.33	2.00	8.27	11.67	20.00
lkha4	0.79	11.13	5.98	5.15	19.79	29.82	51.46
mapk2	0.83	31.93	11.10	6.05	51.61	55.45	60.40
mcr	0.60	17.15	3.62	2.23	30.77	18.09	22.34
met	0.85	63.29	14.22	7.23	92.11	71.08	72.29
mk01	0.83	31.83	9.38	5.83	54.35	46.84	58.23
mk10	0.48	3.85	1.73	1.35	5.97	8.65	13.46
mk14	0.65	11.28	4.15	2.80	17.91	20.76	28.03
mmp13	0.91	37.07	13.19	7.80	56.23	65.91	77.97
mp2k1	0.53	18.33	3.97	2.23	26.83	19.83	22.31
nos1	0.55	4.02	1.20	1.20	4.94	6.00	12.00
nram	0.96	32.13	14.73	8.48	50.00	73.47	84.69
pa2ga	0.62	8.15	3.84	2.33	15.38	19.19	23.23
parp1	0.83	25.43	8.15	5.06	42.30	40.75	50.59
pde5a	0.72	26.16	6.84	4.07	37.28	34.17	40.70
pgh1	0.43	2.07	1.23	0.97	3.67	6.15	9.74
pgh2	0.82	38.97	11.22	6.48	71.91	56.09	64.83
plk1	0.65	2.81	2.06	1.78	4.35	10.28	17.76
pnph	1.00	64.56	19.45	10.00	94.29	97.09	100.00
ppara	0.83	21.48	8.75	5.74	40.61	43.70	57.37
ppard	0.71	11.31	4.67	3.25	21.77	23.33	32.50
pparg	0.79	25.04	8.43	4.84	47.08	42.15	48.35
prgr	0.65	9.58	3.69	2.94	17.61	18.43	29.35
ptn1	0.57	13.98	4.78	3.08	24.66	23.85	30.77
pur2	1.00	54.88	20.03	10.01	100.00	100.00	100.00
pygm	0.51	1.30	3.39	2.08	2.50	16.88	20.78
pyrd	0.76	48.17	10.84	5.77	81.54	54.05	57.66
reni	0.61	18.42	4.24	2.79	27.14	21.15	27.88
rock1	0.56	1.02	1.60	1.40	1.59	8.00	14.00
rxra	0.93	23.12	14.52	8.17	42.86	72.52	81.68
sahh	1.00	55.76	20.07	10.01	100.00	100.00	100.00
src	0.66	11.67	4.35	2.90	17.48	21.76	29.01
tgfr1	0.85	24.15	8.88	6.32	37.21	44.36	63.16
thb	0.89	58.59	15.74	7.97	80.00	78.64	79.61
thrb	0.70	3.69	4.12	3.32	6.20	20.61	33.19
try1	0.76	10.50	6.02	4.37	17.87	30.07	43.65
tryb1	0.66	6.84	2.84	2.50	12.99	14.19	25.00
tysy	0.92	47.11	14.51	7.81	75.00	72.48	77.98
urok	0.83	14.82	7.16	5.50	24.00	35.80	54.94
vgfr2	0.61	6.61	3.67	2.81	10.67	18.34	28.12
wee1	0.99	61.27	19.25	9.61	100.00	96.08	96.08
xiap	0.86	30.26	11.01	6.51	57.69	55.00	65.00
Average	0.74	23.45	7.78	4.69	37.95	38.94	46.98
Stdev	0.15	17.00	4.92	2.50	26.36	24.44	24.78

Reprinted from [112].

Performance of all 102 targets of the DUD-E using mRAISE. The query ligand for aa2ar was not used from DUD-E but downloaded from the PDB for technical reasons. Reprinted from [107] with permission of Springer.

Table A.7.: AUC values for all targets of DUD using different mRAISE modes.

	classic	contact	inclusion
ace	0.91	0.81	0.91
ache	0.78	0.77	0.75
ada	0.83	0.75	0.73
alr2	0.65	0.47	0.61
ampc	0.91	0.79	0.91
ar	0.85	0.81	0.89
cdk2	0.64	0.67	0.67
comt	0.89	0.67	0.85
cox1	0.53	0.56	0.59
cox2	0.95	0.87	0.94
dhfr	0.99	0.91	0.99
egfr	0.96	0.87	0.96
er_agonist	0.94	0.58	0.94
er_antagonist	0.92	0.78	0.92
fgfr1	0.53	0.52	0.54
fxa	0.64	0.61	0.71
gart	0.93	0.82	0.95
gpb	0.92	0.88	0.92
gr	0.61	0.64	0.67
hivpr	0.58	0.68	0.65
hivrt	0.74	0.59	0.64
hmga	0.90	0.88	0.95
hsp90	0.86	0.74	0.80
inha	0.67	0.62	0.58
mr	0.87	0.75	0.85
na	0.99	0.81	0.99
p38	0.47	0.41	0.34
parp	0.74	0.66	0.63
pde5	0.57	0.55	0.61
pdgfrb	0.32	0.40	0.35
pnp	0.99	0.92	0.99
ppar	0.96	0.90	0.96
pr	0.52	0.55	0.71
rxr	0.89	0.76	0.90
sahh	0.97	0.90	0.98
src	0.45	0.52	0.45
thrombin	0.61	0.58	0.68
tk	0.88	0.85	0.88
trypsin	0.79	0.72	0.68
vegfr2	0.42	0.45	0.44

Reprinted from [112].

Table A.8.: EF at one percent for all targets of DUD using different mRAISE modes.

	classic	contact	inclusion
ace	29.85	27.72	27.72
ache	21.06	26.80	22.97
ada	18.35	13.76	9.17
alr2	4.03	8.07	4.03
ampc	25.61	30.73	30.73
ar	25.69	25.69	28.40
cdk2	16.26	14.23	14.23
comt	10.02	10.02	10.02
cox1	17.48	13.11	17.48
cox2	35.95	35.95	35.95
dhfr	33.54	33.54	33.54
egfr	32.96	33.19	32.96
er_agonist	24.10	16.57	24.10
er_antagonist	18.38	18.38	18.38
fgfr1	7.67	10.22	6.82
fxa	4.96	4.26	4.96
gart	5.27	0.00	5.27
gpb	34.65	34.65	32.73
gr	13.16	17.11	17.11
hivpr	3.85	1.92	1.92
hivrt	21.10	15.83	18.46
hmga	36.46	36.46	36.46
hsp90	36.83	36.83	36.83
inha	35.52	29.60	34.34
mr	36.67	36.67	36.67
na	34.46	34.46	34.46
p38	10.99	10.21	10.21
parp	3.05	6.11	3.05
pde5	24.31	18.24	24.31
pdgfrb	7.09	5.80	7.09
pnp	24.19	24.19	24.19
ppar	31.79	31.79	33.06
pr	12.27	16.36	16.36
rxr	31.20	26.00	31.20
sahh	29.55	29.55	29.55
src	1.30	1.30	0.65
thrombin	3.15	3.15	3.15
tk	32.06	32.06	32.06
trypsin	2.41	2.41	2.41
vegfr2	10.87	8.15	10.87

Reprinted from [112].

Table A.9.: Detailed results for all targets of the DUD-E dataset using mRAISE_contact.

