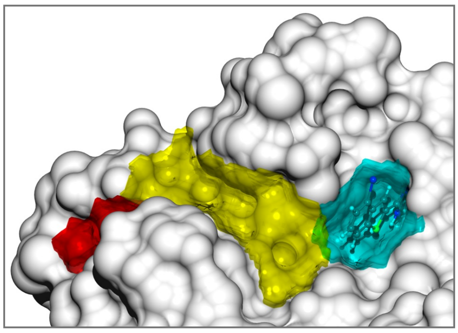
# COMPASITES - Computer-aided active site analysis of protein structures



Cumulative Dissertation
to receive the degree

*Dr. rer. nat.*

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

submitted to the
Department of Computer Sciences of
the University of Hamburg

Andrea Volkamer

born in Merzig

Hamburg, September 2012

1. Reviewer: Prof. Dr. Matthias Rarey
2. Reviewer: Prof. Dr. Ulrike von Luxburg
3. Reviewer: Prof. Dr. Gisbert Schneider

Date of thesis defense: January 14, 2013.

My parents Gertrud and Klaus Volkamer.

# Acknowledgements

Here, I would like to acknowledge the people that supported and encouraged me during my doctoral thesis.
First of all, I want to especially thank my supervisor Prof. Dr. Matthias Rarey, who entrusted me with this interesting project and guided me during the course of this whole study. My thanks extend to Prof. Dr. Ulrike von Luxburg and to Prof. Dr. Gisbert Schneider for reviewing my thesis.
I would also like to thank all my current and former colleagues from the Center of Bioinformatics (which are too many to list here) for the pleasant work atmosphere, the fruitful discussions, the provision of support on divers computational questions as well as for joking around in the coffee kitchen. During my time here, many of them became my friends. Special thanks goes to my former and current "office mates" Sascha Urbaczek and Angela Henzler for the good, although often stressful, times we spend together in this office. In addition, I really appreciated the work of everyone in this group who participated in the creation of a reliable software library, which we were able to base our work on. I especially thank Axel Griewel and Mathias von Behren whose developments became parts of our respective publications.
Furthermore, I want to thank my cooperation partners from Merck KGaA and BioSolveIT, especially Daniel Kuhn, Thomas Grombacher, Friedrich Rippmann, and Christian Lemmen. In this context, I also want to thank Matthias Wirth from Merck Serono in Geneva for a good collaboration.
Furthermore, I thank the BMBF for funding my work and allowing me to present my work on national as well as international conferences.
In addition, I want to thank Florian Lauck, Karen Schomburg, and Carsten Detering for proofreading parts of my thesis.
I'm grateful for my friend and colleague Nadine Schneider who already accompanied me during my studies in Saarbrücken, and who has - as a fellow sufferer - mentally supported me especially during these last weeks of writing this thesis.
Finally, I would like to thank my family and friends for always being there for me when I need(ed) them.

# Kurzfassung

Die vorliegende Arbeit beinhaltet den allgemeinen Teil meiner kumulativen Dissertationsschrift, eingereicht bei der Universität Hamburg im September 2012. Sie beschreibt meine wissenschaftliche Arbeit an Computerstrategien zur strukturbasierten Analyse von aktiven Zentren in der Abteilung für Algorithmisches Molekulares Design, von Juni 2008 bis September 2012. Darüber hinaus beinhaltet meine Dissertationsschrift sechs wissenschaftliche Veröffentlichungen, die in einem gesonderten Literaturverzeichnis aufgelistet sind und mit der Bezeichnung D1 - D6 im Text referenziert werden.

Molekulare Erkennung und Bindung von kleinen Molekülen an ein Protein ist die Basis für die Erhaltung biologischer Systeme. Dabei spielt die dreidimensionale Struktur des Proteins, genauer gesagt die Struktur des aktiven Zentrums, eine elementare Rolle. Die für die Bindung zwischen Protein und Ligand verantwortlichen Wechselwirkungen zu verstehen und zu modifizieren ist für die pharmazeutische und biotechnologische Industrie von großer Bedeutung. Da die Anzahl bekannter Proteinstrukturen immer größer wird, werden effiziente computergestützte Methoden immer wichtiger, um experimentelle Verfahren zu unterstützen und zu ergänzen.

In meiner Arbeit habe ich mich mit verschiedenen strukturbasierten Strategien zur computergestützten Analyse aktiver Zentren befasst. Da das aktive Zentrum eines Proteins der Schlüssel zu seiner Funktion ist, habe ich zuerst einen verläßlichen Algorithmus zur Erkennung von Bindetaschen entwickelt. Verschiedene globale und lokale Deskriptoren können anschließend für diese Taschen berechnet werden und in Kombination mit hierarchischen Gruppierungsverfahren, maschinellen Lernverfahren und der Suche nach nächsten Nachbarn zur Proteinbeschreibung benutzt werden. Die Klassifizierungsmethoden wurden eingesetzt, um Proteine für den Wirkstoffentwurf zu priorisieren und unbekannten Proteinen eine Funktion zuzuordnen. In einer weiteren Studie wurde die sterische Passform der beiden Bindungspartner, Protein und Ligand, untersucht und ihre Komplementarität numerisch erfasst.

Da die Qualität vieler Ansätze unter der starren Modellierung der flexiblen Proteinstruktur leidet, wurde zusätzlich ein Ansatz zum Vergleich von Bindetaschen auf der Basis von Dreiecksdeskriptoren entwickelt, der zumindest kleine Veränderungen in der Bindetasche berücksichtigt.

# Abstract

The thesis in hand comprises the general part of my cumulative dissertation, submitted to the University of Hamburg in September 2012. It describes my scientific work on computer strategies for structure-based active site analysis in the group for Computational Molecular Design of the Center for Bioinformatics from June 2008 till September 2012. Furthermore, this manuscript contains six scientific journal publications, listed in a separate bibliography at the end of this dissertation and cited in the text with D1 - D6.

Molecular recognition and binding of small molecules to a protein is the foundation for the maintenance of biological systems. In this context, the 3D structure of the protein, more precisely, the structure of the active site plays a fundamental role. Understanding and modifying the mechanism of ligand binding is of high practical interest in pharmaceutical and biotechnological research. Due to the growing number of available three dimensional protein structures, efficient computational methods are needed to assist experimental approaches.

In my work, I designed several strategies addressing different parts of the structure-based computer-aided active site analysis cycle. Since the active site of a protein is the key to its function, the first step in my work was the development of a novel algorithm for binding site detection. For predicted sites, several global and local descriptors can subsequently be calculated and used for protein assessment. For protein classification, the descriptors are incorporated into hierarchical clustering, machine learning and nearest neighbor search techniques. These classifiers can be used to prioritize potential disease modifying proteins in drug development processes, and to annotate protein function. In a subsequent study, the shape complementarity requirement for molecular recognition has been explored and numerically registered.

Since many approaches suffer from the rigid modeling of the naturally flexible structure of proteins, I participated in establishing an approach for active site comparison based on triangle descriptors of the active site, accounting for small changes in the binding site.

# Contents

# 1

# Introduction

Bioinformatical software approaches have long entered a broad range of disciplines in genetics and biochemistry. Due to advances in sequencing and structural genomics, the vast amount of data cannot be managed without computer assistance anymore. Besides data organization, computers are established tools in discovery processes like pharmaceutical drug design [1]. Time and cost expenses for experimental methods limit high-throughput applications. Thus, efficient computational approaches are increasingly needed to model such scenarios in order to put biological information to practical use.

It has been known for years that the three dimensional structure of a protein is the key to its function [2]. The temporary complex formation of proteins and small molecules is the basis for the conservation of biological systems. Affecting and modulating these interactions is the main goal of pharmaceutical and biotechnological campaigns. The understanding of the forces driving molecular recognition is still not sufficient. The difficulty of the drug and protein design process results from the complexity of the organisms. Thus, various aspects about the binding scenario have to be known to effectively modify the natural designation of a protein. Such modifications often address the inhibition or the optimization of the function of a protein.

My thesis describes the development of a portfolio of computer methods allowing for a detailed analysis of the active site of a protein based on its structure. Detecting the active site, which is responsible for the reaction of a protein, and describing it with a high specificity was one major goal of my work. The developed software can be used to automatically derive information valuable for protein classification, modification and comparison.

My project was funded by the BMBF[1], grant 0315292A, and was part of the Biocatal-

[1] Bundesministerium für Bildung und Forschung, PTJ, Jülich

ysis2021 cluster. The project was a cooperation between the Center for Bioinformatics, Merck[1] and BioSolveIT[2]. This combination triggered a broad functionality and applicability of the algorithms due to the interdisciplinary nature of the applications in pharmacy and biotechnology. The pharmaceutical partner Merck is highly interested in generating information for the drug development process, amongst others the prediction of the potential to address a disease-related protein by a small molecule. The Biocatalysis2021 cluster investigates topics considering the discovery and efficient yielding of biocatalysts for optimized biotechnological processes. The analysis and optimization of the docking process for rational enzyme design was the duty of the cooperation partner BioSolveIT. In summary, our project was in authority to generate information for rational enzyme design, e.g., the comparison of enzymes and the identification of potential mutations for process optimization.

In the following, the molecular basics to understand the importance of proteins and their interactions to maintain biological functions are introduced first. Second, the scope and the importance of bioinformatics tools in pharmaceutical, biotechnology and biological research is explained. In the third part, the references to and the benefits of my work are motivated. Finally, the structure of this thesis is outlined, containing an explanation of the contributions of my scientific publications to the dissertation.

## 1.1 Proteins, Small Molecules and Interactions

Proteins are the main components of living cells, responsible for a variety of biochemical functions. They are involved in diverse chemical processes like signal transduction, metabolism and energy transfer, operated by binding and conversion of small molecules. Molecules generally consist of atoms from different types of elements connected through covalent bonds [3]. Proteins are large molecules composed of smaller subunits, so called amino acids. A set of twenty standard amino acids exists, consisting of a common part and differing in the side chain which is responsible for their properties. Amino acids are linearly connected to macromolecules by forming covalent[3] peptide bonds. During this process water molecules are released, and the result is an amino acid backbone consisting of a peptide bond and an amino acid residue responsible for a specific physicochemical property like lipophilicity or charge. The different combination and order of amino acids produces a variety of proteins with diverse properties and functions.

[1] Merck KGaA, Merck Serono, Global computational Chemistry, Darmstadt

[2] BioSolveIT GmbH, St.Augustin, supplies a lead identification software assisting in drug design

[3] A covalent bond is a form of chemical binding responsible for the solid coherence of atoms in chemical connections.

The sequence of a protein is encoded by a unique and finite linear composition of amino acids. Due to distance dependent attractive and repulsive forces between the atoms of the amino acids, the protein folds in most cases into a specific conformation. Nevertheless, the 3D structure of the protein is not rigid. Proteins are in constant motion in nature, influenced by ligand binding and environmental changes. Possibilities for protein structure elucidation are crystallization and NMR spectroscopy, whereby a snapshot of the structure can be gathered [4]. This snapshot holds information of the 3D position of each protein atom in the crystal, as well as possible ligand and water atoms, within a specified precision.

If a small molecule, called ligand, binds to a protein, these two binding partners form a complex and react with each other. The position of the protein where the reaction takes place is called active site or binding site[1]. The function of most proteins requires a reversible binding of the ligand. For instance, enzymes are proteins that function as biocatalysts by increasing the turnover rate of chemical reactions. In enzymatic reactions, the ligand, in this scenario called substrate, binds to the enzyme, is subsequently converted into one or more products, which are finally released. Two binding partners have to exhibit some complementarity to be engaged, encoded by their sterical and physicochemical properties. Sterical properties ensure the geometric fit of the two binding partners to each other. Physicochemical properties describe energetic features of atoms or functional groups, like hydrophobicity (water repulsion) and hydrophilicity (water attraction). Furthermore, the interaction between protein and ligand is specific and selective. The forces driving molecular recognition of a small molecule by a protein are not fully understood to date. Two prevalent theories are known for complex formation, the lock-and-key and the induced-fit principle. The lock-and-key principle [2] preexisted, in which the active site of a protein is available in a specific and relatively stable form. In this theory, the substrate has to fit like a key into a lock to bind and interact with the protein. In pursuit of capturing the flexible nature of proteins, the conformational selection concept [5] can be regarded as a continuation of this theory. In this concept, the protein preexists in several conformations, of which the ligand selects the appropriate one. Second, more recently the induced-fit theory [6] gained importance, where the active site of the protein is considered flexible and adapts to the substrate. This strategy explains the binding of ligands that do not properly complement the shape of the unbound protein. In fact, many different intermolecular interactions are responsible for complex formation [7]. One of the driving forces is the hydrophobic effect, which describes the tendency of hydrophobic atoms to

[1] While both terms describe a position on a protein where a small molecule can bind, the term active site more precisely specifies an enzyme's catalytic site.

aggregate in aqueous solution to exclude the energetically unfavorable interaction with water [8]. Besides these undirected hydrophobic interactions, directed interactions, e.g., hydrogen bonds are built between electropositive hydrogen atoms and electronegative atoms. Hydrogen bonds, salt bridges and van der Waals (vdW) interactions are forms of electrostatic interactions, resulting from the proximity of charged particles in an electrical field. VdW interactions are attractive or repulsive forces induced by two dipoles resulting from the distance between two atoms.

## 1.2 Computer-Aided Design and Analysis

Finding novel drugs addressing a disease is the ultimate goal in pharmaceutical research. Modern drug discovery, including experiments and computers, is executed on molecular level. Due to the complexity of the human body, the design problem is subject to high expenses. The estimated time needed for the drug development cycle to complete is 12 years and costs more than 1 billion U.S.$ [9]. Hence, computer-aided drug design (CADD) has become a central and very promising task in pharmaceutical industry. If the structure of the target[1] is known, structure-based otherwise ligand-based approaches are applied [10–12].

The early drug discovery process comprises several steps from target identification to preclinical development [13]. Selecting a target of interest incorporates knowledge about the disease relevance, structural aspects, screening feasibility, selectivity, toxicity, as well as commercial attractiveness. Rapidly and reliably identifying the most promising targets out of a large pool of available structures is a challenging task. In this context, the term druggability prediction has been coined, defined as the general ability of a disease-related protein to be modulated by low-molecular compounds.

Later in the drug development pipeline continues the identification of small molecules, termed hits, modulating the function of a target, and their transformation into leads[2]. Due to time requirements, computational methods are commonly introduced performing high-throughput screenings (HTS). Efficient techniques calculating and optimizing the binding mode[3] of a ligand and estimating binding affinities are investigated, e.g., docking [14, 15] and virtual screening [16, 17]. Furthermore, molecular modeling is used to simulate and study the dynamic behavior of the structure of molecules [18]. Areas of applications for molecular modeling are protein folding, protein stability, molecular recognition and conformational changes [19]. Such simulations provide insight into the

[1] The protein of interest to be addressed by a drug is named target.

[2] The hit identification follows the search for highly effective lead structures (hit-to-lead process).

[3] The binding mode is the orientation of the ligand relative to the protein.

structure-function relationship within proteins.

Furthermore, studying potential drug binding to multiple targets or vice versa is of high practical value. Polypharmacology, adverse effects and drug promiscuity are problems in drug design, which have already been addressed with computational methods [20–23].

As exemplified, virtual methods can help to expand the understanding of the forces that govern molecular recognition and ligand binding. A more efficient search for drugs that bind with high affinity and selectivity would be the positive consequence. While such procedures have been established in pharmaceutical industry for years, biotechnology realized the benefit more recently and introduced computational methods for rational enzyme design [24]. Biotechnology is divided in several sectors covering different fields of action. White biotechnology, which is investigated in the Biocatalysis2021 cluster, is devoted to biotechnological applications in industrial processes [25]. This includes the design and application of enzymes as industrial catalysts to produce useful chemicals. Enzymes in living cells from yeast, molds, bacteria and plants are investigated to synthesize products with optimized parameters, as high yields, easy degradability, less energy requirements and less waste production. Methods used in drug design can be applied to a wide range of biotechnological questions after adaption of parameters. *In silico* prediction of substrate specificity has already been introduced into biotechnology research [26]. By use of docking and virtual screening methods, substrates can be predicted for specific enzymes [27]. Biochemical profiling for large enzyme families has also successfully been employed [28]. Molecular dynamics simulations have become an inherent part in biotechnology, modeling substrate binding and enzyme catalysis [29].

Another challenge of computational biology is protein function prediction. Due to the fast progress in protein sequence and structure elucidation, the available data pool is growing rapidly. Since experimental annotation of function is limited due to time and cost expenses, computational assistance is required to put biological information to practical use [30]. Over 20 billion sequences are deposited in the Uniprot/TrEMBL database [31], and over 84 thousand structures can be accessed from the protein data bank (PDB) [32]. Following the hypothesis of divergent evolution, similarities between protein sequences or structures suggest related ancestors and thus, common function. Computer-based homology-driven protein function predictions can assist in making use of this immense data pool. Although many approaches exist, the failure rate remains high [33], mirroring the still not satisfying understanding of molecular forces. The propagation of erroneous function over databases can only be avoided by the initiation of reliable annotation [34].

The three mentioned scenarios - CADD, rational enzyme design as well as computational biology - imply the need for reliable prediction methods, particularly for binding site detection, analysis and recognition feature encoding [35].

## 1.3 Motivation and Structure of this Work

The computer has become an integral component in many areas of pharmaceutical, biotechnological and biological research. Although many approaches exist covering different parts of the protein assessment pipeline, there is still room for improvements.

A fundamental step for annotation or modification of protein function is a detailed description of the position where the reaction takes place and an understanding of its mode of action. The detection and analysis of specific parts of a protein where small molecules can potentially bind is of high practical interest. This also concerns the detection of sites other than the active site, i.e., allosteric sites. The challenge within this task especially lies in the manifold nature of proteins and binding sites, enhanced by their structural fluctuation. The more granular and precise the description of the binding site, the better is the derived information. Thus, the inspection of possibilities to subdivide this location into smaller pieces may gain further insight into the binding process. These aspects motivated me to further inspect protein binding sites and to develop a novel algorithm for precise binding site detection, presented in my first publication [D1].

The gathering of binding site information is very meaningful in a broad range of applications. Given the situation that several targets of interest participate in executing or blocking a specific disease related effect, ranking these targets by their potential to be inhibited by a small molecular compound is very valuable in drug design. Additionally, if an active substance has been designed, studying drug repurposing and drug promiscuity is of high practical interest [21, 22].

In the biotechnological context, computer tools can similarly be used as idea generator, e.g., predicting mutation sites by comparing the active site of the enzyme of interest with known enzymes producing the required product.

Furthermore, function annotation investigations benefit from a detailed computer-based analysis and comprehensive comparisons. As introduced in the last section, nowadays structures are solved before their function is known, exemplified by several thousands of unclassified structures deposited in the PDB. Although many approaches exist for functional annotation [36–38], the number of not (yet) or miss-classified structures is high [33, 34].

