Methods for the Efficient Comparison of Protein Binding Sites and for the Assessment of Protein-Ligand Complexes

Dissertation

zur Erlangung des Doktorgrades
der Naturwissenschaften
(Dr. rer. nat.)

dem
Fachbereich Pharmazie der
Philipps-Universität Marburg
vorgelegt

von
Diplom-Bioinformatiker
Timo Krotzky
aus
Bad Hersfeld

Marburg/Lahn 2015

Erstgutachter Prof. Dr. Gerhard Klebe
Institut für Pharmazeutische Chemie
Philipps-Universität Marburg

Zweitgutachter Dr. Peter Kolb
Institut für Pharmazeutische Chemie
Philipps-Universität Marburg

Eingereicht am 8.6.2015

Tag der mündlichen Prüfung am 7.8.2015

Hochschulkennziffer: 1180

Die Untersuchungen zur vorliegenden Arbeit wurden auf Anregung von Herrn Prof. Dr. Gerhard Klebe am Institut für Pharmazeutische Chemie des Fachbereichs Pharmazie der Philipps-Universität Marburg in der Zeit von November 2010 bis Mai 2015 durchgeführt.

Meinen Eltern

Deutsche Zusammenfassung

In der vorliegenden Arbeit werden beschleunigte Verfahren für Proteinbindetaschenvergleiche sowie ein erweitertes Bewertungsverfahren für die Beurteilung von Ligandenposen in Proteinbindetaschen vorgestellt. Proteinbindetaschenvergleiche sind ein häufig verwendetes rezeptorbasiertes Verfahren im Frühstadium des Wirkstoffentwicklungsprozesses. Bindestellen anderer Proteine, die der Bindetasche des Zielproteins ähnlich sind, können so bereits vor klinischen Untersuchungen Rückschlüsse auf mögliche Nebenwirkungen des neuen Arzneistoffs zulassen. Darüber hinaus werden Bindetaschenvergleiche angewendet, um Ideen für den möglichen bioisosteren Ersatz einzelner funktioneller Gruppen des neu entwickelten Wirkstoffmoleküls zu erhalten sowie die Funktion bisher unklassifizierter Proteine aufzuklären. Der strukturelle Vergleich von Bindetaschen empfiehlt sich besonders für nur entfernt verwandte Proteine, da hier ein reiner Vergleich auf Ebene der Aminosäuresequenz häufig nicht zielführend ist.

Bewertungsverfahren für Ligandenposen in Proteinbindetaschen werden ebenfalls in der Frühphase der Wirkstoffentwicklung innerhalb sogenannter *Docking*-Programme eingesetzt. Mit ihrer Hilfe versucht man zum einen zu klären, welcher Ligand aus einer ganzen Bibliothek von Molekülen für eine bestimmte Bindetasche am besten geeignet ist und zum anderen, in welcher Konformation sich der Ligand wahrscheinlich in der Bindetasche platziert. Mithilfe dieser Informationen können die Molekülbibliotheken für nachfolgende Affinitätstests vorgefiltert sowie deren molekulare Strukturen hinsichtlich

Affinität und Selektivität optimiert werden.

Im ersten Kapitel dieser Arbeit wird der Einfluss von verschiedenen Methoden der Bindetaschendetektion auf die damit erhaltenen Validierungsdatensätze untersucht. Methoden zum Vergleich von Proteinbindetaschen werden häufig anhand von Datensätzen validiert, die durch ein einfaches Ausschneiden von Proteinatomen um einen gebundenen Ligand erhalten wurden. Dies führt zum einen dazu, dass alle unbesetzten Taschen ignoriert werden und zum anderen, dass eine beachtliche Menge an Informationen über die Gestalt des Liganden auf die ausgeschnittene Tasche übertragen wird. Im Folgenden gestalten sich verschiedene Experimente wie etwa Klassifizierungsexperimente von Taschen, die bestimmte Liganden binden, äußerst trivial und haben nur eine geringe Aussagekraft in Bezug auf die Qualität eines Bindetaschenvergleichsverfahrens. Es werden daher Ergebnisse eines sehr einfachen und stark formbelasteten Verfahrens gezeigt, die belegen, dass eine automatisierte Bindetaschendetektion unbeeinflusst von Ligandeninformationen essenziell ist, um eine zufriedenstellende Vorhersage von möglichen Kreuzreaktivitäten und die Funktionszuweisung von bislang unklassifizierten Proteinen zu ermöglichen.

Als eine erste beschleunigte Bindetaschenvergleichsmethode wird anschließend das Programm LC präsentiert. Um Ähnlichkeiten zwischen molekularen Strukturen zu bestimmen, wird häufig eine Berechnung des größten gemeinsamen Teilgraphen angewendet. Für den Vergleich von Proteinbindetaschen gestaltet sich dieses Vorgehen vor allem deshalb problematisch, da Graphen, die Bindetaschen auf eine detaillierte Weise repräsentieren, sehr groß werden können. In Kombination mit einem NP-schweren Problem wie der Berechnung des größten gemeinsamen Teilgraphen führen große Graphen daraufhin zu einer sehr rechenintensiven Aufgabe. Aus diesem Grund wird für die Repräsentation von Bindetaschen ein gröberes Modell verwendet, das auf sogenannten Pseudozentren basiert. Dies führt allerdings auch zu einem Verlust an strukturellen Informationen, da viele einzelne Atome verworfen werden und keine Information über die Oberflächenform der Bindetasche erhalten bleibt. Im

Cavbase-Modul des Systems Relibase+ wird versucht, dies durch weitere nachfolgende Berechnungen zu kompensieren, die auf zusätzlichen Informationen für die Oberflächenform basieren. Die Gesamtheit dieser Berechnung wird damit sehr aufwendig, was zu einer sehr hohen Gesamtlaufzeit führt. Es wird daher eine neue und effizientere Modellierung vorgeschlagen, die die Größe des Graphenmodells nicht verändert, jedoch deutlich mehr Informationen in den Knoten ablegt als im ursprünglichen Ansatz. So werden zusätzliche Deskriptoren eingefügt, die den Knoten Informationen über die lokale Oberflächenbeschaffenheit hinzufügen. Dies erlaubt einen deutlich schnelleren und dennoch sehr genauen Bindetaschenvergleich.

Basierend auf LC wird im Folgenden die Erweiterung DivLC vorgestellt, in der eine weitere Beschleunigung durch die Verwendung von Graphpartitionierungen erreicht wird. Beide Graphen, die die Bindetaschen für einen Vergleich repräsentieren, werden hierbei vor dem Vergleich in disjunkte Komponenten zerlegt. Die Menge der Pseudozentren wird dafür bezüglich ihrer physikochemischen Eigenschaften zerlegt, was zu sieben sehr viel kleineren Graphen führt. Angewendet auf dieselben Testszenarien wie die LC-Methode führt dieser Ansatz zu einer nochmals deutlich kürzeren Laufzeit, ohne erkennbar an Genauigkeit zu verlieren.

Als dritte und letzte alternative Bindetaschenvergleichsmethode wird schließlich das Programm RAPMAD vorgeschlagen, das hocheffiziente Vergleiche von einzelnen Bindetaschen gegen die komplette Cavbase-Datenbank ermöglicht. Die Proteinbindetaschen werden hierbei als eine Menge von Distanzhistogrammen dargestellt, die sowohl mit linearer Komplexität erzeugt als auch verglichen werden können. Die Effektivität des Verfahrens und die sehr kurze Laufzeit werden in verschiedenen Klassifizierungs- und Abfrageexperimenten demonstriert. Dabei erreicht RAPMAD ähnliche oder sogar höhere Erfolgsraten als der derzeit in Cavbase implementierte Algorithmus sowie andere bisher präsentierte Alternativen, wobei es nur einen Bruchteil deren Laufzeit benötigt. Der praktische Nutzen der Methode wird letztlich anhand einer erfolgreichen prospektiven virtuellen Screening-Studie belegt, die die Identifikation von neuen Inhibitoren des NMDA-Rezeptors anstrebt.

Als Abschluss der Arbeit wird eine Erweiterung des Programms *DSX* vorgestellt, einem Bewertungsverfahren von Ligandenposen in Proteinbindetaschen. Durch die Einbeziehung der Bewertung von Wasserstoffbrückengeometrien konnte eine Verbesserung des Programms mit nur geringen zusätzlichen Laufzeitkosten erreicht werden. Die Erweiterung wurde auf etablierten Testdatensätzen untersucht, was einen umfassenden Vergleich sowohl mit der vorangegangenen Version als auch mit einer Vielzahl von anderen bisher entwickelten Verfahren ermöglicht.

Contents

List of Abbreviations

Å	Ångström (1 Å = 10^{-10} m = 100 pm)
ADR	Adverse drug reaction
AUC	Area under curve
ATD	Amino terminal domain
ATP	Adenosine triphosphate
BK	Bron-Kerbosch algorithm
BLAST	Basic local alignment search tool
DSX	DrugScore eXtended
EC	Enzyme consortium
EF	Enrichment factor
FAD	Flavin adenine dinucleotide
FP	Fingerprint
H-bond	Hydrogen bond
IUPAC	International union of pure and applied chemistry
LC	Local cliques

LPCS	Labeled point cloud superposition
MCS	Maximum common subgraph
NAD	Nicotinamide adenine dinucleotide
NAM	Negative allosteric modulator
NN	Nearest neighbor
PDB	Protein data bank
RAPMAD	Rapid pocket matching using distances
RMSD	Root-mean-square deviation
ROC	Receiver operator characteristics
SAS	Solvent-accessible surface area
SEGA	Semi-global graph alignment
VSEPR	Valence shell electron pair repulsion

List of Figures

List of Tables

List of Publications

Articles

- Krotzky, T.; Klebe, G. Acceleration of Binding Site Comparisons by Graph Partitioning. *Mol. Inform.* (In press)
- Schiebel, J.; Radeva, N.; Köster, H.; Metz, A.; Krotzky, T.; Kuhnert, M.; Diederich, W.; Heine, A.; Neumann, L.; Atmanene, C.; Renaud, J.-P.; Meinecke, R.; Schlinck, N.; Popp, F.; Zeeb, M.; Klebe, G. One Question, Multiple Answers: Biochemical and Biophysical Screening Methods Retrieve Deviating Fragment Hit Lists. *(In preparation)*
- Krotzky, T.; Grunwald, C.; Egerland, U.; Klebe, G. Large-Scale Mining for Similar Protein Binding Pockets: With RAPMAD Retrieval on the Fly Becomes Real. *J. Chem. Inf. Model.* **2015**, *55(1):* 165–179
- Krotzky, T.; Rickmeyer, T.; Fober, T.; Klebe, G. Extraction of Protein Binding Pockets in Close Neighborhood of Bound Ligands Makes Comparisons Simple Due to Inherent Shape Similarity. *J. Chem. Inf. Model.* **2014**, *54(11):* 3229–3237
- Krotzky, T.; Fober, T.; Hüllermeier, E.; Klebe, G. Extended Graph-based Models for Enhanced Similarity Search in Cavbase. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2014**, *11(5):* 878–890

Posters

- Krotzky, T.; Klebe, G. Ultrafast Comparison of Protein Binding Sites Using Distance Histograms. *Gordon Research Conference on Computer Aided Drug Design, West Dover (VT, USA)* **2013**

- Krotzky, T.; Klebe, G. A new Method for Rapid Comparison of Protein Binding Pockets by Capturing Spatial Distributions. *9. German Conference on Chemoinformatics, Fulda (Germany)* **2013**

- Krotzky, T.; Fober, T.; Mernberger, M.; Klebe, G.; Hüllermeier, E. Extended Graph-based Models for Enhanced Similarity Retrieval in Cavbase. *8. German Conference on Chemoinformatics, Goslar (Germany)* **2012**
  Poster Award

Talks

- Using Distance Distributions and Extended Graphs to Speed-up the Comparison of Binding Sites in Cavbase. *4. CCDC Research Partner Forum Meeting, Cambridge (England)* **2015**

- Acceleration of Pocket Comparisons. *Cambridge Crystallographic Data Centre, Cambridge (England)* **2014**

- The Cavbase System. *EuroCUP VII (OpenEye's annual European science meeting), Méry-sur-Oise (France)* **2014**

- Efficient Comparison of Protein Binding Sites Using Distance Histograms. *27. Molecular Modeling Workshop, Erlangen (Germany)* **2013**
  Lecture Award

Part I

Methods for the Comparison of Protein Binding Sites

1

Introduction and Motivation

Proteins are present in large quantities in the body and perform a plethora of different tasks. They are regulating metabolism, catalyze biochemical reactions and establish signaling cascades. Their ubiquitous presence in various regulatory mechanisms turns them also into a prominent target for the treatment of many diseases.

Once the biological molecule which accounts for a specific disease (or just its symptoms) in the human body has been identified, pharmaceutical drug discovery aims at the design of a new and usually smaller molecule than the protein that is able to influence this target in a desired way. A misregulated enzyme, for instance, which is responsible for cleaving peptide chains and thus triggering certain signaling cascades, can be inhibited or activated by the newly synthesized ligand in an allosteric or competitive way.

The prerequisite for most rational drug design efforts is the determination of the 3-dimensional structure of the target molecule. In the past, powerful experimental techniques such as X-ray crystallography have been developed to unravel the arrangement of the amino acids that constitute a protein structure. Increasingly, the structure of novel proteins can also be predicted using sophisticated modeling techniques. Once the structure of the target protein is known and a putative binding site has been identified on its surface, it can be coarsely estimated how a new drug molecule must be shaped in order to fit in the binding pocket like a key that fits in its lock. Most often biochemical reactions take place in such binding pockets or a bound ligand leads to the transmission of signals between cells. Following this strategy, so-called *lead structures* can be developed, which represent the starting point for the discovery of novel and highly potent drug molecules.

During the recent decades a continuously growing number of protein structures has been solved and made accessible to the community by deposition in public databases. A prominent example of such a database is the Protein Data Bank (PDB) [9]. It has been established in 1971 at the Brookhaven National Laboratory by Walter Hamilton and colleagues[1] and has attained a growing importance ever since. Reaching an almost exponential growth in the past, the PDB contains currently more than 107 000 biological structures (April 2015). Due to the increasing amount of structural information, also the pursuit of exploiting the data for pharmaceutical purposes has emerged, which is often tackled by computer-aided techniques.

As a result, the comparative analysis of protein data has become a central task in disciplines such as bio- and chemoinformatics. Prediction of protein function and attempts to explain putative cross-reactivities of novel drug molecules are key objectives in biological sciences in general and pharmaceutical drug development in particular. The cross-binding of a given drug at an unexpected protein target (so-called *off-target effects*) is one of the major explanations for adverse drug reactions (ADRs). As outlined by Sim and Ingelman-Sundberg [138], ADRs account for 7 % of all hospitalizations, 20 % of readmissions to hospital and 4 % of withdrawals of new chemical substances. Hence, they are at least as costly as the drug treatment itself and can be rather ruinous at late-stage drug development. With 100 000 lethal cases per year ADRs are among the leading causes of death in the US [138]. In addition,

[1] Source: `http://pdb.org`. Accessed April 2, 2015

other studies demonstrate that, on average, a single drug molecule interacts with six targets in a cell in addition to the desired one [102]. This underlines the high risk of unwanted cross-reactivities for a newly developed drug. In consequence, the prediction of ADRs by quantifying a drug's risk to interact with unexpected proteins is of utmost importance already at early stages of the drug discovery process, preferentially long before expensive clinical studies are conducted.

The most widely used method in this context is the comparison of sequence data. The alignment algorithms proposed by Needleman and Wunsch [108] or Smith and Waterman [141] are well-established methods for the comparison of sequences. Also the BLAST service [2] is a popular tool for the efficient local alignment of a query to entire databases of amino acid or nucleotide sequences. Yet, not every problem can be solved on the sequence level. While proteins with a sequence identity above 40 % also tend to share a similar function [148], below this threshold a correlation is less obvious and difficult to detect [165]. While prediction of function from sequence can be improved by using more advanced techniques such as hidden Markov profiles [37], the prediction accuracy declines if the sequence similarity falls below a certain threshold [93]. In this context, also the *twilight zone* of sequence alignments has become popular that defines the range between 20 and 35 % sequence identity where it is difficult to decide whether two proteins are really homologous or not [125]. Furthermore, sequence alignments are often not appropriate to identify similar binding sites in unrelated proteins that originated from convergent evolution [22]. As the spatial structure is usually better conserved than sequences for proteins exhibiting similar function, the direct comparison of protein 3D structural data has become increasingly important as an often superior alternative.

Finding similar pockets in the entire protein binding pocket space which can potentially host different ligands will provide valuable information during the lead-optimization phase while improving drug binding in terms of higher affinity and selectivity. In contrast to ligand-based approaches that compare the molecule under consideration with available drug molecules, the approach of binding site comparison is a receptor-based line of action. Following such a strategy, tailored design of promising candidate molecules featuring molecular decorations, incorporating bioisosteric replacements or even pursuing the exchange of novel scaffolds (*scaffold-hopping*) can be used. Moreover, pocket comparisons can also be exploited to successfully annotate biochemical function to orphan proteins [112].

Comparisons of pockets using geometric properties [126, 30], typed triangles or physicochemical features in 3D space [62, 129] have therefore become popular strategies to unravel similarities of protein binding sites. A prominent alignment-free comparison method for protein-ligand binding sites is FuzCav [161]. Cavity fingerprints are defined for binding sites that store information about the presence of pharmacophoric feature triplets as lists of integers. These allow for an ultra-fast comparison in the following, attaining about 1000 calculations per second on a 3.4 GHz processor. Pocket-Surfer which estimates global pocket similarity and Patch-Surfer which also detects local binding site similarities have been introduced by Sael and Kihara [25, 127]. The latter approach represents a pocket as a set of patches described by their shapes, electrostatic potential, degree of burial and hydrophobicity. The comparisons are subsequently carried out by making use of a bipartite matching procedure. Desaphy et al. introduced the pocket description VolSite together with a tool for alignment and comparison called Shaper [32]. The shape and the physicochemical environment of a binding site is stored and then compared via Shaper which aligns pockets by determination of an optimal surface overlap. In PocketAlign, shape descriptors from binding sites are derived which are enhanced by pharmacophoric features [170]. In the comparison step matching pairings of the descriptors are combined into mappings which are subsequently evaluated using different metrics to achieve starting points for reasonable alignments. In order to accelerate binding site comparisons, also geometric hashing has become rather popular. Specific features of binding sites are transformed into a hash table which is consulted in the following comparison step to obtain similarities to other cavities [16, 6, 116].

Zauhar et al. [171] developed a surface-based method, which builds on a technique called *shape signatures* to describe the shape of the bound ligand molecules as well as of the corresponding receptor sites. Here, the volume of the protein binding site is explored by a ray-tracing method. Probability distributions of surface-based descriptors can be derived by this procedure which are subsequently stored in terms of histograms. Subsequently, these histograms are used to test for shape complementary between compounds and receptors. Binkowski and Joachimiak [10] developed a two-step procedure for the comparison of binding site surfaces that consists of a global shape matching based on distance calculations between all unique atom pairs followed by a spatial alignment of the physicochemical texture to identify conserved amino acids. Furthermore, prominent surface-based approaches include methods such as CASTp [11], EF-Site [75], SiteEngine [136] and Cavbase [131, 132]. This group of tools is of special interest in the context of structure-based drug design. Here, the basic idea is to capture physicochemical properties of functional groups that are essential for the interaction of proteins and ligands. The goal of these approaches is to identify portions of proteins (binding sites) that are likely to recognize and thus interact with similar ligand molecules, independent by how much they actually vary in the overall amino acid sequence [164].

An extensive review about methods for the detection of similarity between protein binding sites can also be found in the works of Kellenberger et al. [73] and Vulpetti et al. [155].

The following work, which will be presented in this part of the thesis (Chap. 2–5), has been subject of several publications in scientific journals. A complete listing of the papers that have emerged from the studies can be found in the list of publications on page xix.

2

Binding Site Detection[1]

2.1 Problem Statement

Identifying druggable binding sites is the initial task in receptor-based drug discovery as it has been known for a long time that binding sites occur in most cases in the largest cavity on the surface of proteins [87]. The continuously growing number of available protein structures has increased the desire for automated detection and comparison algorithms to assign putative binding pockets. To accomplish this task, several binding site detection algorithms have emerged to unravel the so-called pocketome [83, 117, 130, 49], which can be basically classified in terms of geometry-based and energy-based methods. In addition, several binding site comparison protocols have been proposed, which are based on the comparative evaluation of structural information. In general, the methods reported in literature to compare proteins fall into three categories: Fold-based, template-based and surface-based [147]. The fold-based

[1] Reprinted (adapted), with permission, from Krotzky, T.; Rickmeyer, T.; Fober, T.; Klebe, G. Extraction of Protein Binding Pockets in Close Neighborhood of Bound Ligands Makes Comparisons Simple Due to Inherent Shape Similarity. *J. Chem. Inf. Model.* **2014**, *54(11):*3229–3237. Copyright 2014 American Chemical Society. The major part of the study was performed by me, Thomas Rickmeyer helped to assemble the used datasets and Thomas Fober performed the necessary comparative calculations using the LPCS approach, which was developed as part of his PhD thesis.

ones include, for example, DALI [64], MC-CE [56], as well as SABERTOOTH [146] and CATHEDRAL [123]. Meanwhile, several fold databases have been developed based on these methods [107, 113, 63]. Template- and surface-based methods do not compare entire proteins but evaluate only parts of their structure. This strategy is based on the assumption that functionally important regions are evolutionarily conserved and, thus, more relevant for the mutual comparison. Template-based methods comprise algorithms such as PINTS [142], JESS [8] or LabelHASH [104].

Usually the implementation of new algorithms for binding-site comparisons should accomplish one of the following three tasks. First, the prediction of putative off-target binding of drug molecules is highly desired, possibly providing an explanation for polypharmacology and adverse drug effects in the early phase of a drug development project [67]. Second, the predictive functional annotation of orphan proteins is of high interest [112]. Third, the discovery of bioisosteric replacements [156, 74] for specific ligand portions by retrieving similar binding sites that accommodate ligands with alternative scaffolds can support drug development. For the third application it is undoubtedly reasonable to focus on pockets (or subpockets) only that have been extracted in close neighborhood of a known ligand. Such pockets along with their bound ligands are successfully exploited, e.g., in KRIPO [167], a method to identify valuable bioisosteric replacements of ligand portions recognized in specific subpockets.

However, many binding site comparison methods, aimed at the other two goals, have been developed and they are subsequently validated by compiling test data sets of binding sites extracted as regions adjacent to bound ligands within a 4–6.5 Å sphere [129, 161, 170, 41, 136]. As a matter of fact such data sets will lack binding sites originating from uncomplexed structures. In consequence, a previously unknown putative binding site cannot be detected as a potential off-target for the drug molecule of interest, unless its pocket was incidentally occupied in the same or highly overlapping region by another ligand during crystallization. Furthermore, it is rather likely that such extracted pockets resemble inflated representations of the ligand shape as only the region close to the accommodated ligand is considered. Thus, with respect to the prediction of drug side effects or functional annotations of orphan proteins, it may be beneficial to apply an automated cavity detection method independent of the presence or absence of a bound ligand. This will be of utmost importance when ligands are studied that address different subpockets of proteins with large binding sites. Several methods have been developed and successfully tested on putative binding cavities [153, 132, 91, 168], extracted independently of the presence or absence of a bound ligand. Hence, they will incorporate pockets of uncomplexed proteins. Nonetheless, any pocket data set extracted solely in the neighborhood of bound ligands will be biased toward intrinsic ligand shape information. This can strongly bias the obtained results as an exaggerated weight is assigned to the ligand-based pocket shape rather than to the exposure of physicochemical properties available to recognize an arbitrary ligand.

Binkowski and Joachimiak [10] alluded to this fact that shape alone cannot be expected as a comprehensive binding-site descriptor, a statement which matches with our assumption. In another study, Kahraman et al. [70] used spherical harmonics to describe binding site shapes. Although they found that the success rate of retrieving similar pockets depends on ligand shape, particularly if rigid host molecules are considered, the success rate declines once increasingly flexible ligands such as ATP, NAD and FAD are subjected to the analysis. The latter ligands involve a large number of rotatable bonds which allow them to adopt multiple conformations of deviating shape (even when bound to members of the same superfamily) [144]. This fact stimulated us to use these cofactor ligands in our evaluation. Moreover, it is suggested that the shapes of the hosting pockets vary more strongly than the accommodated ligands. The more it appears important to assess by how much predefined ligand shape affects the pocket representation and distorts subsequent comparisons if the pockets are extracted as close environments around bound ligands.

In the present chapter, we want to compare the results of a pocket comparison using pockets extracted immediately around the ligands and pockets which result from an unbiased analysis of surface-exposed depressions on proteins (Fig. 2.1). In the latter case we use physicochemical properties to describe the pockets. Furthermore, we examine whether the geometries of ligands and pockets extracted around the bound ligands show high shape-based similarity. To perform these comparisons, we use a very simple geometric approach and describe the extracted binding sites (or bound ligands) in terms of spatial distance distributions of pocket-attributed interaction points (or ligand atoms). With this approach we do not intend to develop a new comparison algorithm but seek for a fast method to compute similarity. A related method was suggested by Binkowski and Joachimiak as the first coarse filtering method in a two-step comparison procedure [10]. They determined the distances between all pairs of pocket surface defining atoms to generate a probability distribution. In our comparison we try an even simpler and thus faster approach by considering all distances to the pocket-describing points with respect to one common center point.

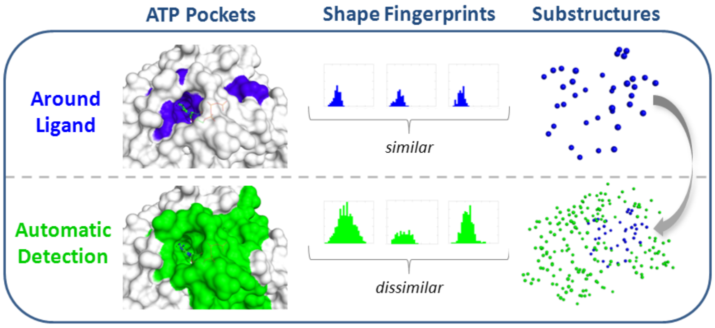
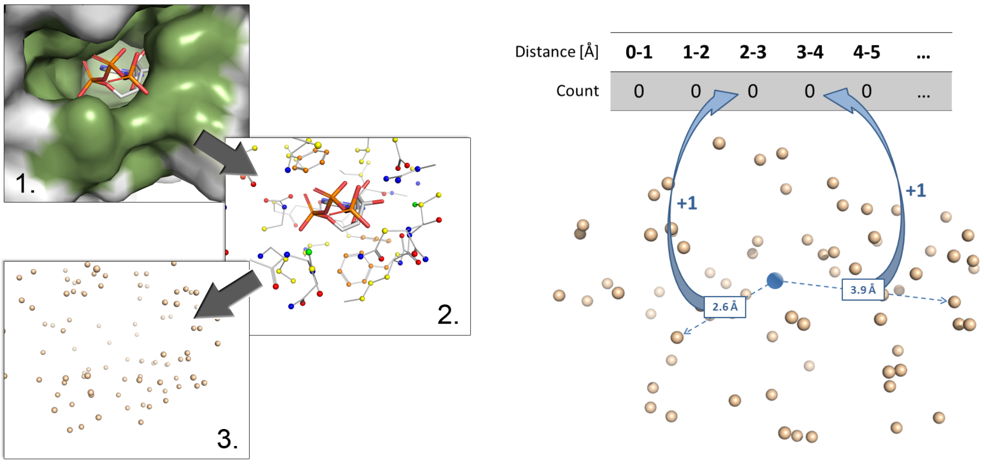


Figure 2.1 In this chapter, we compare automatically detected pockets (green) and pockets that are defined by the position of a ligand (blue). We assume that the shape of the ligand strongly determines the shape of the extracted pockets, which allows a very simple comparison procedure to succeed in the following (middle). Any recognition of subpockets, however, may most likely fail (right).

2.2 Shape-based Comparative Analysis

For the considered data sets the pocket-describing points (or ligand atoms) were obtained following the protocol illustrated in Figure 2.2. Any protein atom approaching an atom of the bound ligand ≤ 6 Å was supposed to be part of the binding pocket. In case of the ligands, we simply considered the composing atoms. Next, we implemented an adjusted version of the program *fconv* [110] to perform an atom-type assignment to all thus defined binding site atoms. Subsequently, they were filtered in terms of represented physicochemical properties: only those atoms were considered that could



(a) Binding site representation only considering shape information. First, all protein atoms approaching any ligand atom ≤ 6Å are defined as binding site (green area). Next (2), all atoms are considered that agree to one of the following physicochemical properties: H-bond donor (blue), H-bond acceptor (red), H-bond doneptor (green), aromatic (orange), or hydrophobic (yellow) character (as classified by the program fconv). In the final step, any differentiation of physicochemical properties is discarded revealing a pure pocket shape description.

(b) Fingerprint generation to capture the spatial distribution of distances of interaction points with respect to their common centroid. All bin counts of the assigned fingerprint are initially set to zero. Next, distances determined between all interaction points (beige) and the centroid (blue), are assigned to the corresponding 1 Å sized bins. Any match to a bin augments the corresponding fingerprint element by one. In a very similar way also the spatial atom distribution of bound ligands was analyzed. Here, the atomic coordinates of the ligands were directly used as input.

Figure 2.2 Illustration of the workflow of the shape fingerprint.

be attributed to groups showing either H-bond donor, H-bond acceptor, H-bond “doneptor” (either being donor or acceptor), or aromatic or hydrophobic character. In the following, this physicochemical information was neglected and solely the spatial location of the retrieved interaction points was used to describe the pocket. Thus, this procedure provided binding site representations solely reflecting shape and no physicochemical information. The comparison of two pockets is then accomplished by the following two-step procedure. First, a fingerprint is calculated for each pocket which captures the distances of all interaction points with respect to their geometric center (centroid). The obtained distances are represented histographically in bins of 1 Å size and the occurrence frequencies of the found distance ranges are compiled (see Fig. 2.2 (b)). After fingerprints have been assigned to all pockets of the data set, the comparative distance between two pockets is calculated by using the Jensen-Shannon divergence. In case two fingerprints vary in length, the shorter one is extended by adding unoccupied bins. For the evaluation of the ligands we proceeded similarly, only taking the composing atoms directly.

2.3 Datasets

Pockets Binding a Particular Ligand

To evaluate our comparison procedures, we used several validation sets of protein binding sites, which will be described in the following and which have been considered by others for the same purpose. As reported by Fober et al. [44], the first applied dataset was originally assembled to examine the discriminative power of the graph-based comparison method GAVEO. A set of 355 binding pockets hosting either the cofactor *nicotinamide-adenine-dinucleotide* or *adenosine-5’-triphosphate* (PDB ligand identifiers `NAD` and `ATP`) was collected. Since a protein may exhibit several binding sites for the ligand, the number of selected pockets per protein was restricted, so that each protein was considered only once in the dataset. Furthermore, the number of binding sites was reduced by calculating the root-mean-square deviation (RMSD) between ligand pairs occupying the binding sites by using the Kabsch algorithm [69]. The RMSD is a frequently used figure of merit in computational chemistry and is calculated with the formula

$$RMSD(u,v) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[(u_{ix}-v_{ix})^2+(u_{iy}-v_{iy})^2+(u_{iz}-v_{iz})^2\right]}\ , \qquad (2.1)$$

where n is the total number of atoms in the molecular structures u and v. The variables u_{ix}, u_{iy} and u_{iz} represent the Cartesian coordinates of the i-th atom in structure u. By defining a maximal RMSD-threshold of 0.4 Å this step filters for pockets only hosting cofactors adopting similar conformations. This value was adjusted in a way as a trade-off between data set size and similarity. Finally, these selection criteria resulted in a set of 141 ATP- and 214 NAD-binding pockets (see Appendix, Tab. A.0.1). This dataset will be referred to as ATP/NAD_{small}.

In addition, we compiled another larger and more comprehensive dataset of ATP- and NAD-binding pockets now omitting the above-mentioned RMSD constraint. In consequence, this set of pockets also hosts ligands with diverse conformations. Moreover, we retrieved a set of binding sites from Cavbase that accommodate *flavin-adenine dinucleotide* (PDB ligand identifier `FAD`), another cofactor related to NAD and ATP that is used by many enzymes in biology. Sets of pockets hosting particular ligands such as ATP, NAD or FAD were chosen to establish challenging comparisons, as Stegemann and Klebe [143] showed that these cofactors are able to bind in various orientations, even though sharing the adenosine diphosphate moiety as common substructure. As sole constraint we decided not to consider binding pockets with a volume greater than 4000 Å^3 in the dataset. Regarding a value of 1.7 Å as van der Waals radius of a carbon atom, such pockets would still provide enough space to host approximately 150 atoms. In consequence, we assume that any larger cavity will certainly display an artificially extracted pocket falsely selected by the LIGSITE algorithm. For the dataset ATP/NAD_{large} (Appendix, Tab. A.0.2) we compiled a sample of 420 ATP- and 402 NAD-binding pockets, now also showing a better balanced ratio between the two class sizes. This provides an additional challenge to our classification experiments, as the success rate of a simple randomized assignment tends to improve in case of an unbalanced dataset with nonequal class sizes [35]. Classification rates that result from a majority voting achieve already $214/355 = 60.28\,\%$ for a simple random assignment using the ATP/NAD$_{small}$ dataset. In the new dataset this voting reduces to $420/822 = 51.09\,\%$.

The FAD dataset was supposed to be even more challenging, since for this cofactor it has been stated that no single protein-based pharmacophore can be derived using binding pocket information [36]. Hence, we expected the set of FAD-binding pockets to be rather diverse with respect to arrangements of residues interacting with the ligand. In this set we furthermore distinguished between a covalently bound and a non-covalently bound FAD. In total, we selected 429 pockets binding FAD non-covalently and 114 pockets hosting the cofactor via covalent attachment (Appendix, Tab. A.0.3). We refer to these datasets as *FAD* and FAD_{cov}, respectively.

The large datasets of ATP, NAD and FAD will be used in the following experiments regarding the difference of shape-biased and unbiased pockets. Due to the involvement of fconv, however, NAD and FAD had to be reduced to 380 and 432 elements, respectively. For the remaining complexes the program was not able to perform a proper separation of the protein and the ligand.

A second data set, suggested by Hoffmann et al. [62], was assembled which comprised pockets accommodating ligands of similar size. This set considered 100 non-redundant proteins with pockets hosting one of the ten ligands of approximately equal size shown in Figure 2.3.

A third dataset was extracted from the PDB (83 000 entries in the release used), using LIGSITE to find putative binding pockets. A database of 451 100 pockets was complied. All pocket atoms were annotated according to *fconv* atom types.

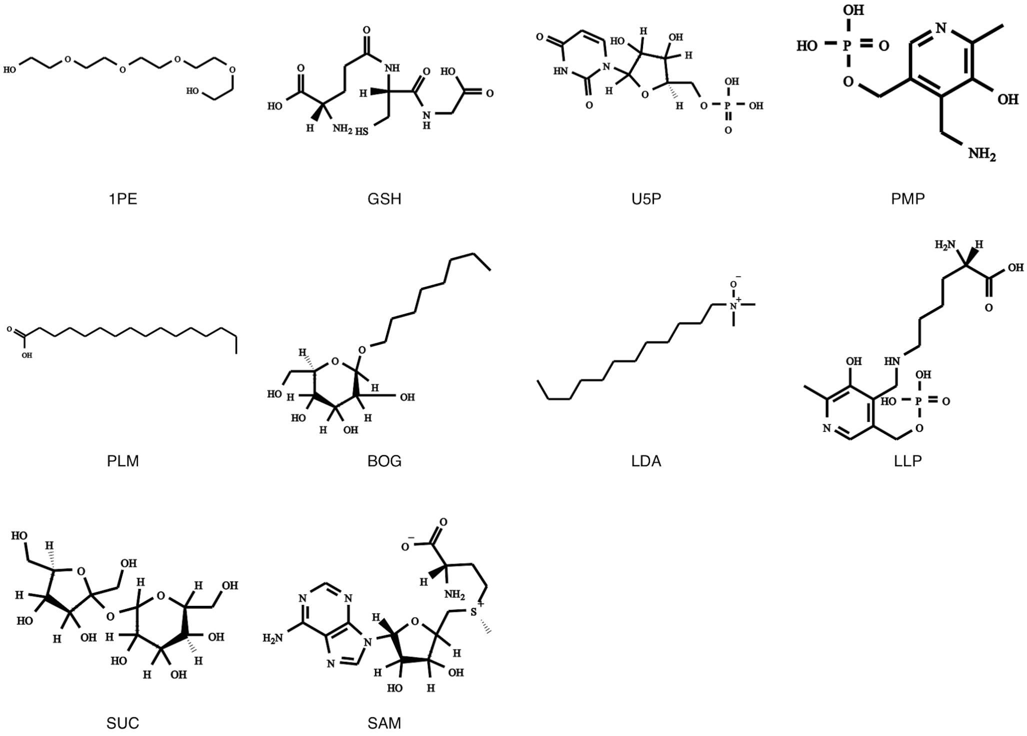


Figure 2.3 The ten ligands that were considered in the comparative study of Hoffmann et al. All structures are labeled with the respective ligand identifier in the PDB.