Target	AUC	$EF_{1\%}$	$EF_{5\%}$	$EF_{10\%}$	$HR_{1\%}$	$HR_{5\%}$	$HR_{10\%}$
aa2ar	0.89	29.75	10.50	6.47	44.83	52.49	64.73
abl1	0.63	15.42	4.84	3.57	25.69	24.18	35.71
ace	0.89	33.07	11.71	7.06	54.39	58.51	70.57
aces	0.61	2.21	1.41	1.52	3.76	7.06	15.23
ada17	0.76	17.51	7.63	5.19	25.62	38.16	51.88
ada	0.95	45.51	14.63	7.64	76.36	73.12	76.34
adrb1	0.65	18.73	6.32	3.97	28.75	31.58	39.68
adrb2	0.55	8.67	3.55	2.47	13.16	17.75	24.68
akt1	0.57	5.13	1.84	1.77	8.98	9.22	17.75
akt2	0.56	12.84	4.96	3.16	21.43	24.79	31.62
aldr	0.79	34.16	9.45	5.47	59.34	47.17	54.72
ampc	0.84	38.57	8.75	5.63	64.29	43.75	56.25
andr	0.66	19.72	5.06	3.23	36.30	25.28	32.34
aofb	0.47	1.64	0.98	0.82	2.86	4.92	8.20
bace1	0.53	2.84	1.06	1.17	4.37	5.30	11.66
braf	0.77	34.53	8.56	4.87	52.00	42.76	48.68
cah2	0.64	1.42	1.42	1.42	2.22	7.11	14.23
casp3	0.61	8.11	3.72	2.36	14.81	18.59	23.62
cdk2	0.71	11.82	5.32	3.46	19.79	26.58	34.60
comt	0.93	32.45	12.71	7.09	34.21	63.41	70.73
cp2c9	0.49	0.00	1.00	0.67	0.00	5.00	6.67
cp3a4	0.50	5.91	2.00	1.12	8.40	10.00	11.18
csf1r	0.69	25.92	7.11	3.98	34.96	35.54	39.76
cxcr4	0.78	40.54	9.52	5.76	47.06	47.50	57.50
def	0.88	39.89	12.59	7.17	70.18	62.75	71.57
dhi1	0.72	11.86	4.79	3.64	19.90	23.94	36.36
dpp4	0.71	14.46	5.10	3.49	18.60	25.52	34.90
drd3	0.41	3.75	1.33	0.75	5.22	6.67	7.50
dyr	0.90	33.77	10.65	6.62	44.83	53.25	66.23
egfr	0.84	37.53	11.52	6.62	57.18	57.56	66.24
esr1	0.84	48.68	12.49	6.79	88.57	62.40	67.89
esr2	0.84	28.68	11.78	7.41	51.22	58.86	74.11
fa10	0.69	12.70	5.10	3.22	33.17	25.51	32.22
fa7	0.82	21.25	7.04	4.66	38.10	35.09	46.49
fabp4	0.79	35.25	7.28	5.54	59.26	36.17	55.32
fak1	0.83	35.32	11.22	6.60	64.81	56.00	66.00
fgfr1	0.47	0.85	0.59	0.65	25.00	17.39	19.15
fkb1a	0.84	11.73	5.05	3.96	22.03	25.23	39.64
fnta	0.78	18.09	7.40	4.83	20.58	36.99	48.31
fpps	0.98	70.64	18.37	9.42	70.59	91.76	94.12

Performance on the DUD-E. Reprinted from [112].

Table A.10.: Continuation of the detailed results for all targets of the DUD-E dataset using mRAISE_contact.

Target	AUC	$EF_{1\%}$	$EF_{5\%}$	$EF_{10\%}$	$HR_{1\%}$	$HR_{5\%}$	$HR_{10\%}$
gcr	0.73	28.38	6.82	3.84	48.03	34.11	38.37
glcm	0.55	3.76	0.74	1.48	5.26	3.70	14.81
gria2	0.78	44.00	12.04	7.03	57.98	60.13	70.25
grik1	0.77	24.93	8.72	5.15	37.88	43.56	51.49
hdac2	0.46	2.18	1.51	0.97	3.85	7.57	9.73
hdac8	0.72	24.75	8.96	5.18	39.62	44.71	51.76
hivint	0.57	6.04	3.00	2.20	8.96	15.00	22.00
hivpr	0.83	13.44	7.02	5.13	19.89	35.07	51.31
hivrt	0.60	14.51	3.97	2.49	25.52	19.82	24.85
hmdh	0.89	43.59	12.84	7.30	83.15	64.12	72.94
hs90a	0.79	28.62	10.03	6.03	51.02	50.00	60.23
hxk4	0.95	36.54	13.72	7.84	70.21	68.48	78.26
igf1r	0.79	18.32	5.55	4.06	28.72	27.70	40.54
inha	0.67	18.95	4.19	2.79	34.78	20.93	27.91
ital	0.32	18.17	3.92	2.03	29.07	19.57	20.29
jak2	0.87	29.92	10.66	6.26	48.48	53.27	62.62
kif11	0.71	33.93	8.80	5.18	56.52	43.97	51.72
kith	0.57	22.32	5.55	3.01	34.91	27.71	30.12
kit	0.38	8.79	3.52	2.11	17.24	17.54	21.05
kpcb	0.82	44.58	13.20	6.75	68.18	65.93	67.41
lck	0.55	6.19	2.05	1.81	9.35	10.24	18.10
lkha4	0.78	8.20	4.10	4.33	14.58	20.47	43.27
mapk2	0.81	33.92	10.71	6.25	54.84	53.47	62.38
mcr	0.71	13.94	4.68	3.09	25.00	23.40	30.85
met	0.89	62.08	14.47	7.59	90.35	72.29	75.90
mk01	0.81	31.83	9.89	5.45	54.35	49.37	54.43
mk10	0.43	3.85	1.35	0.87	5.97	6.73	8.65
mk14	0.59	10.23	3.98	2.66	16.25	19.90	26.64
mmp13	0.89	42.50	12.98	7.20	64.46	64.86	72.03
mp2k1	0.46	18.33	3.97	1.99	26.83	19.83	19.83
nos1	0.45	3.02	1.00	0.70	3.70	5.00	7.00
nram	0.93	29.02	13.30	8.07	45.16	66.33	80.61
pa2ga	0.78	13.24	8.29	5.66	25.00	41.41	56.57
parp1	0.76	24.44	7.64	4.69	40.66	38.19	46.85
pde5a	0.78	26.91	6.53	3.79	38.35	32.66	37.94
pgh1	0.48	2.07	1.13	1.03	3.67	5.64	10.26
pgh2	0.80	38.28	10.90	5.95	70.64	54.48	59.54
plk1	0.66	3.74	2.24	1.78	5.80	11.21	17.76
pnph	0.99	64.56	18.87	10.00	94.29	94.17	100.00
ppara	0.85	19.07	8.37	5.60	36.04	41.82	56.03
ppard	0.80	15.49	6.33	4.42	29.84	31.67	44.17

Performance of all 102 targets of the DUD-E using mRAISE_equality. The query ligand for aa2ar was not used from DUD-E but downloaded from the PDB for technical reasons. Reprinted from [112].