In pursuit of assisting these tasks, I developed novel strategies which can help in protein

classification, i.e., druggability prediction, function annotation and grouping proteins into families, published in [D1, D2, D3, D4].
According to the induced-fit theory, considering protein flexibility when predicting or modeling binding sites is indispensable. Thus, shifting computer methods towards incorporation of flexibility is still one of the main challenges for computational chemistry. Inventing descriptors accounting for a certain degree of structural changes is very promising. The usage of such triangle-based local descriptors for protein comparison and function annotation is addressed in my fifth publication [D5].

During my work, I was able to develop strategies for computer-aided protein analysis exhibiting a novelty value and a performance worth being published in scientific peer-reviewed journals. These publications are part of this cumulative dissertation. The single papers describe subsequent interlocking milestones of my work, resulting in a toolkit for structure-based active site analysis. The workflow showing the individual steps and their association is depicted in Figure 1.1.

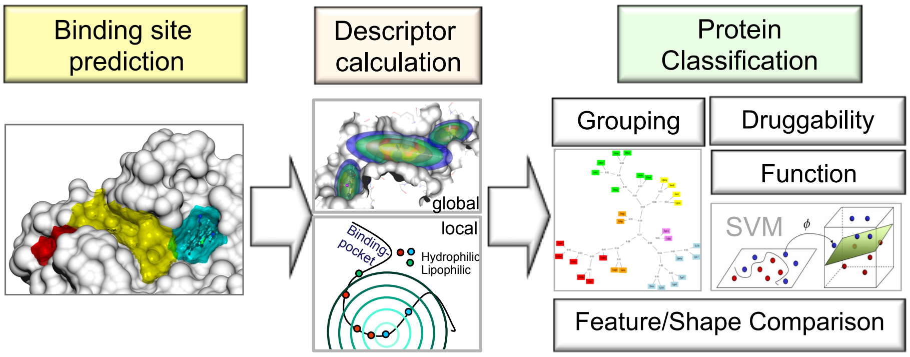


**Figure 1.1:** Outline of the process of my work. Starting from the protein structure, binding sites are identified first. Second, these pockets are described by sterical and physicochemical features, which are finally used for protein classification.

My first paper introduces the novel pocket detection algorithm, DoGSite [D1], a geometric approach using a Difference of Gaussian (DoG) filter. Next, global and local descriptors are calculated from this representation and used for protein classification. In my second paper, the profit of a new method, named DoGSiteScorer, using a support vector machine and a nearest neighbor search for classification, is illustrated in the context of protein druggability prediction [D2]. Furthermore, the complete DoGSiteScorer functionality was ported to a web server. The usability and the benefit of the server are outlined in my third paper [D3]. My subsequent publication exemplifies the application of the classification method in enzyme function prediction [D4], with the enhancement of a stepwise annotation with increasing specificity.

Due to the global character of the developed method, it is prone to small changes in the protein structure. Thus, an enhanced strategy incorporating local triangle descriptors is introduced for binding site comparison. The method and its results are published in my fifth publication contributing to this work [D5].

Finally, in pursuit of a better understanding of driving forces in molecular recognition, the impact of shape complementarity between protein pocket and ligand in the binding process is discussed in my last publication [D6].

The contributions of all authors to these publications are listed in Appendix A.1.

Since the papers describe approaches and applications that build up on each other, my thesis is structured as follows: First, a state of the art of the main addressed topics is provided. Thereafter, the methods section describes the complete strategy from active site detection based on the protein structure to classifications as outlined in Figure 1.1. The diverse evaluations and applications are summarized in the results section. Finally, this thesis finishes with a conclusion of my work and an outlook to possible expansions of the strategies.

# 2

# State of the Art

For a long period of time it has been known that the active site of a protein is the key to its function. Thus, protein binding site prediction and classification are demanding tasks in computational chemistry and biology and have been studied by a multitude of scientific groups. Since the active site is the basis for the diverse classification steps, its detection is addressed first, followed by the introduction of methods for classification scenarios, i.e., protein druggability annotation and binding site comparison with focus on function prediction.

## 2.1 Active Site Determination

The detection and the description of the active site of a protein is the initial step in structure-based drug discovery and rational enzyme design. An excerpt of the many strategies developed for active site prediction will be discussed in the following. The focus lies on approaches belonging to the same category as the method introduced in this work. A detailed list of known approaches can be found in my first publication [D1].

Published approaches for active site detection can mainly be divided in sequence- and structure-based methods. Sequence- or evolutionary-based methods mostly rely on multiple sequence alignments [39] or mapping of phylogenetic information onto the protein surface [40]. Sequences are compared with respect to conservation of residues following the hypothesis that conserved residues encode the function of a protein. In this context, active site profiles as well as homology modeling are used.

As mentioned in the introduction, ligand binding depends on its sterical and physicochemical complementarity to the protein's binding site. Thus, the protein has to exhibit a binding groove with a volume able to assimilate a ligand. Furthermore, the

3D arrangement of the present amino acids provide specific ligand interaction points. Both constraints are not captured in sequence-based methods. Therefore, the work in hand focuses on structure-based approaches, and sequence-based methods will not be further discussed here.

A common strategy within all structure-based approaches is the identification of interesting points on the protein surface, followed by a clustering step assigning these points to pockets. The difference lies in the annotation of the points of interest, which can be divided into the consideration of geometry- and energy-based properties. Geometry-based methods [41–51], which can further be divided into grid- and sphere-based approaches, analyze the shape of the molecular surface to locate cavities solely based on the atomic coordinates of the protein.

Most geometric grid-based methods consider the buriedness of grid points with respect to the protein surface [42, 45, 50]. LIGSITE [45], e.g., maps the protein on a 3D grid with 1.0Å grid spacing. Grid points are subsequently labeled as free or occupied, dependent on their coverage by a protein atom. A cube is placed on each free grid point, and seven lines are spanned through the cube center, traversing its six faces and the eight corners. For each grid point, so called protein-solvent-protein events are counted based on the number of lines enclosed on both sides by protein atoms. The higher this value, the more buried is a grid point. Finally, buried grid points are merged and represent potential cavities on the protein surface. PocketPicker [50] uses a continuative buriedness approach for cavity detection, based on 30 scan rays obtained from the triangulation of an octahedron.

The second subgroup of geometric approaches incorporates spheres. SURFNET [43] and PASS [47] detect pockets by covering the protein surface with spheres. SURFNET, e.g., fits gap spheres between atom pairs and reduces their radii until they are free of clashes with any protein atom. Using Voronoi diagrams (CAST [46]), Delaunay triangulations (APROPOS [44]) or alpha shapes (Fpocket [51]) is another common approach for pocket detection. In Fpocket, the vertices of a Voronoi decomposition of the protein surface are used to place alpha spheres. Subsequently, spheres are pruned based on a size criterion. Solvent inaccessible spheres are discarded by the maximum size cut-off; exposed spheres are subject to the minimum size criterion. The retained spheres describe clefts and cavities of the protein.

Energy-based methods consider the interaction energy of a probe[1] with the protein [48, 52, 53] or use blind and fragment docking for cavity finding [54, 55]. In both forms, regions with favorable energetic responses are annotated as potential pockets. DrugSite

[1] A probe can be an atom or functional group used to sample potential interactions with the environment.

[53] relies on a grid representation of the protein. To calculate the energy potential, a carbon probe is placed on each free grid point and the van der Waals (vdW) energy with the surrounding protein atoms is computed. Next, grid points with unfavorable energy values are truncated and a filter is applied to the grid. Finally, grid points with energy values satisfying a specified cut-off are kept and merged to pockets.

Recently, fused approaches have been introduced to the binding site detection sector, combining several strategies to enhance the prediction power [56–58]. Capra et al. [56] combine evolutionary, sequence and structural information for site prediction. Focusing on structure, SiteMap [57] joins geometric and energy-based information on a grid. Enclosure of a grid point is calculated by a buriedness criterion similar to the one of PocketPicker, but with respect to a higher number of scan directions spanned by 110 rays. To capture grid points able to favorably interact with the protein, the vdW energy, similar to the DrugSite approach, is calculated. Finally, grid points fulfilling both requirements are clustered into groups representing potential binding sites.

Although the recovery rates in retrospective binding site detection studies for the vast amount of methods are high, disadvantages exist for the individual modeling strategies. Methods using a grid representation depend on the protein position and orientation as well as the grid spacing. When using rays for buriedness annotation, methods relating to a higher number of scan directions are less dependent on the orientation of the protein in the grid. Binding site detection methods incorporating spheres encounter problems with wide and open cavities. Energy-based methods depend on the quality of the underlying scoring function in capturing the possible interactions. Further uncertainties lie in the parametrization of the used force fields, the filter procedure or the introduced cut-offs. Another challenge arises from the variety of cavity space. Binding pockets are known from being small to large, forming deep grooves or shallow invaginations, from buried to open ones, spanning over several channels or subpockets. Thus, automatically detecting a perfect pocket for all these shapes is difficult. Additionally, most publications agree in the fact that the annotation of the boundary, especially for open cavities, highly depends on the used algorithm and the question of the true boundary definition can rarely be answered with certainty. Furthermore, small changes in the protein structure may yield large changes in the detected pocket, thus protein flexibility adds to the complexity of the problem.

## 2.2 Protein Classification

This section covers the state of the art of three protein classification scenarios, namely druggability prediction, function annotation, and binding site comparison.

### 2.2.1 Druggability Prediction

The established drug development pipeline suffers from high time and cost expenses [9]. Thus, computer methods assisting and accelerating this process are of high practical interest. One of the multiple parameters involved in target assessment is the *a priori* prediction of target druggability. A prerequisite for investigations in a target is its general ability to be modulated by low molecular weight compounds. This potential to interact with small molecules is differently termed in literature, namely druggability, targetability, ligandability or chemical tractability [59–61]. The terms agree in the fact that the target must be able to bind a molecule, but they disagree in the properties of this molecule, categorized into small, drug-like, binding with high affinity, and being orally bioavailable. Even druggability itself strikes different subclassifications [62–65]. The definition used in the context of this work is restricted to the regulation of a disease-modifying target by orally bioavailable compounds [62].

For over 15 years already, druggability prediction has been actively researched. A variety of methods exist covering experimental as well as computational approaches. Experimental druggability assessment methods like NMR-based screens are commonly and successfully used in pharmaceutical research [61, 64, 66, 67]. The annotation in these experiments is based on the correlation between NMR hit rates and success rates in hit to lead programs. Since computational models require fewer resources, they are used to analyze the nature of ligand binding sites. Similar to experimental NMR-based screening, *in silico* screening methods, calculating success rates of drug-like ligands virtually docked into the pocket, are used to detect the most promising target candidates [60]. Another common proceeding is the identification of specific properties potentially responsible for druggability and the incorporation of these properties into clustering, regression or machine learning techniques.

Due to the availability of sequence prior to structural data, several methods exist, deriving information from sequence [68–70]. Nevertheless, as stated in a druggability review in 2008, sequence-based methods do not exceed accuracies above 68% [9], encouraging the use of structure-based methods [49, 50, 57, 59, 64, 71, 72].

The initial step in structure-based approaches is the specification of the active site, which can be externally precompiled or calculated internally. Next, relevant structural, geometrical and physicochemical features are identified from known protein-ligand complexes, and a scoring function is employed to rate the druggability of a target. The number of features used in the analysis ranges from a couple up to several hundreds [49, 64]. MAP$_{POD}$ [71] studies the binding energy using a structure-based maximal affinity model. The description is reduced to important discrete energy terms and combined with drug-like properties describing oral bioavailability. PocketPicker [50] represents a pocket by a 210-dimensional shape descriptor encoding size and buriedness and classifies a protein by aid of self-organizing maps. A simple but high-performance method is SiteMap [57], rating druggability by a linear combination of three single descriptors. These numerical descriptors include the number of site points encoding pocket volume, the hydrophobicity and the shape of the pocket. A similar method, predicting bindability for all proteins of the PDB is DLID (drug-like density) [59]. The likelihood of a pocket to bind a drug-like ligand is estimated based on the number of pockets binding a drug-like ligand in the local neighborhood. This neighborhood in pocket space is described as a linear combination of the properties volume, buriedness, and hydrophobicity; used parameters were obtained by linear regression. Druggability prediction with Fpocket [72] incorporates physicochemical features of the pocket, normalized by size. These features, comprising local hydrophobicity density (combining size and spatial distribution of hydrophobic agglomerations), general hydrophobicity and normalized polarity, are combined in an exponential function. By aid of a bootstrapping method, those parameters yielding the highest accuracy on the provided test set were chosen. DrugPred [73] uses a partial least-squares projection to derive discriminant features yielding a linear model based on five descriptors. Finally, Perola et al. [74] recently published a rule-based approach providing an intuitive representation of the preferred property space of druggable pockets.

The findings from the approaches based on descriptors introduced above are in good agreement and are summarized in the following. No linear dependency of druggability to one single feature could be observed, thus several combined features are needed for a description. They highly agree that druggable pockets tend to be larger, more complex and exhibit a more hydrophobic character.

Nevertheless, many features are not sufficiently addressed by these descriptors, such as metal or ionic interactions and cofactors[1]. While most of the aforementioned approaches use features describing the complete binding site, more recently published methods focus on local properties. In the druggability study of Schmidtke et al. [72] the effect of local environmental changes in accessible surface area is investigated.

Another difficulty is the noise arising from uncertainties that result from the pocket prediction and definition step, no matter if ligand-based or automatically assigned. Wrong or unspecific pocket or boundary definitions clearly lead to misinterpretations in druggability assignment. Further pitfalls emerge from the available data, in terms of annotation and size. The ambiguous definition of the term druggability makes a clear

[1] Cofactors are helper-molecules that are bound to the protein and are required for its biological activity.

assignment very difficult, and the usage of different data sources questionable. Misleading data strongly affects the prediction power of automatic methods, bearing the risk of learning from wrong examples. Collecting positive data points can be done easily by selecting targets with known marketed orally bioavailable drugs. Contrarily, the assembly of a negative set turns out to be difficult. A pocket should not be annotated as undruggable, solely because no drug is known. Further pharmaceutical investigation or serendipity may uncover drugs to yet undruggable targets. Approaches focusing on bindability are confronted with the same difficulty. Considering ligand free pockets (in the available crystal structure) as not bindable is equally incorrect. Especially high-throughput methods, trained on such data sets, suffer from wrong or miss-annotated data. Despite these inconveniences, several druggable data sets exist. While most data sets were constituted of less than a hundred structures [64, 71], Schmidtke et al. [72] released an elaborated set of over a thousand structures. Recently, Krasowski et al. [73] added a new non-redundant set of druggable and less druggable binding sites, consisting of 115 structures. Such data sets are good starting points for model calibration and evaluation.

### 2.2.2 Function Annotation

With respect to the rising number of solved protein sequences and structures, the need for computational tools for automatic protein function annotation has been growing. A large number of reviews exist, focusing on different parts of the function prediction process [24, 36–38, 75–84], a detailed overview can also be found in my publication concerning this topic [D4]. Similar to approaches for active site prediction, function prediction methods split into sequence- and structure-based groups. Sequence- and evolutionary-based methods will only briefly be covered, followed by the motivation for structure-based approaches which directs the focus of this chapter to structure-based methods.

Over the last decades, sequence-based methods have dominated the *in silico* function annotation field due to the vast amount of available sequence data. Classically, global similarities are employed to transfer information from proteins with well-established function to unknown ones. For this purpose, proteins are compared by aid of multiple sequence alignments and function is inferred with respect to the closest homologue. Methods performing a comparison of the complete sequence are, e.g., BLAST [85] and PFAM [86]. In contrast, BLOCKS [87], PRINTS [88] and PROSITE [89] search locally for function related sequence motifs. Incorporating evolutionary information, e.g., from gene expression data, genomic context, gene ontology, phylogeny and coevolution, has also proven useful for function prediction [82, 90]. Examples for tools processing such information are GoFigure [91], Phydbac [92], SIFTER [93] or FlowerPower [94].

Due to the large and still growing amount of solved protein structures, automatic tools for structure-based protein function predictions have been pushed ahead over the last years. For example, large structural genomic projects reveal new protein structures of which no knowledge about their function is known beforehand. Furthermore, protein structure was found to be more conserved than sequence [95], which motivated the shift towards structure-based protein comparisons. Proteins with low sequence identity can still share functionality through protein fold similarity. Prominent fold comparison tools are SCOP [96], CATH [97] and FSSP/Dali [98]. Similar to sequence comparison, the hypothesis of homology-based information transfer is pursued in structure-based approaches. Two structures are superposed by means of calculated structural alignments; higher compliance suggests higher functional relation. Structural alignments can be built based on the complete structure or on structural fragments, which can be recombined to a complete alignment. Methods based on structural alignments are FATCAT [99], PAST [100], VAST [101], and 3DCOMB [102].

Since the function-specifying reaction takes place in the active site of a protein, structure-based methods put a stronger emphasis on active site comparison decoding potential local similarities between distantly related structures. The presence of specific interaction partners within conserved distances in the active site determines the function of a protein. According to the hypothesis that proteins with similar binding sites share function, all methods performing binding site comparison are capable of predicting function by extracting the most similar sites from large data sets. Since binding site comparison is separately analyzed in this work, this topic is further devised in the subsequent subsection.

One example directly developed for function annotation is the work by Parasuram et al. [34]. Functional sites are described by electrostatic potentials predicted with theoretical microscopic titration curves (THEMATICS). With the aid of a machine learning technique (implemented in POOL) functionally important residues are detected. Thus, uncharacterized proteins can be structurally aligned, and the match with proteins of known biochemical function is used for annotation.

Most of the approaches mentioned so far rely on some kind of alignment or superposition of the protein or the binding site. A different and less frequently used approach is the comparison based on descriptors derived from the protein structure. The idea of descriptor-based approaches is to extract properties of a protein that are specific for one protein class or family. Thus, by observation of these features one can assign the function of a new protein based on its belonging to a specific class. Kontoyanni and Rosnick

[103] used structural, thermodynamic, and geometric attributes of the active site for function annotation. Since using such descriptors depicts a typical classification problem, standard machine learning methods can be used for function annotation [104–107]. Dobson and Doig published an attempt to separate enzymes from non-enzymes using a support vector machine (SVM) [104]. In a subsequent publication, they tried to distinguish on a more granular level, i.e., between members of different enzyme classes [105]. In this study, they calculated 55 attributes representing the complete enzyme structure by crystal structure information, secondary structure content, amino acid composition, surface fractions and bound ligands. Since the enzyme classes are of different size, they chose to use 15 pairwise SVM submodels over one multiple SVM model. Accuracies of 35% for top-ranked predictions and 60% for finding the correct class under the top two ranks have been achieved on a set of 220 non-homologous structures.

Strategies pursuing the hypothesis that proteins with similar biochemical function bind analog ligands are exemplified in the following. In the first approach, ligands from various proteins are compared to the ligand of the protein of interest [108]. The authors calculated the similarities between the bound ligand of a query enzyme and small molecules from BRENDA [109] to assign a function. The similarity in chemical space between compounds is rated by 117 topological descriptors. Another approach, which is based on the similar ligand strategy and independent of structural superposition, uses predicted binding affinities. As early as 1995, Kauvar et al. introduced affinity fingerprints for function prediction [110]. This fingerprint represents the potency for binding of a compound against a small reference set of diverse proteins. Comparison of these fingerprints can detect similarities between known and not yet annotated enzymes.