2.4 Evaluation Techniques

2.4.1 ROC Curves

A commonly used technique in signal detection and medical statistics to test whether a new method is able to distinguish between hits and decoys particularly with respect to retrieval rates are receiver operator characteristic (ROC) curves [58, 14]. ROC curves depict the relative tradeoff between success and failure rates by plotting the number of true positives (TP) against the number of false positives (FP). Correct entries that are falsely recognized as decoys are named “false negatives” (FN) and those that are correctly retrieved as decoys are classified as “true negatives” (TN). A ROC curve plots the true positive rate (TPR, also called recall rate) on the y-axis against the false positive rate (FPR) on the x-axis, where the TPR and FPR are defined as

$$TPR = \frac{TP}{TP + FN}$$

and

$$FPR = \frac{FP}{TN + FP} \; .$$

Each ROC curve starts at the origin (0,0) and a perfect search would result in full retrieval on first ranks (0,1). In this case the graph would possess an infinitely high gradient, resulting in an area under the curve (AUC) of 1. The AUC is a prominent descriptor to rank the quality of a method. A retrieval with random selection of hits will lead to a ROC curve showing unit slope and an AUC of 0.5.

In this work, we will use ROC curves to evaluate the results of our retrieval experiments. For instance, several pockets that bind a particular ligand are compared against the entire Cavbase database. All pockets binding the same ligand will be defined as a hit and all others as decoys. It is generally assumed that a method performing reasonably well in such a retrospective virtual screening is also likely to succeed in a prospective screening scenario [133].

2.4.2 Cross-Validations

While comparing molecular structures such as binding sites, it appears rather difficult to assess calculated values such as similarity scores directly, as these scores do not return a kind of threshold value defining whether two pockets are “similar” or “dissimilar”. To overcome this problem, we make use of an indirect measure by retrieving entries that are closest to the query in terms of distance in score space. Next, a property of interest is extracted from the list of *nearest neighbors* (NN). In the present study as property the name of the bound ligand was compared to that used in the query. We carried out several classification experiments using the datasets described above. The first step of such an experiment is the generation of an all-against-all distance matrix, containing all scores of every pairwise comparison. Subsequently, either a k-leave-one-out or a 10-fold cross-validation is applied on the matrix.

In a k-leave-one-out cross-validation each line of the distance matrix is analyzed with respect to the k-NN structures, and the query structure itself is omitted. The query is then assumed to be member of the same class to which the majority of the next NN belong. After evaluating all lines of the matrix, the total rate of correct classifications can be obtained by comparing the predicted classes of all query structures to the ones assigned by an independent method.

The 10-fold cross-validation is realized in a rather similar way. However, at the beginning of the process the distance matrix is horizontally split into ten equally large portions. Then, a leave-one-out cross-validation using only one NN ($k = 1$) is carried out for every single portion. Ten classification rates are obtained, which are in the following used to calculate the mean and standard deviation. This appears more reasonable than simply evaluating a single classification rate as it is returned in a k-leave-one-out cross-validation.

2.5 Results and Discussion

2.5.1 Cofactor Binding Pockets

As a first example we evaluated the set of cofactor binding pockets accommodating either ATP, FAD or NAD (NADH and NAD+) present with deviating conformations. In this experiment we used the k-nearest-neighbor (k-NN) method for classification, also applied previously with success [82, 44, 45]. We calculated an all-against-all scoring matrix which was subsequently used as input for a 10-fold cross-validation with a k-NN classifier, using $k = 1$. Unexpectedly, the two-class classification experiment of the ATP_{large} and NAD_{large} datasets revealed excellent rating ($96.4 \pm 2.0\%$). Since we considered only unlabeled surface points as pocket descriptors and the analyzed cofactors are of rather different size, it might well be that simply the number of interaction points defining the binding pocket is already responsible for the impressive discrimination. However, using only the number of points for the comparisons leads to a poor rate of only $63.7 \pm 5.3\%$. Also a normalization of the fingerprint distributions (calibrating the area under each curve to 1) to exclude any influence of the total number of points per pocket leads to hardly any change in the above-mentioned classification rate ($95.5 \pm 1.7\%$).

As next experiment we incorporated FAD pockets. FAD and NAD vary evidently less in size than ATP and NAD. While the resulting heat map of the distance scores shown in Figure 2.4 suggests that FAD and NAD pockets are less well discriminated, we still obtained a convincing classification of $94.3 \pm 1.5\%$ in this three-class experiment. Furthermore, also a success rate of $96.2 \pm 2.1\%$ is achieved when only the FAD and NAD pockets are considered in the classification. To estimate the robustness of the obtained results of this three-class experiment, we evaluated the scoring matrix also by another method, which is closely related to the 10-fold cross-validation. We applied the k-leave-one-out cross-validation and varied the number of nearest neighbors k in the range from 1 to 15. As shown in Table 2.1 the obtained rates do not deteriorate rapidly and all excel 90%, which indicates either robustness of our method or simplicity of the used dataset.

Table 2.1 Classification results of the shape FP when the scoring matrix is evaluated by using a k-leave-one-out cross-validation with the number of nearest neighbors k in the range $1 \ldots 15$.

k	**Correct Classifications [%]**
1	94.3
3	93.8
5	93.5
7	92.9
9	92.7
11	91.6
13	90.5
15	90.1

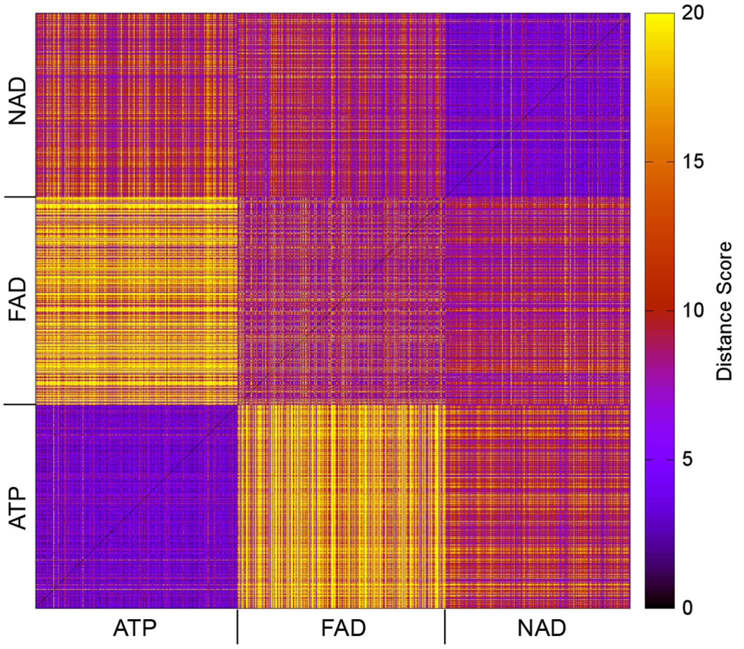


Figure 2.4 Heat map of the scoring matrix resulting from the classification experiment using ATP, FAD and NAD pockets (large datasets). The areas of correctly assigned ATP (lower left corner) and NAD pockets (upper right corner) generally display rather low distance scores among each other, which is indicated by the bluish coloring. They are obviously well separated from the other pockets. The FAD pockets (center) appear to be more similar to the NAD pockets; however, they can also be correctly classified with over 96 % success rate in a two-class experiment that regards FAD and NAD pockets only. The black main diagonal from bottom left to top right indicates the distance values of zero in the cases of a self-comparison.

2.6 Dataset of Equal-sized Ligands

Hoffmann et al. [62] suggested as real challenge for a binding site comparison approach to discriminate pockets accommodating ligands of similar size. To evaluate their comparison method, they compiled a benchmark data set of 100 non-redundant proteins with pockets hosting one of the ten ligands of approximately equal size compiled in Figure 2.3.

For each ligand, ten pockets were extracted by defining the protein atoms in a distance of up to 5.3 Å. The authors hence ended up with a ten-class data set, where each class consisted of ten pockets, which they called a homogeneous data set. In their study, a total of nine pocket comparison methods were tested considering this validation set and the classification rates were analyzed using receiver operating characteristics (ROC) curves. For each method 100 ROC curves were calculated by performing comparisons of each single pocket against the 99 remaining structures. Finally, the average areas under the curve (AUC) of all ROC curves that corresponds to a single method were calculated. An AUC of 0.5 denotes a method which detects hits (pockets of the same class) equally well as a random assignment. On the contrary, a value of 1.0 would be obtained for a method that assigns the highest similarity scores to the nine remaining pockets of its class and achieves perfect classification. The results reported in the above-mentioned study revealed average AUCs between 0.58 and 0.77. We performed a similar analysis of this data sample using our shape-based fingerprint descriptors. Our approach performed surprisingly well, reaching an average AUC of 0.66 (see Fig. 2.5).

2.7 Comparison with Unbiased Pockets

All reported examples demonstrate convincingly well that the success of a comparative binding site analysis is intrinsically given if only shape complementarity next to the bound ligand is used to describe the considered pockets. We therefore applied a ligand-unbiased cavity detection algorithm to extract pockets from protein structures, in order to analyze again our ATP, NAD and FAD data sets. A variety of structure-based methods has emerged to accomplish the task of finding putative binding pockets on the protein surface. They can be divided into either geometry-based methods, such as PASS [15], SURFNET [90], CAST [11], APROPOS [119], SiteFinder [88], fpocket [91], PocketPicker [162], or energy-based approaches, e.g., PocketFinder [4] and SuperStar [150]. Comprehensive overviews of the current binding site prediction methods are provided, e. g., by Pérot et al. [117] and Leis et al. [94]

This time we applied LIGSITE [60], a grid-based method (and thus also a

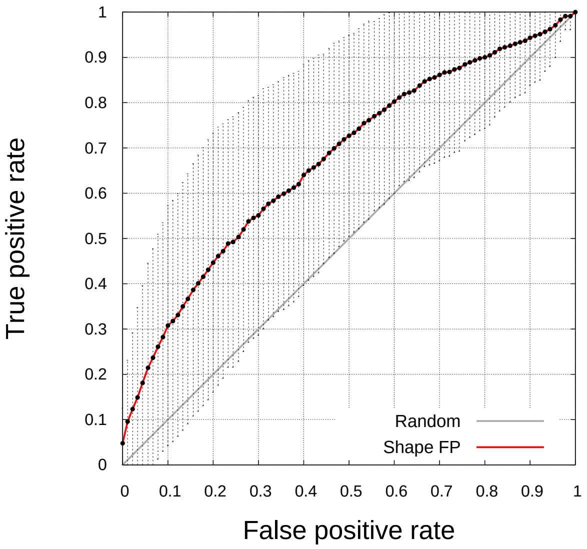
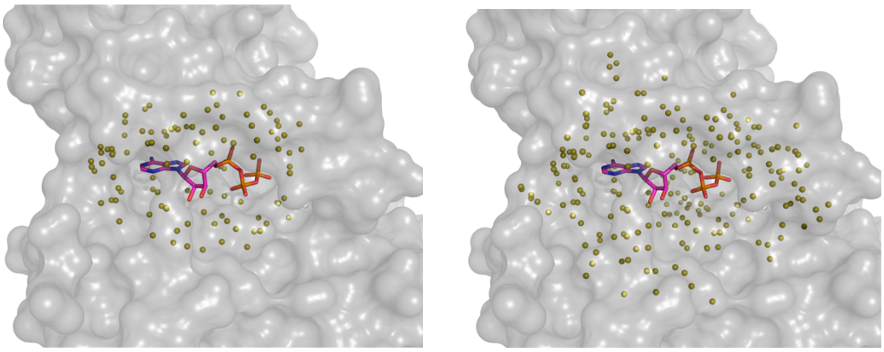


Figure 2.5 Average ROC curve of the shape FP (red) when applied on the homogeneous dataset of Hoffmann et al. Random performance is indicated by a gray diagonal from bottom left to top right. The plot represents the average of all 100 curves that have been obtained and exhibit an average AUC of 0.66 ± 0.16. In addition we also display the standard deviation for each data point shown as black error bars.

structure-based) to detect depressions on protein surfaces. They optionally comprise hosted ligands but usually they extend beyond the actual ligand contact area with the protein. This strategy appears as a less biased protocol to define a binding pocket. To apply LIGSITE, the protein is placed onto a regular grid with a spacing of 0.5 Å. Each grid intersection point is evaluated with respect to its degree of burial. A cluster of at least 320 adjacent buried grid points is then defined as a putative binding site. A detailed description can be found in the original publication [60]. All atoms flanking the thus detected cavities are potentially capable of binding a ligand. The extracted pockets were likewise classified in terms of atom types using *fconv* and subsequently used to construct three new data sets $ATP_{Ligsite}$, $NAD_{Ligsite}$ and $FAD_{Ligsite}$. The LIGSITE pockets are different in shape and generally larger than the ligand-based pockets (on average the number of extracted interaction points is increased by a factor of 2.5), which indicates additional areas competent to recognize a ligand beyond the area actually addressed by the regarded cofactors (cf. Figure 2.6).

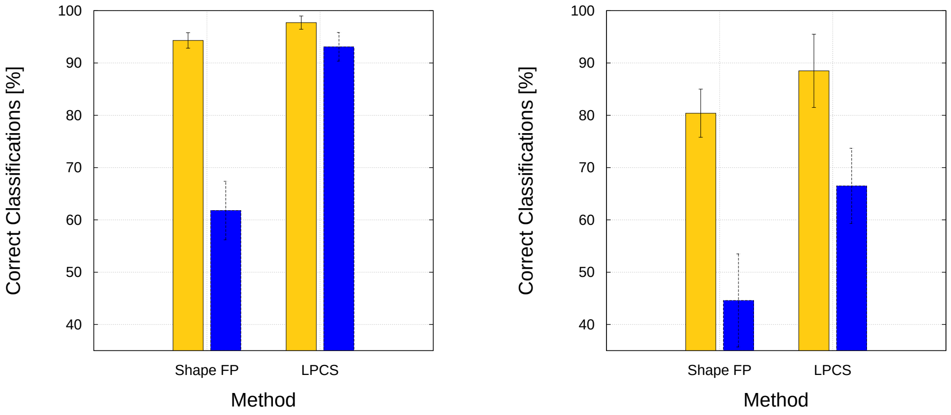


(a) The pocket has been defined by extracting the area of 6 Å around the bound ligand.

(b) The pocket has been defined by the automated cavity detection procedure LIGSITE.

Figure 2.6 Example of an ATP-binding pocket (PDB: `1B38`). The automatically detected cavities are in general much larger than the pockets solely defined by considering the bound ligand.

In contrast to the results obtained with the ligand shape-based pockets of ATP, NAD, and FAD, the success rates of classifying by use of the shape-based fingerprints decrease substantially from 94.3 % to 61.8 %. The actual atom-type assignment for the pocket representation (see workflow in Fig. 2.2 (a)) enabled us to apply a previously presented approach for the comparison of protein binding sites, the so-called Labeled Point Cloud Superposition method (LPCS) [45]. This approach suggested by Fober et al. was applied using the parameter setting recommended by the authors. Applying LPCS in the current case leads to an accuracy of 97.7 ± 1.3 % when it is applied to the ligand-based pockets. Thus, the success rates agree well with the results obtained with our simple fingerprint approach. However, in contrast to the fingerprint approach, LPCS is still able to achieve comparable accuracy of 93.1 ± 2.8 % when it is applied to the larger pockets, extracted by LIGSITE (Fig. 2.7 (a)).



(a) The complete dataset of 1232 pockets was used to obtain the results presented here. The accuracy of the fingerprint approach (left), which is heavily biased by the actual shape information complementarity to the bound ligand, drops by more than 30 % when applied to the LIGSITE extracted pockets. In addition, the standard deviation increases strongly by a factor of almost four. The results of LPCS (right), however, exhibit a deterioration of only 4.6 % and the standard deviation increases minimally by a factor of two among the data sets of differently extracted pockets.

(b) The culled dataset containing 268 pockets was used. In this case, both approaches still show satisfactory success rates when ligand-based pockets are used (yellow). However, the classification rates decrease dramatically when using the LIGSITE pockets (blue), especially if the shape FP is applied. It is no longer able to reach a success rate better than a random classification.

Figure 2.7 Comparison of the ligand shape-based fingerprint and the LPCS approach when subjected to a classification experiment using binding pockets of ATP, NAD and FAD binding proteins. The binding pockets have been defined by either extracting protein atoms in the close neighborhood (≤ 6 Å) about the bound ligand (yellow) or by applying the LIGSITE algorithm to detect cavities in an unbiased way as depressions on the protein surface (blue).

To set up a more challenging task with respect to conformational and structural diversity we culled the three-class dataset to remove sequential redundancies. Therefore the protein sequence culling server PISCES[1] of the Dunbrack lab [157] was employed, where only PDB structures were kept that agreed to the following conditions: sequence identity not exceeding 25 %, the method of structure determination is X-ray crystallography with R-factor ≤ 0.3 and resolution 3 Å or better. As a result 268 elements remained in the dataset (135 ATP pockets, 73 NAD pockets, 60 FAD pockets). An overview of the contained structures is provided in the Appendix (Tab. A.0.4).

When using the culled dataset the actual problem becomes even more apparent. In case of the ligand-based pockets that have been extracted within 6 Å around the bound molecules, both LPCS and the shape FP still achieve satisfactory results (Fig. 2.7 (b)). Although the success rates of FP are worse than those of LPCS, the difference is not significant (80.4 ± 4.6 % *versus* 88.5 ± 7.0 %). When the automatically detected LIGSITE pockets are used, however, the resulting rates decrease substantially by 34 % in case of FP to reveal 44.6 ± 8.9 % correct classifications, which barely deviates from a random assignment (37.8 % in the present example considering the non-equal population of the subsets). The LPCS approach is still clearly better, attaining correct classification rates of 66.5 ± 7.2 % even though it becomes obvious that culling the dataset increases the complexity to the problem of classifying the automatically detected cavities. Thus, this experiment shows once again that the degree of complexity is highly diminished when ligand-based pockets are used for the comparisons instead of automatically detected ones.

2.8 Ligand Atoms vs. Ligand Shape-based Pockets

The minor loss in accuracy indicates that LPCS is obviously quite robust and independent of the actual size and shape of the pocket whereas the

[1] `http://dunbrack.fccc.edu/PISCES.php`. Accessed December 2, 2014

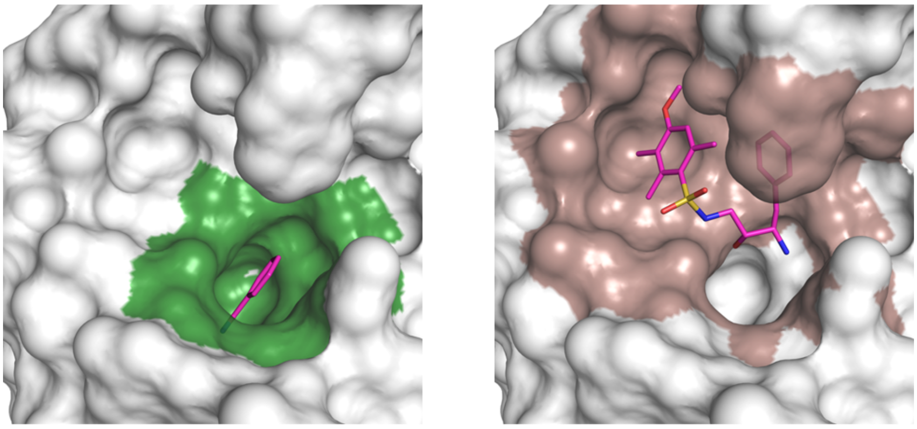
fingerprint approach is strongly affected. Supposedly, the consideration of pockets extracted in close neighborhood of the ligands provides a remarkable advantage in the cavity comparison. In order to examine to what extent the latter pockets resemble just an inflated representation of ligand shape, we performed a comparison of the data sets in which the ligands were used instead of the pockets. Therefore, the ligand atoms were processed in the same way as the pocket atoms beforehand. They were typed using fconv and, subsequently, fingerprints were calculated with the help of the centroid to facilitate a comparison (cf. Fig. 2.2 (b)). In this case a correct classification rate of 98.6 ± 1.0 % was obtained. This is not surprising as the spatial arrangement of ligand atoms is in general less complex than the arrangement of binding site atoms [10]. However, the generation of the scoring matrix enabled us to compare this matrix to the scoring matrices obtained for the pocket comparisons based on either ligand shape-based pockets or surface-exposed pockets (LIGSITE). We calculated the correlation between the ligand scoring matrix and the ligand shape-based pocket scoring matrix.

Alternatively, we faced the ligand scoring matrix to the LIGSITE pockets scoring matrix. To calculate a correlation of two matrices the Spearman's rank correlation coefficients of all matching pairs of rows were determined and, finally, normalized by the total number of row pairs. A high positive correlation of 0.68 was obtained between the ligand matrix and the matrix of ligand shape-based pockets, which underscores the general similarity of ligand shape and pocket shape in this case. On the contrary, there is hardly any correlation between the ligand matrix and the matrix of LIGSITE pockets (correlation coefficient is 0.09), which demonstrates the minor relationship of ligand shapes and the shapes of automatically detected surface-exposed pockets.

2.9 Ligand Pockets vs. Unbiased Pockets

The above-described examples show that the LPCS is obviously still able to extract the relevant information required to match common substructures competent to bind the same ligand when unbiased surface-exposed depressions on proteins are considered in the analysis. As mentioned, the latter approach usually extracts larger pockets as additional areas in the environment, not addressed by the bound ligand, can still provide binding epitopes capable to recognize another ligand. This fact may be responsible for undesired cross-reactivity. To assess whether LPCS outperforms the ligand shape-based pocket fingerprints we carried out another experiment. The PDB contains a significant number of crystal structures determined with the same protein where the bound ligands do not bind to overlapping binding epitopes. This situation can increasingly be expected for fragment binding. In Figure 2.8, the crystal structures of thrombin with benzamidine as an S_1-accommodated ligand and a second fragment, exclusively binding to the S_2-S_4 pocket [65], are displayed. Both ligands address hardly any shared binding region, accordingly an approach extracting binding pockets solely in close neighborhood of bound ligands will likely fail to provide a similarity signature for the two thrombin pockets.

We applied LIGSITE to extract putative binding pockets from the PDB and compiled a database of more than 450 000 pockets. Next, three thrombin query pockets were defined and subjected to this pocket database. First, the structure `3UWJ` was used and all pocket atoms falling within 6 Å next to the accommodated ligand *(N-(Benzylsulfonyl)-D-Leucyl-N-(4-Carbamimidoylbenzyl)-L-Prolinamide*, ligand identifier: `TIF`) were extracted. As this ligand fills the pocket quite extensively, the entire thrombin active site was captured. Second, only the S_1 subpocket of `1DWB` was retrieved by extracting all atoms in a range of 6 Å around the benzamidine. Third, the S_2-S_4 pockets of the PDB entry `2C93` was extracted by using the bound fragment *(N-[(2r,3s)-3-Amino-*



(a) Pocket that has been extracted within 6 Å around benzamidine (PDB code: `1DWB`), in the green S_1 pocket.

(b) Thrombin is shown from the same angle of view with another ligand that accommodates a distinct region of the binding pocket (PDB code: `2C93`, ligand identifier: `C4M`), in the pale S_2-S_4 pocket.

Figure 2.8 Examples of thrombin subpockets. Hardly any overlap of the two pockets is given, if the pockets are extracted next to the bound ligands.

2-Hydroxy-4-Phenylbutyl]-4-Methoxy-2,3,6-Trimethylbenzenesulfonamide, `C4M`, see Fig. 2.8). Subsequently, we performed retrieval experiments based on these three query pockets in order to detect other thrombin cavities in the database. To detect the total number of thrombin entries in our database, we searched for a match with the EC number 3.4.21.5 (thrombin) and the presence of Asp189, a key residue in S_1 for substrate recognition, to guarantee that only the catalytic pockets were captured. In total, we detected 430 thrombin pockets which were used as reference to rank our subsequent retrieval experiments. Figure 2.9 displays the resulting receiver operator characteristics (ROC) curves which were obtained using the three query pockets and either the LPCS or ligand shape-based fingerprint approach.

As mentioned above, ROC curves are widely used to validate retrieval and enrichment results. True positive retrieval rates (y-axis) are plotted against false positive ones (x-axis) and the area under the curve (AUC) indicates the success of the method. As shown in Figure 2.9, LPCS achieves very convincing

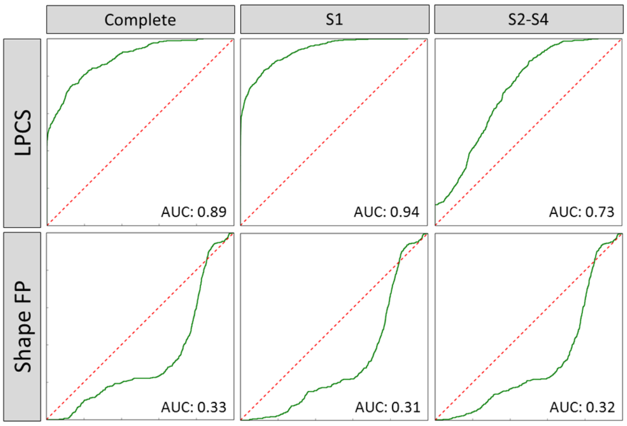


Figure 2.9 ROC curves illustrating the retrieval rates of database screenings based on the LPCS (first row) or the shape fingerprint (second row) approach, respectively. The first column depicts the ROC curves using the complete binding pocket, the second the S_1 subpocket and the third the S_2-S_4 subpocket as a query. The dashed red lines indicate random retrieval rate (AUC = 0.5).

retrieval success (AUCs are 0.94 and 0.89, respectively) and remarkable early enrichment slopes when the complete and the S_1 pocket are subjected as queries. Using the S_2-S_4 pocket as a query results in a somewhat worse ROC curve, though still much better than random (AUC = 0.73). The plots based on the ligand shape-based pocket fingerprint analysis show the unsatisfactory performance of this method when applied to compare any of the query pockets against the database of surface-exposed cavities. All ROC curves exhibit an AUC worse than random retrieval. This result demonstrates that the latter approach is much less robust than the LPCS approach with respect to substructure detection.

2.10 Conclusion

The presented study uncovers the inherent and highly biased shape information of binding sites if they are extracted in close neighborhood of the bound ligands. Simply considering the coarse distribution of potential interaction points in such a ligand shape-based pocket reveals retrieval success rates of more than

95 % in our classification experiments even when ligands of deviating size and conformations are analyzed. Even if high redundancies in protein sequence have been eliminated from the dataset (so-called “culling”) this simple comparison method still achieves high success rates of around 80 %. Any information about the distribution of physicochemical properties across the pockets was neglected and a simple ligand shape-determined fingerprint, assigned to each pocket, was sufficient to accomplish a successful comparison with minimal computational effort (more than 500 000 comparisons per second on a customary computer). We could show that the sole pocket size expressed by the number of interaction points is not discriminative. Thus, the information that enables classifications is stored in the spatial distribution pattern of the interaction points next to the ligands. This pattern is likewise determined as a kind of inflated ligand shape, as the spatial positions of bound ligands were used to extract a binding pocket. The fact, that these pockets can be regarded as size-inflated ligands, is demonstrated by a significant correlation of the distance distributions derived from the ligand atoms and the pocket interactions points defined in close neighborhood of the bound ligands.

An unbiased approach seeking for a cavity comparison of surface-exposed depressions on proteins does not make use of ligand information. Thus, also pockets found in uncomplexed proteins will be extracted and analyzed. The same holds for pockets extracted from the same reference protein which are accommodated by ligands addressing non-overlapping epitopes of the binding pocket. This strongly argues to only analyze and compare automatically detected surface-exposed cavities which are extracted unbiased from any ligand information. Only then surprising results with respect to putative cross-reactivity and functional annotation of orphan proteins can be expected. Most likely, comparative methods seeking similarities between automatically extracted cavities will require more computational effort, since the pockets will be larger and similarity may be detected in terms of subpockets (detecting a subset in pocket A, which is also present in pocket B). As major advantage also putative binding sites of uncomplexed or spatially differently accommodated proteins can be studied, which considerably expands the pocket space. This is of utmost importance in predicting unexpected cross-reactivity of newly developed drugs and will only be of relevance if the evaluation algorithm still detects similarity in subpockets. These criteria match, as convincingly shown in this study, by the LPCS approach, whereas the ligand shape-based fingerprint fails at this challenge.

3

Extended Graph-based Method[1]

A plethora of different methods has been proposed to approach the problem of structural comparison of protein binding sites. The group of surface-based methods (cf. Sec. 2.1) is especially interesting in the context of structure-based drug design in pharmaceutical research. The main idea is to investigate the physicochemical properties of the functional groups involved in the interaction between proteins and ligands, substrates or cofactors and compare binding sites with the goal of identifying proteins that, while different in sequence, are likely to interact with similar molecules. In principle this allows the identification of potential cross-reactivities in early stages of drug design, long before expensive clinical studies are conducted. It can also provide valuable ideas for new ligand scaffolds or how to decorate molecules to avoid undesired cross-reactivity.

As the advantage of using automatically detected protein binding sites was pointed out in Chapter 2, we will now focus on Cavbase [131, 132], a module of the Relibase+ system [61, 57] that was developed for storing information about

[1] Copyright 2014 IEEE. Reprinted, with permission, from Krotzky, T.; Fober, T.; Hüllermeier, E.; Klebe, G. Extended Graph-based Models for Enhanced Similarity Search in Cavbase. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2014**, *11(5)*:878–890. The major part of the study was performed by me, Thomas Fober developed the description and algorithm for the LC heuristic as part of his PhD thesis. Eyke Hüllermeier supervised the computer-science part of the work.

putative protein binding sites. Relibase+ and its modules are distributed and maintained by the Cambridge Crystallographic Data Centre (CCDC). Cavbase exploits the grid-based LIGSITE algorithm [60] to detect cavities exposed to the surface of proteins stored in the Protein Data Bank (PDB) [9]. To be considered in the database, a detected depression must comprise a cluster of more than 320 adjacent grid intersections (grid spacing of 0.5 Å resulting in about 40 $Å^3$) and it must exhibit a certain minimal penetration depth. Next, the putative binding sites are classified in terms of a sparse set of pseudocenters, leading to a compressed representation. The pseudocenters are assigned to atoms or groups of atoms of the amino acids that flank the detected cavity with a distance from the rim closer than 1.1 Å. They encode physicochemical properties that are important for molecular recognition next to the adjacent surface patch, potentially forming an interaction to the bound ligand. In total, seven different pseudocenter types are considered: H-bond donors, H-bond acceptors, H-bond donors/acceptors (which we will call "doneptors"), centers of aromatic rings, centers comprising pi electrons, aliphatic groups and metal ions (see Fig. 3.1 for an example).

One important feature of Cavbase is the option to mutually compare binding sites. While such a similarity consideration of entire proteins is easily performed on a sequence level by applying alignment methods like BLAST [2], a comparison on the structural level becomes computationally a much more demanding but required task. Proteins may exhibit similar functions, even though they are only sharing a low sequence similarity and overall fold [22, 124].

To measure the structural similarity between two protein binding sites, Cavbase follows a two-step procedure in the current implementation: At first, the two sets of pseudocenters are matched to find reasonable mutual superpositions. To this end, the pseudocenter sets are transformed into undirected graphs, which are then used to build a product graph. Afterwards, the Bron-Kerbosch algorithm (BK) [17] is applied on this product graph to

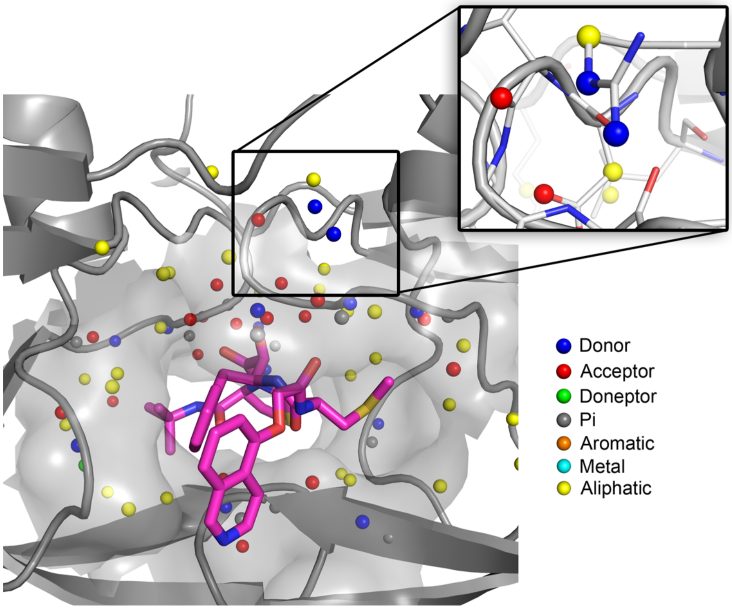


Figure 3.1 Example of a binding site representation of HIV protease in Cavbase (PDB code `1HPX`). A pocket is denoted by a set of pseudocenters in 3D space (colored spheres). Pseudocenters are assigned to specific parts of the amino acids flanking the cavity and representing their physicochemical properties, which is shown as a blow-up inset on the upper right hand side. The ligand (carbon atoms magenta) is shown for clarity. The illustration was prepared using PyMOL [135].

detect the 100 largest cliques, each representing a possible superposition of the two pseudocenter subsets. As the problem of clique detection is known to be NP-complete and therefore computationally quite demanding, Cavbase limits to the detection of this clique subset. To compensate for this simplification, it performs a second computational step to evaluate shape similarity. Every generated superposition is evaluated by considering an expanded set of surface points, which describes the putative contact surface of the protein. This second step is again computationally rather time-consuming. The final similarity score is calculated by assessing the maximal degree of overlapping substructures of two binding sites. This comparison procedure has already been successfully employed for the classification and clustering of protein structures [163, 85, 51].

The comparison procedure is, however, also computationally quite demanding and especially the determination of overlapping surface points represents the bottle-neck of the whole approach. To increase efficiency, we here propose a new approach in which the consideration of the surface points in the second step becomes obsolete. This is achieved by enriching the pseudocenter representation with additional information. With this option we make important geometric information about the binding pocket shape available during the calculations even if the second step of the comparison is not performed. By using the suggested extended representation we can determine similarities between protein binding sites by considering the first of the above mentioned steps only, which is based on clique detection. To enhance efficiency even further, we propose to use a simple but effective heuristic for the detection of cliques in a graph.