Table A.11.: Continuation of the detailed results for all targets of the DUD-E dataset using mRAISE_contact.

Target	AUC	$EF_{1\%}$	$EF_{5\%}$	$EF_{10\%}$	$HR_{1\%}$	$HR_{5\%}$	$HR_{10\%}$
pparg	0.77	26.49	8.18	4.69	49.81	40.91	46.90
prgr	0.65	10.60	3.76	2.80	19.50	18.77	27.99
ptn1	0.49	13.98	4.47	2.46	24.66	22.31	24.62
pur2	1.00	54.88	20.03	10.01	100.00	100.00	100.00
pygm	0.50	0.00	0.26	0.91	0.00	1.30	9.09
pyrd	0.80	49.08	12.10	6.94	83.08	60.36	69.37
reni	0.65	11.64	3.86	2.50	17.14	19.23	25.00
rock1	0.40	2.03	1.00	0.70	3.17	5.00	7.00
rxra	0.88	12.33	10.39	6.88	22.86	51.91	68.70
sahh	1.00	55.76	20.07	10.01	100.00	100.00	100.00
src	0.65	10.52	4.08	2.69	15.76	20.42	26.91
tgfr1	0.83	22.64	9.94	6.62	34.88	49.62	66.17
thb	0.87	49.81	13.99	7.38	68.00	69.90	73.79
thrb	0.66	2.17	2.17	2.00	3.65	10.85	19.96
try1	0.64	4.91	3.39	2.50	8.37	16.93	24.94
tryb1	0.66	4.79	3.79	2.77	9.09	18.92	27.70
tysy	0.92	43.42	15.06	8.17	69.12	75.23	81.65
urok	0.67	11.73	3.33	2.84	19.00	16.67	28.40
vgfr2	0.58	6.61	2.49	2.08	10.67	12.47	20.78
wee1	0.99	61.27	19.64	9.90	100.00	98.04	99.02
xiap	0.95	50.43	16.02	8.71	96.15	80.00	87.00
Average	0.72	22.67	7.37	4.46	36.79	36.96	44.66
Standard deviation	± 0.16	± 17.10	± 4.96	± 2.53	± 27.04	± 24.60	± 25.09

Performance of all 102 targets of the DUD-E using mRAISE_equality. The query ligand for aa2ar was not used from DUD-E but downloaded from the PDB for technical reasons. Reprinted from [112].

Table A.12.: Detailed results for all targets of the DUD-E dataset using mRAISE_inclusion.

Target	AUC	$EF_{1\%}$	$EF_{5\%}$	$EF_{10\%}$	$HR_{1\%}$	$HR_{5\%}$	$HR_{10\%}$
aa2ar	0.88	28.50	10.00	6.23	42.95	50.00	62.24
abl1	0.65	17.08	5.61	3.35	28.44	28.02	33.52
ace	0.87	33.07	11.49	6.99	54.39	57.45	69.86
aces	0.62	2.44	1.50	1.63	4.14	7.51	16.34
ada17	0.75	16.20	7.52	4.94	23.69	37.59	49.44
ada	0.90	41.17	13.34	7.42	69.09	66.67	74.19
adrb1	0.65	16.69	6.08	3.77	25.63	30.36	37.65
adrb2	0.56	8.24	3.46	2.51	12.50	17.32	25.11
akt1	0.56	5.13	1.43	1.50	8.98	7.17	15.02
akt2	0.61	15.41	5.14	3.16	25.71	25.64	31.62
aldr	0.76	34.80	9.07	5.03	60.44	45.28	50.31
ampc	0.81	38.57	10.00	5.63	64.29	50.00	56.25
andr	0.72	18.97	6.70	4.24	34.93	33.46	42.38
aofb	0.45	1.64	0.98	0.82	2.86	4.92	8.20
bace1	0.64	1.77	2.19	1.63	2.73	10.95	16.25
braf	0.78	33.20	8.56	4.87	50.00	42.76	48.68
cah2	0.64	1.63	1.14	1.48	2.53	5.69	14.84
casp3	0.61	5.07	3.22	2.26	9.26	16.08	22.61
cdk2	0.74	10.97	4.77	3.50	18.37	23.84	35.02
comt	0.93	32.45	12.71	7.09	34.21	63.41	70.73
cp2c9	0.55	0.84	1.00	1.00	1.33	5.00	10.00
cp3a4	0.51	5.32	2.35	1.59	7.56	11.76	15.88
csf1r	0.71	27.13	7.11	4.04	36.59	35.54	40.36
cxcr4	0.77	40.54	9.02	5.76	47.06	45.00	57.50
def	0.85	32.91	11.80	6.68	57.89	58.82	66.67
dhi1	0.73	15.51	6.00	4.03	26.02	30.00	40.30
dpp4	0.74	17.28	6.61	4.03	22.22	33.02	40.34
drd3	0.45	3.75	1.25	0.81	5.22	6.25	8.13
dyr	0.84	27.28	9.61	5.59	36.21	48.05	55.84
egfr	0.83	30.14	10.67	6.13	45.92	53.32	61.25
esr1	0.89	50.77	14.68	7.89	92.38	73.37	78.85
esr2	0.87	36.60	13.14	7.49	65.37	65.67	74.93
fa10	0.71	13.08	5.33	3.50	34.15	26.63	35.01
fa7	0.85	26.56	8.97	6.32	47.62	44.74	63.16
fabp4	0.84	35.25	7.70	5.33	59.26	38.30	53.19
fak1	0.81	31.29	9.82	5.80	57.41	49.00	58.00
fgfr1	0.46	0.85	0.59	0.79	25.00	17.39	23.40
fkb1a	0.81	10.83	4.15	3.60	20.34	20.72	36.04
fnta	0.63	1.18	1.62	1.81	1.35	8.11	18.07
fpps	0.98	67.11	16.95	9.18	67.06	84.71	91.76

Performance on the DUD-E. Reprinted from [112].

Table A.13.: Continuation of the detailed results for all targets of the DUD-E dataset using mRAISE_inclusion.