Furthermore, molecular docking is inserted for function prediction. Docking is in general a selection and optimization process, trying to find the best fit of a molecule in the binding site of a protein. For this purpose, the conformational space of a compound is sampled, each conformation is placed into the binding site and ranked according to a scoring function. Based on the estimated binding affinity, potential substrate classes can be identified that preferably bind to the active site and thus infer the function of a protein. By using structure-based docking of high-energy forms of potential substrates, Hermann et al. [111] successfully annotated an adenosine deaminase function to an unknown protein.

Similarly to other areas described in this chapter, the synergistic use of evolutionary, sequence- and structure-based information for protein function prediction has been launched [80, 112–118]. Pierri et al. [80], e.g., provided a combined protocol comprising multiple sequence alignments, binding site prediction, comparative modeling and virtual screening for function annotation. The Enzyme Function Initiative [118] established a multidisciplinary approach. The *in silico* prediction of substrate specificity is in the focus of this strategy including bioinformatics, experimental structural biology, structural modeling, docking, experimental enzymology, microbiology and metabolomics.

Success stories can be told for the vast amount of computational methods and their different lines of action. Nevertheless, still a large amount of structures lack functional annotation or suffer from miss-annotations [33, 34]. One drawback is the contextual definition of biochemical function [24, 119, 120]. Inconsistencies and errors are a consequence of the usage of different functional annotation systems and databases.

Furthermore, the required amount of sequence or structure similarity for high-confidence function transfer is questionable [121]. Clearly, the higher this value the more likely is a shared function. Nevertheless, shared enzymatic functions were occasionally found for non-homologous proteins [90] or vice versa [122]. Another downside of some of the introduced methods is that some ligand binding or class information has to be known beforehand. This concerns approaches using the bound ligand for comparisons [108], the *a priori* substrate selection in docking approaches [111] or the vague knowledge about the functional class for protein selection [34]. Further problems as mentioned in the previous section arise from flexibility of the proteins upon ligand binding.

### 2.2.3 Binding Site Comparison

As annotated in the last section, methods for binding site comparison can generally be applied for function prediction. As summarized in recent reviews [24, 123], site comparison methods consist of three steps: encoding of molecular recognition features, searching for analogies and quantifying these similarities. The recognition feature encoding step serves for reducing the complexity of the comparison problem. Simplified representations are, e.g., 3D coordinates of functional groups, residues or pseudo-centers and characteristic features of the binding site. The number of developed approaches is manifold, the pursued strategies can be divided into alignment-based and alignment-free methods. Alignment-based methods use the encoded features to superimpose the structures of interest. Strategies such as geometric matching (SuMo [124], SiteBase [125], ProSurfer [126]), geometric hashing (TESS [88], SiteEngine [127]) and clique detection (CSC [128], Cavbase [129, 130], eFsite [131], eF-seek [132], IsoCleft [133]) are incorporated for structural alignments. Structural templates [88] or triangles of physicochemical properties [127] are used to uncover analogies between active sites in geometric hashing procedures. Clique detection methods are based on a graph built on the structure [128], on conserved features [129] or molecular surfaces [131]. In Cavbase [129], chemical properties of the binding site are represented by pseudo-centers, calculated by mapping the site residues to a grid. A clique detection algorithm is used to identify shared 3D pseudo-center clusters in two cavities, which can be used to align these cavities.

Since finding a superposition is computationally expensive, fingerprint methods have been introduced avoiding such alignments. In this context, several methods explore specific distances and their distribution in the active site. Some methods use atomic distances [120, 134–137], others separation between fragment pairs [138] or property-encoded shape distributions [139]. For instance, feature vectors describing inter-residue distance patterns are generated by the aid of a cut-off scanning matrix [120] and used for efficient automatic function annotation. Another prominent approach is the comparison of pharmacophoric fingerprints, e.g., FLAP [140], SiteAlign [141] and FuzCav [142]. In SiteAlign, properties are projected onto a sphere triangulated into 80 equal parts. This method can be seen as a meta-approach, since the mapping on the sphere allows for easy alignment. The subsequent development FuzCav is completely alignment-free comparing fingerprints of pharmacophoric triplets with high-throughput. Moment-based pocket representations are rotational invariant and allow for fast comparisons. The protein surface is described by 3D mathematical functions using spherical-harmonics [143] or 3D Zernike descriptors [84]. PocketSurfer and PatchSurfer [84] predict ligand molecules that bind to the query by comparing geometrical shape and physicochemical pocket properties to a database of known binding pockets, allowing a quick real-time scan without pre-alignment. PatchSurfer convinces by means of its local comparison of surface patches, which represent features of local pocket regions accounting for protein flexibility.

Another group of approaches repurposes methods from other research fields for information retrieval, like image or word processing. These methods incorporate spin-images for surface matching [144], visual words descriptors [145] or simple bit strings [146].

Due to the speed of fingerprint-based methods, high-throughput comparisons of up to a million bindings sites are practicable. Nevertheless, this speed entails an absence of interpretability. The outcome of truly fingerprint-based methods is a single number, describing the similarity between two sites. Besides this number, little is known about the features responsible for this similarity. As a consequence recently, much effort has been put into accelerating alignment-based methods, which allow for efficient calculation and easy interpretation of the results. BSAlign [147] is one example for such a combined approach based on clique detection. The recognition features are reduced to residues - together with geometric and physicochemical information - instead of point-based representations. This sparse graph representation, together with the development of an efficient algorithm to circumvent the NP-hard problem of finding the maximum subgraph, allows for high-throughput comparison. Desaphy et al. [148] published another novel bilateral approach, compassing this disadvantage. The new binding site description VolSite is combined with the novel alignment and comparison tool Shaper. Negative images of the binding sites, encoding shape and pharmacophoric properties at regular spaced grid points, are compared. The molecular shape is approximated by smooth Gaussian functions. Thus, the alignment is based on the optimal volume overlap.

The methods mentioned in this chapter show good performances on a variety of different data sets. These data sets vary in their size from a few up to a million sites and are focused on different scenarios as grouping enzymes, classifying proteins into subfamilies or studying polypharmacology. The runtimes of the different approaches are highly divers. A comparison study of various alignment-based, fingerprint and meta-approaches revealed pairwise comparisons in the speed order of several minutes down to a few milliseconds [142]. Concluding, the quality, the runtime and the results of such methods provide valuable information for various application scenarios.

# 3

# Methods

This chapter introduces the algorithms I developed for structure-based analysis of protein binding sites. The individual steps, starting from the detection of the active site, to the derivation of the global binding site descriptors, and their usage for classification, are explained. Finally, the approach based on local triangle descriptors for binding site comparison is introduced.

## 3.1 Protein Pocket Detection

The developed structure-based method DoGSite falls into the category of geometry-based methods relying on a grid. A detailed description of the algorithm is provided in my scientific publication [D1] and is summarized in the following[1].

First, the protein is embedded in a grid, and grid points are labeled dependent on their overlap with the vdW radius of any protein atom. Occupied grid points are set to one, free grid points to zero. In contrast to previous algorithms which mostly rely on buriedness calculation, a strategy for edge detection is borrowed from the image processing field. Edges can be uncovered utilizing a 3D Difference of Gaussian (DoG) filter [149]. For this purpose, the grid is first smoothed by convolution with a Gaussian kernel with radius $\sigma_1$. A second convolution is obtained by blurring the grid with a different radius $\sigma_2$. Subtracting one image from the other preserves positions with drastic shifts, but discards all points that are at continuous areas. Due to the naturally concave character of cavities, pocket grid points have a more drastic response. Next, a specific cut-off for the calculated density values is used to discard non-informative points.

A further innovation of DoGSite is the preservation of subpockets. Clustering the surviving grid points yields small agglomerations, called subpockets. Subsequently,

[1] The original implementation of the DoG filter stems from A. Griewel.

neighboring subpockets are merged to pockets. Both representations can be used depending on the problem to be analyzed. The complete process is depicted in Figure 3.1.

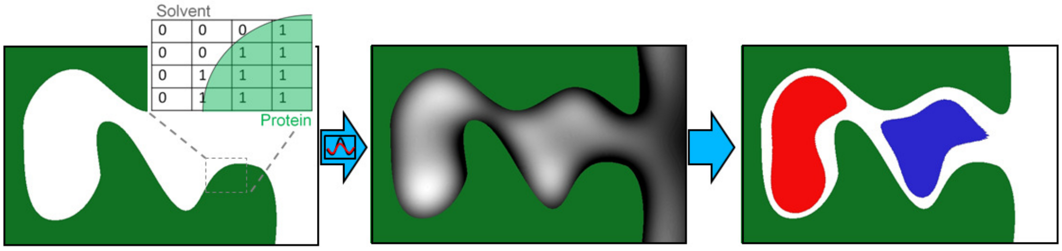


**Figure 3.1:** Simplified representation of the DoGSite pocket detection algorithm. The figure is taken from my publication [D1].

For comparison reason, a geometry- and an energy-based algorithm, LIGSITE [45] and DrugSite [53], have been reimplemented. A rough outline of their functionality has been provided in chapter 2.1, a detailed description can be found in the DoGSite publication [D1].

When evaluating pockets, the criterion defining a correct prediction is ambiguous. In this work, a new criterion is introduced considering the overlap between pocket and bound ligand. Pocket coverage is calculated as the percentage of pocket grid points covered by the VdW radii of the atoms of the ligand. Similarly, the portion of ligand atoms occupied by pocket grid points describes ligand coverage. The better the ligand fits into the pocket and the better it fills the pocket, the higher is the respective coverage and the more precise is the pocket description.

## 3.2 Pocket Description

The method I developed for protein assessment uses global as well as local descriptors of predicted pockets to capture the properties of a protein. A detailed description can be found in my second publication [D2], which is part of this cumulative dissertation; a short overview will be given in this section. Additionally, the shape descriptor for direct ligand and pocket shape comparison is introduced [D6].

### 3.2.1 Global Descriptors

Comparable to other descriptor-based methods, a comprehensive set of properties related to the structure of the pocket is investigated. The global descriptors describe properties of the complete pocket and are all of numerical character. Starting from the grid representation of the pocket, simple properties can be derived by either iterating over the grid points belonging to the pocket, or over the amino acids lining the pocket. Since the molecular recognition depends on shape as well as physicochemical complementarity of two binding partners, these properties are captured in this approach. Features describing the size and the shape of a pocket include volume, surface, depth, and ellipsoids fitted into the pocket (Figure 3.2). For this purpose, pocket grid points are separated with respect to their position in the predicted pocket. A differentiation is introduced between points without contact to a non-pocket grid point (inside), those with liaison to a protein atom (surface) and those in touch with the solvent (solvent exposed).

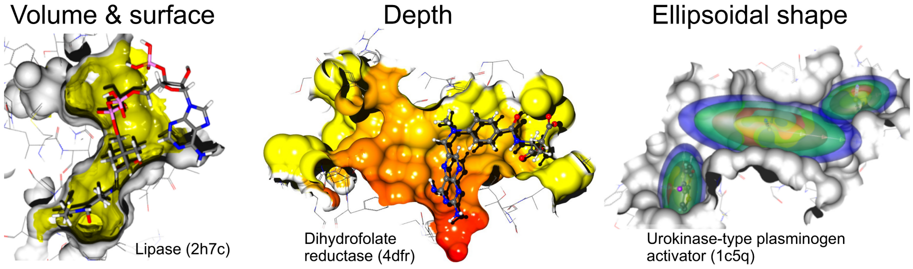

Volume & surface
Depth
Ellipsoidal shape
Lipase (2h7c)
Dihydrofolate reductase (4dfr)
Urokinase-type plasminogen activator (1c5q)

**Figure 3.2:** Extraction of calculated global descriptors. The figure is taken from my publication [D2].

The discrete volume and size of a pocket are calculated by multiplying the number of pocket grid points with the cubed or squared grid spacing, respectively. Depth of a pocket is encoded as the maximum distance between any solvent exposed grid point and an inner grid point, labeled as inside, calculated by means of a breath-first-search. If a pocket is completely buried, its maximal diameter equals its depth.

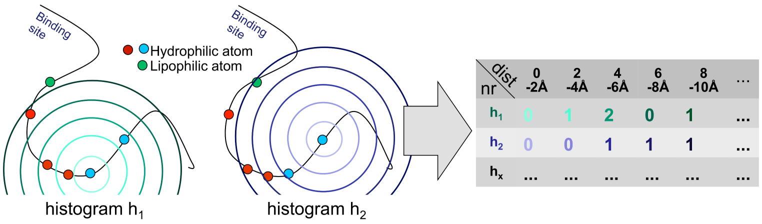
Since shape complementarity is important for molecular recognition, the form of a pocket is elaborated by the main axis of an ellipsoid fitted into the active site. This ellipsoid is derived from the eigenvalues and eigenvectors of the diagonalized covariance matrix over all pocket grid points. To further describe the complexity of a pocket, ratios between several pocket descriptors are calculated. The ratio of surface to volume grid points and the relation between pocket and ellipsoid volume estimate the roughness of a pocket. Furthermore, as a measure of the buriedness of a pocket, the grid points describing the lid of the pocket are related to the grid points describing the hull of the pocket.

Equally important for ligand binding are the physicochemical properties present in the pocket. To capture anchor points and the chemical environment of a pocket, the site lining atoms are investigated. Enumerable properties, like the count of particular

elements presented by these atoms and the number of specific amino acids to which they belong, are calculated. Additionally, the amino acids are grouped into positive, negative, polar and apolar residues. Since metal interactions result in tight bindings, the number of metals present in the pocket is added to the set of descriptors. Counting the number of present donor and acceptor anchor points[1] captures hydrophilic interactions in the pocket. Furthermore, site interaction centers (SIACs), originating from the FlexX interaction model [14], are utilized to describe the interaction profile of a pocket with a potential ligand.
To derive properties specific for proteins but independent of size, most of these descriptors are used in their normalized form. E.g., the fraction of lipophilic centers compared to all centers present in the pocket is used to express the lipophilic character of the pocket; similarly the lipophilic solvent accessible surface fraction is calculated.

### 3.2.2 Local Descriptors

Local properties better suite the ligand profile and can reveal similarities between distantly related proteins not sharing overall structural homologies. Furthermore, specific interaction partners present in the active site within conserved distances dictate the function of a protein. To analyze the local environment of the anchor points present in a pocket, distances between these points are investigated. For this purpose, a distance histogram is calculated for each functional group in the binding site. As potential anchor point each hydrophilic and hydrophobic pocket atom is investigated. A radial search is performed, e.g., starting from a hydrophilic interaction (Figure 3.3).



| nr \ dist | 0 -2Å | 2 -4Å | 4 -6Å | 6 -8Å | 8 -10Å | ... |
|---|---|---|---|---|---|---|
| $h_1$ | 0 | 1 | 2 | 0 | 1 | ... |
| $h_2$ | 0 | 0 | 1 | 1 | 1 | ... |
| $h_x$ | ... | ... | ... | ... | ... | ... |

**Figure 3.3:** Schematic view of the calculation of the distance-dependent histograms. The figure is taken from my publication [D2].

Next, interactions found in subsequent 2Å radii shells are counted. Based on this binning, a histogram arises for each functional atom of a pocket. The first nine bins

[1] Hydrogen bonds can be formed between atoms of different electronegative character. In this case, a hydrogen attached to one atom (donor) is shared with another atom (acceptor).

represent the distances between 0Å and 18Å, the last bin contains all pairs above this distance. The binning effect assists in tolerating small changes within the binding site residues.

### 3.2.3 Descriptor for Pairwise Shape Comparison

Since shape complementarity is an important recognition feature, the descriptor set is extended to allow for a direct protein pocket and ligand shape comparison, which is investigated in my last publication [D6]. Principal moments of inertia (PMIs) approximating the shape of an object are the basis for the comparisons. For pockets, PMIs [150] are derived from the pocket grid points endued with a weight of one. Ligands are processed similarly. For both objects, the moment of inertia tensor is calculated and the matrix is diagonalized. The resulting eigenvalues and eigenvectors are used for the subsequent shape comparisons. The PMIs are further converted into normalized principle moments of inertia ratios (NPRs) [151]. Therefore, the eigenvalues are sorted in ascending order ($I_1 < I_2 < I_3$) and the lower values are divided by the higher one ($npr_1 = I_1/I_3, npr_2 = I_2/I_3$). NPRs do not require a molecular superposition, are independent of the size of an object and describe a finite triangular space. The corners of this triangle are occupied by spherical, discoid, and elongated shapes (Figure 3.4).

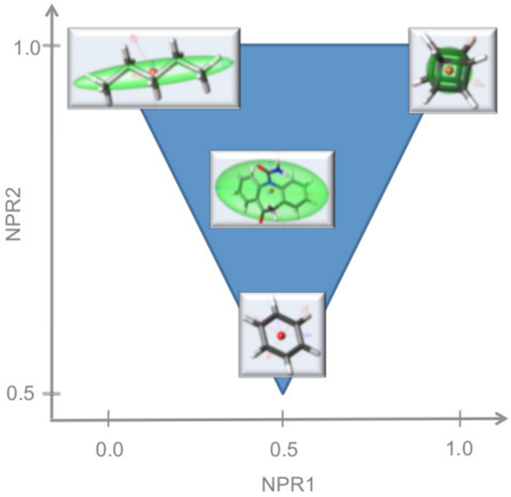


**Figure 3.4:** Overview of the triangular NPR shape space with exemplarily chosen ligand shapes. Fitted ellipsoids are depicted in green color.

By means of this descriptor, pockets and ligands can be positioned in the spanned NPR triangle, exposing information about their shapes. Furthermore, the direct distance in

NPR space can be calculated by the Euclidean distance between the pairs and used as descriptor for shape complementarity between the two binding partners.

## 3.3 Grouping and Classification

Protein binding sites can be described through a variety of properties, as introduced in section 3.2. Nevertheless, annotating the importance of the individual features for ligand binding is a challenging task. It is difficult to capture the variety of binding sites and the complexity of ligand binding by simple descriptors. Therefore, diverse approaches for protein clustering and classification are introduced in the following sections.