3.1 Modeling Protein Binding Sites

In this chapter, we represent protein binding sites in terms of graphs which are derived from the pseudocenter representation of Cavbase. Here, depressions on the surface of proteins from the PDB [9] are detected by using the grid-based LIGSITE algorithm [60]. The geometry of a protein binding site is subsequently represented by a set of pseudocenters. Currently, seven types of pseudocenters are used that account for different types of possible interactions between residues of the protein and a ligand. These pseudocenters are assigned to fragments of amino acids that are flanking the detected cavity surface in a distance closer than 1.1 Å and can be assigned to one of the following types: *donor, acceptor, donor-acceptor, pi, aromatic, aliphatic* and *metal*. For a detailed description on the pseudocenter representation the reader is referred to [132]. To represent such structures, Cavbase makes use of node-labeled and edge-weighted graphs, where a node corresponds to a pseudocenter and is assigned to one of the seven above-mentioned physicochemical properties. To capture geometry, complete graphs are used where an edge is weighted by the euclidean distance between two neighboring pseudocenters. We will also stick to the representation by means of graphs, since they are a prominent data structure in computer science, they allow to capture sufficient geometric and physicochemical information and are rotationally and translational invariant. Moreover, suitable methods for the detection of subgraphs are available. Surface points in Cavbase are generated by the LIGSITE algorithm. They approximate the solvent accessible surface (SAS) of a protein binding pocket. As an example, Figure 3.2 shows an ATP binding site of human mitochondrial NAD(P)+-dependent malic enzyme (`PDB: 1GZ4`). Small dots denote surface points that represent the shape of the protein's surface. All the other larger spheres represent pseudocenters flanking the cavity. We will enrich the pseudocenters by properties using the additional information derived from its neighboring surface points. Thus the size of a graph, which is defined by the number of nodes, remains equal to the number of pseudocenters and the much larger sets of surface points are no longer evaluated. Unlike many of the approaches hitherto developed [13, 47, 45, 100, 43, 163], we thus also consider surface point information. As a first extension, we enrich the information assigned to a pseudocenter by descriptors defining the shape of its

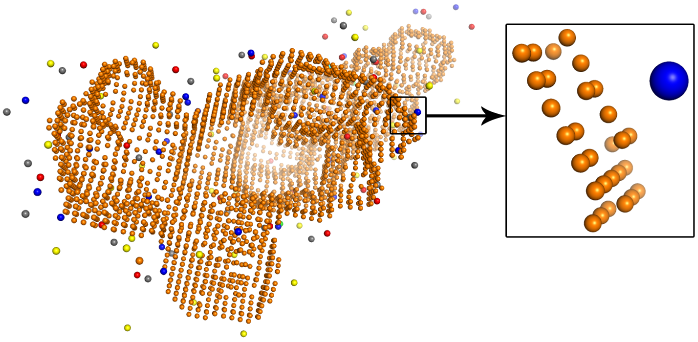


Figure 3.2 Example of an ATP-binding site illustrated by a set of surface points and pseudocenters (`PDB: 1GZ4`). Small orange dots represent surface points that denote the pocket's surface shape. Pseudocenters are depicted as larger spheres. The box on the right hand side shows a close-up view of a surface patch consisting of one pseudocenter and its corresponding set of surface points. The figure was prepared using PyMOL [135].

corresponding surface patch. We therefore examined various methods, such as evaluating the standard deviation of all distances from a pseudocenter to its spatially adjacent surface nodes. However, many tested methods did not perform properly and stimulated us to develop two other methods that turned out to work efficiently in this context. They will be described in the following.

3.1.1 Shape of a Surface Patch by Using a Weighted PCA

We extend a pseudocenter's attribute by two additional properties providing a description of the shape of the surface patch adjacent to the pseudocenter: *convexity* and *concavity* as it has been demonstrated that molecular recognition strongly depends on the shape of a binding pocket [70, 27, 117, 127, 32]. To determine both properties and to summarize them with a real number, a weighted principal component analysis (PCA) was applied to the set of surface points representing a patch next to a pseudocenter. This method has already been presented in an earlier study [81] and has been slightly modified for its application in the present work. Considering an n-dimensional space, a PCA is able to detect up to n perpendicular planes (so-called *principal components*) establishing a new coordinate system. Having detected these principal components, the first one indicates the coordinate in which the points show highest variance, the second one the coordinate with second-highest variance, etc. We evaluate the first two principal components which allow to determine the degree of convexity or concavity by measuring the distances between the axis of the first principal component and the surface points. Such a PCA was performed for every single surface point of a patch. To account for the high diversity of surface patches a *weighted* PCA was employed in order to make the approach more robust against exceptional surface point distributions. Thus, the initial surface point has the largest influence on the position of the principal component; the influences of the subsequent points depend on their distance to the initial one (Fig. 3.3). In detail, for a point q a weight $w = e^{-\frac{1}{2}d(q_i,q)}$ was applied, where $d(q_i, q)$ returns the distance between the initial surface point q_i and q. We chose $\frac{1}{2}$ as scaling factor for the distance to account for the grid space (0.5 Å) of the LIGSITE approach, which generates the surface points to be processed.

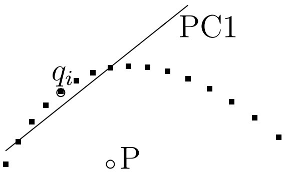


Figure 3.3 Schematic illustration of a pseudocenter (P) and its associated surface points (black squares) after transformation into 2-dimensional space by using a weighted PCA based on a single initial surface point q_i. PC1 denotes the first principal component, which also represents the x-axis of a new coordinate system. The position of the pseudocenter is used to determine a local degree of convexity and concavity for q_i (see Eq. (3.1)). Therefore, the distances of all surface points that are located on the same side of PC1 as the pseudocenter are added to retrieve the convexity score of point q_i. In the same way, all surface points on the opposing side of PC1 account for the concavity score of q_i.

After having calculated the weighted PCA, the coordinates of the pseudocenters also had to be transformed to the new coordinate system. Afterwards, two numbers were calculated returning the degree of convexity and concavity of a surface patch, respectively. If a surface point is located on the same side of the first principal component as the pseudocenter, it accounts for the convexity contribution; otherwise for the concavity contribution. Let Q be the set of surface points of a certain surface patch and P the corresponding pseudocenter. For $q \in Q$ its contribution is determined by the following procedure:

$$q_{convex} = \sum_{v \in Q} \begin{cases} \frac{|v_y|}{ymaxConvex} & \text{, if } sgn(v_y) = sgn(P_y) \\ 0 & \text{, otherwise} \end{cases} \tag{3.1}$$

$$q_{concav} = \sum_{v \in Q} \begin{cases} 0 & \text{, if } sgn(v_y) = sgn(P_y) \\ \frac{|v_y|}{ymaxConcave} & \text{, otherwise} \end{cases}$$

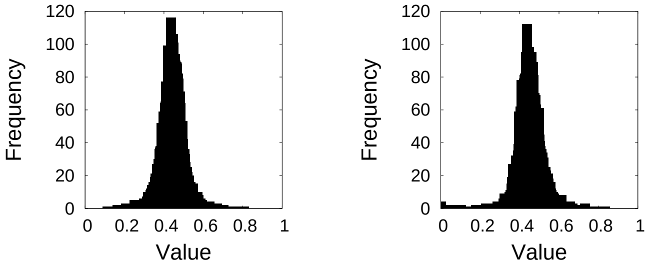
where $sgn(x)$ stands for the signum function returning -1 if $x < 0$, 0 if $x = 0$ and 1 otherwise. P_y and v_y denote, respectively, the y-coordinates of pseudocenter P and the surface point v in the new coordinate system. $ymaxConvex$ and $ymaxConcave$ are the maximum absolute values of those surface points' y-coordinates that contributed to q_{convex} and $q_{concave}$ respectively. In the following, the convexity and concavity scores of a surface point $q \in Q$ were divided by the number of the points $v \in Q$ that have contributed to the respective score (normalization):

$$\begin{aligned} q_{convexNorm} &= \frac{q_{convex}}{\sum_{v \in Q} \begin{cases} 1, \text{ if } sgn(v_y) = sgn(P_y) \\ 0, \text{ otherwise} \end{cases}} \\ q_{concavNorm} &= \frac{q_{concave}}{\sum_{v \in Q} \begin{cases} 0, \text{ if } sgn(v_y) = sgn(P_y) \\ 1, \text{ otherwise} \end{cases}} . \end{aligned} \tag{3.2}$$

In the last step all collected $q_{convexNorm}$ and $q_{concaveNorm}$ numbers were summed up and divided by the total number of points on the patch $|Q|$ to obtain the final convexity score $P_{convexity}$ and concavity score $P_{concavity}$ of the whole patch. As a result of this normalization, the two scores fall into the interval $[0, 1]$. For illustration, the distributions of the convexity and concavity scores of 10 000 randomly chosen patches are given in Figure 3.4 (a) and (b). Both have a maximum in the interval $[0.4, 0.5]$, which is rather densely populated. Furthermore, an example of a rather convex patch is shown in Figure 3.5. In this case, the surface points nicely reflect the convex shape of the patch, which results in a convexity score that is much larger than the concavity score ($P_{convexity} - P_{concavity} = 0.54$).

3.1.2 Shape of a Surface Patch by Using Histograms

Besides using the PCA to derive descriptors for patch convexity and concavity, the spatial distribution of the surface points will be considered by assigning a histogram H to each patch reflecting their distribution. Ballester and Richards



(a) Distribution of convexity scores.

(b) Distribution of concavity scores.

Figure 3.4 Distributions of the convexity and concavity scores of 10 000 randomly chosen patches.

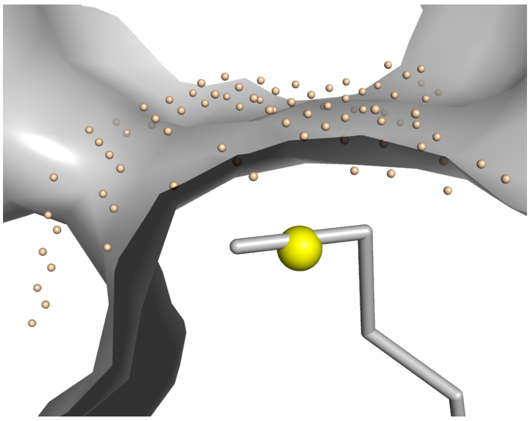


Figure 3.5 Example of a rather convex patch, where the surface points (small dots) nicely reflect the convex shape of the protein surface represented as Connolly SAS surface (shown in gray). Applying the wPCA approach results in a convexity score that is much larger than the concavity score ($P_{convexity} - P_{concavity} = 0.54$).

showed that distance histograms are able to capture important properties of a spatially distributed pattern [7]. To derive such a histogram providing information about the shape of a patch, for each patch the euclidean distances from the pseudocenter to all surrounding surface points were calculated and assigned to bins in the histogram H with a bin size of 0.1 Å. Subsequently, the histogram was smoothed by applying a sliding window averaging to achieve a low-pass filtering in the distribution [29] and thus reduce the impact of outliers or inaccuracies in crystallographic data. The actual realization of this filter is given by

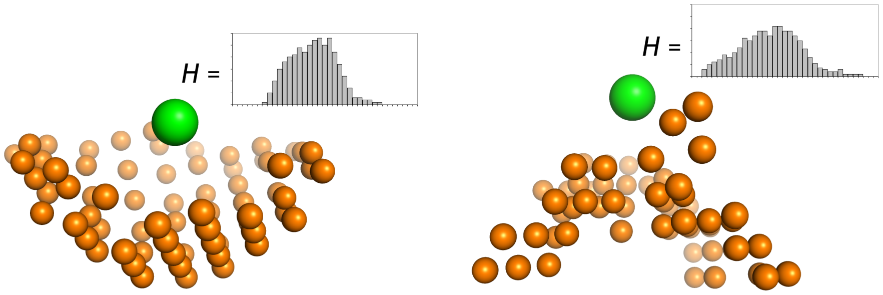
$$H(x) = \frac{1}{k+1}\left[H(x) + \sum_{i=1}^{k/2}\left(H(x-i) + H(x+i)\right)\right] ,$$

where k is restricted to be even. The values of each bin x are first increased by those of their k nearest neighbors and then divided by the number of collected values $k+1$. Thus, strong local peaks and deep depressions are reduced while low frequencies of the distributions remain virtually unaffected. We chose $k = 4$ to account for slight inaccuracies during the detection of surface points. In effect, a graph node thus carried in addition to information about its original properties (i.e. position in 3-dimensional space and physiochemical property) a histogram H summarizing the geometric distribution of the surface points (Fig. 3.6).

Both methods for examining the curvature of a patch, the wPCA and the histogram approach, are independent from pseudocenter types and solely capture information about the surface shape.

3.1.3 Fuzzy Representation of Physicochemical Properties

Furthermore, we considered the representation of a vector, to express in a more discriminating way the physicochemical properties of the environment of a pseudocenter. It replaces the original scalar physicochemical property assigned to each pseudocenter. The vector contains six components, namely *donor, acceptor, pi, aromatic, aliphatic* and *metal*, which are denoted values in the



(a) A rather convex surface patch. (b) A rather concave surface patch.

Figure 3.6 Illustration of two surface patches with a rather different shape (`PDB:1GZ4`). The large spheres represent pseudocenters, the smaller dots indicate the surface points which describe the surface shape next to a pseudocenter. To each pseudocenter a histogram H is assigned which captures the spatial distribution of the patch's surface points. While the surface of the patch shown in (a) is rather convex, the shape of the surface in (b) is more concave, which leads to different distributions of the histograms: The values of the histogram (b) are more widely distributed over its entire distance range.

interval [0,1]. The seventh type *donor-acceptor* that represents in Cavbase a further scalar property is no longer needed as it can be represented by the two vector components donor and acceptor. Earlier studies showed that by using such a vector (instead of a scalar), a fuzzy representation of physicochemical properties of the cavity is obtained [84, 81], which can be more appropriate for the comparison of protein binding sites, especially as related proteins show mutational replacements. To calculate the so-called *patch vector* $\vec{P}$ for a pseudocenter, each surface point $q \in Q$ of the patch was considered and the following vector is calculated for q:

$$\vec{t_q} = \begin{pmatrix} \min\left(1, \sum_{P\in\mathrm{Do}} f(q,P,12)\right) \\ \min\left(1, \sum_{P\in\mathrm{Ac}} f(q,P,12)\right) \\ \min\left(1, \sum_{P\in\mathrm{Ar}} f(q,P,12)\right) \\ \min\left(1, \sum_{P\in\mathrm{Pi}} f(q,P,12)\right) \\ \min\left(1, \sum_{P\in\mathrm{Me}} f(q,P,12)\right) \\ \min\left(1, \sum_{P\in\mathrm{Al}} f(q,P,14)\right) \end{pmatrix},$$

where Do in the first component is for instance the subset of pseudocenters of the binding site which are labeled with the type donor. The function f is defined as

$$f(q, P, t) = \frac{1}{1 + e^{(1+(4d(q,P)-t))}}$$

and is possessing a sigmoid gradient. It is used together with the parameter t to adjust the impact of a pseudocenter P on q, which depends on the euclidean distance d between both points. According to [84, 81] we chose $t = 14$ for aliphatic pseudocenters and $t = 12$ for non-aliphatic ones to account for the different interaction radii of these types (see Fig. 3.7). Note that pseudocenters labeled as "donor-acceptor" contribute to both, the donor and the acceptor component simultaneously, each weighted by 0.7. Finally, to obtain the patch vector, for all surface points $q \in Q$ the vectors $\vec{t_q}$ are averaged, leading to

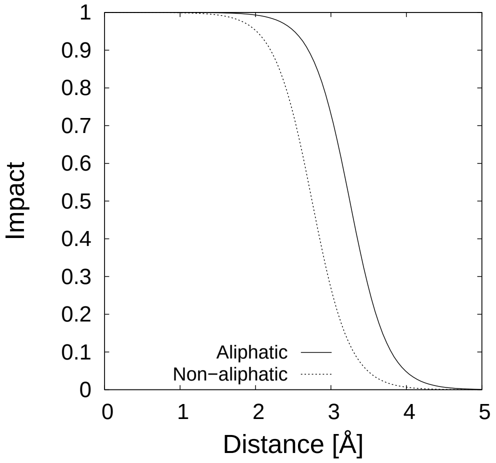


Figure 3.7 Gradients of the function f that is used to weight the impact of aliphatic and non-aliphatic pseudocenters on a patch vector.

$$\vec{P} = \frac{1}{|Q|} \cdot \begin{pmatrix} \sum_{q \in Q}[t_q]_1 \\ \vdots \\ \sum_{q \in Q}[t_q]_6 \end{pmatrix},$$

where the notation $[v]_i$ is used to express that the i-th entry in the vector v is considered. Typically, the component of $\vec{P}$ will have the highest value (close to one) that matches the former definite property of the pseudocenter. However, also the physicochemical environment will now be captured by the other components.

3.1.4 Summary of the Enhanced Node Descriptors

By using the extensions proposed in this paper, the pseudocenters could be enriched by additional valuable information. We still apply a graph $G = (V, E, \ell_V, \ell_E)$. Each pseudocenter is represented by one node. Complete graphs are computed and each edge is labeled with the euclidean distance ℓ_E (cf. Sec. 3.1). However, the function ℓ_V is represented in different ways. It assigns either

- the original attribute giving the physicochemical property,
- the convexity/concavity descriptors,
- the newly introduced patch vector,
- the surface histogram

or several combinations of the list to the node v.

3.2 Comparison of Binding Sites

For the comparison of protein binding sites different techniques to determine similarity have been described, which are all based on the pseudocenter as descriptors [13, 47, 45, 100, 43, 163]. In this paper, we will focus on the extended pseudocenter description and we will apply an established technique to elucidate similarity, namely the maximum common subgraph (MCS) principle. Goal of this article is not to evaluate different strategies to determine similarity but to validate different pseudocenter representations of protein binding sites.

3.2.1 Original Cavbase Measure

The similarity retrieval in Cavbase consists of two steps [132]: At first, the pseudocenter representation of two protein binding sites to be compared is transformed into node-labeled and edge-weighted graphs as described in Section 3.1. In these two graphs the 100 largest common subgraphs are searched by first generating a product graph and then applying the exact BK, however, not with the goal to define similarity but instead to allow a superposition of the surface points. Hence, in the second step, for each detected common subgraph an optimal superposition is calculated by means of the Kabsch algorithm [69], leading to a transformation rule. This transformation rule is finally applied to the surface points and the number of overlapping points is counted, leading to a similarity score. To finally find the best score between the two protein binding sites, the largest from the set of up to 100 scored solutions is considered.

This two-step approach was motivated by the following observations: Finding common subgraphs by means of clique detection [17] is NP-hard. To enable efficient calculations the input should be as small as possible. Here, the pseudocenter description seems to be an ideal input because they generally lead to a very concise binding site representation. Once the MCS has been detected, similarity can be computed in a straight-forward way [20]. However, Schmitt et al. [131] did not follow this approach, because, based on the pseudocenter representation, multiple solutions of deviating clique composition are generated which have to be ranked and compared. Therefore, in a second step, the surface points are considered to rank the obtained clique solutions by calculating a degree of surface overlap.

However, this final scoring is computationally the most demanding step of the whole procedure, as demonstrated in Figure 3.8. This finding is rather obvious, since the set of surface points that describe a pocket shape is much larger than the set of pseudocenters (see Fig. 3.2). It has already been

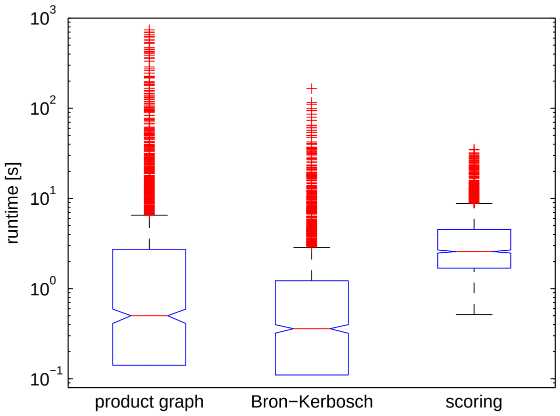


Figure 3.8 Runtime analysis of the different steps for a binding site comparison performed by Cavbase. The final scoring that is based on the surface points is by far the slowest part (average runtime: 4.12 ± 4.45 s.). 1000 randomly selected pairs of binding sites were used for this evaluation.

stated by Schmitt et al. [132] that this step is more than 30 times slower than the preceding clique detection. In contrast to Cavbase, we perform the comparison only at the pseudocenter level to enable a very efficient approach, simultaneously reducing the loss of information by enriching the pseudocenters with shape information derived from the surface points.

3.2.2 Maximum Common Subgraph (MCS)

The MCS measure defines similarity directly in terms of the detected maximum common subgraph. To find a common subgraph, methods based on clique detection are widely applied with success. The central idea of these methods is to exploit the fact that the maximum clique in the product graph $G_{\otimes}$ of two input graphs G and G' corresponds to the MCS of G and G' [20]. In detail, the product graph is defined by node-labeled and edge-weighted graphs $G = (V, E, \ell_V, \ell_E)$ and $G' = (V', E', \ell'_V, \ell'_E)$ as $G_{\otimes} = (V_{\otimes}, E_{\otimes})$, where $V_{\otimes} = \{(v, v') \in V \times V' \mid \ell_V(v) = \ell'_V(v')\}$ and $E_{\otimes} = \{((v, v'), (w, w')) \in V_{\otimes}^2 \mid \|\ell_E((v, w)) - \ell'_E((v', w'))\| \leq \epsilon\}$. For ϵ we chose 2.0 Å as this value is also used in the original Cavbase implementation and cannot be freely varied in the compiled version (it is hard coded). An MCS, which is subject to the constraint that $\|\ell_E((v,w)) - \ell'_E((v',w'))\| \leq \epsilon$, represents a partial alignment, i.e., a one-to-one correspondence of a subset of nodes defining the graphs. Detecting an alignment of nodes allows assigning the alignment of edges indirectly. Obviously, such an alignment can also be used to derive the similarity between two graphs: The larger the MCS with respect to the size of the larger graph, the more similar both graphs will be.

Unfortunately, the detection of the largest clique in a graph is an NP-hard problem, therefore we will also propose a heuristic approach in the following, which is computationally more efficient.

Heuristic for the Detection of Cliques

The problem of maximum clique detection belongs to the class of problems which are difficult to solve even with accelerated versions of the BK [78] or in approximated terms [40]. In this unfortunate situation and given the need to implement a practically tool, it makes sense to apply heuristics. In the literature, many approaches, based on genetic algorithms, relaxations or quadratic programming are proposed. Neither of these techniques guarantees any quality of the solution and they can still become quite inefficient (e. g. a genetic algorithm). Therefore, we decided to use a simple, though efficient approach following the mining of quasi-cliques, as originally proposed by Liu and Wong [96] (Fig. 3.9). To elucidate a local clique LC in a graph $G = (V, E)$, a node $v \in V$ is selected and a neighborhood graph $G^{(v)} = (V_v, E_v)$ is calculated, where $V_v = \{w \in V \mid (v, w) \in E\ \}$ and $E_v = V_v^2 \cap E$. As long as the clique property is not satisfied for $G^{(v)}$, the node in V_v that has the smallest degree together with all its adjacent edges is iteratively removed. In case several nodes are present with equal degree, one node is chosen at random. Once the clique property is fulfilled, an LC is formed that comprises the node v. To test whether a graph G fulfills the clique property criteria, the function *dens* is

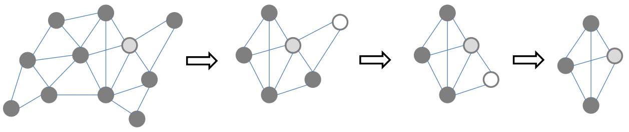


Figure 3.9 Detection of maximum cliques: Each node of the graph is considered once (here the light-gray node) for which the neighborhood graph is constructed (2^{nd} step). This graph is tested for the clique property, which is violated. In the next step the node (white) with the smallest degree is removed and again a test for the clique property is performed. This procedure continues until a clique is obtained.

used, which returns 1.0 if the graph G is a clique. This function is defined as

$$dens(G) = \frac{2 \cdot |E|}{|V| \cdot (|V| - 1)} \ .$$

To generate the set of all local cliques, each node in the graph is considered, hence a set of $|V|$ local cliques is obtained from which the largest one is returned as final result. Due to the NP-hardness of the problem, cliques found in this step will most likely neither be the optimal solution for the maximum clique problem nor will the maximal clique contain v, as illustrated in Figure 3.10. However, we found that this heuristic leads to valuable results when it is applied to the product graph of two graphs representing protein binding sites. Figure 3.11 (a) shows the relative sizes of 1000 common subgraphs, which have been detected by the LC approach, compared to the corresponding MCS

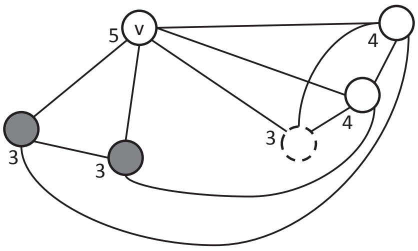
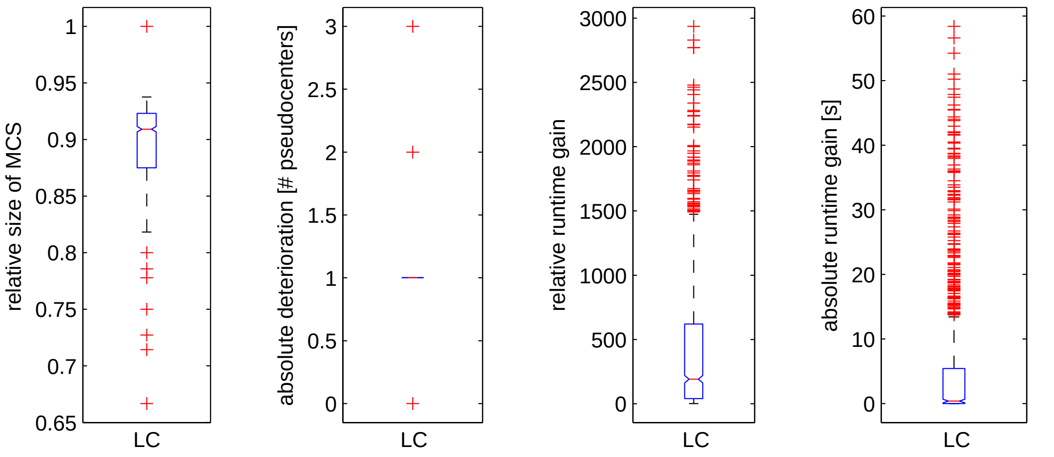


Figure 3.10 Illustrative example for the optimality of the local clique approach: The maximum clique consists of the nodes that are not filled, thus, it has size 4. The dashed and the gray filled nodes have a degree of 3. If the algorithm chooses the dashed node at random, the maximum clique will not be found. The added numbers indicate the degree of connectivity of the different nodes.

found with the exact BK. The average relative size is 0.91, which indicates that the heuristic finds subgraphs that are on average only 9 % smaller than the correct MCS. In (b) the absolute deterioration in terms of the number of pseudocenters is shown (mean = 0.87). For proving the quality of these results, it is worth mentioning that the average size of the 1000 MCS is 16.2 ± 4.3 (average size of the input graphs: 93). Figure 3.11 (c) shows the relative runtime gain of LC, which on average is 417-fold, as well as the absolute gain in (d), which on average is 5.3 seconds. This evaluation confirms that LC is actually able to find the correct or at least a solution close-to the correct MCS in a significantly shorter runtime than the exact BK algorithm.



(a) Relative size of LC's detected MCS compared to the MCS of BK (0.91 ± 0.06).

(b) Absolute deterioration of LC's MCS compared to the MCS of BK (0.87 ± 0.56 pseudocenters).

(c) Relative runtime gain of LC (416.85 ± 521.09).

(d) Absolute runtime gain of LC (5.29 ± 10.09 seconds).

Figure 3.11 MCS sizes and runtimes of LC compared to those of BK by testing 1000 common subgraphs.

3.2.3 Modification of the Product Graph Definition

The set of nodes $V_\otimes$ in the product graph is defined as

$$V_\otimes = \{(v, v') \in V \times V' \mid \ell_V(v) = \ell'_V(v')\} \ , \tag{3.3}$$

hence it cannot be applied directly to our new graph model or, to be more precise, to the new function ℓ_V. Therefore, we introduce a new binary function *equals* and replace the test for equivalence in Equation 3.3 by this function. We distinguished again three different cases, which depend on the concrete realization of ℓ_V. Using our extensions in the graph representations new approaches for comparison have to be defined that take the additional node properties into account.

Comparison of Pseudocenter label

First, we define a function e_1 that is realized in the same way as in Cavbase, i.e., it returns “true” if the physicochemical properties of two nodes match and “false” otherwise.

Comparison of curvature degrees

Convexity and concavity descriptors indicating the shape of the surface points were assigned to patches by a weighted PCA. For a rather convex patch the convexity descriptor is much larger than the concavity descriptor, expressed as $[\ell_V(v)]_2 \gg [\ell_V(v)]_3$. The comparison of the convexity and concavity is achieved by calculating two distances, namely

$$d_2 = |[\ell_V(v)]_2 - [\ell_V(v')]_2|$$

and

$$d_3 = |[\ell_V(v)]_3 - [\ell_V(v')]_3| \ ,$$

where d_2 and d_3 indicate the distances between two convexity and concavity descriptors, respectively. The distances are then combined by calculating the harmonic mean:

$$\Delta(v, v') = \frac{2}{d_2^{-1} + d_3^{-1}} \ .$$

The value returned by this expression indicates the differences between the convexity and concavity descriptors of the nodes v and v'. It thus represents the difference of the curvature of the patches attributed to two pseudocenters. By using a threshold δ it is possible to decide whether the curvatures match to a certain degree, by evaluating $\Delta(v, v') < \delta$, which finally defines a similarity function we call e_2. If one of the four values used (convexity and concavity descriptors of the nodes v and v') is 0, the returned value will also be 0, since no shape comparison can be carried out. In such a case, the function is set to return 'true', thus the similarity scoring is only applied by evaluating the other properties used in the node representation.

Comparison of patch vectors

For comparing the patch vectors, different measures can be applied [34]. We employed the scalar product and required that the result is larger or equal to a threshold of 0.7 to define two vectors as equal; otherwise they are considered unequal. As stated by Kuhn [84] this is to guarantee that also a pair of nodes labeled, respectively, as donor-acceptor and donor or donor-acceptor and acceptor is recognized as a pair of equal nodes. The measure $e_3(\ell_V(v), \ell'_V(v'))$ thus is defined as

$$[\ell_V(v)]_{1..6} \cdot [\ell_V(v)]_{1..6} \geq 0.7 \ ,$$

where the notation $[v]_{i..j}$ is used to express that the elements i up to j are extracted from the vector v to build a new vector v'.

Comparison of histograms

Beside using two real numbers indicating the curvature of a patch, we can also use histograms. Here the node labeling function becomes $\ell_V : V \rightarrow \mathcal{L}_V \times \mathcal{H}$, where the set $\mathcal{L}_V$ is assigned a surface histogram out of the set $\mathcal{H}$. To test two histograms for similarity we used the Jaccard distance [66] that is defined for two histograms h_1 and h_2 as

$$d_J(h_1, h_2) = 1 - \left(\frac{\sum_{i=1}^{\max\{|h_1|,|h_2|\}} \min\{h_1(i), h_2(i)\}}{\sum_{i=1}^{\max\{|h_1|,|h_2|\}} \max\{h_1(i), h_2(i)\}} \right) .$$

In case that one of two histograms covers a larger distribution, the shorter one is extended by empty bins.

The Jaccard distance returns real numbers. To ensure a binary result, the returned values have to fall beyond a given threshold δ_h and thus the comparative function e_4 is defined as $d_J(h_1, h_2) < \delta_h$.

Product graph

To determine whether a new node is inserted in the product graph, two nodes were compared by using conjunctions of the above-defined comparison methods. Besides comparing nodes by only considering the physicochemical property, we used several conjunctions of the comparison methods e_1 to e_4 and the LC approach for the heuristic detection of the maximum clique.

In our first extension we also take the curvature of the surface patch into account. Here, the function ℓ_V becomes a mapping $V \to \mathcal{L}_V \times \mathbb{R}^2$, where the set $\mathcal{L}_V$ contains the physicochemical property and the two real values indicate, respectively, the degree of convexity or concavity. Given two nodes $v \in V$ and $v' \in V'$, the function $equals(\ell_V(v), \ell'_V(v'))$ returns 'true' if the conjunction of $e_1(\ell_V(v), \ell'_V(v'))$ and $e_2(\ell_V(v), \ell'_V(v'))$ is 'true'. We will subsequently examine several values for δ in e_2. In the following sections, this approach will be referred to as LC_{Curv}.

In the second extension, the original label of the physicochemical property is used together with the histogram. To check for equivalence of the physicochemical properties, e_1 is used again. The outcome of this test is connected with the function $d_J(h_1, h_2)$, which is e_4, by a conjunction, eventually leading to another realization of the function *equals* suitable for this kind of node labels. This approach is called LC_{Hist}.

Finally, we used a combination of physicochemical properties, curvature degrees and histograms to evaluate the similarity of two nodes. In this approach, the function *equals* is realized by a conjunction of e_1, e_2 and e_4. Since there are in this case two thresholds to be examined, the evaluation becomes a multi-dimensional optimization problem. We will show the influence of different values for δ and δ_h later, and the evaluation method will be named $\text{LC}_{\text{CurvHist}}$.

In an additional step we replaced the function e_1, which is used to compare the physicochemical labels of two nodes, by the function e_3, which compares two patch vectors. This leads to three additional methods called $\text{LC}_{\text{PatchCurv}}$, $\text{LC}_{\text{PatchHist}}$ and $\text{LC}_{\text{PatchCurvHist}}$. Table 3.1 summarizes the combinations of all methods to define *equals*.

Table 3.1 Summary of the different techniques to build the product graph. The name of each method is given together with its functions to define the final similarity function *equals* which determines whether two nodes are similar or not.

Method	**used functions to realize** *equals*
LC	e_1
LC_{Curv}	$e_1 \wedge e_2$
LC_{Hist}	$e_1 \wedge e_4$
$\text{LC}_{\text{CurvHist}}$	$e_1 \wedge e_2 \wedge e_4$
$\text{LC}_{\text{PatchCurv}}$	$e_3 \wedge e_2$
$\text{LC}_{\text{PatchHist}}$	$e_3 \wedge e_4$
$\text{LC}_{\text{PatchCurvHist}}$	$e_3 \wedge e_2 \wedge e_4$

3.3 Experimental Study

In the experimental part of this study, we investigated the six options to extend the graph model which have been introduced above. To compare the resulting graphs, similarity was derived from the MCS calculated by means of the LC heuristic to enable a very efficient approach. As a reference, we chose the measures *Cavbase* and *mcs* (based on the classical graph-model), the latter being realized by means of the exact BK algorithm or the LC heuristic, respectively. In case of the LC approaches, the similarity score of two pockets is the size of the detected MCS normalized by the size of the larger input graph, respectively. In Cavbase, the detailed similarity score arises from the degree of the corresponding surface overlap.

The assessment of a degree of similarity with respect to bio-molecular structures, such as protein binding sites, is clearly a non-trivial task. In particular, since the concept of similarity is rather fuzzy by itself, it is difficult to evaluate different alternative measures in an objective way. To circumvent this problem, we propose to evaluate similarity measures in an indirect way, namely by means of their performance in the context of a nearest neighbor (NN) classification. The underlying idea is that the better a similarity measure performs, the better the predictive power of an NN classifier (using this measure for determining similar cases) can be assumed. More specifically, we measured performance by means of a 10-fold cross-validation procedure on two-class datasets, to be introduced in the following.

3.3.1 Data

We selected three two-class datasets of binding pockets that each bind the same ligand.

Class one of the first dataset consisted of pockets hosting NAD as a cofactor, where only the variants un-phosphorylated at C2' were considered (ligand identifier in the PDB: `NAD` and no `NAP`). The second class contained ATP. This gives rise to a binary classification problem: Any given protein binding site in the dataset may either bind NAD or ATP. In detail, a set of 355 protein binding pockets has been compiled that host either ATP or NAD as a cofactor (214 NAD and 141 ATP, see Sec. 2.3 for a detailed description). The dataset is termed ATP/NAD$_{small}$.

As a second dataset we used the even larger ATP/NAD$_{large}$ sets (402 NAD and 420 ATP binding pockets). This represents a dataset with diverse ligand conformations as both ligands feature a considerable number of rotatable bonds (11 and 8, respectively) and no RMSD constraint is used this time. Furthermore, a smaller difference of class sizes resulted, which poses an additional challenge in the classification experiments as the success rate of a simple random decision procedure increases according to the imbalance of the class sizes [35].

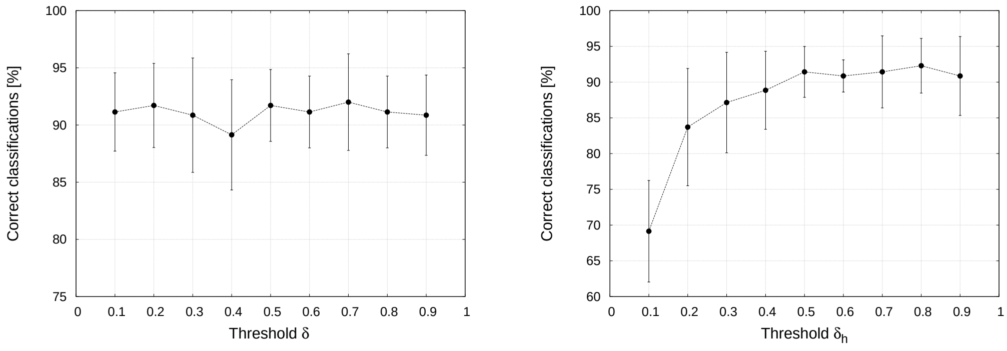
Finally, we used the class of 433 FAD pockets which will be used together with the 402 aforementioned NAD pockets to perform a third two-class classification experiment. FAD has been chosen, since it is more similar in size to NAD than ATP.