Target	AUC	$EF_{1\%}$	$EF_{5\%}$	$EF_{10\%}$	$HR_{1\%}$	$HR_{5\%}$	$HR_{10\%}$
gcr	0.69	26.83	6.82	3.84	45.39	34.11	38.37
glcm	0.53	1.88	0.74	0.56	2.63	3.70	5.56
gria2	0.77	41.45	11.53	6.33	54.62	57.59	63.29
grik1	0.75	28.92	8.33	4.36	43.94	41.58	43.56
hdac2	0.58	4.90	2.60	1.78	8.65	12.97	17.84
hdac8	0.77	29.46	10.61	5.89	47.17	52.94	58.82
hivint	0.56	7.05	3.00	2.10	10.45	15.00	21.00
hivpr	0.83	10.27	6.57	5.13	15.19	32.84	51.31
hivrt	0.62	12.73	4.32	2.55	22.40	21.60	25.44
hmdh	0.88	43.00	13.20	7.36	82.02	65.88	73.53
hs90a	0.78	26.33	10.03	6.14	46.94	50.00	61.36
hxk4	0.92	35.43	13.07	7.29	68.09	65.22	72.83
igf1r	0.77	16.28	5.82	4.06	25.53	29.05	40.54
inha	0.69	18.95	3.73	2.33	34.78	18.60	23.26
ital	0.32	18.17	3.92	2.03	29.07	19.57	20.29
jak2	0.85	28.05	10.66	6.17	45.45	53.27	61.68
kif11	0.75	34.80	10.35	5.87	57.97	51.72	58.62
kith	0.57	22.92	5.67	3.01	35.85	28.31	30.12
kit	0.45	8.79	3.52	2.11	17.24	17.54	21.05
kpcb	0.82	43.84	13.05	6.82	67.05	65.19	68.15
lck	0.55	6.19	2.33	1.98	9.35	11.67	19.76
lkha4	0.81	7.62	4.34	4.57	13.54	21.64	45.61
mapk2	0.91	41.91	13.48	7.63	67.74	67.33	76.24
mcr	0.65	12.86	4.04	2.55	23.08	20.21	25.53
met	0.88	62.68	13.74	7.41	91.23	68.67	74.10
mk01	0.76	31.83	7.61	4.44	54.35	37.97	44.30
mk10	0.54	2.89	2.12	1.73	4.48	10.58	17.31
mk14	0.60	8.85	3.08	2.21	14.05	15.40	22.15
mmp13	0.89	38.82	11.79	6.85	58.89	58.92	68.53
mp2k1	0.47	19.17	3.97	2.15	28.05	19.83	21.49
nos1	0.44	4.02	1.20	0.90	4.94	6.00	9.00
nram	0.89	22.80	11.66	6.84	35.48	58.16	68.37
pa2ga	0.77	32.60	11.53	6.47	61.54	57.58	64.65
parp1	0.75	24.05	7.09	4.41	40.00	35.43	44.09
pde5a	0.74	23.64	6.13	3.39	33.69	30.65	33.92
pgh1	0.44	4.65	1.75	1.23	8.26	8.72	12.31
pgh2	0.78	35.51	10.21	5.59	65.53	51.03	55.86
plk1	0.64	1.87	1.87	1.50	2.90	9.35	14.95
pnph	0.99	65.54	18.29	9.91	95.71	91.26	99.03
ppara	0.87	18.53	8.80	5.90	35.03	43.97	58.98
ppard	0.83	18.01	8.25	5.25	34.68	41.25	52.50

Performance of all 102 targets of the DUD-E using mRAISE_equality. The query ligand for aa2ar was not used from DUD-E but downloaded from the PDB for technical reasons. Reprinted from [112].

Table A.14.: Continuation of the detailed results for all targets of the DUD-E dataset using mRAISE_inclusion.

Target	AUC	$EF_{1\%}$	$EF_{5\%}$	$EF_{10\%}$	$HR_{1\%}$	$HR_{5\%}$	$HR_{10\%}$
pparg	0.78	21.94	8.23	4.77	41.25	41.12	47.73
prgr	0.65	8.21	2.66	2.15	15.09	13.31	21.50
ptn1	0.52	13.21	4.78	3.16	23.29	23.85	31.54
pur2	1.00	54.88	20.03	10.01	100.00	100.00	100.00
pygm	0.49	0.00	0.52	1.30	0.00	2.60	12.99
pyrd	0.84	50.89	12.83	7.03	86.15	63.96	70.27
reni	0.62	28.12	5.98	3.18	41.43	29.81	31.73
rock1	0.44	1.02	0.80	0.60	1.59	4.00	6.00
rxra	0.95	28.51	15.43	8.25	52.86	77.10	82.44
sahh	1.00	55.76	20.07	10.01	100.00	100.00	100.00
src	0.63	9.95	3.93	2.63	14.90	19.66	26.34
tgfr1	0.84	24.15	9.94	6.24	37.21	49.62	62.41
thb	0.91	54.69	15.35	7.87	74.67	76.70	78.64
thrb	0.63	1.95	1.74	1.65	3.28	8.68	16.49
try1	0.66	6.92	3.88	2.81	11.79	19.38	28.06
tryb1	0.59	9.57	2.98	2.16	18.18	14.86	21.62
tysy	0.90	38.80	13.96	7.71	61.76	69.72	77.06
urok	0.68	11.73	3.83	2.66	19.00	19.14	26.54
vgfr2	0.64	6.37	2.79	2.32	10.28	13.94	23.23
wee1	0.99	61.27	19.25	9.90	100.00	96.08	99.02
xiap	0.96	52.45	17.22	9.11	100.00	86.00	91.00
Average	0.72	22.76	7.45	4.45	37.12	37.37	44.60
Standard deviation	± 0.16	± 17.04	± 4.99	± 2.51	± 27.14	± 24.73	± 24.93

Performance of all 102 targets of the DUD-E using mRAISE_equality. The query ligand for aa2ar was not used from DUD-E but downloaded from the PDB for technical reasons. Reprinted from [112].

B Appendix B. Implementation

This section will provide an overview of the implementation of the software tool mRAISE developed during this dissertation. The software has been implemented using the coding language C++ and is available in two different versions:

- mRAISE_cmdline provides a slim command line interface and is ideally suited for large scale screening runs using automated scripts on computer clusters.

- mRAISE provides a GUI, which can also be used for the basic screening functionalities for mRAISE, but is especially designed for the query preparation and visualization of screening results.

While no completely new libraries for the NAOMI library at the ZBH were developed, important contributions were made to the following libraries:

- FastBitIndex

- Trixx

These libraries provide the basic functionalities for descriptor generation, index creation and descriptor matching and are shared between the different screening approaches mRAISE, iRAISE, cRAISE and TrixP.
Besides these two libraries, the mRAISE software tools furthermore depend on further internal components of the NAOMI-library as well as external libraries. Figure B.1 shows all dependencies of both tools to all used libraries.

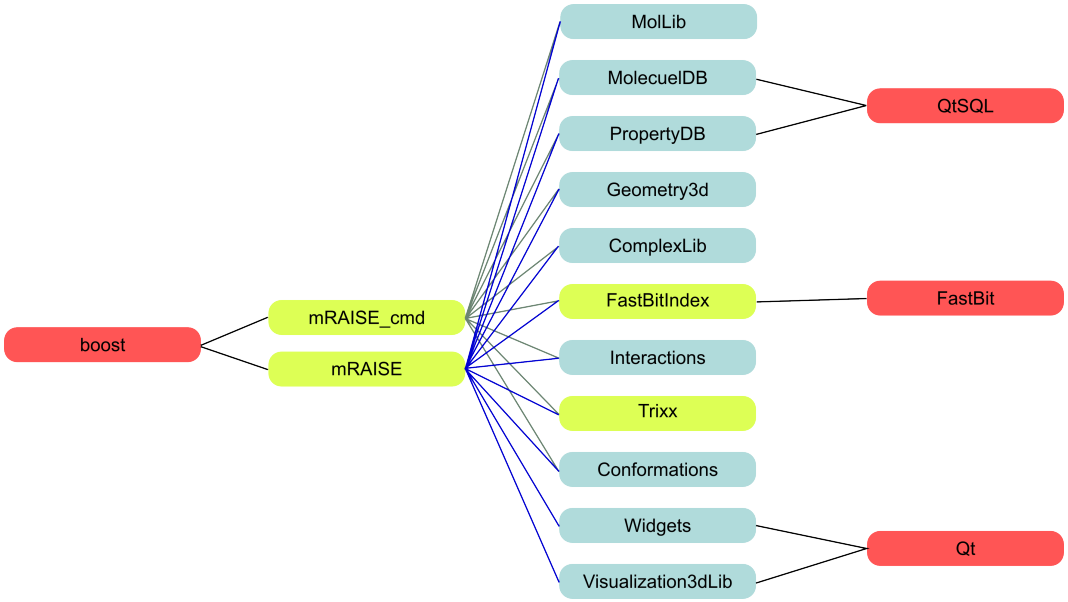


Figure B.1.: Overview of the dependencies of the mARAISE software tools. External libraries are shown in red, Naomi libraries are shown in blue and libraries modified during this work as well as the tools are shown in green.