### 3.3.1 Hierarchical Clustering

Clustering is applied in computational chemistry with the prerequisite of grouping objects by aid of a specific measure of similarity [152]. A first approach for functional annotation during my work was, therefore, the adoption of a hierarchical clustering algorithm [153]. Clustering procedures group a set of objects in a way that objects within one cluster are more similar to each other than objects of the other clusters. The FlexX software is equipped with two hierarchical clustering procedures [154], differing in the linkage criterion used for cluster annotation. Starting from putting all objects into separate clusters, they are consecutively merged based on the closest pair of clusters. This pair is determined by the minimal or maximal distance between two clusters in single and complete linkage approaches, respectively.
To describe the relation between two active sites, the distance between a subset of global descriptors is used as metric. The similarity between two features is calculated by the Tanimoto coefficient [155], which has been widely used in computational chemistry for measuring structural similarity between molecular objects [156]. The non-binary Tanimoto similarity between two pockets $A, B$, where $d_{iA}$ denotes the value of the $i$-th of the $n$ used descriptors for pocket $A$, is calculated as follows:

$$S_{A,B} = \frac{1}{n} * \sum_{i=1}^{i=n} \frac{(d_{iA}d_{iB})}{(d_{iA})^2 + (d_{iB})^2 - (d_{iA}d_{iB})}$$

The used descriptors are normalized and include volume, surface, lipophilic surface fraction, depth, ellipsoidal main axes, SIAC ratios, hydrophilic and lipophilic pair histograms, acidic and basic amino acid ratios and shape distance. The used distance for the clustering procedure is calculated by subtracting the normalized similarity score

$S_{A,B}$ from one.
For protein clustering, thus, a set of descriptors for a number of proteins of interest can be entered. Then, the distance matrix is calculated, the objects are clustered and the algorithm returns the resulting cluster tree. A dendrogram holding the produced cluster information is used to illustrate the arrangement of objects. This allows for easy graphical inspection of the relationship between the investigated proteins.

### 3.3.2 Machine Learning Technique

The method described here was first introduced in the DoGSiteScorer publication [D2] and has further been investigated in a function prediction scenario [D4]. The first paper considered the two-class separation of druggable from undruggable targets. The second function related study implied a separation into multiple classes, e.g., the differentiation into the six enzyme classes. Training and test data sets were separately collected for both experiments.
The global descriptors derived for a predicted pocket, introduced in section 3.2.1, form the basic input for the classification approach. Since a linear separation of proteins into groups or families based on these properties is often not successful, a machine learning technique is used for classification purpose [157]. Several machine learning techniques are available, with advantages and disadvantages depending on the number and nature of the training data and their properties. In a selection study, three different classifiers, namely a Bayesian net, a random forest and a support vector machine (SVM) were compared. Trained and tested on the druggability data set, the SVM outperformed the other two algorithms, and became the method of choice. Throughout this study, the freely available SVM software package `libsvm` [157] was used. The `libsvm` software convinced by its ability to separate data points into multiple classes, the supply of a reliability value for predictions as well as the possibility to incorporate weights into the calculation.
SVMs are widely used to solve regression and classification problems. Non-linear data is transferred into higher dimensional spaces, where the classes can be separated. SVMs are known to be relatively robust to overfitting. Although they are able to handle high numbers of input features, it has proven useful to simplify the models by eliminating irrelevant features [105]. For this purpose, a feature selection procedure has individually been applied to both test scenarios. A shrinkage discriminant analysis (SDA) [158], which accounts for correlations between features, has been utilized as prefilter method. In each case, the total number of global descriptors has been used as input, and those

features best suited for data separation are pursued. In my second SVM-based publication [D4], the built-in `libsvm` feature selection method [157] has been investigated as well, yielding the same top ranking features. Thus, a reduced set of discriminating features from the SDA is returned to the SVM, which can be used without adoptions. Another approach, not tested in this project, is backwards elimination, where features are discarded during the prediction process.
For model building in `libsvm`, the selected features are scaled into the interval $[-1, 1]$. For each of the different classification scenarios, models are built as follows: The data is randomly separated into training and test data, and a model is calculated based on the training data. For kernel parameter selection an internal five-fold cross validation is performed. Subsequently, the test data is used to evaluate the prediction performance of the method.
Two enhancements, available in `libsvm`, have been added to the classification strategy. The output of the SVM models is enriched by reliability values. The SVM returns a normalized vector, representing the probability to which a query pocket belongs to the modeled classes. A clear peak in this vector for one class indicates a high probability of correct annotation. Otherwise, the vector favors multiple classes, therefore those predictions are less reliable.
Furthermore, weights are imposed on the test data sets by way of trial. E.g., in the function prediction scenario, the classes are highly unbalanced. Thus, using weights accounting for the size of the individual sets is also investigated.

### 3.3.3 Nearest Neighbor Search

This method was introduced to capture the similarity between sites, based on specific interaction partners present within conserved distances. The local functional distance histograms are the basis for the nearest neighbor search (see section 3.2.2). The procedure is exemplified here with respect to the two-class problem of separating druggable from undruggable pockets. The procedure is based on the similarity between histograms calculated by means of a histogram distance [159]. This distance is calculated by summing up the absolute values of prefix sums of the difference in each histogram bin. A detailed description of the calculation of the nearest neighbor score can be found in my second publication [D2]. The procedure for finding the nearest neighbor is as follows: As described in the previous section, local distance dependent histograms are calculated for a pocket. To classify the pocket as druggable or undruggable, all histograms are individually compared to two precompiled sets of histograms. These two sets originate from known druggable and undruggable targets, respectively. In each round, the current histogram is compared to all precalculated histograms, the above mentioned histogram distance is calculated, and the score for the most similar histogram from each set is stored. Subsequently, the best druggable score is subtracted from the best undruggable score for each histogram of the query pocket. A score close to zero indicates an unspecific histogram, a high score specifies a histogram exclusively found in druggable pockets. Finally, the maximal absolute value over the scores of all individual histograms of a query pocket represents the nearest neighbor score, and the druggability type of the respective histogram is transferred to the pocket.

## 3.4 Triangle-Descriptor Based Comparison

As described in section 1.3, the quality of binding site comparison tools suffers from small structural changes. Focusing the comparison on local parts of the site and accounting for small variations is, therefore, an important enhancement when comparing sites. In this section, the concept of TrixP[1] [D5] is introduced, a new method for efficient binding site comparison, which is also part of this cumulative dissertation.

As described before [24], methods for binding site comparison have mainly three requirements: a function for the encoding of molecular recognition features, a method for similarity searching and the respective scoring function. The concept of TrixP is derived from TrixX [17], a method for efficient structure-based virtual screening, and adapted to the comparison of binding sites. Recognition features are encoded by triangle descriptors, comprising local physicochemical and shape information of the binding site (Figure 3.5). An index containing precompiled descriptors from known binding sites is built and can be screened unlimitedly. In the following, the characteristics will be explained in more detail, the complete description can be found in the TrixP publication [D5].

The triangle descriptor introduced in TrixX [17] is suited to capture local properties of active sites important for ligand binding. TrixP relies on a geometric matching procedure based on the representation of the binding site by a set of triangle descriptors. A triangle is spanned between each triplet of interaction points present in the binding site. Hydrogen bond donor and acceptor atoms as well as apolar points in the pocket form the corners of the triangles. While apolar points are undirected, hydrophilic interactions sustain a direction modeling the orientation of the attached hydrogen atom or the free electron pair, respectively. Further attributes of the triangle are the lengths

[1] The concept of TrixP was developed in a collaboration with M. v. Behren, who implemented the functionality. Software parts considering index and representation of interactions were provided by A. Henzler, K. Schomburg and S. Urbaczek.

of its edges and the enclosed angles. Finally, a set of constraints is imposed ensuring that only meaningful triangles are considered.

The shape of the binding site is captured by the so-called bulk [160]. Resulting from the refinement of an icosahedron, 80 bulk rays are spanned from the triangle center. The lengths of these rays describe the distance between the triangle center an the protein surface. The bulk mimics the ligand accessible volume. Since the intention within the TrixP development was the projection of local similarities shared between distantly related binding sites, the use of the original bulk descriptor was impracticable. Thus, the idea of using a partial bulk, introduced in an yet unpublished ligand-based method by C. Schärfer, is reused for the binding site scenario. The partial bulk allows to consider a reduced number of rays adjacent in space. Subsets of 25%, 40% and 50% can be generated as well as their respective value when subtracted from 100%. Exemplarily, the annotation of a 25% shape requirement is discussed here (Figure 3.5). Starting from the icosahedron representation, five triangles surrounding a vertex cover 25% of the shape and 20 bulk rays. Furthermore, 12 such sets are possible to encode this shape requirement.

Since the number of calculated descriptors is high and a repetitive calculation of triangle descriptors for similarity searching is impracticable, a bitmap index [160] is used for efficient data management. During index creation, triangle descriptors are vertically partitioned by descriptor type, encoded in the triangle corners. This partition avoids a sequential screening, since triangles differing in their interaction types can be directly excluded from further matching steps. Each triangle descriptor attribute, e.g., corner type, side length, angel, direction and bulk ray, describes one dimension in the bitmap index and is either bit or range encoded. For efficient data storage, the values are additionally discretized in bins.

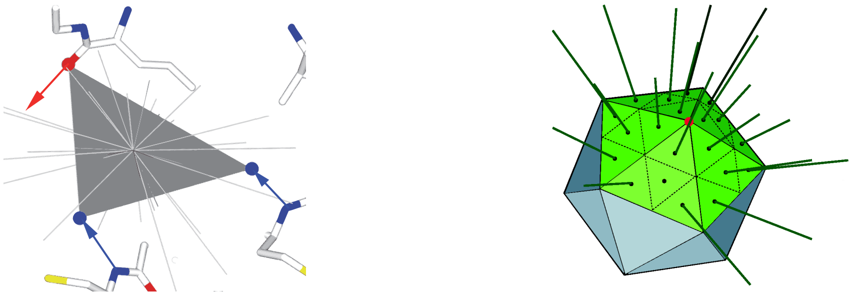


**Figure 3.5:** Example of a triangle descriptor together with its bulk rays, embedded into an active site (left), as well as the partial bulk (right). A similar version of these figures can be found in my publication [D5].

In a binding site comparison scenario, this index is precalculated once for a large amount of known protein data. For similarity searching, this index can be queried in an efficient manner avoiding a sequential screening of all descriptors. Next, descriptors are calculated for the query and the index is scanned using an elaborated query syntax [160]. As a result of this similarity search, matching descriptors are returned from the index. Proteins can subsequently be superimposed based on these triangles (Figure 3.6). For the scoring step, matching binding sites are superposed. To avoid the evaluation of a large number of superpositions and to enhance the quality of such, matching descriptor pairs are clustered.

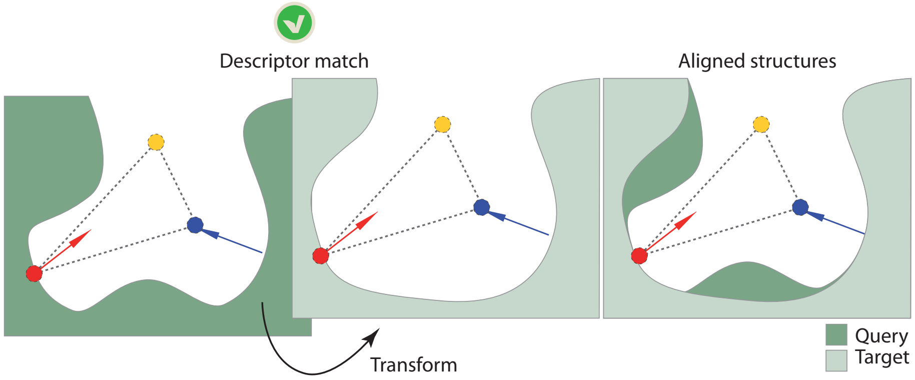


**Figure 3.6:** Schematic view of the structural alignment of TrixP. The figure is taken from my publication [D5].

For the resulting superpositions of clusters, a similarity score is calculated based on the compliance of interaction points of query and matched binding site. Iteratively, each query interaction point is investigated. The score is composed of linear terms rating the similarity of interaction points found within a sphere spanned around the current point of interest. The function rates the matching and mismatching of interaction types within this sphere, the distance as well as the directionality (if annotated). Finally, matching binding sites are ranked by score, the larger the score the higher the detected similarity.

# 4

# Results and Applications

This chapter comprises the evaluation of the developed software parts and the results of the performed studies. The following sections show the broad range of applications, covering pharmaceutical and biotechnological problems, e.g., drug design, biological function prediction, and protein engineering. First, the applications discussed in my scientific publications are summarized. Then, a section highlighting a rational enzyme design experiment, studied in cooperation with a partner from the Biocatalysis2021 cluster, follows.

## 4.1 Protein Pocket Detection

The reliable detection and specification of the active site of a protein is a crucial step in most structure-based computer methods. Many approaches have already been proposed to solve this task (see section 2.1), nevertheless, there is still room for improvements. In pursuit of generating new insights into this field, I developed DoGSite, a novel pocket detection algorithm enabling a precise boundary definition and a division into subpockets. A detailed analysis and discussion can be found in the DoGSite publication [D1], and a short summary will be given here.

The first and obvious demand on a pocket detection algorithm is its ability to consistently recover known binding sites, and to rank the true binding site superior to other detected sites on the protein surface. Parameters that should be considered when evaluating pocket detection methods are the criteria defining a correct prediction, the pocket boundary and its volume.

In a first experiment, DoGSite was evaluated on a set of 48 ligand-bound and unbound structures[1] and was positioned second compared to other published methods tested on this data set. Besides finding the ligand-binding pocket, predicting a pocket that does not overestimate the volume available for the ligand is of crucial importance. In this context, the boundary definition resulting from the pocket algorithm is another criterion influencing the quality. DoGSite convinces by a globular boundary definition, which depicts the ligand accessible volume quite well. In pursuit of detecting pockets that better describe this accessible volume, the terms ligand coverage and pocket coverage were introduced in DoGSite, describing the overlap between the binding partners. DoGSite has been evaluated on two large benchmark data sets, namely PDBbind [161] and scPDB [162], containing 828 and 6754 structures, respectively. Success rates of 91% and above on both data sets, with respect to the criterion of finding the true ligand binding site in the top three ranked pockets, were achieved. Restricting the definition of correct predictions to objects with 25% pocket coverage and 50% ligand coverage lowered the overall success rate but obtained more meaningful pockets.

Furthermore, it could be shown that the consideration of subpockets further improved the specificity of the description. DoGSite allows splitting of pockets at buried bottlenecks as well as partitions between solvent exposed cores. Studying the impact of this more granular description on the PDBbind data set showed that over 60% of the predicted pockets contained subpockets. Such subpockets were proven to better describe the ligand binding regions. Ligands are rarely contained in more than one subpocket, underlining the chemical meaning of the separation. This is also exhibited in higher coverage values of subpockets compared to pockets and thus, improved success rates.

Many algorithms for pocket detection, including DoGSite, are prone to small changes in the protein structure. Pockets detected for different structures of the same protein differ in volume and other properties. A set of 124 HIV-proteases was analyzed with respect to predicted pocket volume and co-crystallized ligand size. A median pocket volume of $810\text{Å}^3$ was found with a standard deviation of $175\text{Å}^3$, suggesting that the impact of the structural changes is not drastic. Although the calculated volumes altered with the bound ligand, this change was not proportional to the size of the bound ligand. In contrast, the coverage of the pocket grew with increasing ligand size. This suggests that the structure of the binding site is relatively stable and the ligand adapts to that form up to a certain extend.

Analyzing different states of the protein in this manner lets one draw conclusions about its flexible structural behavior. Concluding, the new concept enables a good starting point for subsequent descriptor-based analysis and classification scenarios.

[1] The success criterion in the respective study was whether the pocket lies within 4Å of any ligand atom.

## 4.2 Druggability Prediction

Early prioritization of promising drug targets is of high practical interest in the drug development pipeline, and can help to reduce time and cost expenses. I developed a novel approach, named DoGSiteScorer, for fully automatic druggability predictions solely based on the protein structure. The introduced method uses global and local descriptors of detected binding sites incorporated into a support vector machine (SVM) and a nearest neighbor search to rate the druggability of a target. DoGSiteScorer was evaluated in detail in [D2]. A short outline about the main aspects will be given here.
The basis for a successful application of the introduced machine learning approach was a reliable data set to train the SVM models. Fortunately, a druggability data set (DD) consisting of more than one thousand annotated data points was released in a recent publication [72]. Data points were annotated as being druggable, difficult and undruggable. Additionally, a non-redundant version (NRDD) with 70 entries of the DD was compiled. DoGSiteScorer was used to detect pockets for all structures and to describe them by multiple global sterical and physicochemical properties. A correlation between several features regarding size, shape and chemistry of the pockets was observed. Thus, a discriminant analysis was carried out to create a subset of important properties. Analysis of the features revealed that druggable pockets tend to be larger, more hydrophilic and complex when compared to undruggable ones, which was in agreement with previous literature reports [9]. Furthermore, pockets of targets annotated as difficult were found to be more similar to druggable pockets considering size and shape parameters, while closer resembling undruggable pockets in terms of physicochemical features.
For druggability prediction, the discriminating features of the detected pockets based on the NRDD were used to train the SVM. A comparative study to SiteMap [57] and Fpocket [72], two other truly automatic methods, has been performed. Comparison of the relative enrichment curves of the three methods revealed that they all perform well in rating druggable pockets above undruggable ones[1]. DoGSiteScorer slightly led the field, especially when considering subpockets. The trained models were further used to classify the complete set of 1069 DD structures. In this experiment, structures were grouped by the protein family to which they belong, and performance was analyzed based on mean values and standard deviations within the druggability scores of the family members. Considering the mean value, 88% of the families were correctly classified as undruggable, difficult or druggable. Generally, low standard deviations within

[1] Note that a direct comparison is difficult, since the underlying pockets may differ due to the individual pocket detection algorithm of each method.

the scores predicted for the contained family members were observed. Since apo- and holo-structures[1] were present in this data set, as well as bound ligands of different size, the low score deviations showed the robustness of the introduced method with respect to flexibility in the structures upon ligand binding. E.g., the kinase p38 family was represented by 40 structures and the detected pockets span volumes between 450Å$^3$ and 1800Å$^3$, depending on the different crystallized activation states. Despite the variation in size, DoGSiteScorer correctly classified the kinase structures as druggable, because of the combination of other features considered such as high fractions of lipophilic surface area. In contrast, the discrepancy between annotated druggable character and predicted low scores for the carbon anhydrase family exposed the limitations of the global method. The drug binding in carbon anhydrases is dominated by metal binding. Thus, their druggability results from single interactions in the binding site rather than a global property.

This encouraged the consideration of local properties. Therefore, distance dependent histograms, which are irrespective of the global pocket annotation or its boundary, were calculated with DoGSiteScorer. These profiles were incorporated for homology-driven knowledge transfer by means of a nearest neighbor search. Druggable and undruggable pockets were compared with respect to distances between their binding motif histograms. Although, the histograms from both classes were noisy, druggable targets were found to have more short range hydrophilic-hydrophilic and less short range lipophilic-lipophilic interaction partners. Generally, predictions on the DD data set yielded 88% correct annotations.

To improve the outcome and the prediction power of both methods, the combined predictions were investigated on the DD. While local and global predictions mostly agreed in the assigned druggability state, focus was on the cases where they differ. Carbon anhydrases were one example where the global method failed. In contrast, the nearest neighbor method achieved good results modeling the local characteristics of ligand binding.

Nevertheless, some challenges remain, being related to the ambiguous definition of druggability, the uncertainties arising from the pocket prediction step, as well as the discrepancy in set sizes or wrong annotations in the training data, which were further discussed in the publication [D2]. Concluding, the study showed that DoGSiteScorer provides valuable qualitative and quantitative information for target assessment and drug discovery.