3.3.2 Results

For the evaluation of our methods we generated $n \times n$ scoring matrices of the described datasets. To the matrices we applied a 10-fold cross-validation procedure using an NN classifier to calculate the rates of correct classifications. In a preliminary step, we evaluated the influence of the threshold parameters δ_h and δ on the LC extensions LC_{Curv}, LC_{Hist} and $\text{LC}_{\text{CurvHist}}$. Both parameters were considered in the interval $[0.1, 0.9]$ with a step size of 0.1. For evaluation we used different settings in a classification experiment on the smaller and less complex dataset consisting of 355 pockets. We then applied the best performing parameter setting to classification experiments using the larger and more complex datasets.

Classification results using the small dataset

Figure 3.12 (a) summarizes the classification rates for different settings of δ if the model using information about curvature is used together with the LC approach (LC_{Curv}). Regarding the threshold δ, the best classification rate ($92 \pm 4.22\,\%$) is observed for the value 0.7. In consequence we decided to apply this threshold to our larger datasets. In case of the model based on the surface histograms to describe the patches' surface shape and its corresponding threshold δ_h (Fig. 3.12 (b)) we obtained $\delta_h = 0.8$ as best setting which was transferred to the subsequent experiments.



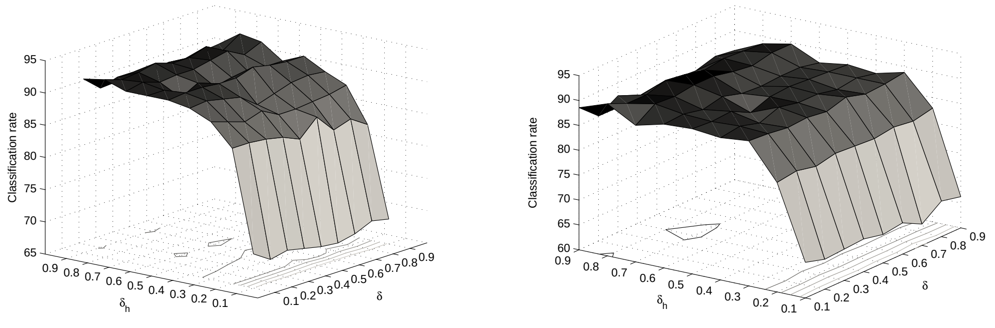
(a) Classification results of $\mathrm{LC}_{\mathrm{Curv}}$ depending on δ.

(b) Classification results of $\mathrm{LC}_{\mathrm{Hist}}$ depending on δ_h.

Figure 3.12 Influence of the δ and δ_h parameter on the classification results using a dataset that consists of 355 ATP and NADH binding pockets, respectively. The tested methods are using e_1 as a part of the *equals* function which assesses similarity between two pseudocenter labels. Error bars indicate the standard deviation, points indicate the mean of the correct classification rates of a 10-fold cross-validation.

In the last approach which uses the physicochemical label in the *equals* function the two extensions that describe the surface shape of a patch (functions e_2 and e_4) were used in combination with e_1 to realize the final function *equals*. As there were two parameters to be adjusted the resulting classification rates led to a 3-dimensional landscape (Fig. 3.13 (a)). This reaches its maximum for $\delta_h = 0.7$ and $\delta = 0.7$. Here, the classification rate achieves $92.57 \pm 4\,\%$, accordingly we used these settings in the following classification experiment for $\mathrm{LC}_{\mathrm{CurvHist}}$.

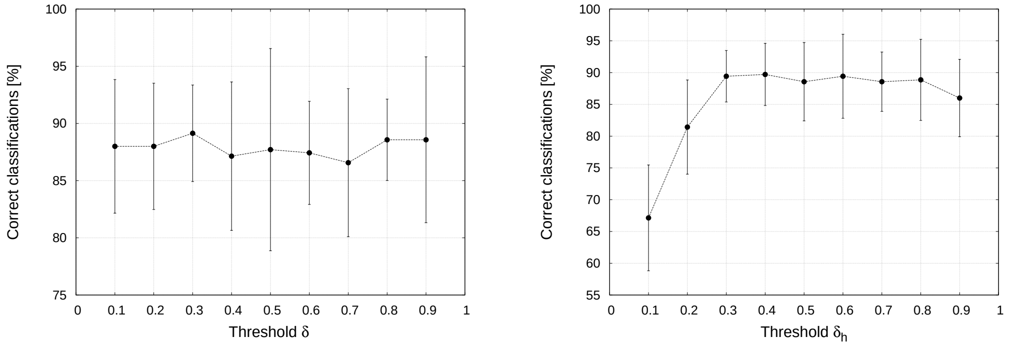
In the next three methods $\mathrm{LC}_{\mathrm{PatchCurv}}$, $\mathrm{LC}_{\mathrm{PatchHist}}$ and $\mathrm{LC}_{\mathrm{PatchCurvHist}}$ the function e_1 is replaced by e_3 since the physicochemical property of a patch was no longer represented by a scalar but by a six component vector. Figure 3.14 (a) shows the results of $\mathrm{LC}_{\mathrm{PatchCurv}}$ using curvature parameters beside the patch vector. For $\delta = 0.3$ the highest classification rates of more than $89\,\%$ can be observed. The method $\mathrm{LC}_{\mathrm{PatchHist}}$ performs equally well and the best



(a) $\mathrm{LC}_{\mathrm{CurvHist}}$ (b) $\mathrm{LC}_{\mathrm{PatchCurvHist}}$

Figure 3.13 Classification results of LC approaches on the small ATP/NAD dataset (355 pockets) depending on δ and δ_h.

results (89.71 ± 4.89) are obtained using $\delta_h = 0.4$. We decided to used the thresholds $\delta = 0.3$ resp. $\delta_h = 0.4$ in the following.



(a) Classification results of $\mathrm{LC}_{\mathrm{PatchCurv}}$ depending on δ.

(b) Classification results of $\mathrm{LC}_{\mathrm{PatchHist}}$ depending on δ_h.

Figure 3.14 Influence of δ and δ_h on the classification results. This time, the tested methods are using e_3 as a part of the *equals* function.

Finally, the function e_3, which determines similarity between two patch vectors, was combined with e_2 and e_4 leading to the method $\mathrm{LC}_{\mathrm{PatchCurvHist}}$. As already mentioned above, there were two thresholds δ and δ_h to be adjusted. The landscape plotted on the basis of the results from this method can be found in Figure 3.13 (b). It shows poor classification rates for low values especially of δ_h, though they are rapidly improving for larger thresholds. The highest value of $90.29 \pm 6.2\,\%$ is achieved for $\delta = 0.4$ and $\delta_h = 0.7$.

In Figure 3.15 we summarize the results for the first classification dataset using the threshold settings that performed optimally. Although Cavbase performs best in this series by reaching 93.6 % correct classifications, the LC extensions also exceed excellent rates of up to 92.3 %. All heuristic approaches achieve at least 89.1 % classification while they require only a fraction of the runtime of Cavbase and BK. Surprisingly, also the approaches LC and BK, which neglect any surface information, perform also very well and lead to rates of more than 90 %.

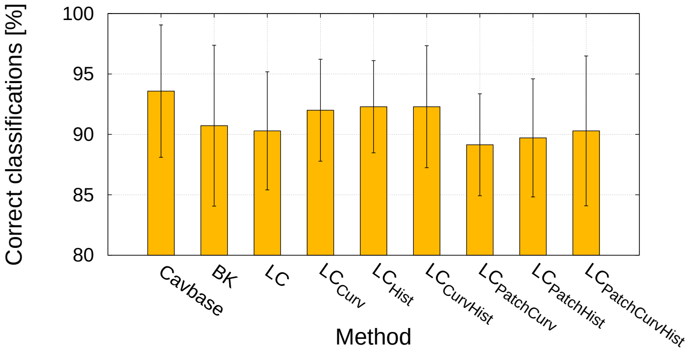


Figure 3.15 Classification rates for the smaller dataset using the LC methods with the best parameter settings. Cavbase and BK reach a classification rate of 93.6 % resp. 90.7 %, but our LC derivatives follow closely. The best of the extensions, LC$_{\text{Hist}}$ and LC$_{\text{CurvHist}}$, reach 92.3 %. LC$_{\text{PatchCurv}}$ performs worst with 89.1 %.

It should be considered that Cavbase and BK were not able to compare 59 input structures since the resulting product graphs for the comparison to any other binding site became too large to fit into memory. First, this is caused by an internal memory limitation of the Cavbase system and second by the generation of the product graphs which tend to become very large when only the pseudocenter label is used to determine whether two nodes v and v' are similar or not. These missing similarity scores have been neglected while calculating the classification rates. In case that these structures would have also been considered and assumed to be randomly classified, the results of Cavbase and BK would decrease substantially to 85.3 % and 82.4 %, respectively. In the LC approaches all binding sites could be classified since there is not such a low memory limit. Furthermore, all LC methods that are working with the extended graph model use more restricted conditions that have to be fulfilled to generate a new product node (see Table 3.1). In many cases this resulted in a substantial reduction of nodes in the product graph as will be shown in Section 3.3.3. Thus, it must be stated that Cavbase and BK omitted the comparisons of the largest and most complex binding sites which further enhances the results of the newly introduced methods.

In general, it could be possible that all the obtained classifications rates are already determined by the different sizes of the examined ligands ATP and NAD. To rule out any bias depending on the pocket sizes, we performed additional experiments using the maximal diameter as well as the number of pseudocenters as the discriminating properties in a classification experiment. In both cases, considerably reduced classification rates were obtained (55.87 ± 15.13 and 53.43 ± 9.8) which definitively prove these features to be of no relevance in the comparison experiments.

Classification results using the large datasets

In the next step, we used the optimal parameter settings derived by the first experiment to classify the larger and more complex datasets, comprising the 822 binding sites (402 NAD, 420 ATP) and 835 binding sites (402 NAD, 433 FAD). Figure 3.16 gives an overview of the achieved classification rates with Cavbase, BK and the seven LC-based methods. We furthermore added a totally different approach for the comparison of cavity structures, called Histograms [43], which is not evaluated in terms of graphs. Here, sets of histograms are generated for every pocket that capture pairwise distances between the pseudocenters.

Almost all heuristic evaluations using the optimized parameter settings

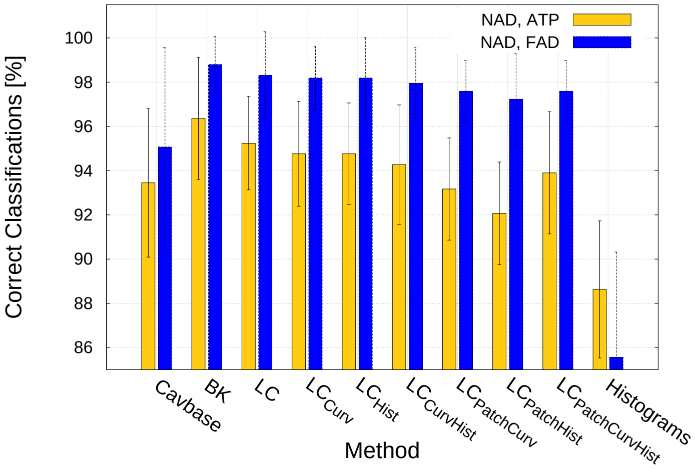
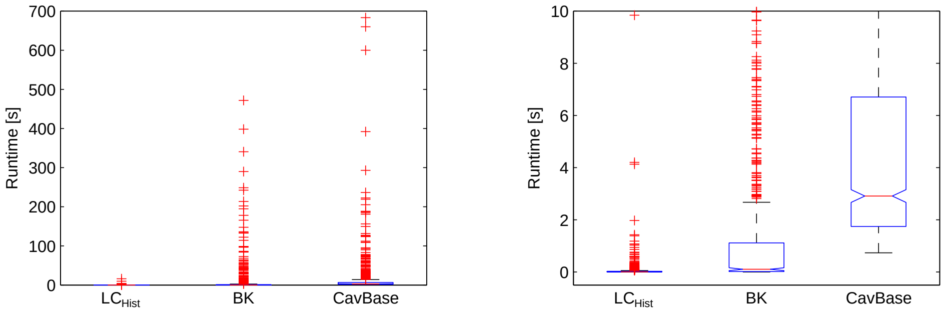


Figure 3.16 Classification rates for the larger datasets (NAD against ATP, and NAD against FAD). The majority of LC and its derivatives performs better than Cavbase in both classification experiments. Surprisingly, BK which does not incorporate surface information performs best, reaching a classification rate of 98.8 % for the classification of NAD and FAD. The Histogram method performs clearly worst in both experiments.

performed better than the original Cavbase implementation, up to 3.1 % improvement for LC_{Hist}. The clique detection approach BK also performs better than Cavbase though it does not take any information about the protein surface into account (96.4 % *vs.* 93.4 %). In the case of comparing NAD against ATP, Cavbase and BK were unable to calculate 58 single scores in the similarity matrix (0.00085 % of the total number of scores). For the current validation this small amount can be neglected.

3.3.3 Runtimes

In order to give an overview of the runtimes, we randomly selected a set of 1000 pairs of protein binding sites from Cavbase. These pairs were compared using the methods Cavbase, BK and LC_{Hist} with $\delta_h = 0.7$. Cavbase and BK were implemented in C++ and all LC approaches in Java. We used the existing implementation of Cavbase to obtain the comparative results, which also includes an implementation of the BK (cf. Sec. 3.2.1). We did not



(a) All collected runtimes.

(b) Details of the distribution of runtimes in the practically more relevant interval $0 - 10$ s.

Figure 3.17 Runtimes of LC$_{\text{Hist}}$ ($\delta_h = 0.7$), BK and Cavbase visualized in terms of boxplots. 1000 pairs of binding pockets where randomly picked from Cavbase and the time needed for each comparison was measured to generate this plots. The average runtime of LC$_{\text{Hist}}$ is 0.078 ± 0.63 s, of BK 6.91 ± 32.59 s and Cavbase needs 12.76 ± 44.90 s. Note that Cavbase does not perform calculations on binding sites exceeding a certain size, hence some longer runtimes are omitted.

measure runtimes of LC$_{\text{Curv}}$ and LC$_{\text{PatchCurv}}$, since computing the curvature descriptors or the patch vectors is computationally not more demanding than the generation of a histogram in LC$_{\text{Hist}}$. Thus we expect very similar results. Hence, it is sufficient to consider the runtime of LC$_{\text{Hist}}$ as a representative for all other extensions. The collected runtimes are depicted in Figure 3.17 as boxplots, and an Intel® Core™ i7-3770 CPU (3.4 GHz) device equipped with 16 GB memory was used as computer platform. Cavbase is the most demanding approach, since it is using the exact BK algorithm to enumerate maximal cliques and in addition performs expensive calculations such as the superposition of surface points. As a result it required on average 12.76 ± 44.9 s for each comparison. The BK algorithm itself performed more efficiently than Cavbase. As it does not take the surface points into consideration, it needed on average only 6.91 ± 32.59 s. The LC heuristic for enumeration of all maximal cliques performed clearly best in this study, which results from its low complexity of $\mathcal{O}(n^4)$. LC took less than one second when it was applied on the original graph model (not shown in the plot). Using the LC_{Hist} method, a binding site comparison was carried out in 0.078 ± 0.63 s on average. The extensions proposed in this paper led to a significant decrease of runtime compared to the original Cavbase implementation since the additional surface information can be calculated rapidly. Moreover, the more stringent conditions for creating product nodes generally lead to smaller product graphs. To demonstrate this, we evaluated the product graph sizes of both LC_{Hist} ($\delta_h = 0.7$) and Cavbase/BK in 62 000 binding site comparisons of our first test dataset (Fig. 3.18). It turned out that the average size in Cavbase and BK was 8852 nodes, whereas 4784 in LC_{Hist}, which is only 54 % of the previous graph. Obviously this accelerates the subsequent clique detection further.

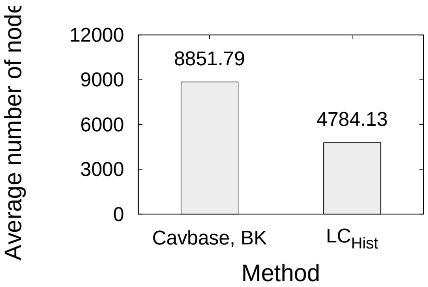


Figure 3.18 Evaluation of the product graph sizes of the methods LC_{Hist} using $\delta_h = 0.7$ and Cavbase resp. BK in 62 000 binding site comparisons. The average size of a product graph in Cavbase and BK is 8852 nodes and 4784 in LC_{Hist}, which is only 54 % in size of the previous graph. Thus, LC_{Hist} needs less time to perform a clique detection on the product graph which results in a further speed-up of a binding site comparison.

The BK algorithm performed almost equally well as Cavbase. However, it has to be stressed that this is only the case for small input cavities. In case of medium or large binding sites the runtime increases rapidly since the problem of clique detection is known to be NP-hard. This effect can be recognized in Figure 3.17 (a) by the large number of cases with rather long runtimes.

3.4 Additional Trials

In addition to the hitherto presented extensions, we tested the influence of two further node properties on the classification results.

1. A feature $[\ell_V(v)]_4$ was assigned to each graph node v which holds the number of surface points that represent the corresponding surface patch. The feature $[\ell_V(v)]_4$ was then considered by a function e_5, which is defined as

   $$\left(1 - \frac{min([\ell_V(v)]_4, [\ell'_V(v')]_4)}{max([\ell_V(v)]_4, [\ell'_V(v')]_4)}\right) \leq \delta_n \ ,$$

   where δ_n is a threshold that defines the maximal allowed difference of the surface point numbers. This property was introduced to enable a comparison of two patches by means of their size.

2. A feature $[\ell_V(v)]_5$ was assigned that holds the standard deviation of the distances from a pseudocenter to the surrounding surface points. The idea behind this approach is that the standard deviation in a rather convex patch is assumed to be much smaller than in a concave patch (Fig. 3.19). Thus, it can also be regarded as another alternative to display the curvature of a surface patch. $[\ell_V(v)]_5$ was considered by the function e_6, which is defined as $|[\ell_V(v)]_5 - [\ell'_V(v')]_5| \leq \delta_s$, where

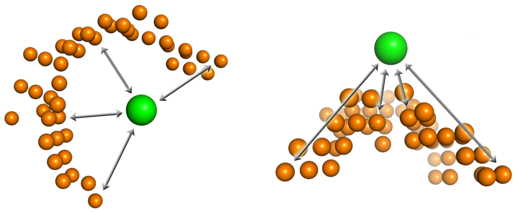
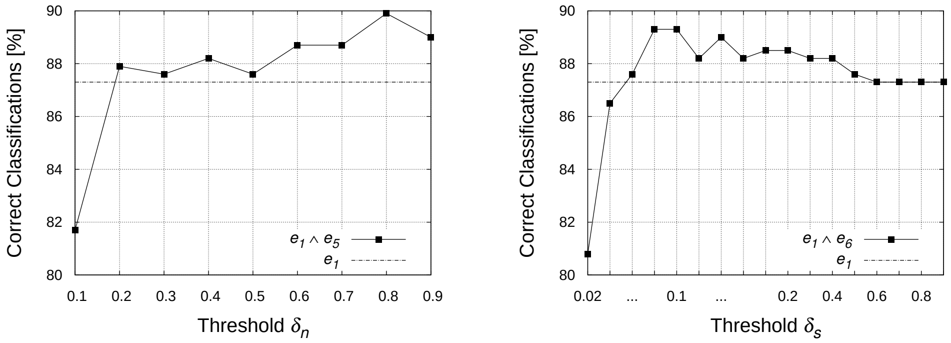


Figure 3.19 Example of a rather convex (left) and a rather concave surface patch (right). For both patches, four selected distances are shown. The standard deviation of the distances between the pseudocenter (green sphere) and the surrounding surface points is likely much higher in case of a concave surface patch.

$[\ell_V(v)]_5$ and $[\ell'_V(v')]_5$ are the standard deviations of the nodes v and v', respectively. δ_s defines a threshold that defines the maximal allowed difference of $[\ell_V(v)]_5$ and $[\ell'_V(v')]_5$ to regard the two patches as similar.

Both mentioned features were then used in the similarity function *equals* together with the label-based method e_1 to implement the comparison of two graph nodes. To be precise, this leads to $e_1 \wedge e_5$ in the first case and $e_1 \wedge e_6$ in the second. These versions of *equals* were subsequently used in a classification experiment on the dataset ATP/NAD$_{small}$, where we compared the success rates obtained by a k-leave-one-out cross-validation with $k = 1$.

When employing the first additional property, which gives information about the patch size, we obtained an increase of up to 2.6 % of the classification success in case of $\delta_n = 0.8$. Only for the very strict threshold of $\delta_n = 0.1$ the rates substantially deteriorate to only 81.7 %. Figure 3.20 (a) displays the overview of the success rates for all investigated values of δ_n. In addition, the reference rate of 87.3 %, which is obtained when using the label comparison e_1



(a) Only for the threshold $\delta_n = 0.1$ the rates improve compared to those that are achieved by e_1 alone. The highest success of 89.9 % is obtained when using $\delta_n = 0.8$, which is 2.6 % higher than the reference rate of e_1.

(b) For very stringent thresholds δ_s below 0.06 the accuracy is lower compared to that of e_1. For all greater values up to 0.5, however, the success rates are improved, reaching a maximum of 89.3 % for 0.08 and 0.1.

Figure 3.20 Overviews of the success rates for the investigated values of δ_n and δ_s in the interval $[0.1, 0.9]$. The reference rate of e_1 at 87.3 % is given as a dashed line in both diagrams.

alone, is indicated as a dashed line.

In case of the second additionally introduced property, which gives information about the surface curvature, an improvement of up to 2 % is reached when using the thresholds $\delta_n = 0.8$ and 1.0 (Fig. 3.20 (b)). The improvements expire for any value ≥ 0.6. As already observed in the case of δ_n, a too stringent threshold leads to clear deterioration of the success rates.

Even though the consideration of these properties turned out to perform fairly well in the presented experiments, we nevertheless think that a comprehensive investigation necessitates some more trials. The tests that have been carried out so far only referred to the ATP/NAD$_{small}$ dataset. More extensive studies which also take the larger datasets into consideration are recommended to underpin the obtained results.

3.5 Conclusions

In this paper we presented novel extended descriptors to improve the pseudocenter model for the representation of protein binding sites. So far, node-labeled and edge-weighted graphs have been used, where edge weights approximate geometry in a very coarse manner; node labels capture physicochemical properties, which is realized by a discrete label. Processing only on this type of graphs leads to a loss of information, especially because detailed information about the surface is not considered. Therefore Cavbase is using the graph-based model only as initial step and subsequently performs calculations on the surface points to obtain a final score.

We showed that the additional step of Cavbase leads to a very inefficient approach. Instead we proposed novel graph models which capture considerably more information on geometry. This information can be exploited during construction of the product graph utilized to detect the maximum common subgraph which is finally used to measure similarity. In a first classification experiment, we adjusted a parameter setting to be applied to further classification experiments using a larger and more complex dataset. The gain of information combined with an efficient heuristic for clique detection led to much shorter runtimes and a comparable or even higher degree of accuracy than the original Cavbase implementation, albeit none of the added properties outperformed any of the alternative choices. A substantial reduction of the product graph size results from our proposed strategy to realize the *equals* function, which also considers the additional information. Thus, comparisons against the entire database of binding sites can now be carried out with a fraction of the previous required time while the accuracy of the results remains mostly unaffected or even improves.

These improvements also pave the ground for further enhancements. Additional descriptors can be imagined to be calculated and stored with the graph nodes, since runtime and memory requirements with novel hardware do no longer impose such strong restrictions. These descriptors could then be used to carry out comparisons that are based on even further information. Together with the node properties that were proposed in this contribution, several alternative combinations of parameters can be evaluated that are either based on shape or other physicochemical data derived from the PDB structures.

4

Accelerated Version of LC[1]

4.1 Introduction

Although the results obtained by Cavbase similarity analyses have been successfully applied for various purposes [132, 85, 51, 164, 163], its extremely long runtime is a limiting factor of being able to subject large-scale studies such as screenings of entire databases to it. As such virtual screenings are an essential part of a drug development campaign though, it is important to search for algorithms that achieve comparable results but require significantly less time at best allowing for an interactive data mining. In the previous chapter, we therefore proposed the Local Cliques (LC) method. We used an extended graph model as well as a heuristic clique detection approach to facilitate the mutual comparisons of Cavbase structures. The nodes of a graph, which represent the pseudocenters of a binding site, have been extended by additional properties efficiently coding the shape of their close-by protein surface and their physicochemical environment. Considering these supplementary features decrease the product graph and leads – together with the simplified clique heuristic – to a significant runtime gain with comparable accuracy.

[1] accepted by *Mol. Inform.* as Krotzky, T.; Klebe, G. Acceleration of Binding Site Comparisons by Graph Partitioning.

In this chapter, we present an additional feature showing that the LC method can be accelerated even further. It is based on the fact that the clique-detection problem is NP-complete and the required time of the algorithms scales with n^2 for every additional graph node. Thus, splitting the graphs to be compared into a reasonable subset of components prior to the clique analysis and performing individual comparisons of these subsets, followed by a merging of all generated partial scores to build the final similarity measure turns out to be significantly faster than the original comprehensive approach. We will present classification experiments using the same datasets as in the former study, and the results show that the accuracy is almost completely maintained. The runtime, however, is decreased to a fraction compared to the original setting.

4.2 Methods

In the Cavbase methodology, pseudocenters represent certain portions of the amino acids that exhibit specific physicochemical properties. They are assigned to coordinates in 3D space and exhibit a physicochemical label $\ell \in \{\text{donor, acceptor, doneptor, aromatic, pi, metal, aliphatic}\}$. A binding site is represented as a graph $G = (V, E, \ell_V, \ell_E)$, where each node $v \in V$ constitutes one pseudocenter. G is an undirected edge-weighted graph, where an edge between two nodes $v, w \in V$ is labeled by the Euclidean distance $\ell_E(v, w)$ between them. Furthermore, $\ell_V(v)$ comprises the pseudocenter label of node v as well as other additional properties such as a surface histogram H, which encodes the shape of a surface patch, or a patch vector $\vec{P}$, which gives information about the physicochemical environment of a pseudocenter. The histogram feature H is of special importance in the present study, as it will be used in several experiments in the following. Here, distances from the pseudocenter to its surrounding surface points are recorded, which serves as a measure for the local surface curvature. When comparing two graph nodes subsequently, also the histograms can be taken into account and compared by using the Jaccard distance (see Chap. 3 for more details).

As cavities stored in Cavbase tend to be very large and on average consist of more than 90 pseudocenters, we were seeking for alternative representations to accelerate the clique detection in the comparative step as follows. While two graphs have been compared as a whole in the LC method, we now propose an alternative method named DivLC in the following, where we divide the pocket graphs which are subjected to the comparison routine into seven partitions, corresponding to the different pseudocenter types. In detail, the set of nodes V in an input graph $G = (V, E, \ell_V, \ell_E)$ is divided into the subsets $V_{1\ldots7}$, where $\forall v, w \in V_i, \ell_{V_1}(v) = \ell_{V_1}(w)$, accordingly all nodes in a subset correspond to the same pseudocenter type. This also leads to $\forall v \in V_i, v \in V$ and $\bigcup_{i=1}^{7} V_i = V$. Based on $V_{1\ldots7}$, seven graphs $G_{1\ldots7}$ are generated, where the sets of edges $E_{1\ldots7}$ are created in the same way as in the original LC method. Subsequently, the seven graphs $G_{1\ldots7}$ of the query pocket are compared in a pairwise manner to the corresponding graphs $G'_{1\ldots7}$ of the probe pocket by determining a clique in every product graph $G_{\otimes_i}$ using the LC method, which leads to the scores $s_{1\ldots7}$ (Fig. 4.1). In the following, the obtained individual scores are merged to one final similarity score S. To combine the subscores $s_{1\ldots7}$ to S, each is normalized with respect to the size of the larger input graph (($|V_i|$ is the number of nodes of type i in the reference graph, $|V'_i|$ is the corresponding value in the probe graph; see Eq. 4.1) and then weighted by using the number of nodes $V_{\otimes_i}$ in the corresponding product graph $G_{\otimes_i}$ (Eq. 4.2). We believe this step is required to ensure that the largest product graph has the highest impact on the final similarity score. S is therefore a weighted mean of all calculated subscores.

$$s_i = \frac{s_i}{\max\{|V_i|, |V'_i|\}} \tag{4.1}$$

$$S = \frac{\sum_{i=1}^{7} s_i \cdot |V_{\otimes_i}|}{\sum_{i=1}^{7} |V_{\otimes_i}|} \tag{4.2}$$

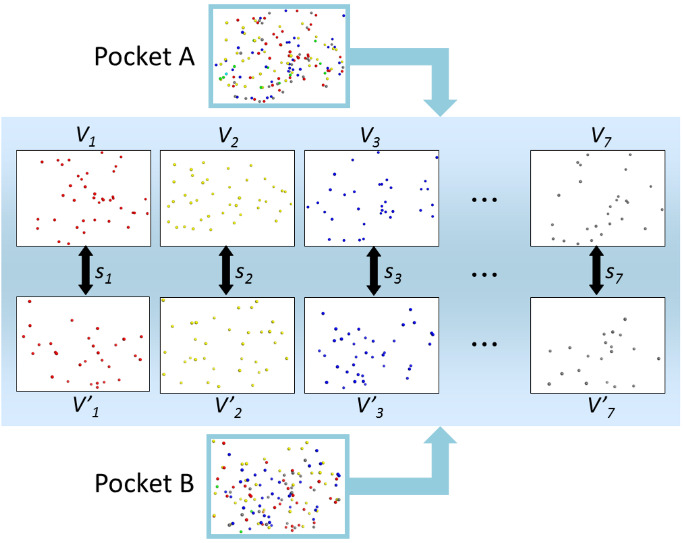


Figure 4.1 Workflow of the calculation of similarities between two Cavbase structures. The sets of pseudocenters are split into seven subsets $V_{1...7}$ depending on their physicochemical type (only four are shown for sake of clarity). The subsets are transformed into graphs and compared in a pairwise manner using LC, which results in the scores $s_{1...7}$. Subsequently, a combination of the scores leads to the final similarity score S.

To compare the results of DivLC with the originally implemented LC method, we performed classification experiments based on the same datasets. We used two sets of cavities that have been co-crystallized with one of the cofactors ATP, NAD (NADH and NAD+), or FAD, which are called ATP/NAD$_{small}$, ATP/NAD$_{large}$ and FAD (see Sec. 2.3). We also used the same evaluation method as in the former publication to achieve the highest degree of comparability: a 10-fold cross-validation was applied on the all-against-all scoring matrix with a k-nearest-neighbor approach as classifier. For any further details, please refer to Section 2.4.2.

In addition, retrieval experiments were carried out to enable a more exhaustive comparison of LC and DivLC. Three binding sites that accommodate one of the cofactors ATP, NAD or FAD were used as query pockets and compared against the entire Cavbase database in its current version (July 2014) using both methods individually. The database comprises 458 070 putative binding sites from 101 379 PDB entries. Rather than searching for specific drug molecules, the use of cofactors provides the opportunity to collect large sets of diverse binding cavities which scatter largely in the conformational space and are broadly distributed over all EC main classes [143]. In order to measure the success of retrieval, the resulting scores were sorted in an ascending way and the occurrences of other binding pockets which host the same ligand were evaluated in terms of ROC curves along with their AUC.

4.3 Results

We compared DivLC in three ways to the original LC approach, first, by investigating the classification accuracy of both methods. Second, by comparing the required runtimes. Third, by evaluating the performance in several retrieval experiments.

4.3.1 Classification Rates

With respect to the results, the rates of correct classifications achieved by the new DivLC method hardly deteriorate compared to those of the original LC approach (Fig. 4.2). In case of the small NAD/ATP dataset (blue bars), LC achieved $90.9 \pm 3.5\,\%$ correct classifications and DivLC performs almost equally well with $90.6 \pm 5.0\,\%$. Also in case of the large NAD/ATP and NAD/FAD datasets (green and yellow bars, respectively) the results of both approaches hardly differ ($94.6 \pm 2.1\,\%$ vs. $92.8 \pm 2.8\,\%$ and $98.3 \pm 1.4\,\%$ vs. $96.8 \pm 1.6\,\%$). Furthermore we used DivLC together with the additional histogram property H and its similarity threshold $\delta_h = 0.7$, which was identified as a good trade-off between speed and accuracy in our last study. This approach will be called $\text{DivLC}_{\text{Hist}}$ and has to be compared to the corresponding results achieved by LC_{Hist}. For both methods we applied the same threshold $\delta_h = 0.7$, as it turned out to be the most suitable setting in the former chapter. Also in the present

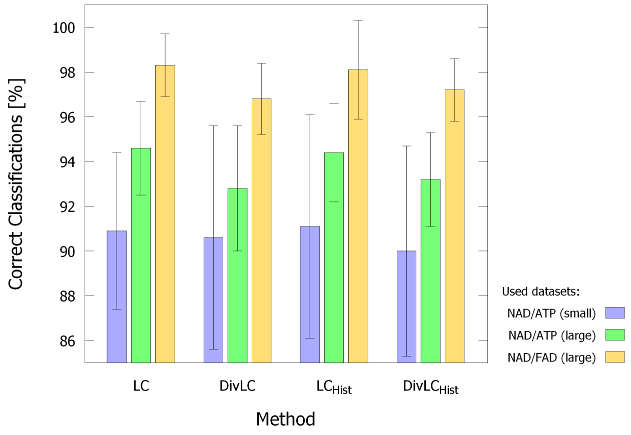
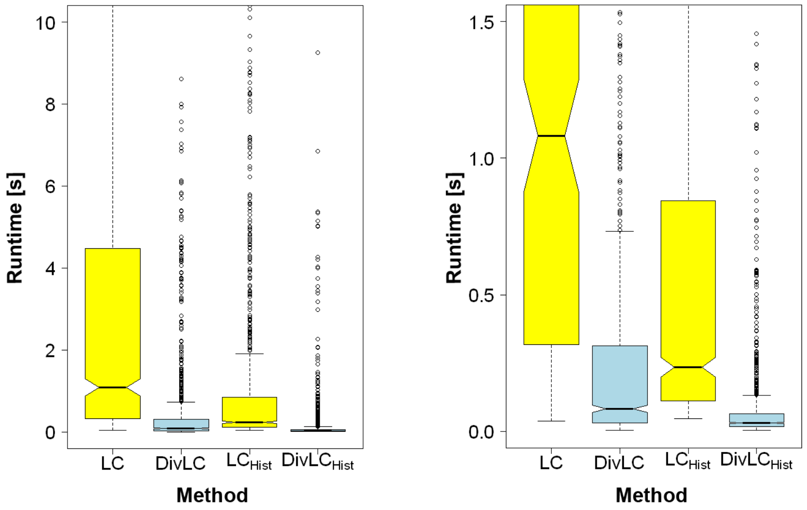


Figure 4.2 Success rates of the methods LC, DivLC, $\mathrm{LC}_{\mathrm{Hist}}$ and $\mathrm{DivLC}_{\mathrm{Hist}}$ when applied on three two-class datasets of protein binding sites. The DivLC approaches perform almost equally well than the original LC methods. Although the mean success rates are slightly worse, no significant deterioration is obtained.

case, the success rates of $\mathrm{DivLC}_{\mathrm{Hist}}$ decrease by only 1 %, indicating almost no loss of accuracy (see the right part of Fig. 4.2).

4.3.2 Runtimes

The partitioning of the input graphs results in a substantial improvement of the runtime, since seven pairwise comparisons of the subsets are carried out much faster than the comparison of conventional binding site graphs. In Figure 4.3, boxplots of the runtimes of the four above-mentioned methods are displayed. To generate the data base for these plots, we used a list of 1000 randomly selected binding-pocket pairs, which were subsequently subjected to each of the comparison methods. All runtimes were measured on a device equipped with an Intel$^{\circledR}$ Core$^{\mathrm{TM}}$ i7-3770 CPU (3.4 GHz) and 16 GB memory. For the runtime measurements, we also considered the time that is needed to create all graphs and product graphs prior to their comparisons. In case of LC, only one large product graph has to be created, whereas DivLC requires seven smaller ones to be calculated. Thus, the required time for constructing the graph is also an important factor for a relevant comparison of the two



(a) Boxplots of the runtimes in the range of 0 to 10 seconds.