B.1. Dependencies to the NAOMI-library

- **MolLib** includes the functionality to initialize molecules from various file formats into the internally used representation. Furthermore, it includes the underlying chemical model and proves also the functionality to write molecules back to file.

- **MoleculeDB** is the utilized SQLite database for the efficient storage of molecules in multiple conformations.

- **PropertyDB** addition to the MoleculeDB to store additional data alongside the molecule entries.

- **Geometry3d** includes three-dimensional objects like spheres and the used icosahedron. Furthermore, this library includes the functionality to calculate the transformations to align point triplets.

- **ComplexLib** provides the internal representation of protein-ligand complexes and has dependencies to the ProLib, which provides the functionality to initialize proteins from files as well as determining an active site.

- **Interactions** calculates directed interactions for molecular structures. The interaction model of this library is the basis for the polar interactions in Trixx.
- **Conformations** provides the latest version of the CONFECT method for the generation of additional conformations for a given molecule.
- **Widgets** includes a variety of ready to use widgets for Qt GUI applications. It has been used primarily to include functionality needed to use a license system in the software.
- **Visualization3dLib** provides the basic functionality to include a widget for 3d visualization of small molecules and proteins alongside other simple three-dimensional objects in a Qt GUI.

B.2. Dependencies to the External-library

- **boost** is a basic library for C++ providing a variety of useful data structures and algorithms.
- **Qt** is a framework for the platform-independent development of GUIs in C++.
- **QtSQL** provides an easy to use interface for SQLite databases. SQLite is an open software library providing an SQL database that does not need an extra server.
- **FastBit** is the library used to create the compressed bitmap index for the triangle descriptors. Since FastBit does not provide support for other systems than Linux at the moment, mRAISE is only available for Linux operating systems.

B.3. Used Modules of the Trixx-library

- **Interactions** includes the creation of interaction points based on the NAOMI interactions and with additional algorithms for the calculation of hydrophobic interaction points.
- **InteractionTriangle** provides the functionality to enumerate triangles based on previously calculated interaction points. This also includes the algorithms for the canonization of descriptors and the creation of the shape descriptor.

- **MoleculePreprocessor** uses both previously mentioned libraries to create a Trixx-Molecule, annotated with all information needed for VS using the TrixX descriptor.

B.4. Used Modules of the FastBitIndex-library

- **TriangleIndexBuilder** creates the descriptor index.
- **TriangleIndexInquirer** matches query descriptors against a previously defined descriptor index.
- **Match** stores the minimally required information of a matching descriptor pair.

Appendix C.

C mRAISE User Guide

After downloading and extracting the mRAISE package, the resulting folder includes both binaries of mRAISE together with the folders ”lib”, ”Licenses”, ”plugins” and ”qml”. These folders include external libraries and other resources required by mRAISE. In the following, the usage of the command line version of mRAISE will be described in detail. For the usage of the GUI version see Appendix D.

C.1. Starting mRAISE_cmdline

mRAISE_cmdline can be directly started from the package folder. For an overview of the available parameters use

```
./mRAISE_cmdline --help
```

or just start it without any parameters. On the command line you now see all parameters with a short description. A list of all parameters can be seen in Table C.1 and Table C.2. Please note that working with mRAISE is divided into three major steps, which can only be used separately. Theses steps represent different use-cases in the LBVS process and are indicated by parameters with a capital letter. mRAISE_cmdline can be used for creating a new descriptor index (I), to screen an existing descriptor index (S) or to evaluate previous screening solutions (E). Each of this steps can be further specified with additional parameters use-case specific and general parameters.

Table C.1.: Overview of all mRAISE_cmdline parameters.

Parameter	Description
General options	
-h [–help]	Prints the help message.
-v [–verbosity] arg (=1)	Regulates the detail of output. quiet(0) / basic(1)=default / detailed(2)
-s [–summary] arg	Summary log file for statistics of index creation or screening runs. The file includes information like the number of generated descriptors, the number of matches, and the run time statistics.
-l [–license] arg	mRAISE uses a license system, with this command a new license can be provided.
-f [–folder] arg	Folder in which a new index can be written or in which a previously calculated index can be found. For the creation of a descriptor index, a new folder will be created with the provided name.
-o [–output] arg	Name of the output file for screening runs. or evaluation.
Indexing options	
-I [–Indexing] arg	Multy-mol2 or -sdf file containing all structures that should be written into a new descriptor index. The index location has to be specified using -f. If the folder already contains an index, molecules will be attached.
-c [–conformations] arg (=0)	Number of conformations that should be generated for each compound.

Table C.2.: Overview of all mRAISE_cmdline parameters.

Parameter	Description
Screening options	
-S [–Screening] arg	File containing a query molecule that should be used for screening an index specified with -f. The provided file needs to be of type .mol2 or sdf.
-t [–matching_type] arg	Matching type paramter for partial bulk comparison. 25% bulk requirement(0) / 50% bulk requirement(1)
-r [–referenceProtein] arg	Reference protein that should be used together with the query molecule in order to derive partial shape constraints.
-p [–partial] arg	Partial Shape Constraint Type as derived from the reference Protein. Inclusion(0) / Contact(1)
-d [–poseDB] arg	Name for a solution database containing the best pose for each conformation. Such a DB is needed to write molecules to sdf during Evaluation. Otherwise it is optional for screening.
Evaluation options	
-E [–Evaluation] arg	Load one or more solution databases (comma separated)
-g [–group] arg (=molecule)	Pick one solution for each 'conformation', 'molecule' or 'name'.
-w [–write_results] arg (=100)	Write down best x conformations to an .sdf file.

C.2. Example Use Cases

In the following, the three different use cases, which can be performed using mRAISE_cmdline are explained with examples for the used parameters. The first use case is the creation of a new descriptor index for a compound library, the second is the screening of an existing index, and the third writes a certain number of top-ranked conformations to a file.

C.2.1. Creating a Descriptor Index

To create a descriptor index, a molecule file containing the screening library needs to be provided. Furthermore, the user has to decide where the index should be created and if additional conformations should be generated for the provided compounds.

Therefore, the tool has to be called with the -I parameter, followed by the molecule file, furthermore, a name needs to be given for the folder the index will be created in. If only a name and no path is given, the directory is created in the current directory. If a folder of this name already exists, the program will terminate.

```
./mRAISE_cmdline -I screeninglibrary.mol2
                 -f indexFolder
```

If the user wants to generate conformations for the compound library, the maximum number of conformations per compound has to be defined using the -c parameter.

```
./mRAISE_cmdline -I screeninglibrary.mol2
                 -f indexFolder -c 200
```

This call is sufficient to create a new descriptor index for all compounds of the file 'screeninglibrary.mol2' in a new folder named 'indexFolder' with up to 200 additional conformations for each compound.
Additionally, the user could specify the name of a log file using -s. After the index creation, this file would contain information like the number of input molecules, the number of generated conformations, the number of generated descriptors and the time needed to fulfill this process.