The complete DoGSiteScorer pipeline starting from active site detection, over calculation of descriptors to the classification based on the trained SVM models, was ported

[1] Apo describes the ligand-free, holo the ligand-bound form of a protein.

to a web server [D3] allowing the community to analyze a protein of interest and export the processed results for further investigation.

## 4.3 Enzymatic Function Prediction

Predicting the enzymatic function of yet unclassified structures is of high practical impact in many research areas. In my work published in [D4], I developed a novel computational protocol aimed at predicting enzymatic function at different levels of granularity. The enzymatic classification (EC) model [163] was used as basis for this approach. The EC scheme, encoded in a four-digit number, specifies the reaction an enzyme catalyzes. The first number divides the enzymes into six main classes. Each main class consists of several subclasses, encoded in the second digit and so on[1].

Developing a pipeline for function prediction with increasing specificity according to this classification scheme was the aim of this part of my work. Although many approaches exist for function annotation, no established EC-based benchmark data set exists to calibrate and evaluate new methods. Using SVMs for classifications requires a sufficiently large data set for model set-up and training. For this purpose, a large set containing all enzymes with annotated EC number present in the PDB, restricted by some quality criteria was assembled. Binding site detection and careful restriction based on the coverage criterion yielded over 26 000 well defined pockets, attached with calculated descriptors. None of the descriptors showed clear trends for class separations by itself, which was not surprising considering the wide range of substrates bound to the EC class members. Nevertheless, some lessons could be learned, e.g., in terms of largest volume (EC1), lipophilic surface fraction (EC1), or negative amino acid ratios (EC3).

For the function prediction cascade, SVM models were built for main class, subclass and substrate-specific sub-subclass predictions, based on a reduced set of discriminating features. A cross validation on randomly selected two thirds test and one third training data yielded 68% accuracy for correct main class annotation. Subsequently, for each class a sub-model was built to discriminate the respective subclasses. The numbers of subclasses per main class varied between 6 and 13, while success rates between 60% and 80% were achieved. Finally, the same procedure was applied to the substrate-specific sub-subclass level. Exemplarily, the performance on the kinase subclass (E.C. 2.7) containing 18 substrate-specific sub-subclasses was outlined and cross-validation studies yielded 58% correct predictions. An interesting enhancement was the indication of a reliability value for each prediction based on the estimated SVM probability. During

[1] BRENDA [109] provides a detailed description about this classification scheme.

evaluation, special attention was paid to potential influences of data redundancy and imbalance between class sizes.
Finally, two structures were studied in a retrospective (PDB code: 1p1m) and a hypothetical (PDB code: 1mjh) function prediction scenario. The function prediction pipeline results were in good accordance with literature and proposals from other tools in this field [111, 127, 164, 165].
Concluding, the study showed the competitiveness of this method with other published approaches and the information gain due to the novel prediction cascade on different levels of granularity.

## 4.4 Triangle-Based Pocket Comparison

To allow for more reliable predictions with respect to structural changes in protein binding sites, I took part in the development of a new locally enhanced approach. TrixP has been developed for structure-based binding site comparison. The usage of the triangle descriptor, enabled by the bulk-imposed shape requirements, in combination with an index-based data management allows for high-throughput applications. A detailed evaluation and several applications of the method were introduced in my scientific publication [D5], and will be summarized in the following.
The experiments are generally executed in a two-step manner, starting with the building of the index followed by the screening step. First, triangle descriptors and bulk rays are calculated for a set of structures with known function. Subsequently, these descriptors are inserted into an index, which can then be queried with any protein of interest.
A vital requirement for binding site comparison tools is the ability to detect similarities between related structures while discarding unrelated ones. A first study was performed on 1331 similar and dissimilar protein structure pairs [142]. Querying with one representative, TrixP was able to recover 81.8% of the respective pairs with a similarity score above the threshold of 0.3. For dissimilar pairs, in 99.5% of the cases the associated pair was scored lower than this cut-off. Besides the ability to discriminate, TrixP convinced with high sensibility in ranking the found structures in accordance with their actual similarity. In a further experiment, an index built from 9802 scPDB structures [162] was queried with structures from four families, e.g., the estrogen receptor family (ER). A detailed analysis of an ER$\alpha$ query showed that 98.5% of the ER structures, included in the index, were found with high similarity scores. Next, the sensitivity in classification was analyzed with respect to TrixP's ability to group structures into subfamilies, exemplified on a kinase set [130]. While the query on the scPDB index was based on

binding sites defined by the co-crystallized ligands, in the kinase study, pockets are predicted with DoGSite for each index and query structure. A clustering based on the similarity matrix calculated by an all-by-all comparison grouped the structures in good accordance with their biological annotation. TrixP especially convinced by its ability to distinguish between different activation states and sub-subfamilies.

Furthermore, quality and runtime comparison studies on a set of eight difficult structural pairs demonstrated the efficiency and the competitiveness of TrixP to other recent and efficient methods in this field.

Summarizing, TrixP convinced in detecting similarities even between distantly related binding sites. Especially, the local character of the triangle descriptor enhanced with partial shape matching, which accounts for a certain degree of flexibility, makes TrixP a very useful tool for structural comparison and function annotation.

## 4.5 Ligand and Pocket Shape Comparison

Understanding molecular recognition is a prerequisite for proper modeling and modification of the binding process. While the driving forces for ligand binding are encoded in chemical and sterical complementary, the impact of shapes was explicitly analyzed in the study published in [D6][1] and the findings will be summarize in the following.

To directly compare ligand and binding pocket shapes, the normalized principle moments of inertia ratios (NPRs), introduced in section 3.2.3, were used. NPRs allow the mapping of shapes into a triangular space, with the extrema describing truly sphere-, rod-, and disk-like shapes.

The analysis was performed on selected scPDB structures [162]. For this set, pockets and subpockets were predicted with DoGSite and represented by calculated NPRs. Similar to findings for small molecules [150], pockets with bound ligands avoid spherical shapes. Nevertheless, spherical shapes were prevalent in small and mostly empty pockets [166]. Furthermore, a direct shape comparison confirmed that pockets are predominately only covered to one third by the co-crystallized ligand [167]. In contrast, subpocket coverage rises to 50%, underlining their more specific and restrictive representation of the ligand available volume. In addition, several shape complementarity parameters were analyzed with respect to pocket coverage. A good shape fit between ligand and pocket is indicated by pairwise shape distances lower than 0.21, center of mass (CoM) distances below 2.4Å and angle deviations between the first principle ellipsoid axes below 28.8°.

[1] This work was performed in collaboration with Matthias Wirth from Merck Serono, Geneva.

Furthermore, bioactivity and binding efficiency of the co-crystallized ligands was analyzed and related to pocket shape on a PDBbind [161] subset. No strong correlation between activity and individual pocket parameters, e.g., pocket coverage, NPR distance, CoM distance or angle deviation could be stated. The best correlation was found for pocket coverage with a Pearson correlation coefficient of 0.4. Nevertheless, by separating the ligands into low, middle and high affinity binders, a trend towards better coverage, smaller NPRs and CoM distances, as well as lower angle deviations could be observed for high affinity binders. Furthermore, activity was analyzed with respect to ligand size, with the goal of deriving ideas about the expected maximal binding efficiency[1] for a given pocket as additional parameter for druggability assessment. Similar to previous studies [168, 169], a decrease in efficiency with increasing molecule size was observed. This information could be enriched by finding a similar decline in efficiency with increasing pocket volumes. Additionally, a range of pocket volumes between 300-700Å was found exhibiting a higher probability of binding ligands with high efficiency.

Finally, the usability of the shape distance as screening filter was analyzed based on the ChEMBL data set [170], screened against four targets. Different distance criteria between selected pocket, co-crystallized ligand and the minimum energy conformation of the ChEMBL compounds were taken into account. The distance between pocket and compound volume proved useful in discarding compounds which were annotated with unfavorable docking scores in a screening experiment with Glide [171].

Concluding the study revealed new insights into the shape fit between protein and small molecule.

## 4.6 Rational Enzyme Design

The Biocatalysis2021 partner Henkel[2] investigated in the directed evolution of an alditol oxidase to enhance its activity towards glycerol. In a crosslink project, potential mutation sites should be suggested by applying the developed binding site comparison software together with molecular modeling[3]. I used the developed software to predict and compare the active sites of alditol oxidases with known glycerol binding proteins. Since only few alditol oxidase structures were available, the idea was to learn from the properties of other sites active to glycerol and detect similarities and differences to the current structures. Typical binding profiles of glycerol binding enzymes shall

[1] Binding efficiency is calculated by dividing the $pK_i$ by the number of ligand heavy atoms.

[2] Henkel AG & Co. KGaA, Düsseldorf, Germany

[3] The molecular modeling experiments were mainly carried out by B. Windshügel.

serve as idea generator for mutations in the active site of alditol oxidases. First, protein structure data was collected for both sets. At the time of the project, five alditol oxidase structures were available in the PDB (PDB codes: 2vfr, 2vfu, 2vfv, 2vfs, 2vft). The search for oxidase structures with bound glycerol yielded 210 structures. Active sites were calculated for all structures using DoGSite [D1]. Descriptors were calculated and the previously outlined complete linkage procedure (see section 3.3.1) was used to group the structures by their descriptor similarity.

The clustering yielded one branch in which all alditol oxidase structures were contained together with 11 glycerol binders. Four of the five alditol oxidase structures belonged to one subbranch shared with two other enzymes, of which the most similar one (PDB code: 1d6z) was used for further investigation. Using molecular modeling software[1], the active site of the amine oxidase (1d6z) was superposed onto the alditol oxidase and analyzed (Figure 4.1). A direct comparison of these structures suggested that mutations in the active site might possibly adapt the function of the alditol oxidase to glycerol binding. Amino acids Glu702 and Lys709 build a salt bridge in 1d6z with a C$\alpha$ distance of 11.5Å. An almost equidistant pair was found in the alditol oxidase between the two amino acids Val250 and Phe275. Furthermore, both amino acids are located in a similar position with respect to the bound glycerol.

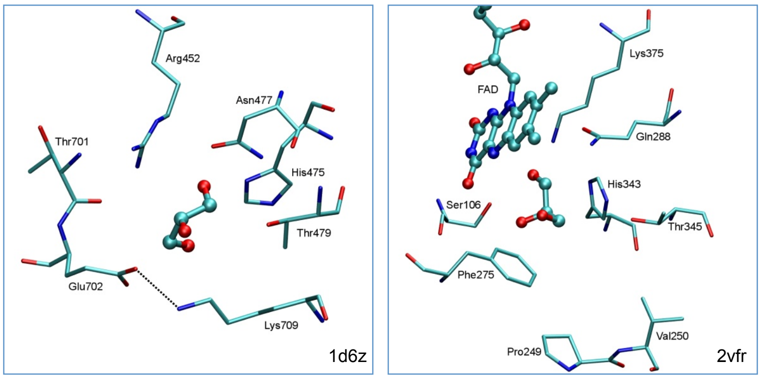


**Figure 4.1:** Comparison of the active sites of the amine oxidase 1d6z and the alditol oxidase 2vfr.

Nevertheless, these mutations may have high impact on the positions of neighboring amino acids and the stability of the structure. Other mutations, e.g. Pro249, may be necessary to create the space for the interactions of the two amino acids. To verify or

[1] MOE (Molecular Operating Environment), `http://www.chemcomp.com/`

falsify these suggestions, nevertheless, detailed modeling experiments should be conducted.

The suggested double mutations together with three other mutations were handed to the cooperation partner Henkel for experimental testing. Unfortunately, the provided mutations were found inactive [172]. Mutants with substitutions in the active site could not be expressed in the cytoplasm and no active oxidase was produced.

Although the software provides valuable information about the active site, information about the stability of the enzyme is missing. Modifications in the active site often prevent the expression of active enzymes and thus make reliable predictions by rational design difficult. A possible next step would be the investigation into molecular dynamics simulations.

# 5

# Summary and Outlook

My work comprises several approaches for structure-based computer-aided active site analysis. The algorithms and their applications showed good results on theoretic and retrospective studies. The approach for pocket prediction introduces a method from image processing into the field of bioinformatics, and convinces by its accurate predictions as well as its capability to predict pockets and subpockets. Global and local descriptors are derived within these pockets and are used for protein classification. The SVM method provides a generic approach, which has been shown to perform well on different scenarios, i.e., druggability and function prediction. Besides good results in both experiments equal or even superior to previous published approaches, the function prediction example stands out especially by its multi-step approach on different granularities. The druggability prediction software is already in use at my cooperation partner Merck, and the provided server for protein assessment also enjoys a growing popularity. Furthermore, the comparison between protein pocket and ligand creates a new understanding of facts important in molecular recognition. The newly developed TrixP approach expands the portfolio of approaches by a method focusing on local properties accounting for small changes in the binding site.

Nevertheless, as mentioned throughout this study challenges remain. In this section, I will suggest possible applications and extensions of the developed methods. A meaningful application for the DoGSite algorithm would be the prediction of protein-protein interactions. For this scenario, the algorithm would have to be reparametrized to the new problem, and adequate training data would have to be collected. The adapted classification software could help to solve questions in the protein-protein interaction research. Another suggestion is its application in the biotechnological context. In a preliminary study, in cooperation with the group of Prof. Liese, I tried to separate polymerization-catalyzing enzymes from other enzymes, with promising results. Further application studies, especially with an experimental evaluation, would be of great value. Another idea is the use of TrixP for the detection of conserved triangles exclusively found within enzyme families. Such positions could give valuable information about the catalytic mechanism of a specific enzyme family or the separation into subfamilies.

In my opinion, the major challenge is, nevertheless, the not covered and not fully understood flexibility of a protein. The approaches developed within this work use one snapshot of the protein structure as input. Small changes in the environment of a protein or its binding partners induce changes in its conformation. Hence, detected pockets of one stage may differ from the pockets detected for another conformation. Deriving descriptors from snapshot representations and using them for annotation bears the risk of tampering the results. Thus, one further enhancement would be the use of a new representation for protein flexibility. Until such representations exist, another circumvention would be the use of protein structural ensembles. The detection of a meta-pocket, based on a superposition of the ensemble structures, and a grid point annotation with density probabilities is one suggestion. The adaption of the global descriptors may be possible but bears the risk of adding noise to the results. Using such a meta-pocket presentation in the TrixP approach is more conducive. The probabilities could be transferred to the respective triangle corners and be incorporated in the scoring process. Furthermore, the development of a strategy to account for large changes in the active site remains a very challenging task.

# Bibliography

[1] G. Klebe. Wirkstoffdesign: Entwurf und Wirkung von Arzneistoffen. *Spektrum Akademischer Verlag*, 2009.

[2] E. Fischer. *Berichte der Deutschen chemischen Gesellschaft zu Berlin*, 27:2985–2993, 1894.

[3] J. M. Berg, J. L. Tymoczko, and L. Stryer. Biochemistry. *WH Freeman and Company New York*, 2002.

[4] W. Gronwald and H.R. Kalbitzer. Automated structure determination of proteins by NMR spectroscopy. *Prog. NMR Spectrosc.*, 44:33–96, 2004.

[5] J. Monod, J. Wyman, and J.-P. Changeux. On the nature of allosteric transitions: A plausible model. *Journal of Molecular Biology*, 12:88–118, 1965.

[6] Koshland D. E. Application of a theory of enzyme specificity to protein synthesis. *Proceedings of the National Academy of Sciences of the USA*, 44(2):98–104, 1958.

[7] H. Gohlke. *Protein-ligand interactions: Methods and principles in medicinal chemistry*, volume 53. Wiley-VCH, 2012.

[8] G. Schneider and K.-H. Baringhaus. *Molecular design: concepts and applications.* Wiley-VCH, 2008.

[9] U. Egner and R.C. Hillig. A structural biology view of target drugability. *Expert Opinion on Drug Discovery*, 3(4):391–401, 2008.

[10] R.S. Bohacek, C. McMartin, and W.C. Guida. The art and practice of structure-based drug design: A molecular modeling perspective. *Medicinal Research Reviews*, 16(1):3–50, 1996.

[11] A.C. Anderson. The process of structure-based drug design. *Chemistry and Biology*, 10(9):787–7973, 2003.

[12] C. M. Song, S. J. Lim, and J. C. Tong. Recent advances in computer-aided drug design. *Briefings in Bioinformatics*, 10(5):579–591, 2009.

[13] K.H. Bleicher, H.-J. Boehm, K. Mueller, and A.I. Alanine. Hit and lead generation: beyond high-throughput screening. *Nature Reviews. Drug Discovery*, 2(5):369–378, 2003.

[14] M. Rarey, S. Wefing, and T. Lengauer. Placement of medium-sized molecular fragments into active sites of proteins. *Journal of Computer-Aided Molecular Design*, 10(1):41–54, 1996.

[15] N. Brooijmans and I.D. Kuntz. Molecular recognition and docking algorithms. *Annual Review of Biophysics and Biomolecular Structure*, 32(1):335–373, 2003.

[16] B.K. Shoichet. Virtual screening of chemical libraries. *Nature*, 432(7019):862–865, 2004.

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

[18] K.I. Ramachandran, G. Deepa, and K. Namboori. *Computational Chemistry and Molecular Modeling - Principles and applications*. Germany: Springer International, 2008.

[19] H. Gohlke, L. A. Kuhn, and D. A. Case. Change in protein flexibility upon complex formation: Analysis of ras-raf using molecular dynamics and a molecular framework approach. *Proteins*, 56:322–327, 2004.

[20] L. Xie, J. Wang, and P.E. Bourne. In silico elucidation of the molecular mechanism defining the adverse effect of selective estrogen receptor modulators. *PLoS Computational Biology*, 3(11):e217, 2007.

[21] B. Stauch, H. Hofmann, M. Perkovic, M. Weisel, F. Kopietz, K. Cichutek, C. Munk, and G. Schneider. Model structure of APOBEC3C reveals a binding pocket modulating ribonucleic acid interaction required for encapsidation. *Proceedings of the National Academy of Sciences*, 106(29):12079–12084, 2009.

[22] S.L. Kinnings, N. Liu, N. Buchmeier, P.J. Tonge, L. Xie, and P.E. Bourne. Drug discovery using chemical systems biology: repositioning the safe medicine Comtan to treat multi-drug and extensively drug resistant tuberculosis. *PLoS Computational Biology*, 5(7):e1000423, 2009.

[23] L. Xie and P.E. Bourne. Structure-based systems biology for analyzing off-target binding. *Current Opinion in Structural Biology*, 21(2):189–199, 2011.

[24] B. Nisius, F. Sha, and H. Gohlke. Structure-based computational analysis of protein binding sites for function and druggability prediction. *Journal of Biotechnology*, 159(3):123–134, 2012.

[25] G. Frazzetto. White biotechnology. *EMBO Reports. Science and Society. Analysis.*, 4(9):835–837, 2003.

[26] A. Macchiarulo, I. Nobeli, and J.M. Thornton. Ligand selectivity and competition between enzymes in silico. *Nature Biotechnology*, 22(8):1039–1045, 2004.