(b) Blow-up of (a) in the range of 0 to 1.5 seconds.

Figure 4.3 Runtimes of the methods LC, DivLC, LC$_{\text{Hist}}$ and DivLC$_{\text{Hist}}$ needed for the comparisons of 1000 random pairs of binding sites. The DivLC approaches (blue) are on average more than ten times faster than the LC approaches (yellow), indicated by the non-overlapping notches of the boxes.

approaches.

The median runtime is 1.08 s for LC, 0.08 s for DivLC, 0.24 s for LC$_{\text{Hist}}$, and 0.03 s for DivLC$_{\text{Hist}}$. We used the medians to compare the runtimes since the mean values are always highly influenced by exceptionally long outliers in the pocket-comparison trials. It becomes apparent that the use of DivLC results in a speed-up by a factor of 13.5 compared to the traditional graph model, and by a factor of 8 when the histogram property H of the extended graph model with $\delta_h = 0.7$ is considered in addition. The non-overlapping notches of the boxplots clearly indicate the significant runtime gains. Thus, using DivLC$_{\text{Hist}}$ for an entire database screening will only take about 3.8 hours. On the contrary, LC$_{\text{Hist}}$ will take a much longer time by requiring around 30 hours and LC even requires 5.6 days, which is hardly acceptable for a virtual screening campaign planned as an idea generator.

4.3.3 Retrieval of Cofactor-binding Pockets

To further examine the performance of $\text{DivLC}_{\text{Hist}}$, we decided to carry out retrieval experiments using pockets which bind the cofactors ATP, NAD and FAD. This is a challenging task due to the high number of rotatable bonds of the cofactor molecules and their high degree of conformational flexibility. Furthermore, the enzymes that host the mentioned ligands are scattered over all six EC main classes which suggests large similarity distances among the proteins in sequence space. In total, protein entries comprising 1181 ATP, 2137 NAD and 2288 FAD binding sites were considered. For each of the three cofactor types we selected one pocket as query pocket and compared this reference against all remaining entries in the database. We decided to use pockets that were well extracted around the ligands of interest and originate from PDB entries that exhibit a resolution of 1.7 Å or better.

For the retrievals of other ATP, NAD and FAD-binding sites we selected the pockets `1KAX.2` (1.70 Å), `1R6D.1` (1.35 Å) and `2PGN.13` (1.35 Å), respectively. The resulting ROC curves of the three retrievals are pictured in the Figures 4.4-4.6. As demonstrated by the plots shown in Figures 4.4–4.6, the results of the retrievals match with those of the previous classification experiments. In all cases, $\text{DivLC}_{\text{Hist}}$ performs only slightly worse compared to LC_{Hist}, however, no substantial difference can be recognized. The gradients of the graphs are nearly identical for each query pocket and the largest difference in the AUCs is 3 % in case of `2PGN.13` (FAD, Fig. 4.6). Table 4.1 summarizes the results for the described retrieval experiments, also showing the AUC for every ROC curve.

Table 4.1 Summary of the AUCs of the retrieval results for ATP, NAD and FAD.

Query	**Cofactor**	**AUC LC_{Hist}**	**AUC $\text{DivLC}_{\text{Hist}}$**
1KAX.2	ATP	0.77	0.75
1R6D.1	NAD	0.72	0.70
2PGN.13	FAD	0.77	0.74

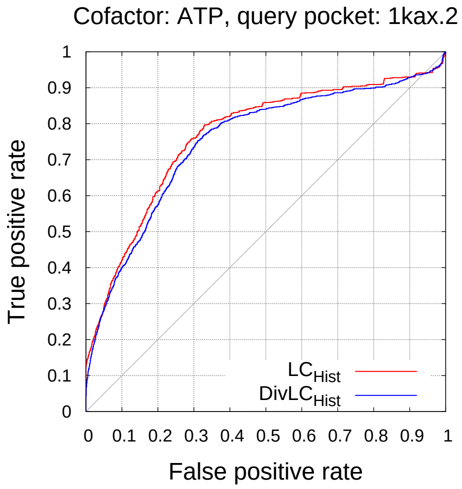


Figure 4.4 Resulting ROC curves of the retrievals of ATP-binding sites. The pocket `1KAX.2` was used as query. LC_{Hist} achieves an AUC of 0.77 and $DivLC_{Hist}$ 0.75.

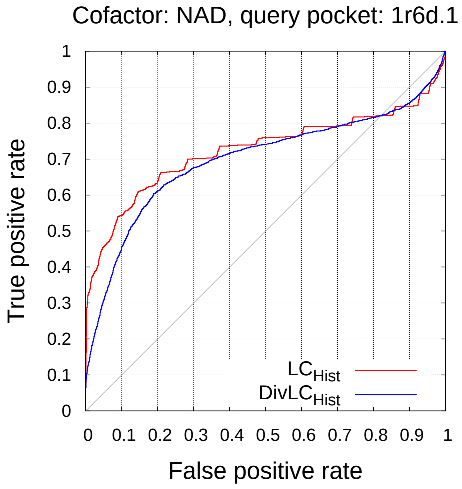


Figure 4.5 Resulting ROC curves of the retrievals of NAD-binding sites. The pocket `1R6D.1` was used as query. LC_{Hist} achieves an AUC of 0.72 and $DivLC_{Hist}$ 0.70.

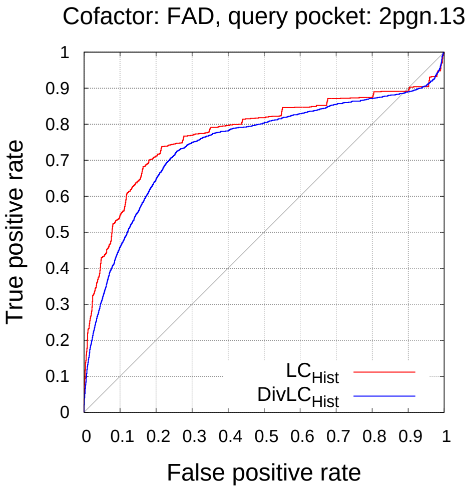


Figure 4.6 Resulting ROC curves of the retrievals of FAD-binding sites. The pocket `2PGN.13` was used as query. $\mathrm{LC}_{\mathrm{Hist}}$ achieves an AUC of 0.77 and $\mathrm{DivLC}_{\mathrm{Hist}}$ 0.74.

In case of ATP, we furthermore carried out retrievals using the most dissimilar pocket of `1KAX.2`, which is `1G21.18`. This even led to an improved performance of $\mathrm{DivLC}_{\mathrm{Hist}}$ compared to $\mathrm{LC}_{\mathrm{Hist}}$ (AUC of 0.81 *vs.* 0.69) which stimulated us to perform a more generalized study of the cofactor retrievals in order to determine the actual accuracy improvement when using $\mathrm{DivLC}_{\mathrm{Hist}}$. Therefore, two further pockets were selected for every considered cofactor, each of which retrieved from a different EC main class. After the pockets have been used to screen the database with both methods, the AUCs were calculated as well as the mean value to quantify by how much the retrieval rates improved or decreased (Tab. 4.2). In agreement with our previous results no substantial deterioration of the accelerated method can be observed. This experiment even suggests that $\mathrm{DivLC}_{\mathrm{Hist}}$ is more accurate than $\mathrm{LC}_{\mathrm{Hist}}$, particularly as there are single query structures (`1ATP.3` and `3W5H.2`) where $\mathrm{LC}_{\mathrm{Hist}}$ has clearly more issues with than $\mathrm{DivLC}_{\mathrm{Hist}}$. To also present a real-life example, we finally want to regard binding sites that host the chemotherapeutic drug

Table 4.2 Deterioration or improvement of the AUC of $\text{DivLC}_{\text{Hist}}$ in percent when three pockets of different EC main classes are used as query for the retrieval of each cofactor.

Cofactor	Query (EC)	LC_{Hist}	$\text{DivLC}_{\text{Hist}}$	Difference
ATP	1KAX.2 (3)	0.77	0.75	−0.02
ATP	1ATP.3 (2)	0.43	0.66	+0.23
ATP	3TUT.2 (6)	0.74	0.74	±0.00
NAD	1R6D.1 (4)	0.72	0.70	−0.02
NAD	2JHF.6 (1)	0.71	0.67	−0.04
NAD	3RIY.6 (3)	0.57	0.56	−0.01
FAD	2PGN.13 (3)	0.77	0.74	−0.03
FAD	3W5H.2 (1)	0.50	0.68	+0.18
FAD	3G5S.4 (2)	0.78	0.71	−0.07

methotrexate. This drug is actually supposed to competitively inhibit the dihydrofolate reductase and thus exert a cytostatic effect. However, also other enzymes have been determined in complex with methotrexate. Thus, a cavity comparison approach will be challenged by a meaningful retrieval task. We chose the pocket of a highly-resolved dihydrofolate reductase structure as query (PDB: `3DAU`, 1.50 Å) to retrieve all other known methotrexate-binding sites and calculated the AUC of the resulting ROC curves. The results confirm the evidence of the former experiments. Leading to an AUC of 0.82, $\text{DivLC}_{\text{Hist}}$ is worse only by 0.05 compared to LC_{Hist}, which achieves 0.87.

4.3.4 Retrieval of Thrombin Active Sites by Using a Subpocket as Query

Since the input graphs are divided and seven disjoint components are employed to facilitate a binding-site comparison, it is likely that pseudocenter patterns which represent local information regarding subpockets are disrupted. In the next experiment, we investigated to what extent this is actually the case when only a subpocket is used as query structure for a retrieval study.

We focused on thrombin, an important serine protease in the blood coagulation cascade. The thrombin structure with the PDB code `3UWJ` was chosen and only the S1 subpocket was selected by extracting all pseudocenters and the corresponding surface points in the range of 6 Å around the benzamidine moiety of the bound ligand (Fig. 4.7). The benzamidine fragment is a frequently used head group featured by inhibitors specific for enzymes of the trypsin-like serine proteases. It forms a salt bridge to Asp189, a key residue in the S1 subpocket for substrate recognition, and a fair number of drug candidates were developed containing this substructure.

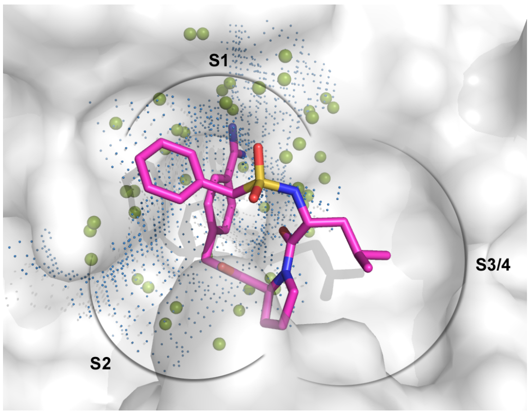


Figure 4.7 Active site of thrombin (PDB: `3UWJ`) with the ligand TIF bound. The benzamidine moiety is directed backwards and addresses the S_1 subpocket. All pseudocenters in a range of up to 6 Å from the benzamidine are shown as green spheres, which are shown with a unified coloring for sake of clarity in this illustration; blue points indicate the corresponding surface points.

Before carrying out the retrieval experiment, also the set of true positives had to be defined. To detect the total number of thrombin active sites in our database, we searched for a match with the EC number 3.4.21.5 (thrombin) and the presence of Asp189. With this strategy, we detected 430 thrombin pockets which were later used as reference set to rank the results of our subsequent comparisons.

$\mathrm{LC_{Hist}}$ was the first method tested and achieved very convincing retrieval success rates (AUC: 0.89). The early enrichment is also remarkable in this case as $\mathrm{LC_{Hist}}$ has detected 59 % of all thrombin active sites after the first one percent of the sorted scoring list. Using $\mathrm{DivLC_{Hist}}$, however, leads to much worse results. This method results in an AUC of only 0.55 which is almost as bad as a random retrieval. The amount of thrombin active sites that have been found after the top one percent of the scoring list is also dramatically decreasing to only 4 % which indicates a drop of performance by a factor of about 15 compared to $\mathrm{LC}_{\mathrm{Hist}}$.

To get an idea of how small the considered subpocket can be to still produce convincing results using both approaches, we carried out another experiment on the thrombin pockets. This time, the thrombin pocket was reduced step by step from the original size (139 pseudocenters) to the size of the S_1 subpocket only (48 pseudocenters). This was accomplished by firstly determining the center c of the S_1 pocket by averaging the coordinates of all joint pseudocenters and, subsequently, performing an iterative discarding of the surrounding pseudocenters depending on their distance to c. Thus, we obtained the originally-sized overall pocket, four gradually smaller subpockets, and finally the S_1 pocket as query structures for six retrieval experiments. The resulting AUCs of these experiments are depicted in Figure 4.8. It is

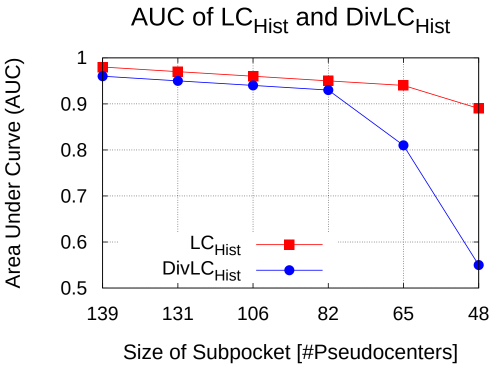


Figure 4.8 AUCs of the ROC curves of $\mathrm{LC}_{\mathrm{Hist}}$ and $\mathrm{DivLC}_{\mathrm{Hist}}$ when retrieving thrombin structures by using a subpocket of `3UWJ.2`. The subpockets have been cut out by defining a certain distance range around the center of the S_1 pocket. The subpocket that consists of 139 pseudocenters (first entry on x axis) represents the original pocket without any restriction; the pocket consisting of only 48 pseudocenters (last entry on x axis) is only the S_1 pocket.

shown that a substantial number of marginal pseudocenters can be discarded before the performance indicated by the AUC of $\text{DivLC}_{\text{Hist}}$ is clearly decreasing compared to the AUC of LC_{Hist}. Only until about half of the original pocket's pseudocenters are neglected, the accuracy of the alternative method begins to drop noticeably.

These results demonstrate that the $\text{DivLC}_{\text{Hist}}$ approach is on the one hand less suited for small substructure detection than LC_{Hist}, whereas it is on the other hand clearly comparable concerning the measurement of global similarity between protein binding sites. However, this observation does not limit the applicability of the faster approach, as the acceleration achieved by DivLC is highest when large pockets are compared and constantly decreases for smaller pockets. Table 4.3 shows the acceleration of DivLC compared to LC when both methods are applied to compare query pockets of different size with respect to a set of 100 random structures. It becomes apparent that the highest acceleration is achieved when a rather large pocket of 139 pseudocenters is used as query. Here, the speed-up of DivLC is nearly a factor of 21. On the other hand, almost no acceleration can be obtained (factor 1.1) when a small pocket of only 19 pseudocenters is used.

Table 4.3 Acceleration factors of DivLC compared to LC when query pockets of different size are compared to a set of 100 random structures.

Size of query pocket	**Acceleration**
139 pseudocenters	20.8
48 pseudocenters	7.2
23 pseudocenters	1.9
19 pseudocenters	1.1

4.4 Conclusions

In this study, we proposed DivLC for the comparison of putative binding sites deposited in the Cavbase database. DivLC is an extension of the method

LC, which we have published in a recent study as a fast alternative for the original Cavbase comparison method. Here, we expand the method by an alternative to significantly accelerate this comparison method even further. By partitioning the original input graphs which represent the binding sites in a set of seven disjoint components, accelerations of around one order of magnitude along with hardly any loss of accuracy in two-class and three-class classification trials could be achieved. We furthermore utilized both methods in a retrieval study of cofactor-binding sites which led to similar results.

By use of a thrombin S_1 pocket as query in the final retrieval experiment, however, we could clearly demonstrate the limitations of the graph-partitioning approach. It could be shown that, even though DivLC provides a real alternative to LC when entire binding pockets are compared, it is advisable to not employ DivLC when only a small subpocket is used as query structure. Most likely, the limited pseudocenter pattern that represents important spatial and physicochemical information of the subpocket is disrupted by the graph partitioning, which leads to a clearly reduced comparison performances in the following. While this issue is clearly less concise when working with large pockets, it becomes a serious problem for very small pockets.

5

RAPMAD[1]

5.1 Introduction

Although some alternative approaches for the comparison of Cavbase input binding site data have been developed [44, 45, 100] and also LC and DivLC turned out to be rather efficient with respect to the hitherto presented methods, the overall comparison remains a computationally intensive task making the use of such approaches in an interactive application scenario impossible. Nonetheless, as cavity comparisons can be used in modeling as a kind of idea generator or a tool to validate working hypotheses, the interactive application is highly desirable. Many methods suffer from the NP-complete optimization problem and even the most efficient ones exhibit a cubic time complexity [45, 100]. This often results from the representation of protein binding sites by means of graphs, which usually requires solutions by least-squares techniques. Even though recently published methods such as SEGA [100], GAVEO [44] or

[1] Reprinted (adapted), with permission, from Krotzky, T.; Grunwald, C.; Egerland, U.; Klebe, G. Large-Scale Mining for Similar Protein Binding Pockets: With RAPMAD Retrieval on the Fly Becomes Real. *J. Chem. Inf. Model.* **2015**, *55(1)*:165–179. Copyright 2015 American Chemical Society. The major part of the study was performed by me, Christian Grunwald and Ute Egerland carried out the computational and experimental work regarding the search for new NMDA receptor inhibitors.

LPCS [45] exhibit impressive accuracy, their long runtime still excludes them from large-scale applications or interactive modeling considerations.

To reduce the runtime of such comparisons, we investigated an alternative representation of Cavbase entries. More efficient approaches employ distance histograms as a medium to describe binding sites. Recently, von Behren et al. [154] presented TrixP, an index-based method which is able to perform fast protein binding site comparisons. TrixP uses descriptors that encode pharmacophoric as well as spatial features to determine binding site similarities by employing a partial shape matching. After this matching step the binding sites are superimposed and, finally, similarity is assessed by means of the overlay of pharmacophoric properties. The binding sites can be provided by either defining a reference ligand or by employing the built-in DoGSite method [152] to predict putative binding sites. Yeturu and Chandra [170] proposed a method called PocketMatch that implements a binding site representation by 90 lists of sorted distances, which are later aligned to accomplish a comparison. The authors were able to show that this representation is appropriate to combine information regarding shape and physicochemical properties. Another example is BSSF, a fingerprint concept, proposed by Xiong et al. [168] The fingerprints are used to store frequencies of spatial distances between the centroids of key interaction portions of the pocket's amino acids. A total of 49 fingerprints, each containing 41 integer values, are generated and subsequently compared by using the Canberra distance measure. Weill and Rognan [161] have developed FuzCav, a method that is also using integer vectors to describe protein binding sites (4833-integer vectors). This fingerprint stores the frequencies of pharmacophoric triplets that were generated from the Cα atoms of amino acids flanking the binding pocket. Comparisons by means of FuzCav can be executed very efficiently due to the ultra-fast comparison of fingerprint data structures. For evaluation, binding sites of the sc-PDB database [101] were exploited. Most importantly, in all three mentioned methods a putative binding site is reduced by considering only the local environment around a given ligand mainly to avoid noise during the automated pocket detection. A distance of 4 Å (PocketMatch), 6 Å (BSSF) and 6.5 Å (FuzCav) is used for the extraction and the comparison procedures are then applied on these pockets well-defined but clearly biased by an initially bound ligand. In contrast we will present a histogram method that is instead based on the entirely automatic detection of binding pockets in Cavbase, unbiased by the prerequisite to show a bound ligand in the input structure.

We propose the description of binding pockets by a set of distance histograms, which can be rapidly generated and compared, since both tasks are conducted with linear complexity. A similar although much more reduced

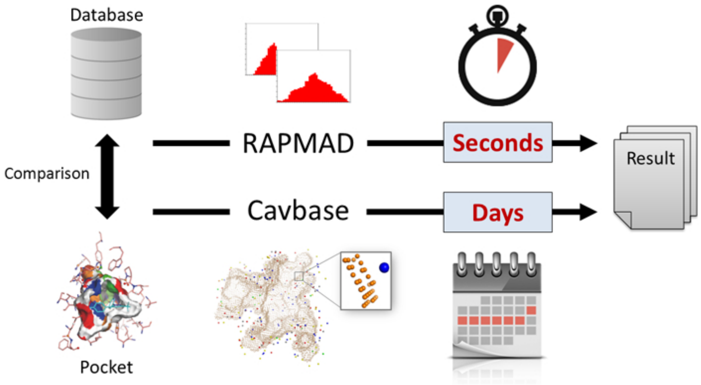


Figure 5.1 Illustration of the RAPMAD approach and its required time compared to the original method currently implemented in Cavbase.

representation has already been introduced in Chapter 2 and was successfully applied for the comparison of binding sites that have been extracted by employing the positions of ligand atoms. In general, such a representation allows for an ultra-fast binding site comparison, where geometric shape and the physico-chemical features of the cavities can be considered simultaneously (Fig. 5.1). The combination of both descriptors is highly relevant to characterize binding sites [70].

5.2 Methods

5.2.1 Histogram Representations of Binding Sites

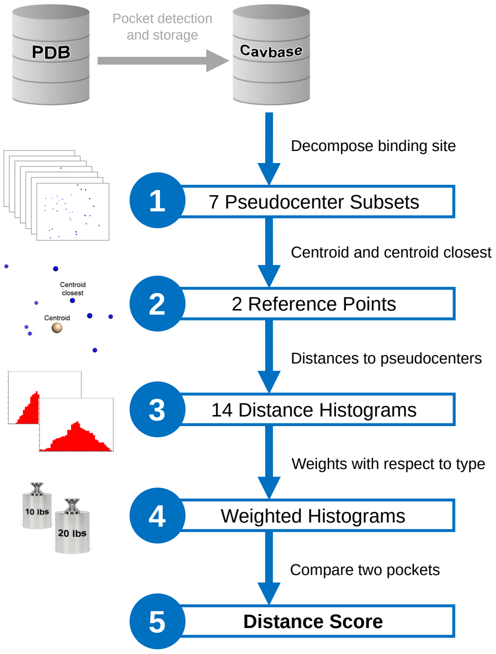


Figure 5.2 Workflow of the RAPMAD approach (highlighted in blue). In a first step, the set of pseudocenters that represents a binding site is divided into seven subsets, based on their physicochemical properties. Subsequently, two reference points (centroid and centroid closest) are determined for every subset (step 2). In step 3, distances from these reference points to all neighboring pseudocenters are calculated and summarized in individual histograms. The histograms are subsequently weighted with respect to the relative pseudocenter frequencies in Cavbase (step 4). Finally, the histograms are used to accomplish a pairwise pocket comparison (step 5).

The workflow of the RAPMAD approach is shown in Figure 5.2 and will be explained in detail. Firstly, the assigned set of pocket pseudocenters P is split into seven subsets $P_1 \dots P_7$, according to the pseudocenter classification scheme based on the physicochemical properties (step 1). Next, two spatial reference points are defined for each subset, namely the centroid and the centroid closest (step 2). Such points have already been used by Ballester and Richards [7] to perform efficient comparisons of small molecules. Osada et al. [114] also used the centroid to derive shape distributions of various 3D objects and facilitate fast subsequent comparisons. The centroid c represents the geometric center of all pseudocenters in a subset P_n and its coordinates c_x, c_y and c_z are defined as

$$\begin{aligned} c_x &= \frac{\sum_{p \in P_n} p_x}{|P_n|} \\ c_y &= \frac{\sum_{p \in P_n} p_y}{|P_n|} \\ c_z &= \frac{\sum_{p \in P_n} p_z}{|P_n|} \ . \end{aligned}$$

While c is an assigned point in space not necessarily coinciding with any atom or pseudocenter of the molecule, the centroid closest cc is one pseudocenter of the subset that shows the smallest Euclidean distance to c.

For every subset the distances from these two reference points to all other pseudocenters $p \in P$ are calculated and summarized in terms of histograms using a bin size of 0.4 Å (step 3). This way, a spatial distribution profile of all pseudocenters in the pocket as well as the relative position of a subset within the total set of pseudocenters are captured. To find the best performing bin size, we examined all values in the interval [0.1, 2.0] Å with a step size of 0.1 Å and found 0.4 Å to work most properly. Figure 5.3 shows RAPMAD's accuracies when using varying bin sizes in the interval $[0.1, 2.0]$ Å with a step size of 0.1. The values represent the success rates in a classification experiment on a test set of pockets that bind either ATP or NAD called ATP/NAD$_{small}$ (see Sec. 2.3). Apart from c and cc as representative reference points, we also examined other definitions of reference points, e. g. the centroid furthest pseudocenter. As a result the selection of c and cc turned out to be the most information rich one with respect to efficiency.

In total, a set of 14 histograms (2 histograms for each of the 7 pseudocenter types) is generated to describe pockets. In addition, we assign a weight to each histogram to consider the relative occurrence frequency of the respective

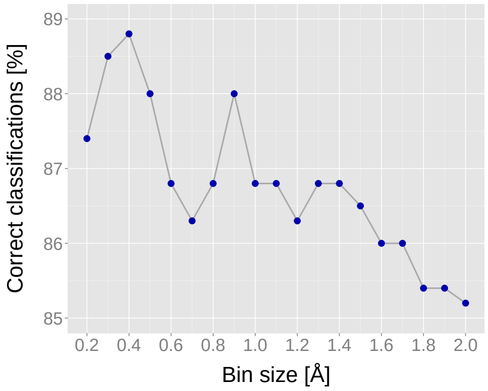


Figure 5.3 Variation of RAPMAD's accuracy when changing the histogram bin sizes in the interval $[0.1, 2.0]$ Å. The best classification result is obtained with a bin size of 0.4 Å. Increasing the bin size leads to a nearly continuous deterioration of the success rates. The worst value is, however, obtained when using a bin size of 0.1 Å (75.6 %, not shown in the diagram).

pseudocenter type in Cavbase (step 4). According to Xiong et al. [168], who stated that donor, acceptor, doneptor and aromatic functionalities are enriched in binding sites, we selected a weighting scheme that accounts for this observation. All 25 million pseudocenters in Cavbase were evaluated. As shown in Figure 5.4, the most frequent pseudocenters (29.9 %) are assigned to an aliphatic functionality and will hence receive the least weight. Also donors and acceptors are quite commonly populated with 20.2 % and 24.5 %, respectively. The smallest fraction is the set of metal pseudocenters, which only accounts for 0.1 % and thus is too small to be visible in the diagram.

Weights are determined in a way that is often referred to as the inverted Boltzmann method [133]. It is supposed to weight those pseudocenters highest that occur least in the given data sample, as it is assumed that they are carrying most of the discriminating information. The weight w for every histogram is thus calculated as

$$w = log(\frac{1}{f}) \ ,$$

where f denotes the relative frequency of a certain pseudocenter type.

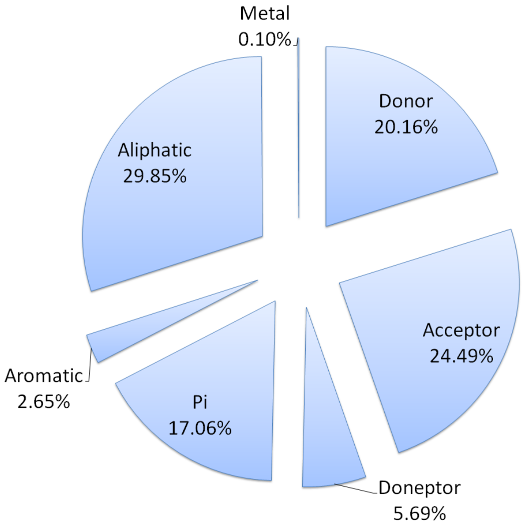


Figure 5.4 Frequency of the different pseudocenter types in Cavbase entries. We evaluated the entire database, consisting of 25 452 124 pseudocenters out of 275 097 putative binding sites. The vast majority of all pseudocenters are hydrogen-bond acceptors, hydrogen-bond donors and aliphatic centers. Taken together, they account for almost 75 %. Clearly less frequent are pseudocenters representing pi centers and H-bond doneptors. Also the set of aromatic pseudocenters is rather sparsely populated (2.65 %). The least frequent, however, is the type metal centers that only accounts for about 0.1 %.

We also calculated the average histograms for each pseudocenter type in Cavbase (14 histograms in sum) to find the areas with highest variation. Bins in the range from 20 to 50 entries that represent distances from 8 to 20 Å exhibited the highest standard deviations, which were thus incorporated to accomplish another bin-wise weighting. However, it was not possible to further improve the results by this procedure and we therefore decided to remain with the above-mentioned histogram weighting.

5.2.2 Pocket Comparison

The comparison of two pockets A and B (step 5) is performed by mutually facing two sets of the 14 histograms. Each pair of histograms assigned to the same pseudocenter type is matched, which leads in total to 14 pairwise comparisons. We examined a large variety of different histogram comparison methods to calculate the distance score S between two pockets A and B, such as the

- Minkowski distance (also known as L_p-norm) [3] with $p = 1$:

$$S(A, B) = \sum_{i=1}^{14} \sum_{j=1}^{|H_{Ai}|} (|H_{Ai}[j] - H_{Bi}[j]| \cdot w_i)$$

- Minkowski distance with $p = 2$:

$$S(A, B) = \sum_{i=1}^{14} \left(\sum_{j=1}^{|H_{Ai}|} |H_{Ai}[j] - H_{Bi}[j]|^2 \right)^{\frac{1}{2}} \cdot w_i$$

- Kolmogorov-Smirnov test [79, 140]:

$$S(A, B) = \sum_{i=1}^{14} \max_{0 \le j \le |H_{Ai}|} (|H_{Ai}[j] - H_{Bi}[j]|) \cdot w_i$$

- Chi square test:

$$S(A, B) = \sum_{i=1}^{14} \left(\sum_{j=1}^{|H_{Ai}|} \frac{(H_{Ai}[j] - H_{Bi}[j])^2}{H_{Ai}[j] + H_{Bi}[j]} \right) \cdot w_i$$

- Distribution moments, where the function *skew* returns the skewness, *var* the variance and *avg* the average of a histogram. This method was successfully applied by Ballester and Richards in Ref. 7:

$$\begin{aligned} S(A, B) = \sum_{i=1}^{14} (|skew(H_{Ai}) - skew(H_{Bi})| \\ + |var(H_{Ai}) - var(H_{Bi})| \\ + |avg(H_{Ai}) - avg(H_{Bi})|) \cdot w_i \end{aligned}$$

- Jaccard distance [66]:

$$S(A, B) = \sum_{i=1}^{14} \sum_{j=1}^{|H_{Ai}|} \left(1 - \frac{min(H_{Ai}[j], H_{Bi}[j])}{max(H_{Ai}[j], H_{Bi}[j])} \right) \cdot w_i$$

- Canberra distance [89]:

$$S(A,B) = \sum_{i=1}^{14} \sum_{j=1}^{|H_{Ai}|} \frac{|H_{Ai}[j] - H_{Bi}[j]|}{H_{Ai}[j] + H_{Bi}[j]} \cdot w_i$$

- Hellinger distance [59]:

$$S(A,B) = \sum_{i=1}^{14} \frac{1}{\sqrt{2}} \left(\sum_{j=1}^{|H_{Ai}|} (\sqrt{H_{Ai}[j]} - \sqrt{H_{Bi}[j]})^2 \right)$$

- Kullback-Leibler divergence:

$$S(A,B) = \sum_{i=1}^{14} \sum_{j=1}^{|H_{Ai}|} log(\frac{H_{Ai}[j]}{H_{Bi}[j]}) \cdot H_{Ai} \cdot w_i$$

- Jeffrey's divergence:

$$S(A,B) = \sum_{i=1}^{14} \sum_{j=1}^{|H_{Ai}|} (H_{Ai}[j] - H_{Bi}[j]) \cdot (log(H_{Ai}[j]) - log(H_{Bi}[j])) \cdot w_i$$

- Euclidean distance:

$$S(A,B) = \sum_{i=1}^{14} \sqrt{\sum_{j=1}^{|H_{Ai}|} (H_{Ai}[j] - H_{Bi}[j])^2 \cdot w_i} \ .$$

Here, $H_{Ai}[j]$ and $H_{Bi}[j]$ stands for the j-th value in the i-th histogram of pocket A and pocket B, respectively. w_i is the weight which is assigned to histogram number i. To facilitate the pairwise comparison of two histograms, the smaller one is always extended with empty bins. Furthermore, any value b in a fraction $\frac{a}{b}$ is set to one to avoid a division by zero.

Finally, we decided to use the Jensen-Shannon divergence [28] to carry out the required comparisons, as this method provided the highest accuracy for all investigated datasets and can be calculated efficiently. For two histograms H_A and H_B the Jensen-Shannon divergence is based on the Kullback-Leibler divergences [86] of (H_A, M) and (H_B, M), where M is the median distribution of H_A and H_B. The value of a bin j in M is calculated as follows:

$$M[j] = \frac{H_A[j] + H_B[j]}{2} \ .$$

The Kullback-Leibler divergence KLD of this median distribution M and a histogram H is then defined as

$$KLD = \sum_j log\left(\frac{H[j]}{M[j]}\right) \cdot H[j] \cdot w \ ,$$

where w returns the weight of H. If the value of any bin j of either H or M is zero, it is discarded. Having calculated KLD_A for (H_A, M) as well as KLD_B for (H_B, M), the final Jensen-Shannon divergence JSD can be computed as:

$$JSD = \frac{KLD_A + KLD_B}{2} \ .$$

After the JSD was calculated for every pairwise histogram comparison, all 14 values are summed up to determine the final distance score S of the two binding sites:

$$S = \sum_{i=1}^{14} JSD(H_{Ai}, H_{Bi}) \ .$$

Obviously, the lowest distance scores are obtained if a pocket is compared with itself, which leads to a value of zero. In general S falls in the interval $[0, \infty)$, since a maximal distance of two histograms in terms of the JSD cannot be defined. We like to mention that this distance function is by no means a metric, even though it may appear as such at first glance. Only two of the four conditions of a metric will always be fulfilled, namely non-negativity ($S(A, B) \geq 0$) and symmetry ($S(A, B) = S(B, A)$). The identity of indiscernibles ($S(A, B) = 0$, if and only if $A = B$) and the triangle inequality ($S(A, C) \leq S(A, B) + S(B, C)$), are not necessarily fulfilled in every case.

We also investigated the influence of normalizing and smoothing the histograms previous to the comparisons. In a normalization the integral of every histographic distribution is calibrated to 1. It has frequently been applied to reduce the impact of the numbers of elements in the distribution (numbers of pseudocenters in a pocket) on shape comparisons [70]. Smoothing functions, e. g. Gaussian smoothing [109], triangular smoothing [52] or sliding window averaging [82], are often applied to counteract the discontinuous behavior of conventional histograms [137] as well as crystallographic uncertainties. It turned out, however, that both methods were not able to improve our results which was why we rather omitted these additional steps and achieved another runtime gain instead.

5.2.3 Datasets

Datasets Based on EC Classes

In addition to the datasets of ATP, NAD and FAD pockets, which have been introduced earlier (see Sec. 2.3), we considered two datasets comprising protein binding sites classified by the EC class numbers the corresponding enzymes have been assigned to. Firstly, we used a diverse dataset of 502 pockets compiled by Glinca and Klebe which spreads over all 6 main EC classes [51]. Table 5.1 provides an overview of this compilation, which we name *EC dataset*. A detailed list of all pockets can be found in Table A.0.5 in the Appendix.

In addition, we used an EC class dataset which consists of 1028 binding sites of serine proteases. The set was compiled in-house by two of our collaborators (Strickert and Fober, personal communication). As pointed out by Schomburg and Rarey [134] serine proteases possess a highly conserved overall sequence and structural similarity. In particular, the catalytic triad of serine, histidine and aspartate that represents the key sequence motif in the active site is present in all considered structures. As stated by the authors, any sequence-based classification methods will therefore most likely fail on this target class. Furthermore, Schomburg and Rarey underscore the importance of the serine proteases in drug development by stressing that the selective inhibition of thrombin and factor Xa is of high pharmacological interest with respect to a direct anticoagulant therapy.