C.2.2. Screening a Descriptor Index

Once an index has been created, is can be screened as often as desired. In this step, the user has to choose parameters according to the kind of query he wants to use.

The query ligand that should be used has to be provided either in mol2 or sdf format using the -S parameter. Furthermore, the descriptor index that should be screened has to be defined using again the -f parameter and the output file showing the scores for all compounds has to be provided with -o.

```
./mRAISE_cmdline -S queryMol.mol2 -f indexFolder
                 -o ranking.csv
```

If the provided query ligand has been created using the mRAISE GUI, the query mode defined in the file, i.e. 25% shape matching, 50% shape matching or manually selected partial shape constraints provided in the file annotation are used.
However, if the ligand does not contains mRAISE query information, the matching type has to be defined using -t and can be either 25%(0) or 50%(1) shape matching.

```
./mRAISE_cmdline -S queryMol.mol2 -f indexFolder
                 -o ranking.csv -t 1
```

Alternatively, a reference protein can be provided using -r and complex-derived partial shape constraints are derived according to the simultaneously provided -p parameter. For example

```
./mRAISE_cmdline -S queryMol.mol2 -f indexFolder
                 -o ranking.csv -r queryProtein.pdb
                 -p 0
```

starts a screening run using complex derived inclusion queries and matching ligands are likely to fit into the respective binding site.

An optional parameter during the screening procedure is again -s to create a log file with the given name including information like the number of query descriptors, the number of matches and run time measurements. Additionally -d can be used to write the best scored alignment poses into a database for later evaluation using the GUI or this tool with the -E parameter as shown in the next section.

C.2.3. Evaluation of Screening Results

If a screening run has been performed and a solution database has been created using the -d parameter, the database can be processed to write the best scored poses to a new molecule file. If the compound library has been split to screen the different parts of it simultaneously, the resulting solution databases can be combined and evaluated at once.

Using the -E parameter, one or multiple solution databases can be selected and the number of solutions provided with the -w parameter are written into a new file provided with -o.

```
./mRAISE_cmdline -E screening.db -w 50 -o top50.sdf
```

By default, this writes the best 50 solutions into the top50.sdf file with only one solution per unique molecule. Alternatively, the -g parameter can be used to instead choose to write the best solution per molecule name or per conformation.

D Appendix D. mRAISE GUI User Guide

The second version of mRAISE provides a fully functional GUI. Using this interface the user can create and screen descriptor indices, visualize query ligands and descriptors, define manual shape queries, and visualize screening results.
The following user guides introduces the different functionalities of the mRAISE GUI.

D.1. Starting mRAISE

The mRAISE binary can be started directly and does not require any additional parameters.

```
./mRAISE
```

D.2. Screening Preparation

LBVS runs can be prepared in mRAISE using the screening tab (see Figure D.1). This tab shows a list of prepared queries and indices and if an index and a query are selected, a screening run can be started directly using the 'Start Screening' button. In order to create a new index the 'Create Index' button has to be clicked. In the following pop-up dialog, a name, a directory and a molecule file containing the compound library can be selected. If additional conformations should be generated for the screening library, the maximum number of conformations per compound can be selected as well. After accepting the dialog, the descriptor index is created

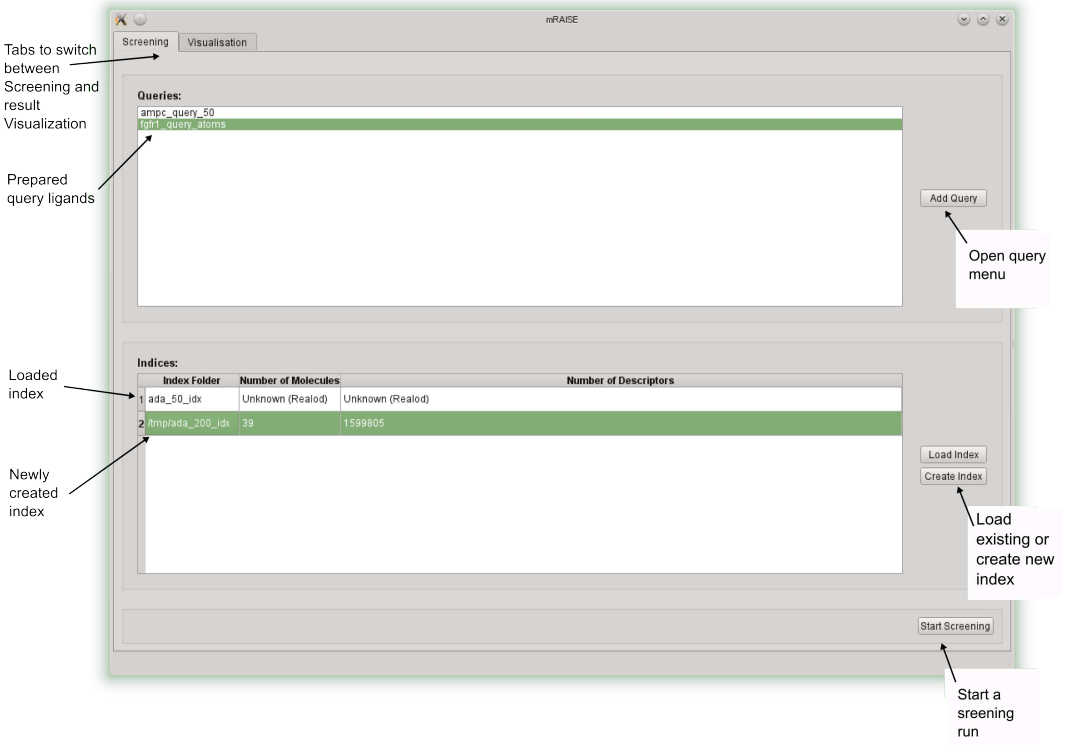


Figure D.1.: Screenshot of the screening tab with explanatory text.

accordingly. To define a query based on a new ligand, the query definition dialog can be started by clicking on the 'Add Query' button.