[27] C. Kalyanaraman, K. Bernacki, and M.P. Jacobson. Virtual screening against highly charged active sites: identifying substrates of alpha-beta barrel enzymes. *Biochemistry*, 44(6):2059–2071, 2005.

[28] S. Tyagi and J. Pleiss. Biochemical profiling in silico–predicting substrate specificities of large enzyme families. *Journal of Biotechnology*, 124(1):106–16, 2006.

[29] P. Oelschlaeger and J. Pleiss. Hydroxyl groups in the betabeta sandwich of metallo-beta-lactamases favor enzyme activity: Tyr218 and Ser262 pull down the lid. *Journal of Molecular Biology*, 366(1):316–329, 2007.

[30] S.D. Brown, J.A. Gerlt, J.L. Seffernick, and P.C. Babbitt. A gold standard set of mechanistically diverse enzyme superfamilies. *Genome Biology*, 7(1):R8, 2006.

[31] R. Apweiler, A. Bairoch, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C. O'Donovan, N. Redaschi, and L.-S.L. Yeh. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 32(Database issue):D115–9, 2004.

[32] F.C. Bernstein, T.F. Koetzle, G.J. Williams, E.F.J. Meyer, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The protein data bank: a computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, 112(3):535–542, 1977.

[33] A.M. Schnoes, S.D. Brown, I. Dodevski, and P.C. Babbitt. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Computational Biology*, 5(12):e1000605, 2009.

[34] R. Parasuram, J.S. Lee, P. Yin, S. Somarowthu, and M.J. Ondrechen. Functional classification of protein 3D structures from predicted local interaction sites. *Journal of Bioinformatics and Computational Biology*, 8 Suppl 1:1–15, 2010.

[35] S. Perot, O. Sperandio, M.A. Miteva, A.-C. Camproux, and B.O. Villoutreix. Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. *Drug Discovery Today*, 15(15-16):656–667, 2010.

[36] D. Ringe, Y. Wei, K.R. Boino, and M.J. Ondrechen. Protein structure to function: insights from computation. *Cellular and Molecular Life Sciences : CMLS*, 61(4):387–392, 2004.

[37] J.D. Watson, R.A. Laskowski, and J.M. Thornton. Predicting protein function from sequence and structural data. *Current Opinion in Structural Biology*, 15(3):275–284, 2005.

[38] A. Godzik, M. Jambon, and I. Friedberg. Computational protein function prediction: are we making progress? *Cellular and Molecular Life Sciences : CMLS*, 64(19-20):2505–2511, 2007.

[39] P. Aloy, E. Querol, F.X. Aviles, and M.J. Sternberg. Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *Journal of Molecular Biology*, 311(2):395–408, 2001.

[40] A. Armon, D. Graur, and N. Ben-Tal. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *Journal of Molecular Biology*, 307(1):447–463, 2001.

[41] C.M. Ho and G.R. Marshall. Cavity search: an algorithm for the isolation and display of cavity-like binding regions. *Journal of Computer-Aided Molecular Design*, 4(4):337–354, 1990.

[42] D.G. Levitt and L.J. Banaszak. POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *Journal of Molecular Graphics*, 10(4):229–234, 1992.

[43] R.A. Laskowski. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *Journal of Molecular Graphics*, 13(5):323–30, 1995.

[44] K.P. Peters, J. Fauck, and C. Frommel. The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *Journal of Molecular Biology*, 256(1):201–213, 1996.

[45] M. Hendlich, F. Rippmann, and G. Barnickel. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *Journal of Molecular Graphics & Modelling*, 15(6):359–63,, 1997.

[46] J. Liang, H. Edelsbrunner, and C. Woodward. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Science: a Publication of the Protein Society*, 7(9):1884–1897, 1998.

[47] G.P.J. Brady and P.F. Stouten. Fast prediction and visualization of protein binding pockets with PASS. *Journal of Computer-Aided Molecular Design*, 14(4):383–401, 2000.

[48] A.T.R. Laurie and R.M. Jackson. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics (Oxford, England)*, 21(9):1908–1916, 2005.

[49] M. Nayal and B. Honig. On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins*, 63(4):892–906, 2006.

[50] M. Weisel, E. Proschak, and G. Schneider. PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chemistry Central Journal*, 1:7, 2007.

[51] V. Le Guilloux, P. Schmidtke, and P. Tuffery. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics*, 10:168, 2009.

[52] P.J. Goodford. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of Medicinal Chemistry*, 28(7):849–857, 1985.

[53] J. An, M. Totrov, and R. Abagyan. Comprehensive identification of ‘‘druggable’’ protein ligand binding sites. *Genome Informatics. International Conference on Genome Informatics*, 15(2):31–41, 2004.

[54] J. Ruppert, W. Welch, and A.N. Jain. Automatic identification and representation of protein binding sites for molecular docking. *Protein Science: a Publication of the Protein Society*, 6(3):524–533, 1997.

[55] S. Zhong and A.D.J. MacKerell. Binding response: a descriptor for selecting ligand binding site on protein surfaces. *Journal of Chemical Information and Modeling*, 47(6):2303–2315, 2007.

[56] J.A. Capra, R.A. Laskowski, J.M. Thornton, M. Singh, and T.A. Funkhouser. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Computational Biology*, 5(12):e1000585, 2009.

[57] T.A. Halgren. Identifying and characterizing binding sites and assessing druggability. *Journal of Chemical Information and Modeling*, 49(2):377–89, 2009.

[58] B. Huang. MetaPocket: a meta approach to improve protein ligand binding site prediction. *Omics : a Journal of Integrative Biology*, 13(4):325–330, 2009.

[59] R.P. Sheridan, V.N. Maiorov, M.K. Holloway, W.D. Cornell, and Y.-D. Gao. Drug-like density: a method of quantifying the ‘‘bindability’’ of a protein target based on a very large set of pockets and drug-like ligands from the Protein Data Bank. *Journal of Chemical Information and Modeling*, 50(11):2029–2040, 2010.

[60] R.A. Ward. Using protein-ligand docking to assess the chemical tractability of inhibiting a protein target. *Journal of Molecular Modeling*, 16(12):1833–1843, 2010.

[61] F.N. Edfeldt, R.H. Folmer, and A.L. Breeze. Fragment screening to predict druggability (ligandability) and lead discovery success. *Drug Discovery Today*, 16(7-8):284–287, 2011.

[62] A.L. Hopkins and C.R. Groom. The druggable genome. *Nature Reviews. Drug Discovery*, 1(9):727–730, 2002.

[63] A.L. Hopkins and C.R. Groom. Target analysis: a priori assessment of druggability. *Ernst Schering Research Foundation Workshop*, 42:11–17, 2003.

[64] P.J. Hajduk, J.R. Huth, and C. Tse. Predicting protein druggability. *Drug Discovery Today*, 10(23-24):1675–1682, 2005.

[65] M.K. Sakharkar, K.R. Sakharkar, and S. Pervaiz. Druggability of human disease genes. *The International Journal of Biochemistry & Cell Biology*, 39(6):1156–1164, 2007.

[66] P.J. Hajduk, J.R. Huth, and S.W. Fesik. Druggability indices for protein targets derived from NMR-based screening data. *Journal of Medicinal Chemistry*, 48(7):2518–2525, 2005.

[67] M. Pellecchia, I. Bertini, D. Cowburn, C. Dalvit, E. Giralt, W. Jahnke, T.L. James, S.W. Homans, H. Kessler, C. Luchinat, B. Meyer, H. Oschkinat, J. Peng, H. Schwalbe, and G. Siegal. Perspectives on NMR in drug discovery: a technique comes of age. *Nature Reviews Drug Discovery*, (9):738–745.

[68] C.J. Zheng, L.Y. Han, C.W. Yap, Z.L. Ji, Z.W. Cao, and Y.Z. Chen. Therapeutic targets: progress of their exploration and investigation of their characteristics. *Pharmacological Reviews*, 58(2):259–279, 2006.

[69] C. Zheng, L. Han, C.W. Yap, B. Xie, and Y. Chen. Progress and problems in the exploration of therapeutic targets. *Drug Discovery Today*, 11(9-10):412–420, 2006.

[70] L.Y. Han, C.J. Zheng, B. Xie, J. Jia, X.H. Ma, F. Zhu, H.H. Lin, X. Chen, and Y.Z. Chen. Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness. *Drug Discovery Today*, 12(7-8):304–313, 2007.

[71] A.C. Cheng, R.G. Coleman, K.T. Smyth, Q. Cao, P. Soulard, D.R. Caffrey, A.C. Salzberg, and E.S. Huang. Structure-based maximal affinity model predicts small-molecule druggability. *Nature Biotechnology*, 25(1):71–75, 2007.

[72] P. Schmidtke and X. Barril. Understanding and predicting druggability. a high-throughput method for detection of drug binding sites. *Journal of Medicinal Chemistry*, 53(15):5858–5867, 2010.

[73] A. Krasowski, D. Muthas, A. Sarkar, S. Schmitt, and R. Brenk. DrugPred: A Structure-Based Approach To Predict Protein Druggability Developed Using an Extensive Nonredundant Data Set. *Journal of Chemical Information and Modeling*, 51(11):2829–2842, 2011.

[74] E. Perola, L. Herman, and J. Weiss. Development of a rule-based method for the assessment of protein druggability. *Journal of Chemical Information and Modeling*, epub ahead of print, 2012.

[75] E.V. Koonin, R.L. Tatusov, and M.Y. Galperin. Beyond complete genomes: from sequence to structure and function. *Current Opinion in Structural Biology*, 8(3):355–363, 1998.

[76] T. Hawkins and D. Kihara. Function prediction of uncharacterized proteins. *Journal of Bioinformatics and Computational Biology*, 5(1):1–30, 2007.

[77] D. Lee, O. Redfern, and C. Orengo. Predicting protein function from sequence and structure. *Nature Reviews. Molecular Cell Biology*, 8(12):995–1005, 2007.

[78] T. Hawkins, M. Chitale, and D. Kihara. New paradigm in protein function prediction for large scale omics analysis. *Molecular BioSystems*, 4(3):223–231, 2008.

[79] R. Rentzsch and C.A. Orengo. Protein function prediction–the power of multiplicity. *Trends in Biotechnology*, 27(4):210–219, 2009.

[80] C.L. Pierri, G. Parisi, and V. Porcelli. Computational approaches for protein function prediction: a combined strategy from multiple sequence alignment to molecular docking-based virtual screening. *Biochimica et Biophysica Acta*, 1804(9):1695–1712, 2010.

[81] M. Brylinski and J. Skolnick. Comparison of structure-based and threading-based approaches to protein functional annotation. *Proteins*, 78(1):118–134, 2010.

[82] R.D. Sleator and P. Walsh. An overview of in silico protein function prediction. *Archives of Microbiology*, 192(3):151–155, 2010.

[83] F. Xin and P. Radivojac. Computational methods for identification of functional residues in protein structures. *Current Protein & Peptide Science*, 12(6):456–469, 2011.

[84] L. Sael, M. Chitale, and D. Kihara. Structure- and sequence-based function prediction for non-homologous proteins. *Journal of Structural and Functional Genomics*, 2012.

[85] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

[86] M. Punta, P.C. Coggill, R.Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E.L.L. Sonnhammer, S.R. Eddy, A. Bateman, and R.D. Finn. The Pfam protein families database. *Nucleic Acids Research*, 40(D1):D290–D301, 2012.

[87] J.G. Henikoff, E.A. Greene, S. Pietrokovski, and S. Henikoff. Increased coverage of protein families with the blocks database servers. *Nucleic Acids Research*, 28(1):228–230, 2000.

[88] A.C. Wallace, N. Borkakoti, and J.M. Thornton. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Science: a Publication of the Protein Society*, 6(11):2308–2323, 1997.

[89] C.J.A. Sigrist, L. Cerutti, N. Hulo, A. Gattiker, L. Falquet, M. Pagni, A. Bairoch, and P. Bucher. PROSITE: a documented database using patterns and profiles as motif descriptors. *Briefings in Bioinformatics*, 3(3):265–274, 2002.

[90] D.E. Almonacid, E.R. Yera, J.B.O. Mitchell, and P.C. Babbitt. Quantitative comparison of catalytic mechanisms and overall reactions in convergently evolved enzymes: implications for classification of enzyme function. *PLoS Computational Biology*, 6(3):e1000700, 2010.

[91] S. Khan, G. Situ, K. Decker, and C.J. Schmidt. GoFigure: automated Gene Ontology annotation. *Bioinformatics (Oxford, England)*, 19(18):2484–2485, 2003.

[92] F. Enault, K. Suhre, and J.-M. Claverie. Phydbac ‘‘Gene Function Predictor’’: a gene annotation tool based on genomic context analysis. *BMC Bioinformatics*, 6:247, 2005.

[93] B.E. Engelhardt, M.I. Jordan, K.E. Muratore, and S.E. Brenner. Protein molecular function prediction by Bayesian phylogenomics. *PLoS Computational Biology*, 1(5):e45, 2005.

[94] N. Krishnamurthy, D. Brown, and K. Sjolander. FlowerPower: clustering proteins into domain architecture classes for phylogenomic inference of protein function. *BMC Evolutionary Biology*, 7 Suppl 1:S12, 2007.

[95] K. Illergard, D.H. Ardell, and A. Elofsson. Structure is three to ten times more conserved than sequence–a study of structural response in protein cores. *Proteins*, 77(3):499–508, 2009.

[96] T.J. Hubbard, A.G. Murzin, S.E. Brenner, and C. Chothia. SCOP: a structural classification of proteins database. *Nucleic Acids Research*, 25(1):236–239, 1997.

[97] C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells, and J.M. Thornton. CATH-a hierarchic classification of protein domain structures. *Structure (London, England: 1993)*, 5(8):1093–1109, 1997.

[98] L. Holm and C. Sander. The FSSP database: fold classification based on structure-structure alignment of proteins. *Nucleic Acids Research*, 24(1):206–209, 1996.

[99] Y. Ye and A. Godzik. Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics (Oxford, England)*, 19 Suppl 2:ii246–55, 2003.

[100] H. Taubig, A. Buchner, and J. Griebsch. PAST: fast structure-based searching in the pdb. *Nucleic Acids Research*, 34(Web Server issue):W20–3, 2006.

[101] J.F. Gibrat, T. Madej, and S.H. Bryant. Surprising similarities in structure comparison. *Current Opinion in Structural Biology*, 6(3):377–385, 1996.

[102] S. Wang, J. Peng, and J. Xu. Alignment of distantly related protein structures: algorithm, bound and implications to homology modeling. *Bioinformatics (Oxford, England)*, 27(18):2537–2545, 2011.

[103] M. Kontoyianni and C.B. Rosnick. Functional prediction of binding pockets. *Journal of Chemical Information and Modeling*, 52(3):824–833, 2012.

[104] P.D. Dobson and A.J. Doig. Distinguishing enzyme structures from non-enzymes without alignments. *Journal of Molecular Biology*, 330(4):771–783, 2003.

[105] P.D. Dobson and A.J. Doig. Predicting enzyme class from protein structure without alignments. *Journal of Molecular Biology*, 345(1):187–199, 2005.

[106] C.Z. Cai, L.Y. Han, Z.L. Ji, and Y.Z. Chen. Enzyme family classification by support vector machines. *Proteins*, 55(1):66–76, 2004.

[107] H. Strombergsson and G.J. Kleywegt. A chemogenomics view on protein-ligand spaces. *BMC Bioinformatics*, 10 Suppl 6:S13, 2009.

[108] S. Izrailev and M.A. Farnum. Enzyme classification by ligand binding. *Proteins*, 57(4):711–724, 2004.

[109] M. Scheer, A. Grote, A. Chang, I. Schomburg, C. Munaretto, M. Rother, C. Sohngen, M. Stelzer, J. Thiele, and D. Schomburg. BRENDA, the enzyme information system in 2011. *Nucleic Acids Research*, 39(Database issue):D670–6, 2011.

[110] L. Kauvar. Predicting ligand binding to proteins by affinity fingerprinting. *Chemistry & Biology*, 2(2):107–118, 1995.

[111] J.C. Hermann, R. Marti-Arbona, A.A. Fedorov, E. Fedorov, S.C. Almo, B.K. Shoichet, and F.M. Raushel. Structure-based activity prediction for an enzyme of unknown function. *Nature*, 448(7155):775–779, 2007.

[112] A. Gutteridge, G.J. Bartlett, and J.M. Thornton. Using a neural network and spatial clustering to predict the location of active sites in enzymes. *Journal of Molecular Biology*, 330(4):719–734, 2003.

[113] R.A. Laskowski, J.D. Watson, and J.M. Thornton. ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Research*, 33(Web Server issue):W89–93, 2005.

[114] D. Pal and D. Eisenberg. Inference of protein function from protein structure. *Structure (London, England: 1993)*, 13(1):121–130, 2005.

[115] N.V. Petrova and C.H. Wu. Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties. *BMC Bioinformatics*, 7:312, 2006.

[116] E. Youn, B. Peters, P. Radivojac, and S.D. Mooney. Evaluation of features for catalytic residue prediction in novel folds. *Protein Science: a Publication of the Protein Society*, 16(2):216–226, 2007.

[117] R. Alterovitz, A. Arvey, S. Sankararaman, C. Dallett, Y. Freund, and K. Sjolander. ResBoost: characterizing and predicting catalytic residues in enzymes. *BMC Bioinformatics*, 10:197, 2009.

[118] J.A. Gerlt, K.N. Allen, S.C. Almo, R.N. Armstrong, P.C. Babbitt, J.E. Cronan, D. Dunaway-Mariano, H.J. Imker, M.P. Jacobson, W. Minor, C.D. Poulter, F.M. Raushel, A. Sali, B.K. Shoichet, and J.V. Sweedler. The Enzyme Function Initiative. *Biochemistry*, 50(46):9950–9962, 2011.

[119] D. Devos and A. Valencia. Practical limits of function prediction. *Proteins*, 41(1):98–107, 2000.

[120] D.E.V. Pires, R.C. de Melo-Minardi, M.A. dos Santos, C.H. da Silveira, M.M. Santoro, and W.J. Meira. Cutoff Scanning Matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns. *BMC Genomics*, 12 Suppl 4:S12, 2011.

[121] B. Rost. Enzyme function less conserved than anticipated. *Journal of Molecular Biology*, 318(2):595–608, 2002.

[122] M.Y. Galperin, D.R. Walker, and E.V. Koonin. Analogous enzymes: independent inventions in enzyme evolution. *Genome Research*, 8(8):779–790, 1998.

[123] E. Kellenberger, C. Schalon, and D. Rognan. How to measure the similarity between protein ligand-binding sites? *Current Computer-Aided Drug Design*, 4(3):209–220, 2008.

[124] M. Jambon, A. Imberty, G. Deleage, and C. Geourjon. A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins*, 52(2):137–145, 2003.

[125] A. Brakoulias and R.M. Jackson. Towards a structural classification of phosphate binding sites in protein-nucleotide complexes: an automated all-against-all structural comparison using geometric matching. *Proteins*, 56(2):250–260, 2004.