The binding sites in this *serine protease dataset* are distributed over 40 subclasses of the EC class 3.4.21 (see Tab. A.0.6 in Appendix for a detailed

Table 5.1 The EC dataset comprises 502 protein binding sites that were arranged in 16 subsets, which cover all 6 main EC classes.

EC class	Name	Number of pockets
1.1.1.21	Aldose/Xylose reductase	62
1.1.1.42	Isocitrate dehydrogenase	21
1.1.1.62	Estradiol 17β-dehydrogenase	16
1.14.13.2	Hydroxybenzoate-monooxygenase	30
2.7.1.37	Cyclin-dependent kinase 2	46
2.7.1.112	C-Src tyrosine kinase	20
2.7.4.9	Thymidylate kinase	35
3.4.21.5	Thrombin	41
3.4.23.16	HIV-1 protease	48
3.4.24.86	TNF-α converting enzyme	16
4.1.1.23	COMP-decarboxylase	36
4.2.1.1	α-carbonic anhydrase I, II, III, IV	70
5.3.1.5	Xylose isomerase	13
5.4.2.1	Phosphoglycerate mutase	5
6.3.2.1	Pantoate-β-alanine ligase	27
6.3.4.4	Adenylosuccinate synthase	16

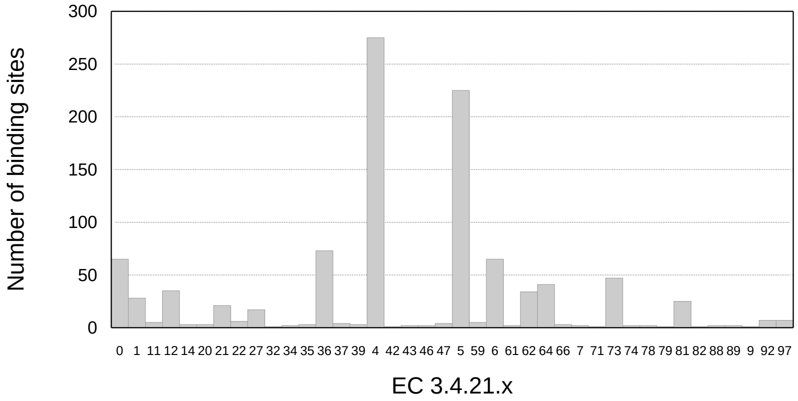


Figure 5.5 Distribution of binding sites in the serine protease dataset. 1028 structures are arranged in 40 subclasses of EC class 3.4.21, which are listed along the x-axis. Subclass 0 comprises all the binding sites (65 entries) that could be annotated to any particular subclass of 3.4.21.x. The highest populated subclasses 4 and 5 contain pockets of trypsin and thrombin structures, respectively.

list of all used pockets). This set also includes a subclass “0”, to which all pockets were artificially assigned that were so far not attributed to any of the other classes (Fig. 5.5). As for all datasets used in the present study, each PDB entry was restricted to one single pocket and therefore it appears only once in the set.

Proteases

Finally, we used another dataset previously studied by Glinca and Klebe [51], which comprises 90 binding sites of proteases and is termed *protease dataset* (see Tab. A.0.8 in Appendix). Also for this dataset binding sites were selected that comprise the catalytic centers of enzymes. The authors used this dataset as a challenge for a clustering on the Merops clan level, which worked well using the similarity measure of Cavbase, but completely fails when using approaches solely based on sequence comparisons. We can therefore rule out any remarkable redundancies with respect to sequence space similarity in this dataset. By using the latter set of binding sites, we will be particularly able to assess our comparison procedure with regard to conserved three-dimensional substructures. Merops [122] clans have been predominantly assigned to capture such information via the recognition of shared substrate motifs. Thus, a comparison with respect to the Merops classification will enable further assessment of our novel comparative approach. More precisely, we will use the Merops clans as templates for another classification test. Subsequently, this will allow to check to what extent our results agree to the independently assigned Merops classification.

5.2.4 Inhibitors for the NMDA Receptor

For the treatment of various neurological disorders, such as depression, Alzheimer’s disease, or Parkinson’s disease, new potent NMDA receptor inhibitors are searched [72, 105]. Present investigations are focused on the detection of negative allosteric modulators (NAMs) rather than competitive ligands. NAMs may allow some kind of fine tuning whereas competitive ligands at the glutamate binding site can lead to complete channel blocking causing severe side effects. Ifenprodil is the prototype of modulating compounds which bind at the NR1/NR2B interface of the amino terminal domain (ATD) of NMDA receptors [21, 105]. Ifenprodil and similar derivatives are potent non-competitive NMDA antagonists but also display a number of off-target activities, e. g. interaction with adrenergic and sigma receptors and hERG blockade. Thus, there are many efforts to discover new compound classes with improved selectivity and similar or enhanced NMDA potency [105].

In this context we have applied our approach RAPMAD. Recently, crystal structures of GluN1/GluN2B ATDs in complex with ifenprodil and Ro 25-6981, another potent and selective NMDA blocker, have been deposited in the PDB [72, 71, 92] (PDB codes: `3QEM, 3QEL, 4PE5, 4TLL, 4TLM`). Their pockets, automatically extracted by the LIGSITE algorithm, were compared against the entire Cavbase database and thus used as starting point of a virtual screening campaign in order to identify other potent inhibitors of the NMDA receptor. Ligands of the most similar pockets were used as a first idea for possible new inhibitors. By several filtering steps such as docking and visual inspection better insights were gained into the binding modes of the retrieved candidates and the set could be further reduced to the most reasonable structures. Finally, the top-ranked and commercially available ligands were experimentally tested by determining IC_{50} values using an NMDA-binding assay developed for the binding of ifenprodil. A detailed protocol of the experimental procedure can be found in the Supporting Information.

5.3 Results and Discussions

5.3.1 Classification Tests on Two-Class Datasets

To evaluate our newly developed comparison method, we start with a classification experiment using the well-established ATP/NAD$_{small}$ dataset. This dataset has already been the subject of many previous comparative studies to evaluate new cavity matching approaches with respect to the original Cavbase implementation. We therefore regard this dataset as a reference standard benchmark for the evaluation of binding site comparison methods. Relying on the work of Fober et al. [44] and Mernberger et al. [100], we can compare RAPMAD to a variety of other methods. In particular, we compare our method with the results obtained by the clique comparison algorithm originally implemented into Cavbase. Therefore, we first calculate the $n \times n$ distance matrix and apply a k-leave-one-out cross-validation subsequently. According to the results reported in the above-mentioned studies, we used different values for k ranging from 1 to 9. Remarkably, many algorithms perform better than the original Cavbase implementation, although they do not take any information about the local shape of the cavity surface into account. Table 5.2 lists the results of various comparison methods. Presented values indicate the percentage of correct classifications, when a varying number of nearest neighbors was used as classifier. The graph alignment approach (GA) was proposed by Weskamp

Table 5.2 Comparison of the classification results of RAPMAD with respect to the ATP/NAD$_{small}$ dataset evaluated by several other binding site comparison methods, including the original Cavbase implementation.

k	**Cavbase**	**GA**	**GAVEO**	**LPCS**	**SEGA**	**BK**	**RAPMAD**
1	81.7	76.6	78.9	93.5	91.6	83.4	88.7
3	83.1	71.8	76.6	91.6	92.4	82.8	87.3
5	83.1	72.4	78.0	89.0	91.3	81.4	85.4
7	81.1	71.8	78.6	88.5	91.6	80.6	84.8
9	79.4	71.3	76.6	86.2	-	81.4	83.9

et al. [163] GAVEO, SEGA and the Bron-Kerbosch algorithm (BK) are also graph-based methods. LPCS is a geometrical approach in order to find optimal superpositions of labeled point clouds in 3D space. RAPMAD achieves higher classification rates than the original Cavbase implementation and many of the other methods presented previously. Only LPCS and SEGA achieve better results, but requiring significantly longer runtime. The experiment is carried out by RAPMAD in around 5.6 seconds on a single core. On the contrary, SEGA and LPCS need 2.7 and 3.3 days instead, and Cavbase takes even more than one week to finalize. This comparison indicates already that the novel RAPMAD approach makes binding site comparisons feasible interactively on the fly. A more detailed analysis regarding the runtime will be demonstrated later in our study.

Since the coenzymes ATP and NAD clearly vary in size, it could well be that already a simple size criterion assigned to the pockets differs sufficiently and can be adequately captured by the histogram approach. To rule out the suspicion that convincing results of RAPMAD are solely based on pocket size differences with respect to their maximal diameter, we performed another classification experiment based on the same dataset. This time, the distance score has been determined solely based on the maximal diameter of a pocket as discrimination criterion. The maximal diameter can be easily obtained by calculating all pairwise distances of the pseudocenters in a pocket. The new distance score S_d was then defined as the difference of two maximal diameters D_{max}, which is calculated as

$$D_{max} = \max_{i=1\ldots|P|, j=1\ldots|P|} d(p_i, p_j) \ ,$$

where P represents the set of pseudocenters of the pocket and $d(P_i, P_j)$ returns the Euclidean distance between the pseudocenters P_i and P_j. S_d is finally calculated as

$$S_d = |Da_{max} - Db_{max}| \ ,$$

where Da_{max} and Db_{max} represent the maximal distances in pockets A and B, respectively. Using this criterion very poor results are obtained as shown in Table 5.3. Also the average number of pseudocenters that comprises a binding pocket is unlikely a sufficiently reliable discriminating feature across the dataset, as it amounts to 184.89 ± 168.94 for ATP and 144.73 ± 93.55 for NAD. Thus, the two distributions do not differ significantly. Even considering the number of type-specific pseudocenters is also not sufficient to discriminate the pockets successfully, as the correct classifications decrease to an average rate of 71.2 % in this case. If neither pocket size nor the number of pseudocenters is the important information that is captured by the histograms, it must be coded in the spatial distribution of the pseudocenters.

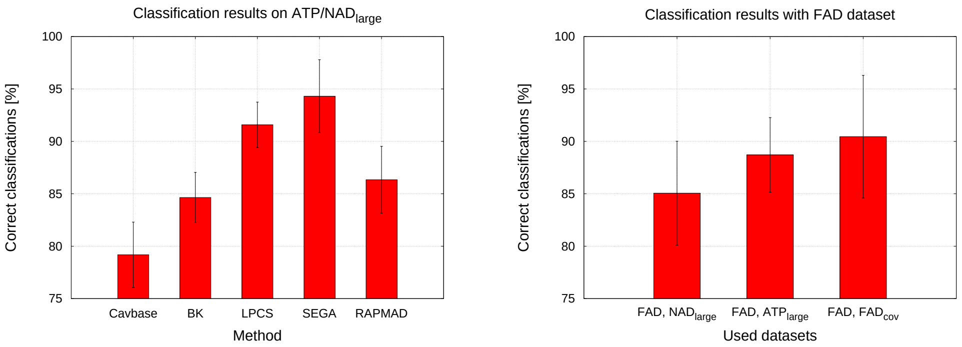
Table 5.3 Classification results of RAPMAD on the ATP/NAD$_{small}$ dataset when using the difference between the maximal diameters of two pockets as a measure of dissimilarity.

k	**Classification rate [%]**
1	51.8
3	57.2
5	56.3
7	55.2
9	56.3

To evaluate the robustness of our method with respect to pocket flexibility or crystallographic errors or uncertainties (e. g. assignment of donor/acceptor properties in Asn, Gln or His) we assessed the sensitivity of our comparison procedure by applying a random perturbation of the assigned pseudocenter positions. In detail, the x, y and z coordinates of all pseudocenters were randomly modulated within local boundaries of -2 to +2 Å before calculating the histograms, which led to an average RMSD (see Eq. 2.1 on page 15) between the original and the perturbed pseudocenter sets of 1.99 Å. In fact the classification rates were only deteriorated on average by 4.7 % for $k = 1 \ldots 9$, indicating significant robustness with respect to these perturbations.

To further evaluate the discriminative power of RAPMAD compared to the original Cavbase implementation, we focused next on the larger and more complex dataset ATP/NAD$_{large}$. Moreover, we switched to the 10-fold cross-validation as evaluation technique, since it also provides the standard deviation of the classification rates as additional information.

As already mentioned in an earlier study, the Cavbase implementation achieves 79.2 ± 3.1 % using this dataset [82]. Using the BK algorithm to determine the maximum common subgraph of two graphs representing the binding sites to be compared achieves 84.6 ± 2.4 %. Also here LPCS and SEGA, which also provided very good classification results for ATP/NAD$_{small}$, perform best with 91.6 ± 2.2 % and 94.3 ± 3.5 %. Nevertheless, RAPMAD still excels BK and Cavbase, attaining 86.4 ± 3.2 % (see Figure 5.6 (a)). It is important to note that the results obtained by LPCS and SEGA are based on individually optimized parameters for the given datasets.



(a) Classification results of different approaches for the ATP/NAD$_{large}$ dataset. Cavbase performs worst, leading to only 79.18 ± 3.12 %. LPCS and SEGA perform almost equally well, attaining 91.58 ± 2.17 % and 94.31 ± 3.48 %, respectively. The BK algorithm returns considerable rates of 84.64 ± 2.40 % as well and RAPMAD, which still achieves rather impressive rates of 86.34 ± 3.19 %.

(b) Discriminative power of RAPMAD by performing classification tests on the datasets FAD vs. NAD$_{large}$, FAD vs. ATP$_{large}$ and FAD vs. FAD$_{cov}$ (FAD includes only non-covalent pockets). In all experiments shown, our method leads to convincing results, especially when comparing the pockets binding FAD non-covalently to those which bind it covalently (90.45 ± 5.85 %, right).

Figure 5.6 Classification results regarding the datasets ATP/NAD$_{large}$, FAD and FAD$_{cov}$.

We complete our classification tests considering the cofactor datasets which incorporated FAD binding sites (Fig. 5.6 (b)). In all cases our novel implementation achieves considerably good results. The comparison between pockets of the datasets FAD and NAD$_{large}$ was expected to be rather challenging due to the similarity in molecular structure of the two cofactors. Nevertheless, RAPMAD reaches $85.1 \pm 5.0\,\%$ correct classification rate in this case. Pockets hosting FAD or ATP can even be differentiated more easily. Here, results of $88.7 \pm 3.6\,\%$ are attained. Before running the experiment that concerns the FAD and FAD$_{cov}$ datasets, we randomly reduced the number of pockets in the FAD set from 429 to 114 in order to achieve two equally populated sets of pockets in both cases in order to avoid classification of a highly biased dataset. Remarkably, the best results are obtained on these datasets, where the mean rate of correct classifications exceeds $90\,\%$ ($90.5 \pm 5.9\,\%$). Also in this case the discriminating factor reflected by the shapes of the histograms must be coded in the spatial distribution of the pseudocenters because the pockets in the two datasets FAD and FAD$_{cov}$ consist of an almost identical number of pseudocenters (170 ± 55 and 177 ± 45). In this case, we attribute the good results to the different overall shape of the binding sites in FAD$_{cov}$, as these structures tend to host the cofactor in a much more buried fashion.

5.3.2 Classification Tests on a Multiple-Class Protease Dataset

The small protease dataset, which is based on the Merops classification, was utilized for the next evaluation. It includes 84 pockets split into three clans, namely serine, cysteine and metalloproteases. For assessment purposes, we also calculated the scoring matrix by using the Cavbase implementation. As the portions would be way too small to apply a 10-fold cross-validation on this matrix, we used the k-leave-one-out cross-validation for evaluation. Table 5.4 shows the classification rates of this experiment when either RAPMAD or the comparison approach originally implemented in Cavbase is used. [1] As classifier the k-NN method was used with $k = 1 \ldots 9$. Cavbase clearly deteriorates with increasing k, while the rates of RAPMAD improve. In general, however, the results of RAPMAD are more stable and do not vary as much as the Cavbase rates depending on the used classifier (standard deviation 2.3 *vs.* 5.1). Moreover, Cavbase omits about 17 % of the comparisons due to an overflow of its internal memory limitations. In RAPMAD, such problems are very unlikely and have never been observed in any of the trials executed in this study. As mentioned earlier the most challenging classification when using the k-NN classifier is given when all the used classes comprise equal size. By randomly discarding seven entries from the metallo protease set and 20 examples from the serine proteases we succeeded to obtain three equal-sized datasets (19 pockets in each subset). For this experiment the results do not change remarkably to the previous ones considering the larger and less balanced dataset. They now range between 80.7 % and 84.2 % for $k = 1 \ldots 9$ which indicates no substantial bias regarding the size differences of the used protease classes. Also recalculating the classification rates only on the common subset (as Cavbase has omitted 17 % of the comparisons) does not provoke remarkable changes of the results.

Table 5.4 Classification results of Cavbase and RAPMAD on the protease dataset.

k	**Cavbase**	**RAPMAD**	**RAPMAD***
1	86.9	79.8	84.5
3	83.3	82.1	84.5
5	79.8	85.7	88.1
7	77.4	84.5	86.9
9	73.8	83.3	85.7

In addition, RAPMAD was able to detect two clear outliers in the serine protease set (see red arrows in Figure 5.8), as they exhibited metal ions in their

[1] The column RAPMAD* indicates the results after removing artificially added metal ions from the structures 1Q3X and 2QXJ.

catalytic sites. In one structure (PDB code 1Q3X), a sodium ion was found, which was picked up from the buffer medium during the crystallization. In the second structure (PDB code 2QXJ), two copper ions are present that have been experimentally added to investigate their enhancing or inhibiting influence on enzyme activity (see Fig. 5.7). After removal of the metal pseudocenters

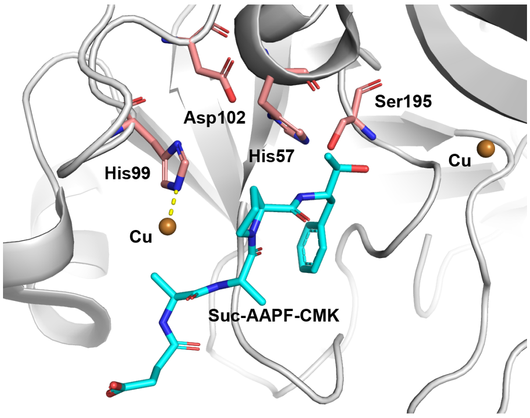


Figure 5.7 Catalytic site of human tissue kallikrein 7 (PDB code 2QXJ). The binding site is shown in complex with the inhibitor *succinyl-Ala-Ala-Pro-Phe-chloromethyl ketone* (Suc-AAPF-CMK). Two copper ions are shown as brown spheres, where the ion bound to His99, which is close to the catalytic triad (Asp102, His57, Ser195), accounts for a noncompetitive inhibition [31].

from the cavity descriptions of these two binding pockets, the accuracy of our method improved even further, which is shown in the column denoted by RAPMAD* (Table 5.4). On the contrary, the results obtained by the Cavbase implementation did not change remarkably. In the corresponding heat map of RAPMAD's scoring matrix (Fig. 5.8), the regions of high similarity across the serine and the cysteine proteases can be easily detected (upper right and lower left corner). Also for the metalloproteases a darker region can be identified, albeit less discriminating. The heat map suggests furthermore a close structural relationship between the active sites of cysteine and serine proteases, which is demonstrated off diagonal by the reddish areas in the upper left and lower right corner. For both enzyme species the mechanism follows very similar principles. It is known that serine proteases can be successfully morphed

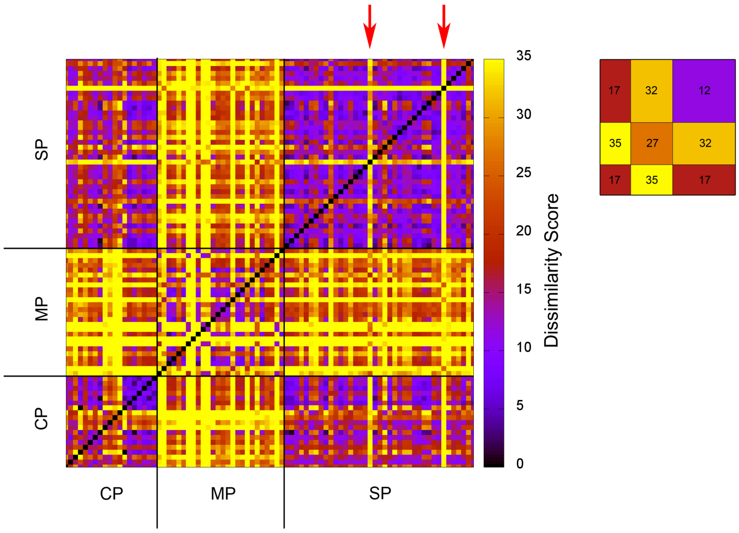


Figure 5.8 Heat map of the $n \times n$ dissimilarity matrix for the protease dataset. Next to the detailed heat map a smaller version is pictured that shows the average distances scores of all important areas. Areas of high similarities (violet) are obtained for the comparisons of pockets assigned to the same class. In the upper right corner of the matrix serine proteases (SP) are compared with each other with a mean distance score of 12, in the lower left corner cysteine proteases (CP) with a mean similarity score of 17. The region where metalloproteases (MP) are compared with each other remains more fragmented (mean score of 27), but nevertheless it discriminates well from the other protease pockets. Off from the main diagonal the detailed heat map also indicates structural similarity between SP and CP, which is highlighted by the darker regions in the lower right and upper left corner. Red arrows identify SP pockets from `1Q3X` and `2QXJ` that are obviously rather dissimilar to all the remaining SP structures.

into cysteine proteases and vice versa by exchanging the catalytic cysteine and serine residues [38] and there is also a Merops clan PA which contains serine proteases as well as cysteine proteases. Nevertheless, we observed that RAPMAD is able to obtain correct retrieval rates more than 80 % if only cysteine and serine proteases are subjected to a classification experiment.

5.3.3 Classification Tests on EC Datasets

As last classification experiment, we considered the EC and the serine protease datasets. With respect to the EC dataset we succeeded in 94.2 ± 2.4 % correct classifications. Thus, the differentiation of binding sites of proteins with respect to independently annotated EC numbers, RAPMAD apparently works very successfully.

Nonetheless the EC dataset is too small to allow for general conclusions. We thus decided to consider a larger serine protease dataset, which consists of twice as many binding sites (in total 1028 entries). In the following experiment, we removed the pockets which have not been annotated to any specific EC number yet, namely those artificially assigned to subclass “0” (65 entries) and, moreover, discarded all subclasses comprising only one pocket entry, as classification with the k-NN classifier requires at least two pockets. Subsequently, we created three subsets from the remaining subclasses for classification tests. In a first step, all 33 remaining subclasses were used and RAPMAD achieves an assignment accuracy of $86.63 \pm 2.94\,\%$. To avoid the above-mentioned problem of trivial majority voting, which results from classifying attempts based on imbalanced population, we decided to use in the next step only the classes comprising between 15 and 75 pockets. 10 classes remained and the accuracy of RAPMAD improved to $92.37 \pm 5.61\,\%$. Finally, only the two largest subclasses 4 and 5 (trypsin and thrombin), which include 275 and 225 pockets, respectively, were used to perform a two-class assignment experiment. In this case our algorithm performs best, succeeding in $96.8 \pm 1.69\,\%$ correct classifications. These experiments demonstrate the discriminative power of our method to correctly assign even closely related proteins, such as binding sites of serine proteases of the trypsin family that only differ in the last digit of the EC number.

We were also interested to which EC classes RAPMAD assigns the above-mentioned entries of subclass “0”. A listing of all 65 pockets together with the most similar structures of the remaining dataset shows a remarkably high success rate. In many cases, pockets of either the same proteins or proteins with a very high sequence similarity were matched (nine and eleven cases, respectively). In ten cases, our method furthermore discovered the true EC class of elements that were so far assigned to subclass 0. Here, the most similar pocket affiliated to a different subclass than 0 and a subsequent consulting of the PDB confirmed that also the query pocket actually belongs to the same subclass. These findings are even more considerable as all the relationships were detected by solely using putative binding sites, which are usually quite small portions of the entire protein surfaces.

5.3.4 Retrieval of Enzymes Catalyzing the Same Reactions

To assess the retrieval power of RAPMAD we tried to discover cavities of enzymes that catalyze the same chemical reactions. This search was based on the assumption that enzymes catalyzing the same reactions will share 3D similarity of their catalytic centers. We have chosen a binding pocket of a carbonic anhydrase II (EC number 4.2.1.1) as query structure (PDB code 1OQ5) and matched this entry against all cavities with EC annotation in the cavity database. Thereby a retrospective functional annotation study was conducted. This resulted in a total number of 161 822 comparisons, including 801 cavities assigned to the same EC annotation as the query. These structures were defined as correct hits. To face the RAPMAD comparison with the former method, we also evaluated this query with the original Cavbase comparison procedure. The results of both retrievals are illustrated by the ROC curves in Figure 5.9. At the very beginning Cavbase achieves a slightly better enrichment than RAPMAD, approaching an enrichment after one percent of the sorted scoring list of 0.46, where RAPMAD reaches 0.41. But, after the comparison of about 7 % of the database, the results of RAPMAD outperform Cavbase and lead to a higher AUC of 0.75, whereas Cavbase achieves 0.69. This difference also has to be assessed with respect to the runtimes of both approaches. When RAPMAD has scanned the entire database, the Cavbase implementation has actually processed less than 0.01 % of all entries.

Similar experiments were conducted using a trypsin and a subtilisin pocket as reference for a protease query. As pointed out in Ref. 132 trypsin and

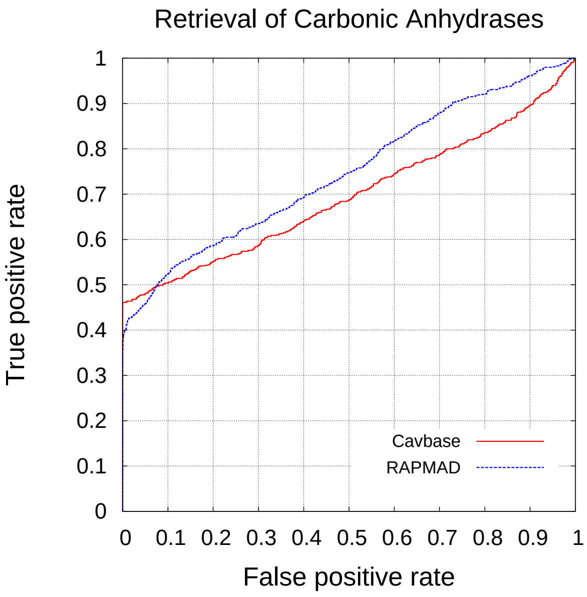


Figure 5.9 ROC curves illustrating the retrieval results of carbonic anhydrase pockets using RAPMAD and Cavbase. A binding site of a carbonic anhydrase (PDB code `1OQ5`, EC number 4.2.1.1) was used as query and compared against all EC-annotated pockets in Cavbase. The plots illustrate the detection of cavities with the same EC number as the query structure.

subtilisin exhibit the same biochemical function by executing the same enzyme mechanism but they share no sequence or fold homology. We used the pocket of `1TPO` for trypsin and `2PRK` for subtilisin as queries. In a first figure-of-merit, the performance of retrieving pockets that belong to the same family as the query pocket was evaluated. The retrieval based on trypsin as reference pocket achieves both for Cavbase and RAPMAD the same AUC of 0.68. In case of subtilisin as reference Cavbase performs slightly better, reaching an AUC of 0.75 whereas RAPMAD succeeds in 0.71. Subsequently, we removed all pockets of the same family apart from the query from the list of results and analyzed the performance of the pocket used for retrieval with respect to the other class. While using the trypsin pocket as query to retrieve subtilisin example Cavbase and RAPMAD attained comparable AUCs of 0.69 and 0.68, respectively. Using the subtilisin pocket to extract trypsin binding sites the retrieval even resulted in identical AUCs of 0.65 for Cavbase and RAPMAD. Taking all five experiments together, RAPMAD achieves on average an AUC of 0.694, while Cavbase achieves 0.692, which demonstrates once more equal performance of RAPMAD compared to the currently implemented approach in Cavbase. The results are summarized in Table 5.5.

Table 5.5 Overview of the retrieval results to detect similar binding sites validated in accordance to the annotated EC number.

Retrieving...	**AUC Cavbase**	**AUC RAPMAD**
carbonic anhydrases using 1OQ5.1	0.69	0.75
trypsins using 1TPO.1	0.68	0.68
subtilisins using 2PRK.1	0.75	0.71
subtilisins using 1TPO.1	0.69	0.68
trypsins using 2PRK.1	0.65	0.65
Average AUC	0.69	0.69

5.3.5 Retrieval of Proteins Binding a Particular Ligand

In the next step, we will focus on binding sites present in different proteins that bind the same ligand. This time we will focus on pharmaceuticaly relevant drugs that are present in a fairly large set of PDB entries. As a first example, we performed a retrieval experiment on pockets that bind methotrexate, a chemotherapeutic drug, which is also used for the treatment of autoimmune diseases. By searching the PDB we found 30 structures that bind methotrexate and are also stored in our current version of Cavbase. We then used the protein culling server PISCES [157] to reduce the putatively given redundancy in sequence space of this dataset. The sequence identity threshold was set to 20 % in order to consider only proteins beyond the so-called twilight zone threshold with respect to sequence [125], the minimal resolution was asked to be below 2.5 Å and the highest acceptable R-factor to 30 %, which resulted in four remaining structures: a pteridine reductase (`1E7W`), a thymidylate synthase (`1AXW`) and two dihydrofolate reductases (`3DFR,` 3DAU). The flexible structure comparison method FATCAT [169] subsequently returned that all selected structures do not exhibit significant similarity except the dihydrofolate reductases 3DFR and 3DAU, which motivated us to discard one of these two structures. We decided to remove the entry with lower resolution. Thus, we finally considered the three structures 1E7W, 1AXW, 3DAU. Next, we consulted Cavbase to retrieve all pockets accommodating methotrexate in these proteins, which suggested five pockets overall, since we found two independent binding pockets for 1E7W and 1AXW. These binding pockets also accommodate in addition the cofactor NADPH dihydro-nicotinamide-adenine-dinucleotide phosphate or 2'-deoxyuridine 5'-monophosphate, respectively. Nonetheless, all comparisons were performed based on the entire binding pocket instead of limiting solely to the regions hosting methotrexate. Thereby we can assess the capability of RAPMAD to deal with cavities that are significantly larger than the actual ligand-binding region of interest. For any of the pockets we performed mutual comparisons against the entire cavity database and plotted ROC curves showing at which stage the remaining methotrexate pockets were retrieved (Fig. 5.10). To define a hit set we retrieved all cavities from Cavbase that actually accommodate methotrexate (50 in total). By visual inspection six structures were subsequently omitted because LIGSITE was obviously not able to detect the methotrexate binding site properly (e. g. volume of more than 4000 Å^3 was assigned or two independent binding sites were joined), which finally resulted in 44 pockets.

In the following, we paid special attention to the pockets of the sequentially and structurally unrelated proteins, which are marked as triangles, circles and rectangles in Figure 5.10. As can be seen in the ROC plots, RAPMAD has no difficulties to detect pockets originating from the unrelated proteins, since they are all discovered among the first 10 % of the retrieval list (marked by 0.1 on the x-axis). This result appears very satisfactory, as it underlines that pocket similarity is only defined by the position of the pseudocenters and does not result from inherent sequence or structural folding similarity. Moreover, the

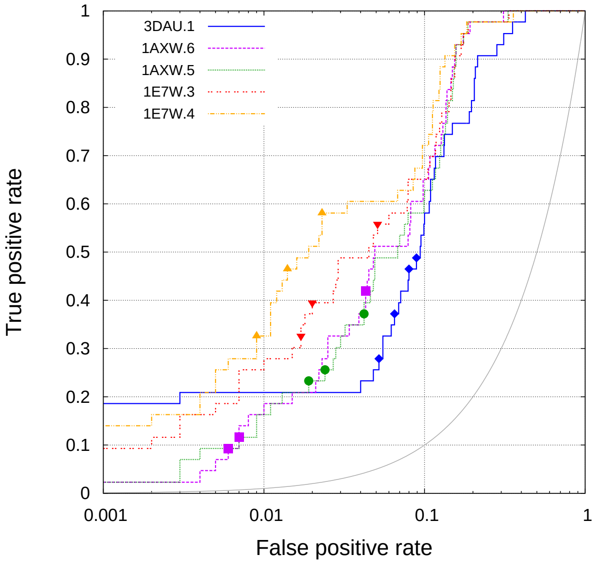


Figure 5.10 Logarithmically scaled ROC curves for the retrieval of protein binding sites that host the drug methotrexate. Five pockets of the sequentially and structurally unrelated proteins `3DAU, 1AXW` and `1E7W` were used as queries and compared against the entire database. To plot the ROC curves, a set of 43 additional methotrexate-binding pockets have been defined as a hit. In addition, the detection levels of the pockets corresponding to the other two unrelated proteins are highlighted as triangles, circles and rectangles. A random performance is indicated by the gray line. All ROC curves exhibit an average AUC of 0.92.

ROC curves show that independent of the actually selected query pocket at least half of the stored methotrexate-binding pockets are retrieved among the first 10 %, which indicates a rather promising early enrichment. The curves also exhibit a very convincing average AUC of 0.92.

As a second example, we used the binding pockets that host pemetrexed, a chemotherapy drug for the treatment of pleural mesothelioma and cell lung cancer. Again, we searched the PDB for proteins that bind pemetrexed and found six structures: four thymidylate synthases (`1JUJ, 1JU6, 3K2H, 4FQS`), a folate receptor (`4KN2`) and a pteridine reductase (`2X9G`). By culling this set of proteins using PISCES in order to reduce sequential redundancy, it was limited to three entries (`3K2H, 4FQS, 2X9G`) and by applying FATCAT we ended up with the two structures 2X9G and 4FQS. Next, the cavity database was exploited to identify the binding sites of these proteins that actually accommodate the desired ligand. In doing so, four binding sites for 2X9G and two for 4FQS could be retrieved from the database. The hit set of pemetrexed-binding sites that could be found in Cavbase comprised 19 pockets. The resulting ROC curves are shown in Figure 5.11, where detections of pockets that correspond to sequentially and structurally unrelated proteins are particularly highlighted. Also for this example, the ROC curves indicate that pockets of unrelated

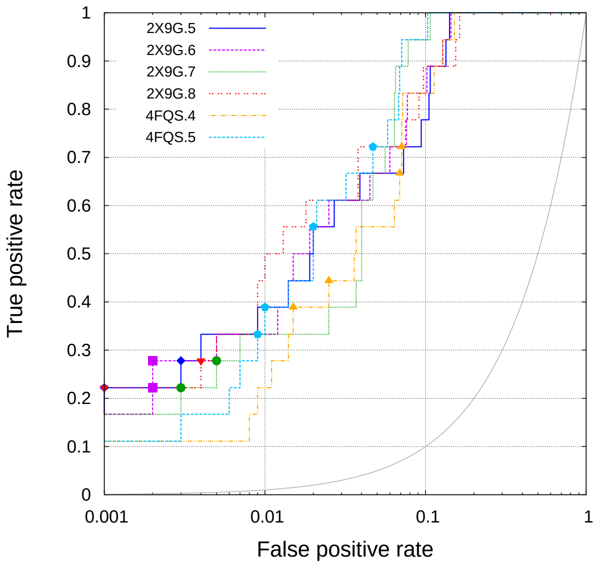


Figure 5.11 Logarithmically scaled ROC curves for the retrieval of protein binding sites that host the drug pemetrexed. Six pockets of the sequentially and structurally unrelated proteins 2X9G and 4FQS were used as queries and compared against the entire database. To plot the ROC curves, a set of 18 other pemetrexed-binding pockets have been defined as a hit. In addition, the detection levels of the pockets corresponding to the other unrelated protein are highlighted as triangles, circles and rectangles. A random performance is indicated by the gray line. The ROC curves exhibit an average AUC of 0.96.

proteins are retrieved on very early ranks, e. g. the first pteridine reductase example is on average detected after 1.2 % of the dataset, the first thymidylate synthase even already after 0.17 %! The curves in this experiment achieve an even higher average AUC, which is 0.96 and clearly demonstrates that RAPMAD can be used as a tool to seek for putative cross-reactivity.