D.3. Query Definition

Figure D.2 shows the query definition dialog. Here, molecules or protein-ligand complexes can be loaded from input files to create queries for screening.
Once a ligand or complex is loaded, it can be visualized by clicking on its entry in the respective lists. For a displayed structure, interaction points and descriptors can be shown by clicking on the respective checkboxes.
By clicking on the 'Create Query' button, the currently selected ligand or complex is used to derive a new query for screening. For a ligand, the query can be defined by selecting one of the options 'Match 25%', 'Match 50%' or 'Atom selection'. In order to derive partial shape constraints based on an atom selection, first atoms have to be selected by clicking on them as shown in Figure D.3. In case a complex is selected when clicking on 'Create Query', the user can decide if inclusion of contact constraints should be derived from the complex. Defined queries are displayed in

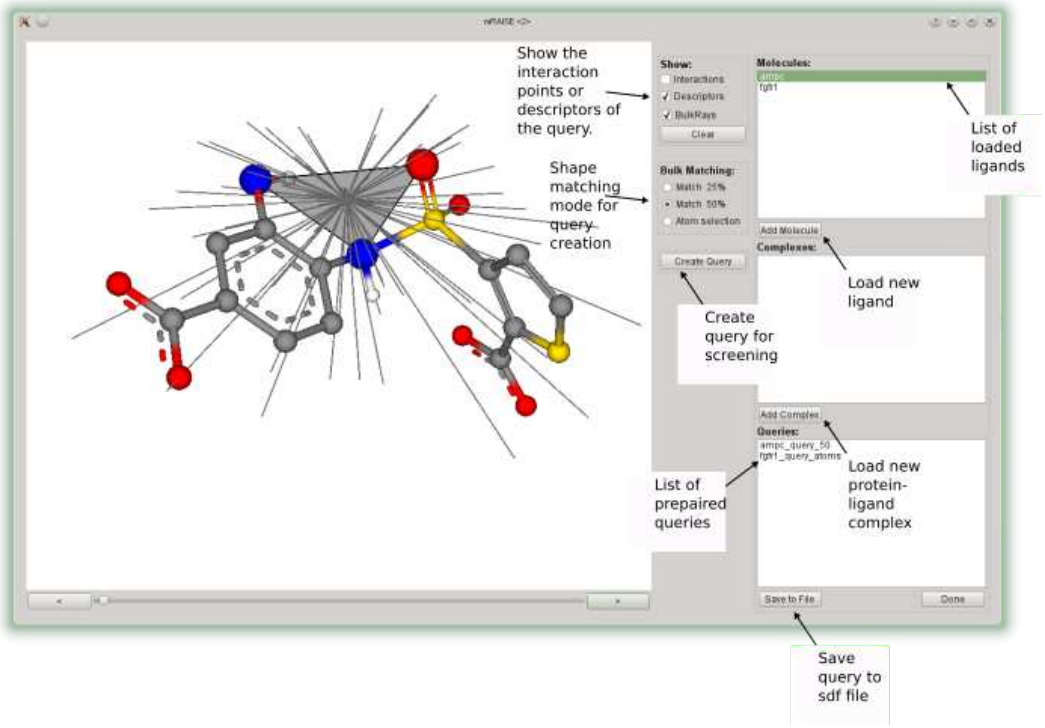


Figure D.2.: Screenshot of the query definition dialog with explanatory text.

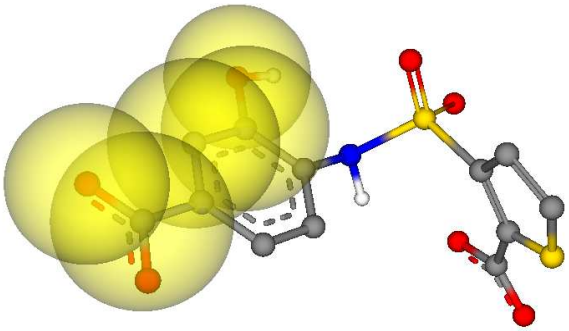


Figure D.3.: Screenshot of a ligand with selected atoms indicated by yellow spheres in the query definition dialog.

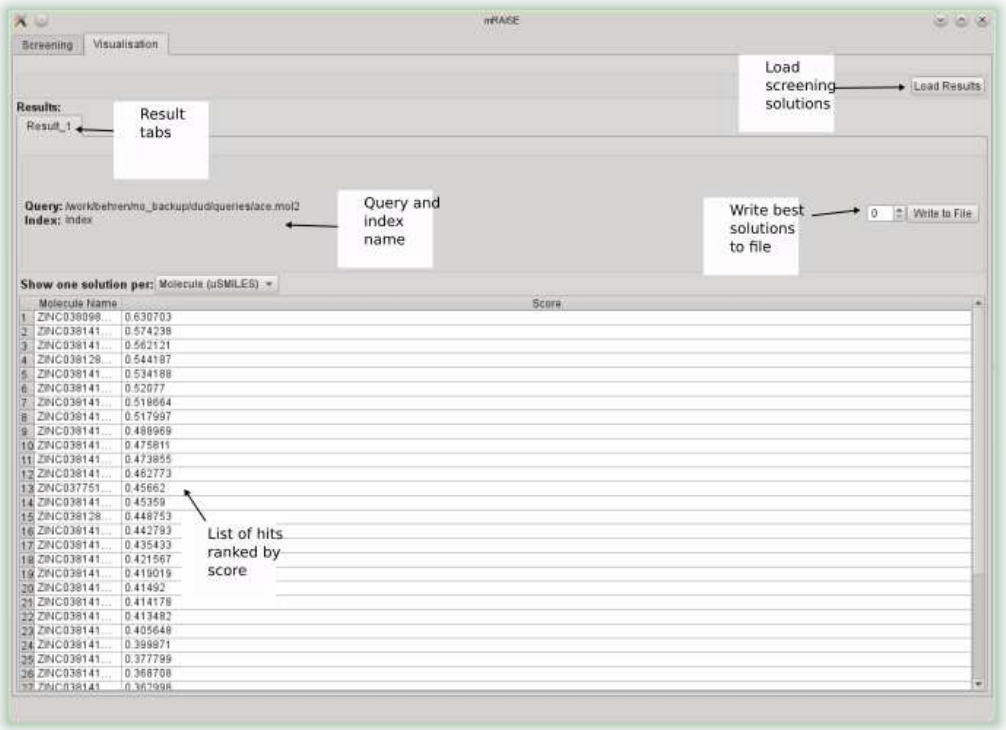


Figure D.4.: Screenshot of the solutions visualization tab with explanatory text.

the 'Queries' list of the query definition dialog as well as in the respective list of the screening tab when the dialog is closed.
Additionally, defined queries can be saved to sdf files using the 'Save to File' button. This way, even queries based on atom selections can be saved and afterwards used in the mRAISE_cmdline version.

D.4. Screening Solution Visualization

If a screening run is performed using the GUI, the results are automatically displayed in the visualization tab (Figure D.4). Alternatively, screening solutions of previous screening runs stored in a solution database can be loaded by clicking on 'Load Results', providing the path to the database as well as the used query ligand. In the bottom half of the window, a sorted list of found hits together with the respective similarity score are shown and by clicking on the 'Show one solution per' dropdown menu the user can select what kind of solutions should be displayed. By clicking on the 'Write to File' button, a selected amount of top ranked solutions can be saved to a molecule file.

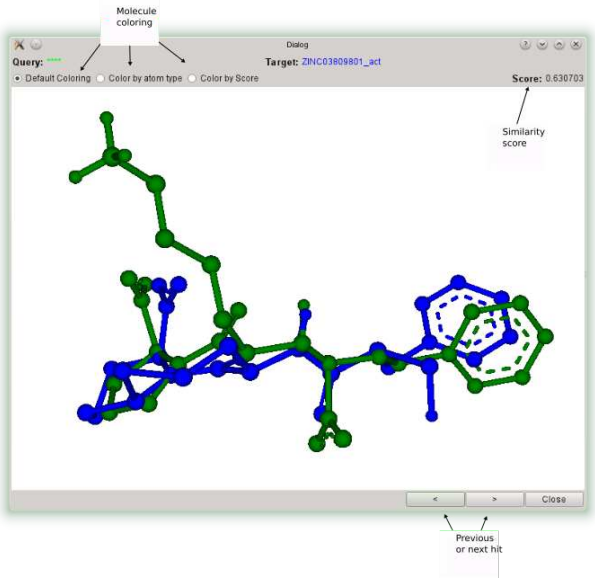


Figure D.5.: Screenshot of the alignment visualization of a screening solution with explanatory text.