[126] R. Minai, Y. Matsuo, H. Onuki, and H. Hirota. Method for comparing the structures of protein ligand-binding sites and application for predicting protein-drug interactions. *Proteins*, 72(1):367–381, 2008.

[127] A. Shulman-Peleg, R. Nussinov, and H.J. Wolfson. Recognition of functional sites in protein structures. *Journal of Molecular Biology*, 339(3):607–633, 2004.

[128] M. Milik, S. Szalma, and K.A. Olszewski. Common Structural Cliques: a tool for protein structure and function analysis. *Protein Engineering*, 16(8):543–552, 2003.

[129] S. Schmitt, D. Kuhn, and G. Klebe. A new method to detect related function among proteins independent of sequence and fold homology. *Journal of Molecular Biology*, 323(2):387–406, 2002.

[130] D. Kuhn, N. Weskamp, S. Schmitt, E. Huellermeier, and G. Klebe. From the similarity analysis of protein cavities to the functional classification of protein families using cavbase. *Journal of Molecular Biology*, 359(4):1023–1044, 2006.

[131] K. Kinoshita, J. Furui, and H. Nakamura. Identification of protein functions from a molecular surface database, eF-site. *Journal of Structural and Functional Genomics*, 2(1):9–22, 2002.

[132] K. Kinoshita, Y. Murakami, and H. Nakamura. eF-seek: prediction of the functional sites of proteins by searching for similar electrostatic potential and molecular surface shape. *Nucleic Acids Research*, 35(Web Server issue):W398–402, 2007.

[133] R. Najmanovich, N. Kurbatova, and J. Thornton. Detection of 3D atomic similarities and their use in the discrimination of small molecule protein-binding sites. *Bioinformatics (Oxford, England)*, 24(16):i105–11, 2008.

[134] J.S. Fetrow, A. Godzik, and J. Skolnick. Functional analysis of the Escherichia coli genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. *Journal of Molecular Biology*, 282(4):703–711, 1998.

[135] J.S. Fetrow. Active site profiling to identify protein functional sites in sequences and structures using the Deacon Active Site Profiler (DASP). *Current Protocols in Bioinformatics*, Chapter 8:Unit 8.10, 2006.

[136] K. Yeturu and N. Chandra. PocketMatch: a new algorithm to compare binding sites in protein structures. *BMC Bioinformatics*, 9:543, 2008.

[137] T.A. Binkowski and A. Joachimiak. Protein functional surfaces: global shape matching and local spatial alignments of ligand binding sites. *BMC Structural Biology*, 8:45, 2008.

[138] B. Xiong, J. Wu, D.L. Burk, M. Xue, H. Jiang, and J. Shen. BSSF: a fingerprint based ultrafast binding site similarity search and function analysis server. *BMC Bioinformatics*, 11:47, 2010.

[139] S. Das, A. Kokardekar, and C.M. Breneman. Rapid comparison of protein binding site surfaces with property encoded shape distributions. *Journal of Chemical Information and Modeling*, 49(12):2863–2872, 2009.

[140] M. Baroni, G. Cruciani, S. Sciabola, F. Perruccio, and J.S. Mason. A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for Ligands and Proteins (FLAP): theory and application. *Journal of Chemical Information and Modeling*, 47(2):279–294, 2007.

[141] C. Schalon, J.-S. Surgand, E. Kellenberger, and D. Rognan. A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins*, 71(4):1755–1778, 2008.

[142] N. Weill and D. Rognan. Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites. *Journal of Chemical Information and Modeling*, 50(1):123–135, 2010.

[143] R.J. Morris, R.J. Najmanovich, A. Kahraman, and J.M. Thornton. Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. *Bioinformatics (Oxford, England)*, 21(10):2347–2355, 2005.

[144] I. Merelli, P. Cozzi, D. D'Agostino, A. Clematis, and L. Milanesi. Image-based surface matching algorithm oriented to structural biology. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(4):1004–1016, 2011.

[145] B. Pang, N. Zhao, D. Korkin, and C.-R. Shyu. Fast protein binding site comparisons using visual words representation. *Bioinformatics (Oxford, England)*, 28(10):1345–1352, 2012.

[146] J. Ito, Y. Tabei, K. Shimizu, K. Tomii, and K. Tsuda. PDB-scale analysis of known and putative ligand-binding sites with structural sketches. *Proteins*, 80(3):747–763, 2012.

[147] Z. Aung and J.C. Tong. BSAlign: a rapid graph-based algorithm for detecting ligand-binding sites in protein structures. *Genome Informatics. International Conference on Genome Informatics*, 21:65–76, 2008.

[148] J. Desaphy, K. Azdimousa, E. Kellenberger, and D. Rognan. Comparison and druggability prediction of protein-ligand binding sites from pharmacophore-annotated cavity shapes. *Journal of Chemical Information and Modeling*, 52(8):2287–2299, 2012.

[149] D. Marr and E. Hildreth. Theory of edge detection. *JProceedings of the Royal Society of London. Series B, Biological Sciences*, 207(1167):187–217, 1980.

[150] M. Wirth and W.H.B. Sauer. Bioactive molecules: Perfectly shaped for their target? *Molecular Informatics*, 30:677–688, 2011.

[151] W.H.B. Sauer and M.K. Schwarz. Molecular shape diversity of combinatorial libraries: a prerequisite for broad bioactivity. *Journal of Chemical Information and Computer Sciences*, 43(3):987–1003, 2003.

[152] G. M. Downs and J. M. Barnard. Clustering methods and their uses in computational chemistry. *Reviews in Computational Chemistry*, 18:1–40, 2003.

[153] S. C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 2:241–254, 1967.

[154] 2.02 FTrees. Biosolve IT GmbH: An der Ziegelei 75, 53757 St. Augustin, Germany.

[155] T. T. Tanimoto. An elementary mathematical theory of classification and prediction. *IBM Internal Report*, 1958.

[156] P. Willett, J.M. Barnard, and G.M. Downs. Chemical similarity searching. *Journal of Chemical Information and Computer Sciences*, 38(6):983–996, 1998.

[157] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:1–27, 2011.

[158] M. Ahdesmaki, V. Zuber, and K. Strimmer. *sda: Shrinkage Discriminant Analysis and CAT Score Variable Selection*, 2011. R package version 1.2.0.

[159] S. Cha and S. Srihari. On measuring the distance between histograms. *Pattern Recognition*, 35(6):1355–1370, 2002.

[160] J. Schlosser and M. Rarey. Beyond the virtual screening paradigm: Structure-based searching for new lead compounds. *Journal of Chemical Information and Modeling*, 49(4):800–809, 2009.

[161] R. Wang, X. Fang, Y. Lu, and S. Wang. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *Journal of Medicinal Chemistry*, 47(12):2977–2980, 2004.

[162] E. Kellenberger, P. Muller, C. Schalon, G. Bret, N. Foata, and D. Rognan. sc-PDB: an annotated database of druggable binding sites from the protein data bank. *Journal of Chemical Information and Modeling*, 46(2):717–727, 2006.

[163] Webb E.C. Enzyme nomenclature 1992: recommendations of the nomenclature committee of the international union of biochemistry and molecular biology on the nomenclature and classification of enzymes. *San Diego: Published for the International Union of Biochemistry and Molecular Biology by Academic Press.*, ISBN 0-12-227164-5, 1992.

[164] E. Krissinel and K. Henrick. Protein structure comparison service fold at european bioinformatics institute. Accessed July 9, 2012, from `http://www.ebi.ac.uk/msd-srv/ssm`.

[165] T.I. Zarembinski, L.W. Hung, H.J. Mueller-Dieckmann, K.K. Kim, H. Yokota, R. Kim, and S.H. Kim. Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics. *Proceedings of the National Academy of Sciences of the United States of America*, 95(26):15189–15193, 1998.

[166] S. Sonavane and P. Chakrabarti. Cavities and atomic packing in protein structures and interfaces. *PLoS Computational Biology*, 4(9):e1000188, 2008.

[167] A. Kahraman, R.J. Morris, R.A. Laskowski, and J.M. Thornton. Shape variation in protein binding pockets and their ligands. *Journal of Molecular Biology*, 368(1):283–301, 2007.

[168] I. D. Kuntz, K. Chen, K. A. Sharp, and P. A. Kollman. The maximal affinity of ligands. *Proceedings of the National Academy of Sciences of the United States of America*, 96(18):9997–10002, September 1999.

[169] C.H. Reynolds, B.A. Tounge, and S.D. Bembenek. Ligand binding efficiency: trends, physical basis, and implications. *Journal of Medicinal Chemistry*, 51(8):2432–8, May 2008.

[170] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J. P. Overington, 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 Research*, 40(D1):D1100–D1107, 2011.

[171] R.A. Friesner, J.L. Banks, R.B. Murphy, T.A. Halgren, J.J. Klicic, D.T. Mainz, M.P. Repasky, E.H. Knoll, M. Shelley, J.K. Perry, D.E. Shaw, P. Francis, and P.S. Shenkin. Glide: A new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *Journal of Medicinal Chemistry*, 47(7):1739–1749, 2004.

[172] S. Gerstenbruch, H. Wulf, N. Mussmann, T. O'Connell, K.H. Maurer, and U.T. Bornscheuer. Asymmetric synthesis of D-glyceric acid by an alditol oxidase and directed evolution for enhanced oxidative activity towards glycerol. *Applied Microbiology and Biotechnology*, epub ahead of print, 2012.

# Bibliography of this Dissertation's Publications

[D1] A. Volkamer, A. Griewel, T. Grombacher, and M. Rarey. Analyzing the topology of active sites: On the prediction of pockets and subpockets. *Journal of Chemical Information and Modeling*, 50(11):2041–2052, 2010.

[D2] A. Volkamer, D. Kuhn, T. Grombacher, F. Rippmann, and M. Rarey. Combining global and local measures for structure-based druggability predictions. *Journal of Chemical Information and Modeling*, 52(2):360–372, 2012.

[D3] A. Volkamer, D. Kuhn, F. Rippmann, and M. Rarey. DoGSiteScorer: A web server for automatic binding site prediction, analysis, and druggability assessment. *Bioinformatics*, 28(15):2074–75, 2012.

[D4] A. Volkamer, D. Kuhn, F. Rippmann, and M. Rarey. Predicting enzymatic function from global binding site descriptors. *Proteins: Structure, Function and Bioinformatics*, 81(3):479–89, 2013.

[D5] M. v. 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*, Accepted January 2013.

[D6] M. Wirth, A. Volkamer, F. Rippmann, V. Zoete, O. Michielin, M. Rarey, and W. H. B. Sauer. Protein pocket and ligand shape comparison and its application in virtual screening. To be submitted January 2013.

# Appendix A

# Publications and conference contributions

## A.1 Publications in scientific journals

This section summarizes the publications of the author in scientific journals and explains the authors' contributions.

D1 **A. Volkamer**, A. Griewel, T. Grombacher, and M. Rarey. Analyzing the topology of active sites: On the prediction of pockets and subpockets. Journal of Chemical Information and Modeling, 50(11):2041-2052, 2010.
Based on a Difference of Gaussian filter implemented by A. Griewel, a new method for pocket detection was designed. The author of this thesis, A. Volkamer, developed the novel DoGSite method and performed the validation studies listed in the paper. T. Grombacher and M. Rarey supervised this work.

D2 **A. Volkamer**, D. Kuhn, T. Grombacher, F. Rippmann, and M. Rarey. Combining global and local measures for structure-based druggability predictions. Journal of Chemical Information and Modeling, 52(2):360-372, 2012.
On the basis of the developed method for pocket detection (D1), the author of this thesis, A. Volkamer, designed and implemented the new approach for druggability prediction DoGSiteScorer. D. Kuhn assisted the evaluation studies.
T. Grombacher, F. Rippmann and M. Rarey supervised the work.

D3 **A. Volkamer**, D. Kuhn, F. Rippmann, and M. Rarey. DoGSiteScorer: A web server for automatic binding site prediction, analysis, and druggability assessment. Bioinformatics, 28(15):2074-75, 2012.
Based on the developed method for druggability prediction (D2), the author of this thesis, A. Volkamer, set up a web server for pocket detection, description and druggability prediction of new protein structures. D. Kuhn, F. Rippmann and M. Rarey supervised this work.

D4 **A. Volkamer**, D. Kuhn, F. Rippmann, and M. Rarey. Predicting enzymatic function from global binding site descriptors. Proteins: Structure, Function and Bioinformatics, 81(3):479-89, 2013.
The author of this thesis, A. Volkamer, designed the method, collected the data sets and performed the studies for enzymatic function prediction. In collaboration with D. Kuhn the test sets listed in the paper were evaluated. F. Rippmann and M. Rarey supervised the study.

D5 M. v. 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, accepted January 2013.
M. v. Behren and A. Volkamer jointly established the idea of the new binding site comparison method TrixP. M. v. Behren implemented the method, assisted by A. Volkamer, the author of this thesis. Together, they designed and performed the evaluation studies. Fundamental TrixX functionality was provided by A. M. Henzler, K. T. Schomburg, S. Urbaczek. M. Rarey supervised this work.

D6 M. Wirth, **A. Volkamer**, V. Zoete, F. Rippmann, O. Michielin, M. Rarey, and W. H. B. Sauer. Protein pocket and ligand shape comparison and its application in virtual screening. To be submitted January 2013.
The idea of the pocket-ligand shape comparison studies evolved from a collaboration between M. Wirth and A. Volkamer, the author of this thesis. The used moment-of-inertia shape representation stems from a previous work of M. Wirth, the pocket prediction and description from the previous work of A. Volkamer (D1, D2). V. Zoete, F. Rippmann, O. Michielin, M. Rarey, and W. H. B. Sauer supervised the study.

## A.2 Conferences

This section first lists the author's oral presentations and finishes with the presented posters.

### A.2.1 Talks

1. **A. Volkamer**, T. Grombacher, F. Rippmann, C. Lemmen and M. Rarey. COMPASITES: A Tool for Automated Active-Site Based Structure-Function Analysis. Biocat2010, Hamburg, 29. August-2. September 2010

2. **A. Volkamer**, A. Griewel, T. Grombacher, F. Rippmann and M. Rarey. Automatically Predicted Subpockets Pave the Way for Descriptor-based Druggability Studies. EuroQSAR, Rhodes, 19.-24. September 2010
   → Short presentation due to poster price

3. **A. Volkamer**, A. Griewel, T. Grombacher, F. Rippmann and M. Rarey. Combining global and local measures for druggability predictions. 9th International Conference on Chemical Structures (ICCS), Noordwijkerhout, June 5.-9. 2011

4. **A. Volkamer**, T. Watolla, F. Sonnenburg, C. Lemmen, D. Kuhn, F. Rippmann and M. Rarey. Combining Automatic Active Site Analysis and Docking for Structure-Based Protein Function Prediction. Dechema 7th Status Seminar Chemical Biology, Frankfurt, 5.-6. December 2011

5. **A. Volkamer**, D. Kuhn, F. Rippmann and M. Rarey. Exhaustive Computer-Aided Active Site Analysis for Structure-Based Protein Function and Druggability Predictions. 10th Swiss Snow Symposium, Lenk, Swiss, February 2012

6. **A. Volkamer**, D. Kuhn, F. Rippmann and M. Rarey. Combining global and local measures for structure-based druggability predictions. ACS Spring Meeting, San Diego, California, 25.-29. March 2012

7. **A. Volkamer**, T. Watolla, F. Sonnenburg, C. Lemmen, D. Kuhn, F. Rippmann and M. Rarey. Combining automatic active site analysis and docking for structure-based protein function predictions. ACS Spring Meeting, San Diego, California, 25.-29. March 2012

### A.2.2 Posters

1. **A. Volkamer**, A. Griewel, T. Grombacher and M. Rarey. Where are the boundaries? Automated Pocket Detection for Druggability Studies. German Conference on Cheminformatics, Goslar, 8.-10. November 2009
   → Poster price

2. **A. Volkamer**, A. Griewel, T. Grombacher and M. Rarey. Pockets are made of Sub-pockets: Automated Detection of Ligand Binding Sites for Structure-based Function Analysis and Druggability Studies. CHI - Structure-Based Drug Design, Boston, 23.-25. June 2010

3. **A. Volkamer**, A. Griewel, T. Grombacher, F. Rippmann and M. Rarey. Automatically Predicted Subpockets Pave the Way for Descriptor-based Druggability Studies. EuroQSAR, Rhodes, 19.-24. September 2010
   → Poster price

# Appendix B

# Working with COMPASITES

The algorithms developed in my work are implemented in a software named COMPASITES (Computer-aided Active Site analysis). In the following, a short tutorial about the usage of the software will be given. The software can be applied in several use cases:

- Prediction of potential binding site(s) of protein structures
- Calculation of various descriptors of the active site
- Clustering of active sites based on the descriptors
- Annotation of druggability (using the SVM-based method)

The program is embedded into the FlexX Toolkit. Therefore, in this section only the functionality of the COMPASITES functions is listed. For the detailed basic FlexX functionality, please refer to the FlexX user guide (available at `http://www.biosolveit.de`).

## B.1 Starting COMPASITES

Please ensure that you have access to the config.dat, static data and a valid license key before starting the program. The COMPASITES program can be started from command line by typing `./bin/CompaSite`. If paths and license are set correctly, the program starts and the prompt `FLEXX>` appears. By pressing return you get an overview about the provided functionality. Generally, different output levels can be chosen for user info. Depending on the level of detail of the process that the user wants to see, the user can change the granularity by typing `SET VERBOSITY`.

## B.2 Menu Navigation

In the following, it is explained which steps are needed to address the various use cases, starting from the FLEXX>-prompt. Each functionality is provided in a separate submenu.

### B.2.1 Load a protein - RECEPTOR

To load a protein, change into the receptor menu (type RECEPTOR ) and read in a PDB structure of your desired protein (READ [path/]protein.pdb). Make sure that you either specify the complete path to the file of the protein structure or that the path is set in your configuration file.

### B.2.2 Load a ligand - LIGAND

This step is optional, if you have a co-crystallized ligand and you want to include it for binding site revision, you can load the ligand by changing into the ligand menu (LIGAND) and reading in the ligand (READ [path/]ligand[.mol2]). In this step, only one molecule at a time (in mol2 or sdf format) can be read in. For input of multi-mol2 ligand files, containing multiple ligands, please refer to A.2.3.2.

### B.2.3 Pocket detection and analysis - COMPASITE

To start the prediction functionality, change into the COMPASITES submenu (COMPASITE). The COMPASITE submenu holds the commands for pocket detection (POCKET), pocket and ligand coverage calculation (LIG_CHECK), descriptor calculation (DESCRIPT), drawing (DRAW) and output of the results (WRITE_PDB).