5.3.6 Identification of New Ligands for the NMDA Receptor

Stimulated by the last example we wanted to apply RAPMAD to a predictive case study. In a collaboration with BioCrea[2] our goal was the identification of potential NAMs for the NMDA receptor. The search was initiated with a retrieval for the most similar pockets to the ifenprodil-binding sites of 3QEL and 3QEM. As these are homodimers LIGSITE detected two pockets for ifenprodil (Fig. 5.12 (a)) in each of the structures. In order to limit background noise and to focus on the essential information in the search process the four pockets were reduced to the pseudocenters in an area of 8 Å around the ligand prior to the subjection to comparative searches. Next, we used the methods RAPMAD and Local Cliques (LC), a graph-based approach which we presented previously (Chap. 3), to compare the four query pockets against all entries in the cavity database. These individual screening runs revealed eight ranking lists. In each case, the top 250 entries were extracted and merged to one single hit list and, subsequently, any duplicate pocket entries were removed. For the remaining 1562 pockets we extracted the accommodated ligands and reduced this collection further by visual inspection. Molecules that did not match with our predefined criteria regarding size and druglikeness were discarded, resulting in a set of 97 remaining ligands. Subsequently, the ligands were docked into the target pocket of 3QEL using the program GOLD, which yielded 16 promising hits. Of these, 15 structures – or at least similar derivatives – were referenced in the ZINC database and finally we purchased five compounds for testing in a binding assay (assay protocol can be found in Appendix, Sec. A.0.7). The remaining hits were already present in the in-house database of tested NMDA binders of BioCrea, thus no repetitive test was required. In case an exact

[2] BioCrea GmbH, Meissner Strasse 191, 01445 Radebeul, Germany.

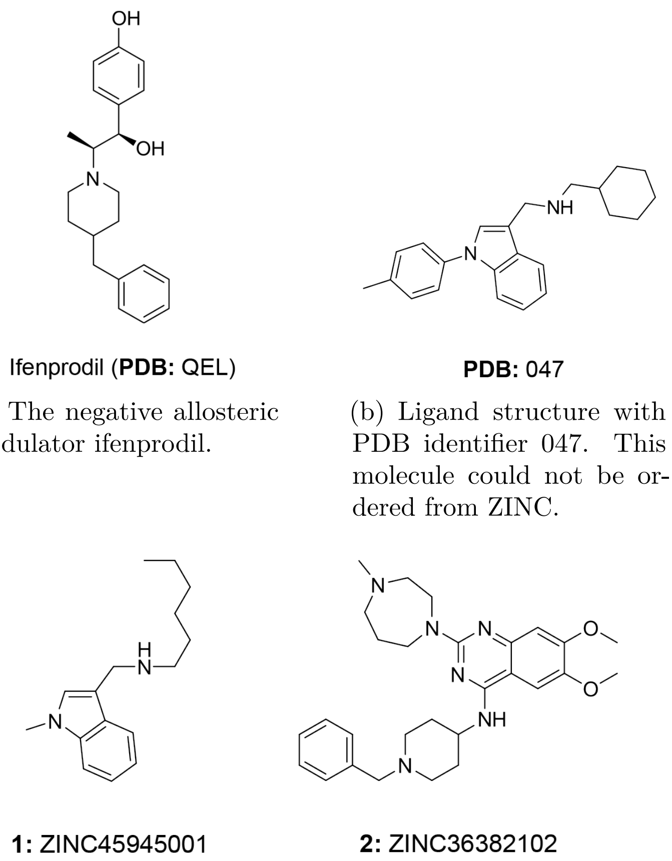
match of the retrieved structure was not possible in the ZINC database, the most similar structure was ordered.

Two of the tested compounds showed significant binding affinity to the NMDA receptor (Fig. 5.12 (c)). Compound 1 exhibits an IC_{50} of 1051 nM. It is a similar compound to the ligand with PDB identifier `047` (see Fig. 5.12 (b)) as the latter could not be ordered from ZINC. In the initial screening step (pocket comparisons) the binding site hosting this ligand was found by the method LC on rank 51, which is within the top 0.19 ‰ of the ranking list, when the pocket `3QEM.6` was used as query. Compound 2 achieves an IC_{50} of 1560 nM and was found by RAPMAD on position 133 (top 0.48 ‰) using the pocket `3QEL.8` as query. In a follow-up step the hit compounds could be modified synthetically by decorating them with additional functional groups, or specific moieties could be substituted by bioisosteric portions to further increase their binding affinities. This follow-up study will afford additional resources in synthesis but this would be by far out of the scope of this feasibility study. Nonetheless this predictive example shows that RAPMAD can be successfully applied to retrieve new ligands via a comparative binding pocket approach.

5.3.7 Runtimes

The major advantage of RAPMAD compared to all previously developed methods is its extremely short runtime without sacrificing any accuracy. The explanation for this impressive speed-up is the sole evaluation of distance histograms that are based upon certain reference points and can both be constructed and compared with linear complexity $\mathcal{O}(n)$. This is an utmost improvement, since the fastest hitherto known methods work with at least cubic complexity $\mathcal{O}(n^3)$. The runtime of RAPMAD will now be analyzed in detail.

Building histograms: Let n be the number of pseudocenters in a protein binding site. At first, seven subsets are generated from the total set of pseudocenters, which requires n calculations. After that the coordinates of the



Ifenprodil (**PDB:** QEL)

(a) The negative allosteric modulator ifenprodil.

PDB: 047

(b) Ligand structure with PDB identifier 047. This molecule could not be ordered from ZINC.

1: ZINC45945001

2: ZINC36382102

(c) Hits of the binding assay that exhibit substantial affinity to the NMDA receptor. Compound 1 shows an IC_{50} of 1051 nM and compound 2 an IC_{50} of 1560 nM.

Figure 5.12 Results of the virtual screening for novel NMDA receptor inhibitors.

centroid and the centroid closest are calculated for each subset ($2n$ calculations in total). Finally, for each of the seven subsets the distances from the centroid and the centroid closest to all pseudocenters are measured, which requires $7 \cdot 2 \cdot n$ calculations. In sum, $n + 2n + 7 \cdot 2 \cdot n = 17n$ calculation steps are performed, which is written as $\mathcal{O}(n)$.

Comparison of pockets: Let $m = \max(|H_A|, |H_B|)$ be the maximum number of bins of two histograms that are used in a pairwise comparison. For all the seven subsets of pseudocenters the mean distribution M of the two input histograms has to be computed (m calculations). Subsequently, the

Kullback-Leibler divergences KLD_A and KLD_B are determined in $2m$ steps. Thus, the comparison of two protein binding sites requires $7 \cdot (m + 2m) = 21m$ calculation steps, which is again in linear complexity $\mathcal{O}(m)$.

To demonstrate the actually achieved speed-up, we compared the runtime of RAPMAD to those of the previously developed programs. The runtimes of GA and BK were adopted from previous studies [46, 100]. While determining the runtimes of the original Cavbase implementation, LPCS, SEGA and RAPMAD we used a state-of-the-art computer equipped with an Intel® Core™ i7-3770 CPU (3.4 GHz) Processor and 16 GB memory. All times have been measured using only a single core of the device and a benchmark set of 1000 randomly chosen pairs of protein binding pockets were compared. Table 5.6 lists the mean runtimes and the standard deviations. Obviously, RAPMAD requires exceptionally short runtimes that are more than 40 000 times faster than those of LPCS or SEGA (all Java implementations), while sacrificing only a minor amount of accuracy with respect to the classification. Most importantly, the currently implemented binding site comparison approach in Cavbase (C++ implementation) is outperformed by more than five orders of magnitude. Figure 5.13 illustrates the required runtimes on a logarithmic(!) scale which emphasizes the magnitude of the difference. The diagram points out that a comparison of one pocket against the entire database (about 275 000

Table 5.6 Mean runtimes in seconds and standard deviations of several binding site comparison approaches and RAPMAD using 1000 randomly chosen pairs of binding sites for retrieval.

Method	**Runtime**
Cavbase	12.76 ± 44.90
GA	121.74 ± 418.02
GAVEO	584.51 ± 2199.02
LPCS	2.24 ± 1.27
SEGA	1.85 ± 2.27
BK	6.91 ± 32.59
RAPMAD	0.000045 ± 0.00012

pockets) requires less than 13 seconds when using RAPMAD. The next faster methods SEGA and LPCS in contrast require 5.9 and 7.1 days, respectively, to fulfill this task. This advantage clearly moves binding pocket comparisons with RAPMAD into the scope of interactive modeling.

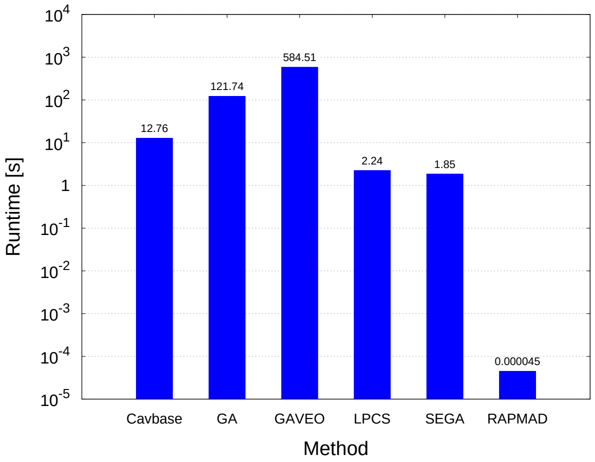


Figure 5.13 Average runtimes in seconds which are required for a single binding site comparison by all previously developed methods. The runtime is plotted on the y-axis in logarithmic scale. The ultra-fast RAPMAD clearly sticks out, since it is more than four orders of magnitude faster than the fastest hitherto developed algorithms. The algorithm presently implemented in Cavbase is outperformed by more than five orders of magnitude.

5.4 Conclusions

The detection of similar binding sites along with their occupants can greatly assist the drug development process particularly by suggesting bioisosteric molecular portions or novel scaffolds for optimization and predicting putative sites of cross-reactivity. In the present study, we described the development, evaluation and application of the novel similarity retrieval engine RAPMAD to detect putative protein binding sites, which are available from the Cavbase database. Our approach represents binding sites in terms of a set of translational and rotational invariant histograms. Undoubtedly on first sight, the histogram representation implies a loss of information about the cavity shape, as no surface points are considered. But as shown in previous studies, most of the more recently developed comparative matching algorithms of Cavbase entries also operate without regarding surface points [44, 45, 100]. They solely consider pseudocenters and achieve impressive retrieval rates. Our new method RAPMAD accomplishes results of similar quality in both classification and retrieval that are comparable to the achievements of other state-of-the-art methods, albeit, importantly enough, in a fraction of time. It returns a distance score that can be interpreted as a measure of dissimilarity subsequently used to rank the compared pockets with respect to the query. Although a pass/fail-decision whether two pockets are similar or not may be desirable in some cases, a graded ranking can be beneficial for other applications. However, this feature is shared by RAPMAD with most of the other binding-site comparison methods which all are based on the respectively implemented scoring functions.

In classification tests on datasets of pockets that bind a specific cofactor, our procedure achieves substantially better results than the originally implemented algorithm in Cavbase. We showed furthermore that different types of proteases can be successfully discriminated with respect to a totally independent Merops classification. A set of serine proteases can be correctly classified with respect to the fourth digit of the assigned Enzyme Commission number. Also enrichment experiments confirm the accuracy by correctly retrieving the active sites of target proteins on first ranks. Most importantly, all comparisons could be performed with hardly any loss of accuracy but at higher speeds of more than five orders of magnitude faster compared to the algorithm currently implemented in Cavbase. When comparing the runtime of RAPMAD to the information published by the authors of other recently developed fast comparison methods, the efficiency of our appraoch can be underlined by the following figures. In one second and using a single CPU core, RAPMAD is able to perform about 20 000 pocket comparisons and thus clearly outperforms methods such as BSSF, FuzCav, PocketMatch and TrixP.

Large-scale studies such as screenings of the entire Cavbase database, which comprises almost half a million cavities in its latest version (August 2014), become feasible within seconds using RAPMAD, making the tool applicable as interactive idea generator. This was demonstrated by eleven database screenings that were carried out for query pockets binding to either the cancer drug pemetrexed or methotrexate. The results of these retrievals are convincing and stress that our method displays a valuable and extremely fast alternative to current binding site comparison procedures. In our study, we focused on comparison procedures for putative protein binding sites that use the internal Cavbase format and descriptors. We appreciate that other powerful descriptor schemes to detect and compare pockets are available (such as FuzCav, PocketMatch or BSSF), however, a direct 1:1 comparison of RAPMAD with these tools will be difficult as they built on distinct descriptor sets and cavity extraction algorithms. Apart from the search performance in terms of speed and retrieval rate, the success and relevance of such methods highly relies on the strategy applied to define similarity. Very different concepts have been used in literature but comparing them would address a completely different aspect, which is not intended by this study.

Combining the presented method with other approaches for the comparison of Cavbase entries could help to further improve the accuracy by implementing a kind of consensus scoring. There is a variety of ways to realize such a combination (e. g. rank-by-number, rank-by-rank, rank-by-vote [158]) and it has been shown that consensus scoring in general delivers more robust results than a single method [151]. A combination of two or more highly efficient methods applied subsequently after the histogram approach will still be a lot faster than a single scoring as used in the current Cavbase implementation. In our study we implemented a consensus scoring comprising RAPMAD and the more elaborate graph-based LC method. By using this combination as a virtual screening and data mining tool we were able to identify promising hits in a binding assay regarding the affinity to the NMDA receptor. Testing five compounds revealed two ligands that exhibit IC_{50} values of about 1000 and 1500 nM, respectively, which corresponds to a hit rate of 40 %.

Although our method leads to very good results in general, limitations become apparent when two binding sites are matched only in a certain region. RAPMAD estimates global similarity between binding sites and currently runs into problems when pockets hosting the same ligand deviate rather strongly in size and shape. This became obvious in the screenings for methotrexate-hosting pockets, where pockets hosting methotrexate only are found less successfully than those accommodating methotrexate along with a cofactor such as NAD, NDP or UMP in the dehydrogenases or synthases.

The screening for binding pockets along with the retrieval of bound occupants as an idea generator is a typical modeling approach in drug design where a sketchy defined search query has to be fitted and optimized according to the obtained results. If the database screen requires days to be accomplished however, any interactive optimization of the search query is impossible. With RAPMAD this very typical and appropriate working strategy becomes feasible.

Finally, our method is not restricted to protein binding sites alone but can easily be exploited for the comparison of other labeled point sets in 3D space, e. g. pharmacophoric points across ligands. With regard to its linear complexity and the ultra-fast runtime, it might also be applied easily to much larger point sets, which will be impossible for graph-based methods due to their exceptionally long runtimes and high memory requirements. Further examinations could focus on exploring the diverse areas of applications.

Part II

Improvement of the Knowledge-based Scoring Function DSX

6

The Program DSX

6.1 Introduction and Motivation

After the target protein has been identified, the binding site detected and a lead structure discovered, iterative optimization steps are usually applied to optimize the novel ligand molecule. Lead structures already show a convincing potency but are still not effective enough for therapeutic application. To increase binding affinity, a large variety of properties can be improved which strengthen the binding to the target protein, adjust the duration of action, optimize bio-availability and minimize side effects. However, it may not be reasonable to just randomly add substructures or functional groups, as the properties that make a molecule likely orally active are bound to specific physicochemical limits. These have been defined for instance by Lipinski's *rule of five* [95], in order to avoid early undesired metabolic degradation already in the first pass through the liver.

Once a new candidate structure has been designed and the crystal structure of the protein-ligand complex determined, a quantitative structure-activity relationship (QSAR) analysis can be performed to provide the basis for further modification hypotheses that aim at an increase of binding affinity. Computational methods are frequently applied to estimate binding affinity and the optimal placement of a ligand in the binding pocket. For this purpose, plenty of *molecular docking* tools have been developed. Examples include the programs DOCK [39], FlexX [121], AutoDock [54], eHits [173], Glide [48] and OpenEye's FRED [99]. Moreover, also public web services emerge, such as the "1-Click Docking" of mcule[1]. While the actual process of docking just aims to find reasonable placements of the ligand in a receptor pocket regarding its shape, the evaluation of these poses is mostly performed by a scoring function. Most often docking programs provide one or more scoring functions. Although many of them succeed in the generation of geometries that are close to the experimentally observed crystal structure, the native pose of the ligand is nevertheless rather arbitrarily contained within the list of top-ranked docking solutions [120], but not necessarily placed on the first rank.

In this chapter, we describe the improvement of *DSX*, a knowledge-based scoring function for the assessment of protein-ligand complexes, by introducing a new term for geometry validation. Knowledge-based scoring functions calculate the score by summing up a number of statistical potential contributions derived from a structural database [98, 26, 23, 80, 55, 106]. In case of DSX, the used data are extracted from the Protein Data Bank (PDB) and the Cambridge Structural Database (CSD) [1], respectively. DSX is an extended version of *DrugScore*, which has been presented by Gohlke et al. in 2000 [52] and has undergone several improvements ever since. In its first version, the heavy atoms of the protein and the ligand that participate in complex formation are assigned to one of 17 Sybyl atom types (a set of atomic descriptors devised by Tripos Inc.), as only for this subset sufficient data was found in the PDB. Subsequently, all pairs of protein and ligand atom types with a distance of less than 6 Å are assessed by using the statistical potentials derived from the distributions in the PDB. In consequence, the total score can be calculated as the sum of all pair scores. In 2005 Velec et al. also

[1] Source: `https://mcule.com/apps/1-click-docking`. Accessed March 2, 2015

derived statistical potentials from the CSD [149]. The CSD contains a multiple of the number of structures found in the PDB, which are moreover better resolved in general. Consequently, this leads to a more balanced distribution of atom types and potentials of superior statistical significance. Experimental results confirmed the gain of accuracy, which could be achieved by the usage of CSD data in DrugScore. Finally, Neudert and Klebe presented the version *DrugScore eXtended* (DSX) in 2011. In this release the set of atom types has been remarkably expanded to the much larger set of fconv atom types. In total, 158 types are defined of which 64 were used in DSX in order to avoid a too detailed representation of ligand and protein interface contacts and to regard only those occurrence frequencies with sufficient statistical evidence.

However, all of the hitherto existing versions of DrugScore solely consider atom pair *distances* to assess the pose of a ligand in a binding site. Even though DSX contains the option to additionally evaluate torsion angles formed in the complex, other very important geometric properties such as hydrogen bond (H-bond) angles are largely neglected. Any angular constraints in DrugScore and DSX are only indirectly imposed in case the distance information of regarded atom-atom pair potentials is locally highly over-determined. H-bonds are however regarded as the most important specific interactions in complex formation by far [12]. As stated by Williams and Ladbury, optimization of local H-bonds is the prime approach for favorably enhancing the interactions between pairs of molecules [166]. H-bonds are very important non-covalent interactions and of great interest in drug design. H-bonding involves an electronegative donor group X, which is mostly O, N or S, and is able to withdraw electrons from the proton in a covalent X–H bond. The hydrogen is thus partially positively charged which results in the capability of another electrostatic interaction between H and a second electronegative group Y. This is a simple electrostatic view of H-bonding originally based on the work of Pauling in 1938 [115]. It is nevertheless still comprised in most computational models of bio-molecular systems and has finally been standardized by IUPAC in 2011:

> “The hydrogen bond is an attractive interaction between a hydrogen atom from a molecule or a molecular fragment X–H in which X is more electronegative than H, and an atom or a group of atoms in the same or a different molecule, in which there is evidence of bond formation [5, 33].”

Although significant research efforts have been spent on H-bond energies of different types, the results have been mostly reported in relation to the gas phase. It is therefore inappropriate to simply extrapolate them to the aqueous phase [118, 76]. However, also several studies were published where H-bond data were determined in a protein environment [77]. For this purpose, the change in molar Gibbs free energy of binding ΔG is calculated in both, the complex considering the wild-type protein structure and a second one considering a mutated variant of the protein that does not allow for a specific H-bond formation:

$$\Delta G = -RT \cdot ln \left[(\frac{K_{cat}}{K_m})_{mutant} / (\frac{K_{cat}}{K_m})_{wild-type} \right] \ .$$

This enables an estimation of the energy contribution of individual H-bonds. The results indicate that an H-bond between uncharged partners accounts for about 0.5–1.5 kcal/mol [42] and studies on charge-assisted H-bonds and salt bridges suggest that these types contribute between 2.4 and 4.8 kcal/mol to binding affinity [53]. The benefit of the usage of a two-dimensional H-bond potential in a scoring function has already been reported by Zheng et al. In their study the authors obtained substantial improvements when they compared the results to the previous version of their scoring function wise that solely considers distances [172]. Thus, there is justified evidence to believe that the success rates of DSX can be further improved when also considering H-bond angles in the scoring step. We therefore developed a new version of DSX which also contains an assessment of possible H-bonds formed between receptor and ligand atoms.

6.2 Methods

Since a detailed description of the theory behind DSX can be found in references 52 and 109, as well as in the thesis of Gerd Neudert [111], the methodical part of this work will be restricted to the information which is necessary to comprehend the extensions that will be presented in the following.

6.2.1 Statistical Potentials

The basic idea of a knowledge-based scoring function such as DSX is to make use of the Boltzmann law to transform observed frequencies into statistical potentials. More precisely, the observation of an atom-type pair p (protein atom) and l (ligand atom) with distance d is scored by comparing it to the probability to find p and l in distance d with respect to the overall atom-type pairs that exhibit the very same distance to each other. Let ρ and ρ_{ref} therefore be two normalized distributions. $\rho(p, l, d)$ defines the probability to observe p and l in distance d and it is calculated as

$$\rho(p, l, d) = \frac{N(p, l, d)}{\sum_{d=0}^{6} N(p, l, d)} \ , \tag{6.1}$$

where the function $N(p, l, d)$ returns the frequency of atom pair p_l found in distance d. ρ_{ref} is the mean pair distribution function, which is also called *reference distribution*, and is calculated as the average of all distributions $\rho(p, l, d)$. The considered distances cover a range from 1 to 6 Å with a step size of 0.1 Å in case of DrugScore and 0.01 Å in case of DSX. Following Sippl's approach [139], a score $s(i)$ for each atom-type pair in the protein pocket of the receptor-ligand complex is calculated as

$$s(i) = -ln\left(\frac{\rho(i)}{\rho_{ref}}\right) \ . \tag{6.2}$$

Every protein atom and ligand atom can be used several times to form an atom-type pair, depending on how many counterpart atoms are found in the considered distance range. The total score S for the entire complex is finally computed by summing up the scores of all individual atom-type pairs as $S = \sum_i s(i)$.

6.2.2 Potentials for Hydrogen-Bond Angles

In the past it has already been discovered that there are clear deviations from an optimal angle of the H-bonds present in protein-ligand complexes [128]. Hence, we picked up the approach of statistical potentials also for this type of contact geometry, which has been successfully employed for atom pair distances in all previous versions of DrugScore. While the interactions of hydrophobic molecular portions and their counterparts can be adequately assessed by inspecting the distance only, features like H-bonds need to meet additional geometrical requirements. H-bond distance and angle distributions have already been extracted from the PDB or CSD in earlier studies and were subsequently used in scoring functions of protein-ligand complexes [172, 103], protein-protein complexes [80], protein-RNA interfaces [23] and proposed for the application in protein structure refinement [55]. Even three-dimensional H-bond probabilities have been collected from the CSD and were proposed to be used to calculate putative hot spots for ligand interactions [103]. In order to match the approach of DrugScore and to process on the most current data, we will however derive new statistical potentials for H-bond angles from both the PDB and the CSD. We extracted data from structures resolved by X-ray crystallography only, which tends to be rather poor at locating H-atom nuclear positions (positions of hydrogens can only be recognized in structures with a resolution better than 1 Å). Therefore, we calculated the optimal positions of hydrogens for each atom of interest with respect to the valence shell electron pair repulsion (VSEPR) model [50] and defined this as the optimal geometry for an H-bond.

To collect any available H-bond geometries from experimental data depositories, we first of all had to identify the pairs of atom types that are actually able to form an H-bond. Therefore, we extended the atom types which are assigned by the fconv routine by another property

$$pharm\_group \in \{acceptor, aromatic, doneptor, donor, hydrophobic, n/a\} \ ,$$

which represents the pharmacophoric type of an atom and has originally been introduced in the program *HotspotsX* [111]. The value n/a stands for an unknown pharmacophoric property. Only those atoms types that match with the pharm_group H-bond *acceptor*, *donor* or *doneptor* are considered for the further evaluations. Table 6.1 lists the atom types in every considered pharm_group. A more detailed description of the entire set of atom types can be found in the Appendix of Ref. 109.

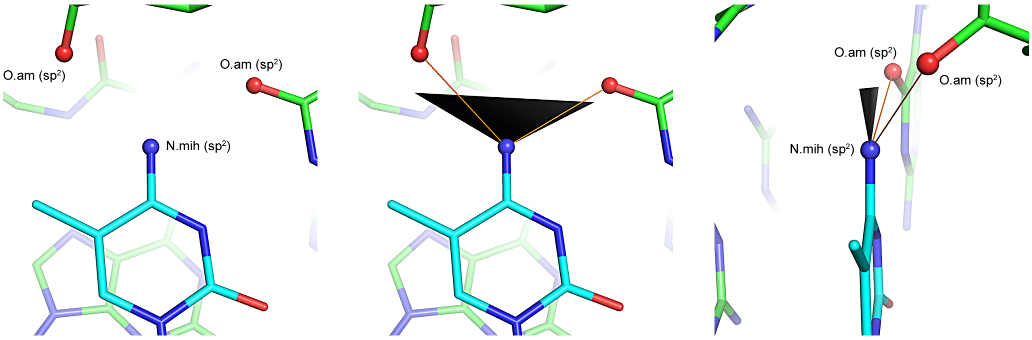
pharm_group	**fconv atom types**
Acceptor (31)	N.1, N.2n, N.2s, N.3t, N.aat3, N.ar2, N.oh, O.2co2, O.2es, O.2hal, O.2p, O.2po, O.2s, O.2so, O.3es, O.3et, O.3eta, O.3po, O.3so, O.am, O.carb, O.co2, O.n, O.o, O.r3, P.3, S.2, S.3, S.r3, S.s, S.thi
Donor (12)	N.4h, N.amp, N.ams, N.ar3h, N.ar6p, N.arp, N.guh, N.ims, N.mih, N.ohac, N.samp, N.sams
Doneptor (16)	N.2p, N.3s, N.3n, N.3p, N.aap, N.gu1, N.gu2, N.mi1, N.mi2, N.r3, O.3ac, O.3oh, O.h2o, O.noh, O.ph, S.sh

Table 6.1 Overview of the 59 atom types which belong to one of the pharm_groups H-bond acceptor, donor or doneptor.

Since the assessment of distances is already implemented in DSX, we decided to only derive additional potentials for the bond angles. H-bonds can be formed between a donor (atom) type and an acceptor type, between donor and doneptor, as well as between acceptor and doneptor. As it is shown by this list of considered atom types, we focus on strong and very strong H-bonds and neglect the weaker interactions which could be formed between hydrogens attached to electron-deficient carbon atoms that act as donor groups. Additionally lone pairs of electronegative atoms and delocalized Π-orbitals of unsaturated or aromatic systems can act as acceptors [166]. In addition, we only wanted to process geometric information of possible H-bonding partners in reasonable distances. Considering the exhaustive study on various typical bond lengths (covalent and non-covalent ones) carried out by Bissantz et al., the distance between the two heavy atoms forming an H-bond is usually 2.6–3.2 Å (see Ref. 12, Tab. 2). We expanded the maximum range to 3.5 Å to regard possibly given crystallographic uncertainties, which could range on average up to 0.2–0.4Å (personal communication with Dr. Johannes Schiebel), in order to include all H-bonds in less-resolved structures. All remaining pairs will not be regarded by our H-bond scoring scheme in the following.

In the next step, the optimal position of the interaction partner is calculated in accordance with the VSEPR model. Therefore, atom type information like the element type, pharm_group and hybridization state is employed. In the following, some examples for computing the optimal geometries will be presented, which can be regarded as representatives for all other considered fconv atoms types.

In case of a double-bonded, sp^2-hybridized nitrogen atom which is covalently attached to one non-hydrogen atom (fconv atom type N.mih), two values have to be calculated in order to determine the angle scatter with respect to a putative partner atom. In Figure 6.1 (a) the spatial positions of two possible H-bond partners are shown (oxygens, illustrated as red spheres), which are both in a distance of 3.0 Å from the donor nitrogen. The optimal direction of the H-bond partners can be estimated by calculating the positions of the lone pairs, each of which is spanning a 120° angle with the bond of the nitrogen and the ring carbon atom. These two lines are represented in Figure 6.1 (b) by the legs of an isosceles triangle. In the next step, the actually observed shortest contact geometry (shown by the two orange lines) is perpendicularly projected onto the plane that is represented by the triangle. In other words, the shadow of the orange line thrown on the triangle is calculated by using an imaginary light source which shines from perpendicular to the triangular plane. Using this projection the deviation from the shortest contact direction is determined,



(a) A nitrogen atom and two oxygens as putative H-bond partners.

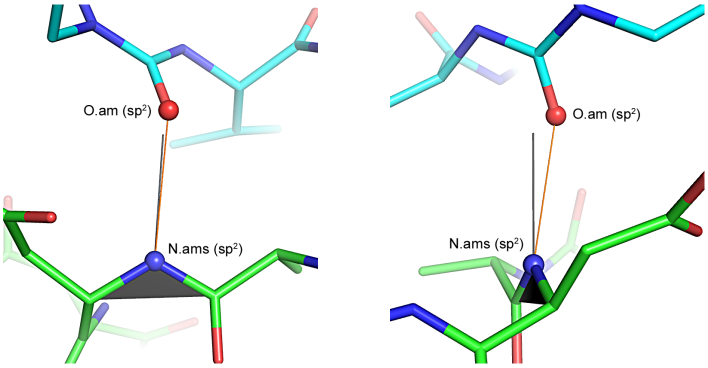
(b) Observed H-bond geometries (shortest contact direction, orange lines) and optimal geometries that match with the assumed N–H bond vectors (edges of the triangle).

(c) Rotated view to illustrate deviations from the optimal plane, spanned by the two assumed N–H bond vectors.

Figure 6.1 Example of the H-bond angle deviations for an sp^2-hybridized nitrogen atom (type N.mih) together with two oxygens as H-bond partners of type O.am. The receptor structure is displayed with green carbons and the ligand with cyan carbons.

which is thus called *direction deviation*. In addition, a second deviation is calculated that defines the deviation from the optimal plane (Fig. 6.1 (c)). This is achieved in a rather similar way, namely, by computing the angle between the observed shortest contact direction (orange line) and the calculated projection on the plane. According to the introduced terminology this value will be called *plane deviation*. These two computed values are sufficient to exactly define the position of every point on a hemisphere around the central origin atom. We do not need to define the position in the entire sphere as we assume local symmetry of the distribution and hence the frequency of observations will be the same on both sides. This assumption is also applied in other geometry statistics systems such as IsoStar [18].

In case of an sp^2-hybridized nitrogen atom, which is bound to two non-hydrogen atoms (type N.ams), the calculation of the plane deviation is similar to the procedure described in the previous case. The calculation of the direction deviation is however slightly different to the approach mentioned above. Since there is only one putative position of a hydrogen atom left, the N–H bond vector oriented along the projected line bisecting the C–N–C bond angle of the considered fragment (gray line in Figure. 6.2 (a)). The direction deviation is thus calculated by means of the projection of the observed direction on the plane and the shortest contact direction.



(a) Observed and shortest contact geometries shown by the orange line and gray line & plane, respectively.

(b) Rotated view to illustrate deviations from the optimal plane.

Figure 6.2 Example of the H-bond angle deviations for an sp^2-hybridized nitrogen atom (type N.ams) together with an oxygen as possible H-bond partner of (type O.am).

For an sp^3 hybridized oxygen atom only one deviation value has to be calculated. The optimal direction to place the putative O–H bond vector is found by creating a cone with a 109.5° opening angle along the extended oxygen–carbon bond. As the H-bond is freely rotatable in this case, the optimal position can be illustrated as anywhere on the cone above the oxygen atom (Fig. 6.3). Here, the direction deviation is computed by simply calculating the angle between the central sp^3 oxygen, its covalently attached carbon and the possible H-bond partner atom and, finally, subtracting 109.5°. The final subtraction is needed to obtain just the deviation from the estimated optimal geometry.

All calculations of spatial deviations described above are carried out for

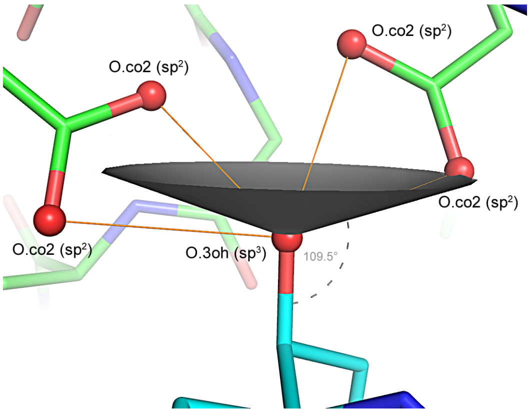


Figure 6.3 Example of H-bond angle deviations for an sp^3-hybridized oxygen atom (type O.3oh). Four possible H-bond partners of type O.co2 are in a suitable distance of 2.6–3.0 Å. Since the optimal direction for an H-bond partner is freely rotatable around the O–C bond, it is illustrated by a cone (gray) that forms an angle of 109.5° to this bond. The observed geometries of all possible partners are shown as orange lines.

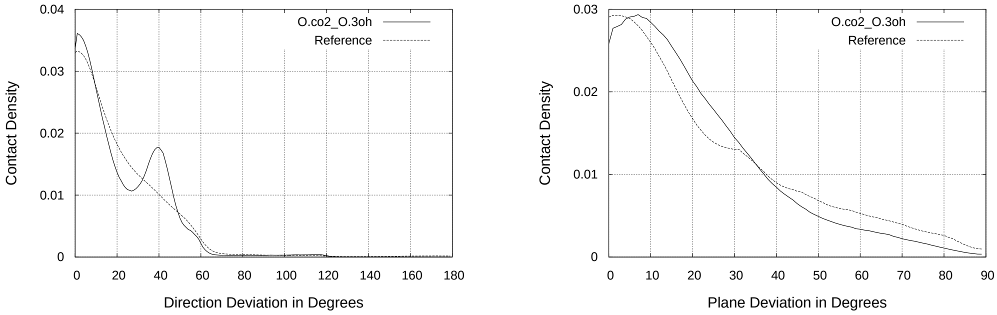
both partners of an interaction pair, either in the direction from protein atom to ligand atom and vice versa.

6.2.3 Data Collection

For the generation of statistical potentials for the H-bond deviations, we collected data from the PDB and the CSD. The used version of the PDB (May 2014) contained 101 046 structures in total, of which 43 694 were X-ray structures with a resolution of 2.5 Å or better. These structures were subjected to the program fconv, which disassembled them to complexes comprising the host protein and the corresponding ligands. Only ligands with six or more non-hydrogen atoms of the element type C, N, O, S or P were considered for further processing. Subsequently, the angle deviations of 742 978 possible H-bonds were calculated as described above.

In case of the CSD, we used the ConQuest software [19] to retrieve structures from the database. As basis of the search, a single carbon atom was used. In addition, the following filters were applied to ensure that only structures of sufficient correctness were retrieved: a minimum R factor of 0.05 was chosen and disordered, erroneous or polymeric structures were discarded. Thereby a total of 401 424 small-molecule structures were collected. We then used fconv to generate the crystal packings so that every molecule was fully embedded in a set of surrounding crystal mates. As there can be more than one symmetry-equivalent molecule in a CSD entry, we thus produced 1 475 299 small-molecule complexes, which were likewise subjected to our angle deviation calculation. In total, 875 425 possible H-bonds were evaluated.

To make the extracted data available to the DSX formalism, all measured deviations were recorded in histograms with a bin size of 1° to obtain discrete distributions for further processing. The histogram for direction deviations consisted of 180 bins, the histogram for plane deviations of 90 bins (since 180° and 90° resp. are the maximum measurable directional or plane deviations). Given the sets of 59 considered atom types (see Tab. 6.1) we ended up with 2120 possible interaction pairs. We restricted any distribution containing less than 2000 values from further consideration, which resulted in 79 remaining distributions for the PDB and 144 for the CSD (lists of respective pairs are shown in Appendix, Sec. A.0.9). The distributions were afterwards normalized by calibrating the area under the curve to 1 in order to obtain a density distribution and straightened by applying a Gaussian smoothing. The value of σ, which is the parameter determining the width of the Gaussian function, was set to 3. Finally, the distributions were transformed into statistical potentials by calculating the reference distributions for the direction and plane deviations and applying the workflow described in Section 6.2.1. Figure 6.4 shows the density distribution of the direction deviations (a) and the plane deviations (b) of the highest populated interaction type O.co2_O.3oh as solid lines. These define the measured geometry deviations from the atom type O.co2 to atom type O.3oh. In addition, the calculated reference distributions are shown as dashed lines. As an example, Figure 6.5 moreover shows the resulting plane potential for this interaction type.