Furthermore, by clicking on an entry in the list, a new dialog is opened showing the respective molecular alignment. This is shown in detail in the next section.

D.5. Alignment Visualization

For a selected hit the molecular alignment is displayed in a separate window (see Figure D.5). By default, the query is shown in green and the aligned compound is shown in blue. However, by clicking on 'Color by atom type' or 'Color by Score' the color of the ligands can be changed to display other information. In the right corner the score of the respective alignment is displayed and with the buttons at the bottom the visualization can switch to the previous or next hit of the sorted list.

Appendix E.

E mRAISE Dataset

The mRAISE dataset has been composed as benchmarking set for future comparison studies. It is therefore available on the homepage of the Center for Bioinformatics Hamburg. An overview of the dataset can be seen in Table E.1

Table E.1.: List of all PDB–IDs of the alignment validation dataset with corresponding HET codes of the ligands.

PDB	HET	PDB	HET	PDB	HET	PDB	HET
1C1P	BAI	2F18	GB1	3MHM	J75	4IPW	1G7
1C1Q	BAI (Fragment)	2F1B	GB3	3MHO	J43	4IQ9	1GB
1C5Q	ESI	2F7O	MSN	3MMF	D9H	4KTF	1TM
1C5S	ESX	2F7P	2SK	3MNA	DWH	4KTJ	KTJ
1C5T	ESP	2F7R	SK3	3MYQ	E27	4KTK	KTK
1GHZ	120	3BLB	SWA	3MZC	S6I	2AVV	MK1
1GI0	BMZ	3DX2	MZB	3N0N	P9B	2F80	017
1GI6	124	3DX3	YTB	3N3J	WWV	2QCI	065
1GJ6	132	3DX4	GOO	3OIK	WZB	3PWR	ROC
1O2I	655	3EJP	HN2	3OYQ	OYQ	3QAA	G04
1O2N	762	3EJQ	HN3	3P3H	84A	3SAB	F78
1O2P	972	3EJR	HN4	3QYK	IE2	3TH9	9Y9
1O2Q	991	3EJS	HN5	3RJ7	RCS	3TOG	079
1O2R	CR9	1RMZ	NGH	3RYJ	RYJ	4HDB	G52
1O2U	847	2HU6	37A	3RYV	RYV	4KB9	G79
1O33	801	3F16	HS3	3RYY	RYY	3U5L	08K
1O35	802	3F17	HS4	3RYZ	RYZ	3ZYU	1GH
1O37	653	3F18	HS5	3S71	EVD	4BW1	S5B
1O3D	780	3F19	HS6	3S74	03T	4BW3	9BM
1O3J	334	3F1A	HS7	3S8X	E59	4CFL	8DQ
1O3L	678	4GR0	R4B	3S9T	E49	4F3I	0S6
2AYW	ONO	4GR8	R4C	3SAX	E50	4HXM	1A8
3A8A	4FZ	4H76	10B	3SBI	E90	4HXN	1A7
3LJO	11U	1GZ8	MBP	3V5G	0F3	4HXS	1A3
3NKK	JLZ	2R3F	SC8	3ZP9	9TH	hHB4	SCV
4AB9	VXQ	2R3H	SCE	4BF1	9FK	1OBN	ASV
4ABA	SW1	2R3I	SCF	4DZ7	D02	1QJF	ACS
4ABD	SW2	2R3Q	5SC	4DZ9	ID4	1UZW	CDH
4ABE	913	2R3R	6SC	4FPT	0VZ	1W04	HCG
1C5N	ESI	4EK4	1CK	4KAP	1QV	1W3V	MDZ
2CF8	ESH	4GCJ	X64	3FVP	UB2	1W3X	W2X
2CN0	F25	2AW1	COX	3QGO	0A9	2BU9	HFV
2ZFF	53U	2FOQ	B15	3T74	UBY	2IVI	ACW
3P17	99P	2NNG	ZYX	3T87	UBZ	2IVJ	BCV
3RM0	S54	2NNS	M25	3T8D	UBV	2JB4	A14
3U8O	PRD_000940 (DTH,DPN,PRO,NH2,DAR)	2QO8	3CC	3T8F	UBU	2VBB	VAZ
3U8R	PRD_001093 (DPN,PRO,NH2,DAR,ILE)	2QP6	MB1	3T8G	UBT	2Y60	M8F
3U98	BJA	2WEG	FBV	3T8H	UBS	3ZKU	HCV
3UWJ	TIF	2WEO	FBW	4D9W	X32	3ZKY	WT4
3VXE	DPN	3DCW	EZL	3G5H	YTT	4BB3	KKA
1TQS	SSO	3IBU	O48	4G48	PZB		
1TQW	BLT	3M96	E38	4IPS	1G4		

Ensembles are separated by horizontal lines and listed in the following order: Trypsin, Thrombin, ALPHA-MANNOSIDASE II, Matrix metalloproteinase-12 (MMP-12), CDK 2 Kinase, Carbonic Anhydrase II, Thermolysin, CYP121, HIV Protease, Bromodomain-containing protein 4, Isopenicillin N Synthase
Reprinted from [107] with permission of Springer.

F Appendix F. Publications

In this appendix, the scientific contributions of the author are listed.

F.1. Publications in Scientific Journals

1. **M. M. von Behren, S. Bietz, E. Nittinger, and M. Rarey, "mRAISE: an alternative algorithmic approach to ligand-based virtual screening," *J. Comput. Aided Mol. Des.*, vol. 30, pp. 583–594, Aug 2016.**

2. **M. M. von Behren and M. Rarey, "Ligand-based virtual screening under partial shape constraints," *J. Comput. Aided Mol. Des.*, Manuscript submitted for publication.**

3. M. M. von Behren, A. Volkamer, A. M. Henzler, K. T. Schomburg, S. Urbaczek, and M. Rarey, "Fast protein binding site comparison via an index-based screening technology.," *Journal of chemical information and modeling*, vol. 53(2), pp. 411–22, 2013.

F.2. Publications in Scientific Books

1. A. Volkamer, M. M. von Behren, S. Bietz, M. Rarey, "Prediction, Analysis and Comparison of Active Sites" *Chemoinformatics, Basic Concepts and Methods*, Manuscript submitted for publication.

F.3. Conference Posters

1. **M. M. von Behren, M. Rarey, An Index-based Virtual Screening Technology and its Application for Ligand-based Screening and Compound Library Design, Gordon Research Conference, 2015, Boston, USA**

Eidesstattliche Versicherung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form in einem Verfahren zur Erlangung eines akademischen Grades vorgelegt. Es wurde an keinem anderen Fachbereich ein Antrag auf Eröffnung eines Promotionsverfahrens gestellt.

Hamburg, den 30.11.2016

Mathias Michael von Behren