#### B.2.3.1 Predict binding pockets - (M)POCKET

- POCKET
  With the command POCKET, the pocket prediction can be started. Per default the complete protein structure is analyzed. If the monomeric prediction option is chosen (SET COMPAS_IP_MONOMER 1), additionally the chain for which the pocket shall be predicted has to be chosen (for example: POCKET A), otherwise the first chain is chosen per default. If the user output level is set to zero, the program only outputs the number and the size of the predicted pockets. If the output level is set to 4 (SET VERBOSITY 4), you get detailed information about the single steps:
  a. Grid size

b. Energy, buriedness or DoG calculation (based on chosen algorithm)
c. Subpocket calculation

- MPOCKET
  This command enables the annotation of a pocket around a user provided ligand (`MPOCKET [path/]lig.mol2`). This procedure has been implemented to allow the user to use the complete `COMPASITES` functionality for the pocket of interest.

#### B.2.3.2 Pocket and ligand coverage calculation - LIG_CHECK

This command requires the input of a ligand (`LIG_CHECK [path/]lig.mol2`). Firstly, all ligands specified in the input file are checked one after the other if they lie in the predicted binding pocket and secondly, the ligand and pocket coverage is calculated. Nevertheless, the ligands are not globally stored in the program (differing from the `READ LIGAND` option shown in A.1.2.2) and cannot be further used. If you want to visualize the ligand change back into the ligand menu and read in the ligand. Note that the coverage calculation is only possible if pockets have been calculated before, the program outputs a warning if no pockets have been predicted so far.

#### B.2.3.3 Calculate binding pocket descriptors - DESCRIPT

After choosing the `DESCRIPT` command, all descriptors are calculated and displayed on the screen. If you want to send the output to a file, use the FlexX `SELOUTP` command and specify the name of your file. All calculated descriptors are output in form of a table, one for the predicted pockets and one for sub pockets, respectively. This description already contains a measure for druggability annotated as *simple score*, which is calculated based on linear combination of three pocket properties, describing size, compactness and hydrophobicity of the pocket. Herein, a short overview of the contained values is given:

- Volume and shape descriptors:
  *poc*: pocket number; s_poc: subpocket number (poc_spoc)
  *score*: simple druggability score based on linear combination of three properties volume, se_/h_gps, and siac_ratio
  *lig*: ligand number, if multiple ligands are loaded (-1 if no ligand found)
  *lig_cov*: percentage of ligand covered by the predicted pocket
  *poc_cov*: percentage of the pocket covered by the provided ligand
  *B/SE*: buried (=1) or solvent exposed (=0) pocket

  *se_cl*: number of solvent exposed clusters, corresponds to the number of solvent exposed sites of the pocket
  *energy*: total energy of the pocket, summed up over all grid points (only calculated if energy based pocket detection is enabled)
  *#s_atms*: number of surface atoms lining the pocket boarder
  *depth*: depth of the pocket in Å
  *volume*: pocket volume in Å$^{3}$
  *surface*: solvent accessible pocket surface in Å$^{2}$
  *lipo_surf*: lipophilic surface
  *nof_gps*: number of grid points contained in the pocket
  *nof_surf_gps*: number of surface grid points in Å$^{2}$
  *se_/v_gps*: ratio of number of solvent exposed and total number of grid points
  *se_/h_gps*: ratio of number of solvent exposed and surface grid points
  *s_/v_gps*: ratio of number of volume and surface grid points
  *ell_vol*: quotient of fitted ellipsoid to pocket volume
  *ell_a, ell_b, ell_b*: ellipsoid main axis, with $a > b > c$

- Amino acid and element composition descriptors:
  *chain*: number of chains pocket belongs to
  *element type*: C, N, O, S or other (X)
  *ALA, ASN,...*: 3-letter code of 20 amino acid types
  *H-don*: number of hydrogen bond donors
  *H-acc*: number of hydrogen bond acceptors
  *Met*: number of metals
  *Hphob*: number of hydrophobic contacts
  *ratio*: relative number of hydrophobic SIACs (Site InterAction Centers)
  *aa_apol/pol/pos/neg*: relative number of apolar, polar, positive, and negative amino acids

This description can also be printed on the screen using the command `DESC_INFO`.

#### B.2.3.4 Automated process - AUTO

By typing `AUTO [ligand]`, the program automatically carries out the actions `POCKET`, `LIG_CHECK`, and `DESCRIPT`. Please ensure that you provide a ligand, which is needed for the `LIG_CHECK` step.

#### B.2.3.5 Visualization - DRAW

In this submenu context, the `DRAW` routine is only suitable to draw the detected pockets. For drawing receptor and ligand please change into the respective menus and refer to the FlexX user guide. For pocket visualization, you can choose between different coloring schemes and presentations. You are asked line by line what you want to have drawn.
First you can choose which pockets you want to draw, remember pockets are sorted by size. Second, you can choose if you want to draw the pocket in the margin cube presentation or in the grid point presentation. Either pockets, or their division into subpockets can be annotated by different coloring. In the grid point mode, the coloring scheme can be further defined by several pocket properties. Pockets can be colored to represent the shell, buriedness (psp), energy, and tightness of the environment. Furthermore, you can enable the drawing of the underlying grid density values (DoG) in mcube representation. If the DoG grid is drawn, the user has the possibility to change the DoGSite cut-off parameter interactively. This can be done by using the slider in the FlexV menu (FlexV is the graphic user interface opened after the drawing dialog is finished in the prompt). Only values below zero give reasonable results. The closer the value is to zero the larger is the grid cloud, the lower the value, the more restrictive is the pocket algorithm. This allows for inspection of the cores of detected pockets.
By typing `GO` the pocket information is send to FlexV and a new window appears containing the predicted pockets.

#### B.2.3.6 Clear data - DEL_POC and DEL_DESC

These two commands can be used to manually delete the specific information. Nevertheless, whenever new pockets are predicted (`POCKET`) or new descriptors are calculated (`DESCRIPT`), the deletion of the old information is done automatically by the program.

#### B.2.3.7 Output - WRITE_PDB, WRITE_DESC, WRITE_SVM

These commands can be used to write the generated output to user defined files.

- `WRITE_PDB`
  This command can be used to write PDB files of the predicted pockets or binding sites. You have to provide the name of the output file (without file ending ".pdb") and the number of pockets you want to output. Next, you can choose between two output variations, either the pocket volume in terms of grid points (0) or the binding site residues (1); the choice will be written to the file. The

program generates one output file per predicted pocket and attaches the pocket number to the filename (e.g. 4dfr_P0.pdb, 4dfr_P1.pdb,...), respectively one for each subpocket if chosen (e.g. 4dfr_P0SP0.pdb, ...).

- *Pocket output*: Due to the fact that the pockets may be large, only the surface of the pockets is output. Furthermore, the option of a reduced surface can be enabled, which prints a less granular grid representation of the pocket surface.
  Each line in the PDB file specifies one pocket surface grid point. The regular PDB format is used and each line consists of the following entries: *ATOM, gp_nr, atm, resName, chainID, s_poc_nr, x, y, z, spec_value.* Explanation: *gp_nr* specifies the number of the grid point in the pocket (note: the numbering is not continuous, since only surface grid points of the pocket are output); *atm, resName and chainID* are set per default to CA, ALA and A respective; *s_poc_nr* specifies the subpocket to which the grid point belongs and x, y and z are the coordinates of the grid point. Spec_value specifies the buriedness (0), the vdW_energy (1) or the dog_value (2) of the grid point, depending on the pocket detection algorithm that is used.
- *Binding Site output*: If this option is chosen, all pocket lining binding site residues are printed to the user specified file in PDB format.

- `WRITE_DESC`
  This command writes the descriptor information to a user defined output file. Again, you can specify, if descriptors for pockets, subpockets or both representations shell be written to the file.
  The counter command `READ_DESC` can be used to reenter descriptors to the program, which can be of interest when using the clustering procedure, explained in A.2.4.

- `WRITE_SVM`
  This command can be used to store the descriptor data in the libsvm required format for druggability or function predictions. For further instructions concerning druggability prediction read section A.2.5.

### B.2.4 Clustering by binding site descriptors - CLUSTER

To compare proteins by pocket descriptors and to group them into families, an adapted version of the FlexX `CLUSTER` command is provided. Note that you either have to pre-process your set of structures (e.g, with the `COMPASITES` functionality `AUTO`), which will make the descriptors globally available for the program, or you can reenter precalculated descriptor information using the READ_DESC command. After changing into the submenu CLUSTER, the AUTO command can be used to cluster the set of input structures by their distance in descriptor space. Per default, a complete linkage clustering is performed, but the user can change this option to single linkage as well. Finally, using the command DENDRO writes a dendrogram file, which holds the information to graphically display the generated clusters.

### B.2.5 Druggability prediction

For druggability predictions, the software version must provide a folder containing the support vector machine (SVM) executable of libsvm, as well as the pre-trained SVM-models. Furthermore, you first need to store the descriptors of your query protein in the libsvm format by using WRITE_SVM (see A.2.3.7). Next, the script contained in the same directory, named DoGSiteScorer.sh, can be called with the following three parameters: Query descriptor file, SVM-model file and path to the SVM directory (all without suffix).
Caution: For pockets and subpockets two different SVM models have been trained. Make sure that your query and model levels are consistent. After submission, scores are calculated and output to a file. An example of this proceeding can be found in the scripting subsection.

### B.2.6 COMPASITES parameter setting - SET

The following COMPASITES parameters can be change by typing the name followed by the desired value, for example SET COMPAS_IP_POC_ALG 0. The value behind the specific parameter describes the default value; in brackets the options or the value range is specified. (IP stands for integer parameter, DP for floating point parameter)

- COMPAS_IP_POC_ALG 3 [0-3]: Defines the algorithm by which the pockets are calculated.
  0: Geometric-based algorithm (LIGSITE reimplementation)
  1: Energy-based algorithm (DrugSite reimplementation)
  2: Combination of 0 and 1 (Not maintained any longer)
  3: Edge-detection algorithm (DoGSite)

- COMPAS_IP_NEIGHBOR 5 [0-26]: Number of neighbors that a point must have to belong to the final predicted binding site point cluster. (Geometry-based method).

- `COMPAS_IP_VDW_CUTOFF 8 [4-12]`: Van der Waals cut-off in Å for energy calculation. Maximal distance between the carbon probe that is rolled over the surface and the surrounding atoms. (Energy-based method)
- `COMPAS_DP_NUM_FILTER 1.75 [1-15]`: Depending on the chosen algorithm (1,3): Energy-based method (1): Number of filter steps if average space filter is used; or value for sigma if Gaussian filter is used. If DoGSite (3) is chosen, it specifies the value of sigma for difference of Gaussian calculation.
- `COMPAS_IP_FILTER_TYPE 2 [1;2]`: (Energy-based method)
  1: Average space filter
  2: Gaussian filter
- `COMPAS_DP_GRID_DELTA 0.4 [0.4-1.2]`: Grid spacing.
- `COMPAS_DP_MAX_VDW -0.4 [-0.8-0.0]`: Van der Waals cut-off. All grid points with a higher value are re-set to the cut-off for the filter procedure.
- `COMPAS_DP_MAP_CUTOFF 3 [0.-4.]`: Threshold for calculation of the map contour level: $CL = Mean * Threshold - Standard\ deviation.$
- `COMPAS_IP_RANK 2 [0;2]`: Generally pockets are ranked by size (0), when DoGSite algorithm is used, ranking by a slightly different size criterion (2) performs better which only considers the original kernels of the pockets.
- `COMPAS_IP_MONOMER 0 [0;1]`: Defines the part of the structure to be analyzed.
  0: Complete protein structure (all chains) is used to predict the pockets
  1: Only one chain (which has to be specified, see 4.1) is used to predict the pockets
- `VERBOSITY 4 [0;3;4]`: Set Output level.
  0: No output
  3: Normal user output
  4: Detailed user output

## B.3 Scripting

Instead of typing each command by hand, you can use a script, containing the abovementioned options. Three examples are stored in the script folder and are attached to this section. The first script can be used to read in a protein, the co-crystallized ligand, perform the binding site prediction, get the descriptors and output the predicted pockets as PDB files. You can run it from the command line by typing `./bin/CompaSites -b example_single.bat`. Alternatively, it can be accessed during the program execution by typing `SCRIPT example_PA3_single`. The second script can be used to read in a list of protein structures specified in a text file (list.txt). The calculations are performed for all structures in the list and output to a new text file. The command is analog to the one for the single run. The third script describes the use of the druggability predictor.

- Script for processing a single protein.

```
#####################################################
# COMPASITES - binding site detection algorithm #
# written by Andrea Volkamer, 2012              #
# script: example_single.bat                    #
# Script to predict pockets and to calculate    #
# descriptors for protein structure 1c5q.pdb    #
# Additionally, the largest 3 pockets of the    #
# structure are output to a pdb file.           #
#####################################################

set verbosity 0
ligand
  selinit *
# menu receptor, read receptor
receptor
  read ./example/1c5q.pdb
# change to menu composite,
# to calulate pockets and descriptors
compasite
  pocket
  lig_check ./example/1c5q_H
  descript
  write_pdb 1c5q_pred 3
end
```

- Script for processing a list of proteins with subsequent clustering.

```
##################################################
# COMPASITES − binding site detection algorithm #
# written by Andrea Volkamer, 2012               #
# script: example_multi.bat                      #
# Script to predict pockets and to calculate     #
# descriptors for all protein structures         #
# contained in the file list.txt                 #
# Additionally, the largest 3 pockets of each    #
# structure are output to a pdb file.            #
# Pockets can be clustered using the cluster     #
# menu and a dot file with the clustering can    #
# be output.                                     #
##################################################

SETVAR $(pdbdir) ”./example/”
SETVAR $(workdir) ”./”
SETVAR $(ligdir) ”./example/”

seloutp $(workdir)/example_PA3_multi.txt a
set verbosity 3
  ligand
    selinit *
  FOR_EACH $(FILENAME) IN $(pdbdir)/list.txt
    # menu receptor, read receptor
    receptor
      read $(pdbdir)/$(FILENAME).pdb
      # change to menu compasite,
      # to calulate pockets and descriptors
    compasite
      pocket
      lig_check $(ligdir)/$(FILENAME)_H
      descript
      write_pdb $(FILENAME)_pred 3
      del_poc
    # delete ligand and receptor
    main
      delall y
  END_FOR
  cluster
    auto
    dendro
end
```

- Script for processing a single protein and calculate its druggability.

```
####################################################
# COMPASITES - binding site detection algorithm #
# written by Andrea Volkamer, 2012               #
# script: example_svm.bat                        #
# Script to predict pockets and to calculate     #
# descriptors for protein structure 4dfr.pdb     #
# Additionally, the druggability of the          #
# predicted pockets and subpockets is estimated.#
####################################################

#load receptor
receptor
  read tmp/4dfr.pdb
#change to compasite submenu
compasite
  # calculate pockets
  pocket
  # calculate ligand coverage
  lig_check tmp/4dfr.mol2
  #calculate descriptors
  descript
  #write svm file pocket
  write_svm 0 myLibsvmPoc.txt
#external call off svm script
!./svm/DoGSiteScorer.sh myLibsvmPoc svmModelPoc ./svm
  #write svm file subpocket
  write_svm 1 myLibsvmSpoc.txt
#external call off svm script
!./svm/DoGSiteScorer.sh myLibsvmSpoc svmModelSPoc4 ./svm
end
```

# Appendix C

# Software Architecture

The COMPASITES software has been implemented in C and is integrated in the FlexX library. Figure C.1 shows the structure of the software.

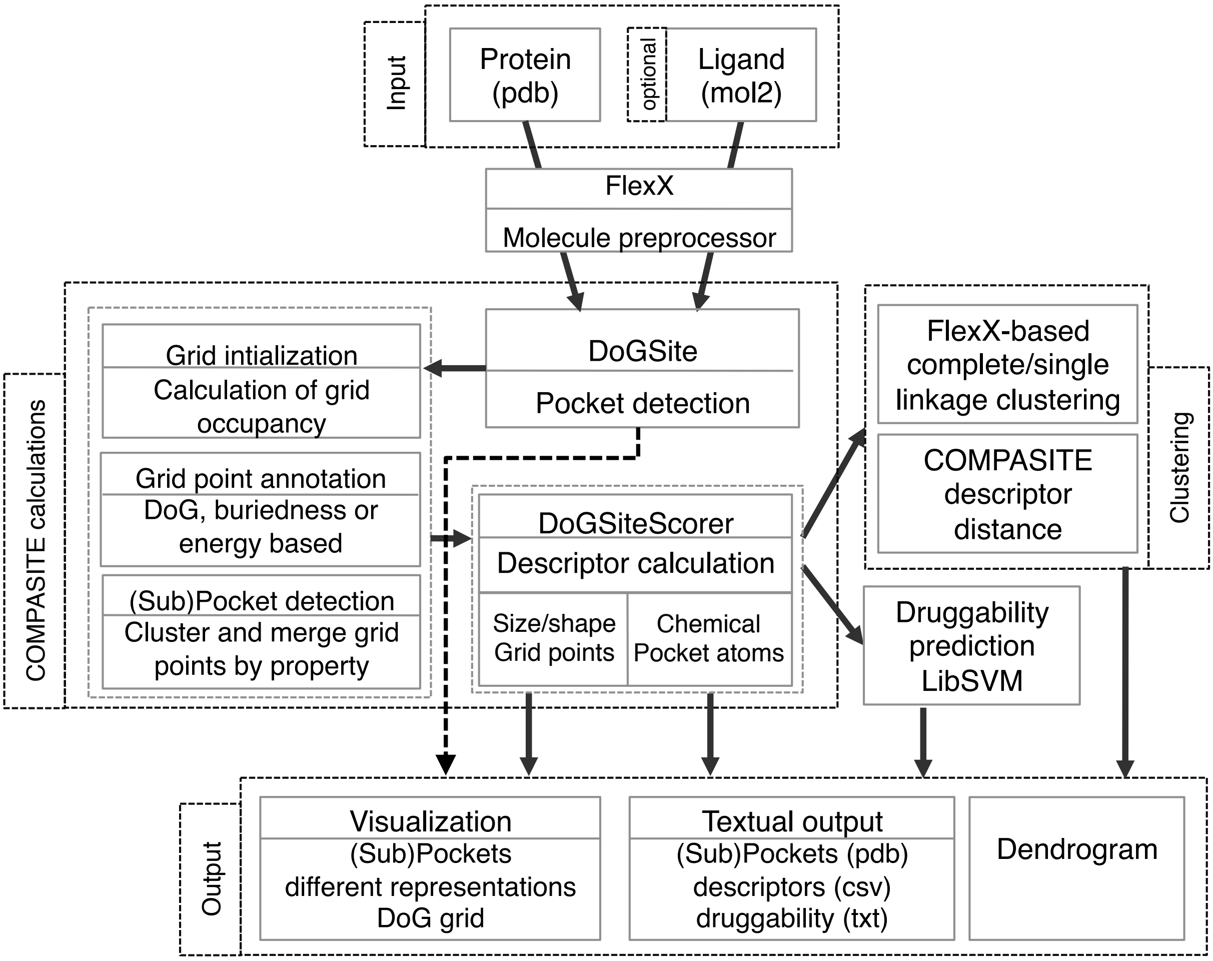


**Figure C.1:** COMPASITES software structure.