(a) Direction distribution.

(b) Plane distribution.

Figure 6.4 Density distributions of the most populated interaction type in the PDB (O.co2_O.3oh) for which 83 526 H-bond geometries were evaluated.

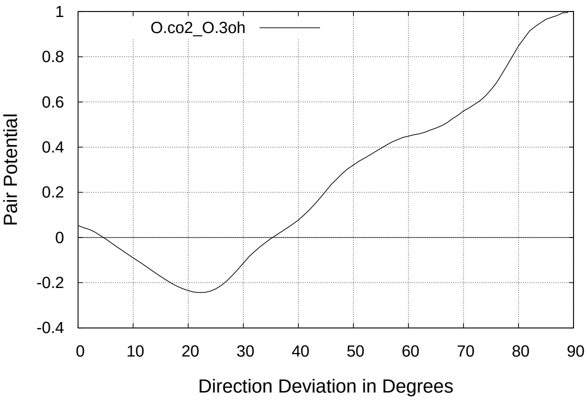


Figure 6.5 Potential for the plane deviation of type O.co2_O.3oh.

6.2.4 Score Calculation

The original distance score S_d for a receptor-ligand complex is computed as the sum of the potential values for all detected atom pairs with a distance of up to 6 Å. The newly added H-bond scores will be calculated likewise. Here, all H-bonds in the complex will be scored individually and subsequently summed up to the final H-bond score S_{hb}. More precisely, it is calculated as

$$S_{hb} = \sum_{i=0}^{N_{hb}} \frac{P_d(A_i\_B_i) + P_p(A_i\_B_i) + P_d(B_i\_A_i) + P_p(B_i\_A_i)}{N_i} , \qquad (6.3)$$

where N_{hb} is the total number of H-bonds in the complex. The two involved atoms of an H-bond i are denoted as A_i and B_i. The function P_d returns the potential value of the measured direction deviation; P_p returns the potential value of the plane deviation in case it can be detected for this interaction. Finally, N_i is the number of available potential values that have been summed up (between zero and four). If N_i is zero, the entire fraction is set to zero as well.

6.2.5 Test Datasets

In order to compare the results of the former DSX version to the newly extended variant, we performed a validation on the same test dataset as considered in the former study. The test set is taken from an investigation of Cheng et al. [24], who carried out an exhaustive comparison of 16 popular scoring functions. The authors collected a diverse set of 195 protein-ligand complexes by applying the following protocol: The PDBbind database[2] [159, 160] (version 2007) was employed which contained more than 3000 binding affinities of receptor-ligand complexes collected from the PDB in the respective release. By applying filters regarding resolution (2.5 Å or better), knowledge of binding data (K_d or K_i value must be present), binding mode (only non-covalently attached ligands), structural symmetry (no multiple ligands in one structure), elements (no uncommon elements in the ligand) and molecular weight (ligand not heavier than 1000 Da) the set was refined to 1300 complexes. To remove redundancies, this set was clustered by using BLAST [2] with a similarity threshold of 90 %. 65 remaining clusters were obtained, each of which covering one specific target. Subsequently, the complex with highest affinity, lowest affinity and an affinity close to the mean were extracted. Thus, a total of $65 \cdot 3 = 195$ complexes was retrieved. In accordance to the work of Cheng et al. and the previous studies on DrugScore, this dataset will be called the *primary test set*. For each of the complexes a large number of highly diverse decoy poses was computed with

[2] `http://www.pdbbind.org`

respect to their RMSD by employing several popular docking programs, which can be used in the following for validation purposes.

Moreover, four additional test sets were compiled by Cheng et al. containing 112 HIV protease complexes, 73 trypsin complexes, 44 carbonic anhydrase complexes and 38 thrombin complexes. All of these complexes were taken from the refined dataset and were chosen as these four target classes are the highest populated ones.

6.2.6 Validation

In the works of Cheng et al. and Neudert and Klebe the quality of a scoring function is determined by means of *docking power*, *ranking power* and *scoring power*.

Docking power assesses the method's ability to correctly rank a given set of docking poses with respect to their RMSD to the known crystal structure. If the crystal structure is among the set of decoy poses itself, it should hence be ranked on top of the list.

Ranking power describes the ability to correctly rank a number of docking poses with respect to their experimentally measured binding affinities. In case of the primary test set described above, there are three complexes with measured affinities for each protein target. Thus, there are six possible ways to rank these structures, however, only one of them is correct.

Scoring power is finally the most challenging merit. It describes to what extent the computed scores for the three complexes of each protein target correlate with the experimentally measured binding affinities, which is achieved by calculating the Pearson correlation coefficient.

To activate the assessment of H-bond geometries an additional switch `-T4 w` was added to the DSX program, where w defines the weight of the H-bond score contribution to the final score. In the standard setting of the hitherto existing DSX version only the weight of the distance scoring (`T0`) is set to 1.0 and `T1` (= torsion angles), `T2` (= intramolecular clashes) and `T3` (=

SAS evaluation) are by default weighted with 0. We will call this program *DSX::Pair* in the following. Whenever the assessment of H-bond geometries is activated in our experiments, the weights of atom-pair distances and H-bonds are set to 1.0, unless otherwise stated. Accordingly, this version will be called *DSX::PairHb*. All H-bond potentials were used in version 02/15 and the remaining potentials (pair distances, torsion angles, SR) were used in version 05/11.

Docking Power

In a first experiment the newly developed H-bond geometry assessment was combined with the assessment of atom pair distances by performing a similar study as it was carried out in the mentioned studies of Cheng et al. (see Ref. 24, Fig. 3) and Neudert and Klebe (Ref. 109, Tab. 1). This trial assesses the scoring function's docking power by scoring all 195 receptors together with the ligand of the crystal structure as well as the generated decoy poses. An ordering of the complexes on the basis of the scores answers the following questions, the corresponding results are given in Table 6.2:

- Column "Top 1 rank" (left): How often is the crystal pose found on rank 1?
- Column "Top 5 ranks" (left): How often is the crystal pose found on rank 1–5?
- Column "Top 1 rank" (right): How often is a pose found on rank 1 that deviates at most 2 Å from the crystal pose?
- Column "Top 5 ranks" (right): How often is a pose found on rank 1–5 that deviates at most 2 Å from the crystal pose?

In the first part of the table, success rates of plenty of hitherto published scoring functions are shown, in the second part DSX results without the

	Crystal pose on		**≤ 2.0 Å pose on**	
Scoring function	Top 1 rank	Top 5 ranks	Top 1 rank	Top 5 ranks
DS::Jain	1.5	15.4	44.8	79.2
DS::LigScore2	17.9	49.7	71.6	92.9
DS::LUDI2	9.7	29.2	57.4	83.6
DS::PLP1	40.5	56.4	75.4	97.3
DS::PMF	19.5	44.1	43.7	67.2
GOLD::ASP	36.9	71.8	82.5	95.6
GOLD::ChemScore	17.9	50.8	70.5	86.9
GOLD::GoldScore	8.2	28.7	68.9	89.6
GlideScore::SP	18.5	50.3	73.2	93.4
SYBYL::F-Score	21.5	49.2	64.5	90.7
X-Score1.2	32.3	64.6	67.2	91.3
X-Score1.2::HMScore	30.3	57.9	68.3	90.7
DrugScore$^{\mathrm{CSD}}$::Pair	50.3	79.5	58.5	94.0
DrugScore$^{\mathrm{CSD}}$::PairSurf	44.6	80.0	54.1	95.6
DrugScore$^{\mathrm{PDB}}$::Pair	40.0	73.8	74.3	93.4
DrugScore$^{\mathrm{PDB}}$::PairSurf	39.5	74.9	74.3	95.1
DrugScore$^{\mathrm{PDB}}$::Surf	3.6	20.0	32.8	80.3
DSX$^{\mathrm{CSD}}$::Pair	50.8	77.4	83.6	95.6
DSX$^{\mathrm{CSD}}$::PairSR	51.3	79.0	84.7	96.2
DSX$^{\mathrm{CSD}}$::PairTors	52.3	77.4	84.2	95.1
DSX$^{\mathrm{CSD}}$::PairTorsSR	52.8	77.9	85.2	96.2
DSX$^{\mathrm{PDB}}$::Pair	50.3	78.5	84.2	95.6
DSX$^{\mathrm{PDB}}$::PairSR	51.8	77.9	84.7	95.6
DSX$^{\mathrm{CSD}}$::PairHb	51.9	81.4	84.7	96.2
DSX$^{\mathrm{CSD}}$::PairSRHb	50.3	80.3	84.7	96.2
DSX$^{\mathrm{CSD}}$::PairTorsHb	53.0	79.8	84.2	95.1
DSX$^{\mathrm{CSD}}$::PairTorsSRHb	53.6	80.3	85.3	96.2
DSX$^{\mathrm{PDB}}$::PairHb	51.9	80.9	84.7	95.6
DSX$^{\mathrm{PDB}}$::PairSRHb	53.0	80.3	84.7	95.6

Table 6.2 Overview of the success rates of several scoring methods in percent.

assessment of H-bond geometries are listed (all cited from the work of Neudert and Klebe [109]). They can be directly compared to the results of DSX that take also the angles into consideration, which are listed in the third part.

As in the former studies, the success rate S was calculated as

$$S[\%] = \frac{s-5}{195-5-7} \cdot 100 \tag{6.4}$$

where s is the total number of success cases. This adjustment is necessary since there are five complexes in the primary test set where all decoys have an RMSD ≤ 2.0 Å and another seven complexes where the RMSD is > 2.0 Å for all decoys.

The results of DSX do not change substantially, however, it becomes clear that the consideration of H-bond geometries is able to generally improve the results. The highest increase of the success rates can be obtained for the second column (crystal pose on top 5 ranks), which almost all raise to more than 80 % whenever H-bonds are involved. This points out the principally better assessment of the native pose with respect to the decoy poses.

Scoring Power

For the evaluation of scoring power the experiment of Neudert and Klebe [109] was repeated in which the Pearson correlation coefficients between the computed scores and experimentally measured affinities were calculated. This was done for the primary test set as well as for the four additional test sets. At first the calculations on the primary test set are analyzed, the results are summarized in Table 6.3. The predictive power of DSX::Pair and DSX::PairHb does hardly vary independently of the used potentials (PDB or CSD). Also in case of the additional data sets the involvement of H-bond potentials does not clearly cause an improvement of the correlation coefficients (Tab. 6.4). Although there are target proteins such as the HIV protease for which the evaluation of H-bond geometries causes a slight loss of accuracy, this effect can be by no means generalized. In case of thrombin, for instance, the H-

Scoring Function	**on original complex structures**	**on optimized complex structures**
DSX$^{\mathrm{PDB}}$::Pair	0.567	0.576
DSX$^{\mathrm{CSD}}$::Pair	0.597	0.608
DSX$^{\mathrm{PDB}}$::PairHb	0.567	0.575
DSX$^{\mathrm{CSD}}$::PairHb	0.596	0.607

Table 6.3 Pearson correlation coefficients between DSX scores and experimentally measured affinities for the primary test set.

Scoring Function	**HIV protease**	**trypsin**	**carbonic anhydrase**	**thrombin**
DSX$^{\mathrm{PDB}}$::Pair	0.196	0.754	0.463	0.672
DSX$^{\mathrm{CSD}}$::Pair	0.217	0.767	0.481	0.701
DSX$^{\mathrm{PDB}}$::PairHb	0.191	0.759	0.463	0.669
DSX$^{\mathrm{CSD}}$::PairHb	0.212	0.767	0.481	0.708

Table 6.4 Pearson correlation coefficients between DSX scores and experimentally measured affinities for the additional test sets.

bond evaluation improves the correlation to a similar extent when the CSD potentials are applied.

Ranking Power

The performance of our novel enhancement was finally evaluated with respect to ranking power. Using PDB potentials, the original version DSX::Pair orders 50.8 % of all complex triplets correctly with regard to their binding affinities. DSX::PairHb is able to improve this rate to 52.3 %. However, no improvement can be observed when using CSD potentials. Here, both versions achieve the same success rate of 49.2 %.

In this study, we want to refrain from another evaluation using a combined scoring function as this would need the application of weights of the individual scoring terms. In the most current study on DSX the examined combinations of scoring terms have not been applied with different weights either and only the values 0.0 and 1.0, respectively, were used [109]. An optimization will lead to a rather complex problem because five qualitatively rather different potentials are employed. We nevertheless attempted an optimization procedure of the used weights which is described in the next section.

Optimization of Scoring-term Weights

The optimization was approached by using the *SPOT* package for the program *R* [**?**]. This abbreviation stands for sequential parameter optimization toolbox which provides a framework for parameter tuning, based on modeling techniques, design of experiments (DoE) and statistical methods. The experiments on ranking power and on docking power were employed to optimize the five weighting parameters $w_{\{0...4\}}$, where SPOT was tasked to find the set of five weights $w_i \in [0, 1]$ which leads to the best result. Therefore, numerous optimization cycles were carried out, each of which contained seven different sets of parameter values. We chose seven parallel runs as the procedure was carried out on an eight-core CPU. The remaining core was used for the program R itself (the main process). For all calculations CSD potentials were used as no torsion potentials exist from the PDB database.

After a reasonable number of optimization cycles (689), the following weights were found to achieve the best result regarding the ranking power experiment: distances = 0.622, torsions = 0.742, intramolecular clashes = 0.173, SAS = 0.93, H-bonds = 0.392. With this parameter setting a success rate of 58.5 % is reached. In this case, the rate of correctly ordered triplets increases again by 6.2 % compared to the one reached by DSX::PairHb, which is as high as the best overall measured success rate in this experiment (X-Score::HSScore achieved 58.5 % too in the study of Cheng et al.) [24].

In the second experiment, which is about finding the true crystal poses within sets of decoys, SPOT proposed the weights distances = 0.676, torsions = 0.376, intramolecular clashes = 0.784, SAS = 0.649 and H-bonds = 0.17

after 647 optimization cycles. This setting leads to 56.8 % of all cases, where the crystal ligand was ranked on the top position, which is 3.2 % higher than the best hitherto obtained value achieved by DSX$^{\mathrm{CSD}}$::PairTorsSRHb. Using this weighting scheme to calculate all the success rates of this experiment – as already presented in Table 6.2 – leads to the results shown in Table 6.5. Here, also the best hitherto achieved results of the non-optimized version are shown for comparison.

	Crystal pose on		**≤ 2.0 Å pose on**	
Scoring function	Top 1 rank	Top 5 ranks	Top 1 rank	Top 5 ranks
DSX$^{\mathrm{CSD}}$::PairTorsSRHb	53.6	80.3	85.3	96.2
DSX$^{\mathrm{CSD}}$::All_Opt	56.8	80.9	86.3	96.2

Table 6.5 Results of DSX in the docking-power experiment using the optimized weights for all scoring terms (DSX$^{\mathrm{CSD}}$::All_Opt). The results are compared to the best hitherto obtained results which have been attained by DSX$^{\mathrm{CSD}}$::PairTorsSRHb.

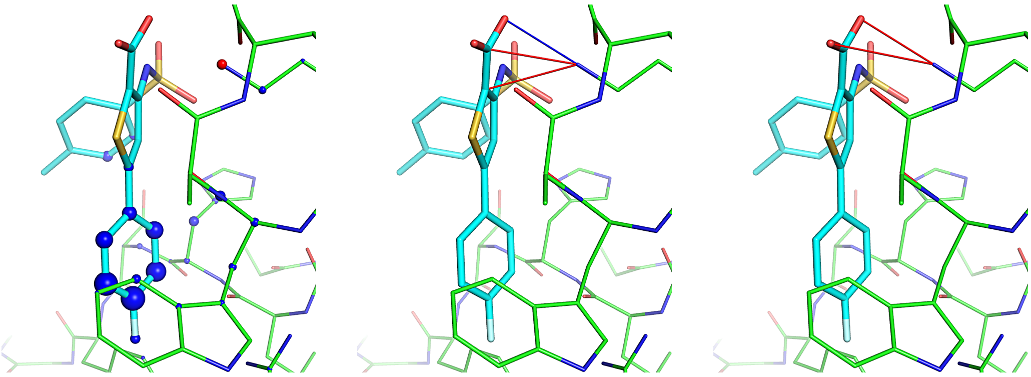
The two parameter settings that have been detected in the mentioned experiments are fairly dissimilar, which can be explained by the rather different tasks in both experiments. While the first challenge regards the estimation of binding affinities, the second is about finding native poses and near-native poses of the ligands. However, as there are clearly more results available for the docking-power experiment, it might be beneficial to use the weights proposed by the second optimization run.

Runtimes

Another very important point of the comparison of different scoring methods is the runtime. In the original DSX publication it was shown that the usage of specific potentials substantially increases the runtime of the program. Here, DSX::PairSR required a 4.8-fold longer time than DSX::Pair. We hence carried out a similar study to quantify the additional time that is needed by DSX::PairHb compared to DSX::Pair. We scored all 16 554 complexes of the primary test set, for which DSX::Pair needed 53.6 seconds. DSX::PairHb was only slightly slower, requiring a time of 62.5 seconds (+ 16.6 %), and is thus still applicable for any kind of large-scale studies.

Visualization

Aside from the extensions presented so far on the score calculation, a display of bad H-bond geometries was implemented according to the visualization in the previous version of DSX. By providing the switch `-v` in the program call a visualization file is automatically created which can be opened in PyMOL. So far, the size of all potentials on every considered atom are shown as red and blue spheres, as well as good or bad contact distances as red or blue lines (see Fig. 6.6 (a) and (b), respectively). Moreover, the visualization of bad torsion angles is offered which inspired us to provide a similar opportunity for H-bonds. In the new version of DSX also the H-bonds with a score over



(a) Values of the potentials are displayed as blue (= negative) and red (= positive) spheres.

(b) Preferred (= blue) and unfavored (= red) interaction distances.

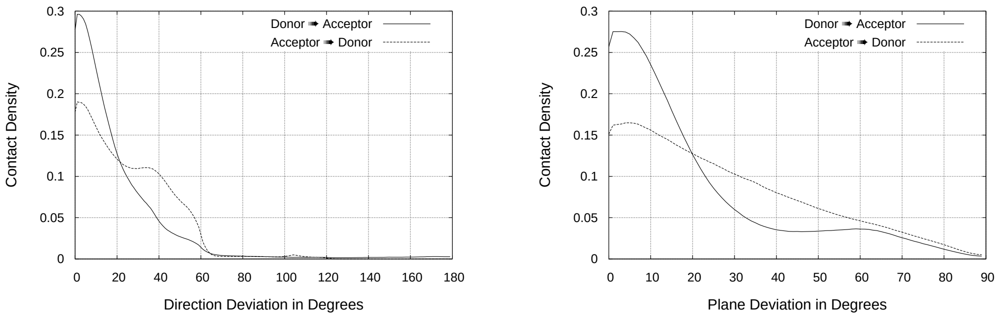
(c) Unfavored H-bond geometries.

Figure 6.6 Several parameters that can be visualized by DSX. Sizes of the potential values on every atom are displayed as blue and red spheres (a). Preferred and unfavored interaction distances are denoted as lines between receptor and ligand atoms (b). Unfavored H-bond geometries are displayed as red lines between receptor and ligand atoms (c). The images show a docking pose of the ligand `FIH` in hepatitis C virus RNA-dependent RNA polymerase (PDB: `2D3Z`).

a certain threshold can be shown as red lines which indicate bad interaction angles. In the present example, H-bonds with a score greater than 0.5 were displayed (Fig. 6.6 (c)).

Further Findings

Deriving potentials for the angles of H-bonds also allowed us to take a more generalized look on the preferences of certain atom type groups that participate in H-bond formation. More precisely, we calculated the average deviation densities for the set of all H-bond acceptor atoms as well as for all donor atoms that were found in the PDB complexes. In doing so, 868 distributions of acceptors were summarized that interact with a donor on the other side (containing 613 127 measured values), and 564 distributions of donors (550 068 values). As a result, we found that a higher deviation from the optimal geometry is much more unlikely on the donor side. This is shown in Figure 6.7 by a direct comparison of the two density distributions for direction and plane deviations of acceptors and donors. It can be clearly seen that the graphs representing the deviations which originate with donor atoms (solid lines) start considerably higher and drop with a steeper gradient in the following. A minor deviation from the optimal H-bond angle is hence clearly more often observed



(a) Direction deviations.

(b) Plane deviations.

Figure 6.7 Comparison of direction- and plane-deviation density distributions of the H-bond acceptors and donors.

for both the direction and the plane around donor atoms. This finding may be helpful for further improvements on the H-bond geometry assessment in DSX or general future studies on H-bonds.

Summary and Future Research

In this work, improved and accelerated methods for the comparison of protein binding sites are presented. Prior to the development of such approaches, however, the importance of detecting relevant and significant pockets automatically has been demonstrated (Chap. 2). It could be shown that the extraction of binding pockets in close proximity of the bound ligands makes comparisons trivial due to the inherent shape similarity, which is transferred from the ligands to the extracted pockets. Binding site comparisons are thus rather trivial, even in datasets that hardly contain redundancies in the sequence information. Applying this strategy in the pocket-extraction step, unoccupied pockets that are likely to contain important information about the putative accommodation of yet unknown ligands will remain unconsidered.

Subsequently, an extended graph-based model for enhanced similarity search in Cavbase was presented in Chapter 3. We proposed a novel and efficient modeling formalism that does not increase the size of the graph model used in Cavbase, but leads to graphs containing considerably more information assigned to the nodes. More specifically, additional descriptors considering surface characteristics are extracted from the local surface and attributed to the pseudocenters. Combined with a heuristic for the efficient detection of maximum common subgraphs, these properties are evaluated as additional node labels in the program LC, which leads to a gain of information and enables much faster but still very accurate comparisons between different structures. Moreover, the acceleration DivLC was discussed in Chapter 4 which makes use of graph partitioning. Therefore, graphs are divided into disjoint components prior to their comparisons. The pseudocenter sets are split with regard to their assigned physicochemical type, which leads to seven much smaller graphs than the original one. Applying this approach on the same test scenarios results in another significant speed-up without sacrificing accuracy. The graph partitioning approach only revealed weaknesses when small subpockets were used for the mutual comparisons.

In Chapter 5, a method for large-scale mining of similar protein binding pockets was introduced. A program called RAPMAD (RApid Pocket MAtching using Distances) was developed, which allows for ultra-fast similarity comparisons as protein binding sites are represented by sets of distance histograms that are both generated and compared with linear complexity. Thus RAPMAD attains a speed of more than 20 000 comparisons per second, which makes screenings across large datasets and even entire databases easily feasible. The practical use of the programs RAPMAD and LC was proven by a successful prospective virtual screening study that aimed at the identification of novel inhibitors of the NMDA receptor.

Finally, an extension of the program DSX, a scoring method for protein-ligand complexes, was introduced (Chap. 6). By adding the assessment of hydrogen-bond geometries an improvement of the program could be achieved along with only little increase in runtime. The extended version was tested on well-studied test datasets, which enables an exhaustive comparison with the previous version as well as a plethora of other hitherto developed approaches.

6.3 Future Work

Despite the improvements that have been presented in this work, there is still space for further enhancements. Firstly, I want to point out the algorithmic workflow of the methods LC and DivLC. During the generation of the product graph we chose a value of 2.0 Å for the parameter ϵ. This parameter defines the maximum distance between two nodes to be inserted as a new product node in the product graph. Thus, it is significantly responsible for the size and quality of the product graph. In our trials, we decided to use 2.0 Å as this matches the parameter setting in Cavbase. However, it is very likely that varying ϵ to smaller or greater values will also lead to different binding site comparison results. To our knowledge, this value has been fixed in the workflow of Cavbase without a real rational derivation. It is therefore reasonable to calculate the classification results for a broad range of different values, e. g. in the range of $[0.1 \ldots 3.0]$, in order to optimize the threshold ϵ.

The scoring function DSX holds remarkable potential for improvements as well. Although a moderate amendment of the scoring results could be obtained after implementing the evaluation of H-bond geometries, there are many more geometrical features that are reasonable to be assessed when scoring a receptor-ligand complex. For instance, also halogen bonds and the mutual positions of aromatic rings Π-stackings or edge-to-face configurations could be taken into consideration. It was shown that specific halogen bonds can even contribute as much to binding affinity as hydrogen bonds [68, 97]. Furthermore, Taylor has shown the high potential of especially iodine to form strong halogen bonds to nitrogen and oxygen by performing an exhaustive study of crystal structures in the CSD (see Ref. 145, Tab. 5). Lu et al. moreover proposed an ordering of the strengths of halogen bonds, which is $\mathrm{H} \cdots \mathrm{I} > \mathrm{Br} > \mathrm{Cl}$ [97].

Furthermore, examinations could be held that consider also water molecules in the scoring process. Due to the novel consideration of H-bonds this could lead to another performance gain of DSX. In addition, one could enhance the implementation of the H-bond scoring even further such that even water molecules in more remote locations are involved which are part of the so-called *second solvation shell*.

In the present study on DSX, we computed the positions of hydrogen atoms by using the VSEPR model, as the electron density does not disclose the positions of hydrogens properly in X-ray structures with a resolution well below 1.0 Å. However, future studies could exclusively process input structures for the generation of statistical potentials that already contain H positions. Aside from highly resolved X-ray structures, the data deposited in the CSD contains in many cases experimentally determined H atom positions. Moreover, structures that have been resolved by neutron diffraction comprehend well-determined H positions.

A

Lists of Datasets

A.0.1 Small Dataset of 141 ATP and 214 NAD Pockets

ATP

1a0i.4 1b76.4 1b8a.4 1c0g.4 1csn.1 1dej.6 1dv2.3 1dy3.1 1e2q.1 1e4g.1 1e8x.10
1esq.4 1esv.3 1f9a.10 1fmw.3 1g21.18 1g64.2 1gn8.1 1gol.3 1gtr.6 1gz3.7 1gz4.18
1h1v.4 1h1w.1 1hi1.6 1hlu.2 1i7l.4 1ijj.3 1j09.1 1j1z.10 1j7k.2 1jwa.2 1kax.2
1kay.1 1kaz.1 1kh2.6 1kj8.2 1kj9.2 1kp2.2 1kp3.2 1kvk.4 1kxp.7 1l2t.4 1lhr.7
1m83.3 1ma9.8 1mau.2 1miw.1 1n56.5 1n75.1 1nge.1 1ngf.1 1ngg.1 1ngh.1 1nsf.2
1nyr.9 1o93.4 1o9t.5 1obg.1 1ojl.10 1p8z.4 1pk8.22 1px2.6 1q97.8 1qhx.1 1qmz.7
1qrs.4 1qrt.3 1qru.4 1r8b.1 1rgi.7 1rys.4 1s9j.2 1tqp.1 1tyq.18 1u5r.1 1uev.1
1v1b.6 1w7a.12 1x01.2 1xdn.2 1xdp.14 1xef.5 1y8q.12 1yid.9 1yun.2 1zao.1 1zyd.3
2a3z.3 2a40.9 2a41.4 2a42.2 2aqx.2 2aru.1 2bek.1 2bup.1 2c96.2 2c9c.2 2cch.6
2cci.8 2cjm.8 2ddo.2 2dra.2 2dto.3 2dxt.3 2eww.1 2f02.4 2faq.4 2fgj.9 2gnk.1
2gwj.1 2hix.3 2hmp.3 2hmw.1 2i4o.8 2idx.6 2iyw.1 2j9c.3 2j9e.1 2npi.10 2ogx.5
2olq.1 2p55.2 2p9k.14 2p9s.15 2pbd.1 2phk.1 2pze.4 2pzf.2 2q0d.7 2q31.2 2q66.1
2q7g.1 2qb8.5 2qk4.3 2r6x.1 2r7l.1 2vhq.7 3c4w.10 3c4x.9 3cjc.6

NAD

1a9y.2 1a9z.1 1ahh.2 1ahi.2 1b14.3 1b8u.1 1bdb.1 1bmd.1 1bpw.4 1bw9.3 1bxg.7
1bxk.4 1c1d.2 1c1x.3 1cdo.1 1cer.5 1cw3.24 1cwu.5 1d1s.5 1d1t.5 1d4f.15 1d7o.2
1dbv.10 1deh.5 1dhr.1 1dhs.3 1dir.4 1dli.3 1dqs.6 1e3s.3 1e3w.4 1ee2.3 1efl.8
1ej2.1 1ek5.1 1eno.2 1eny.2 1enz.2 1ez4.9 1fdv.12 1fk8.4 1fmc.2 1gad.1 1gae.2

1geg.1 1geu.2 1giq.3 1gt2.3 1hdr.1 1hdx.5 1hdy.6 1hdz.6 1hex.1 1hku.1 1hl3.2
1hld.4 1hlp.4 1ht0.2 1htb.5 1hwy.10 1hzj.2 1i3k.2 1i3l.2 1i3n.2 1ib0.4 1ie3.5
1j0x.4 1j5p.3 1jq5.1 1ju9.5 1jvf.3 1jw7.3 1k0u.12 1k4m.5 1kvq.1 1kvr.1 1kvu.1
1kyq.1 1ldg.3 1ldy.4 1lj8.2 1lrj.2 1lrk.2 1lrl.2 1lsj.5 1lvl.2 1m76.3 1m8f.1
1m8g.1 1m8j.1 1m9h.1 1ma0.4 1mc5.5 1mg0.7 1ml3.7 1mp0.4 1mx3.2 1nah.1 1nai.2
1nff.2 1nfr.3 1nr5.5 1o6z.5 1o9j.17 1obb.6 1oc4.5 1og3.1 1ojs.2 1p45.4 1pj3.17
1psd.5 1qs2.4 1rlz.3 1roz.1 1rqd.4 1sby.3 1sg6.2 1t24.3 1t2d.3 1tae.9 1teh.1
1u5c.2 1u7h.1 1u8f.4 1uda.1 1udb.1 1udc.1 1uwk.5 1uwl.5 1uxj.3 1uxk.2 1uxt.2
1v59.5 1vbi.2 1vc2.1 1vi2.3 1vjp.4 1vjt.3 1vko.6 1vm6.4 1w1u.5 1wze.9 1x14.6
1x1t.1 1x7d.1 1x87.5 1xag.1 1xah.4 1xaj.3 1xal.3 1xel.1 1xwf.7 1yba.24 1yc2.4
1yl7.16 1ywg.4 1z2i.8 1z45.5 1zbq.14 1znq.7 1zrq.5 2a9k.1 2b36.10 2b69.1 2bkj.2
2c20.4 2c54.4 2c59.4 2c5a.3 2c5e.3 2d37.1 2d8a.1 2dc1.5 2dld.2 2dt5.1 2dvm.7
2ed4.1 2eer.2 2ep7.1 2fnz.7 2fr8.4 2fzw.4 2g76.4 2gag.11 2gah.11 2gwl.1 2h7l.2
2h7m.2 2hae.3 2hu2.2 2i9p.2 2ixa.1 2ixb.1 2nad.10 2o2s.7 2ome.11 2oxi.5 2pd3.3
2pd6.5 2pzm.2 2q1t.1 2q1u.1 2q1w.8 2qjo.8 2udp.2 3b6j.1 3bts.7 3dbv.10 3had.5
3hdh.9 3hud.3 4mdh.3 5mdh.3 6adh.5

A.0.2 Large Dataset of 420 ATP and 402 NAD Pockets

ATP

1a0i.4 1a82.1 1aq2.2 1asz.13 1atn.4 1atp.3 1ayl.3 1b38.1 1b39.3 1b76.4 1b8a.4
1bcp.25 1c0f.5 1c0g.4 1csn.1 1d4x.4 1dej.6 1dy3.1 1e24.3 1e2q.1 1e4g.1 1e8x.10
1ee1.3 1esv.3 1fin.3 1g5t.1 1g64.2 1gn8.1 1gol.3 1gtr.6 1gz3.19 1gz4.20 1h1v.4
1h1w.1 1h3e.3 1hck.3 1hlu.2 1i7l.4 1ijj.3 1j09.1 1j1z.10 1j21.3 1j7k.2 1jst.6
1jwa.2 1kax.2 1kay.1 1kaz.1 1kh2.7 1kj8.2 1kj9.2 1kmn.13 1kp2.2 1kp3.2 1kvk.4
1kxp.7 1lcu.4 1lhr.7 1lot.10 1m83.3 1ma9.8 1mau.2 1maw.15 1mb9.9 1mdu.6 1mjh.3
1n75.1 1n77.5 1nge.1 1ngf.1 1ngg.1 1ngh.1 1nlv.3 1nm1.3 1nmd.3 1nsf.2 1o93.4
1obd.2 1ol6.1 1os1.2 1p8z.4 1phk.1 1pj4.17 1pk8.24 1px2.6 1q24.5 1q97.8 1qhx.1
1ql6.1 1qmz.7 1qrs.4 1qrt.3 1qru.4 1qz5.2 1qz6.2 1r0x.5 1r0z.3 1r10.6 1rdq.3
1rdw.2 1rfq.9 1rgi.7 1s9j.2 1svm.17 1t44.2 1tf7.29 1tid.8 1til.7 1tqp.1 1tyq.18
1u5r.1 1u9i.23 1ua2.11 1uf9.3 1v1b.9 1v3s.2 1vc9.1 1vjc.1 1vjd.1 1wkl.2 1wua.2
1x01.2 1xdn.2 1xf9.2 1xfa.2 1xkv.6 1xmi.4 1xmj.1 1y64.7 1y8p.5 1yag.4 1yfr.6
1yid.9 1yp3.11 1ytm.6 1yun.2 1yvn.3 1yxq.1 1z0s.5 1zao.1 1zfn.4 1zp9.3 1zyd.3
2a3z.3 2a40.9 2a41.4 2a42.2 2a5y.10 2a84.2 2aqx.2 2aru.1 2bbo.2 2bek.1 2biy.3
2btf.2 2bu2.5 2bup.1 2c8v.2 2c96.2 2c9c.2 2cch.6 2cci.10 2cjm.8 2cv1.4 2d1k.3
2ddo.2 2dto.3 2dxi.3 2dxt.3 2e5y.2 2e89.9 2eww.4 2f02.4 2ff3.2 2ff6.3 2fgj.9
2fxu.3 2gbl.29 2gnk.1 2gwj.1 2gwk.1 2hf4.2 2hix.3 2hmp.3 2hmu.1 2hmw.1 2hs0.1
2i4o.8 2idx.6 2ivp.1 2iyw.1 2j9c.3 2j9e.3 2jax.3 2npi.10 2nt8.1 2o0h.5 2oan.4
2ogx.5 2olq.1 2olr.2 2ooy.14 2p55.2 2paa.4 2pav.2 2pbd.1 2pc9.11 2phk.1 2pxz.2
2q0d.7 2q0u.2 2q31.2 2q36.2 2q7g.1 2q97.5 2qk4.3 2qkm.13 2qrd.12 2qui.5 2r6x.2
2r7l.1 2r86.4 2r9v.2 2rd5.1 2v51.1 2v52.2 2vcp.1 2vhq.7 2vt3.2 2x15.1 2xbp.1
2xcw.6 2xjd.3 2xje.5 2xti.11 2xul.5 2xzw.4 2y27.6 2ych.2 2yj4.2 2yj5.7 2yje.2
2yjf.10 2yw2.3 2yww.3 2yxu.7 2z02.2 2z08.1 2z1u.2 2zan.1 2zdq.3 2zhz.3 2zsf.4
3a5m.4 3a5n.3 3a5o.4 3a8t.1 3a8w.3 3ab8.7 3am1.1 3amt.1 3aqn.1 3att.2 3b2q.7
3blq.7 3bu5.2 3buz.1 3c16.5 3c4w.10 3c4x.9 3c9r.3 3chw.3 3ci1.2 3ci4.3 3ci5.5
3cip.3 3cjc.6 3cqd.2 3crc.2 3daw.3 3dkc.3 3dnt.6 3dv3.1 3dvl.22 3dy7.2 3e7e.4
3e8n.1 3ehg.1 3eks.2 3eku.2 3el2.2 3eps.12 3eqb.1 3eth.1 3ffk.13 3fjq.2 3fkq.1
3g59.2 3gah.3 3gaj.2 3gni.3 3gqk.3 3h1q.3 3h39.4 3h8v.2 3hav.4 3hbt.1 3hmn.2
3hne.9 3hqo.20 3hrc.1 3hrf.1 3hy2.3 3hzi.3 3i7v.1 3ibq.1 3ie7.2 3ikh.3 3inn.1
3iq0.2 3jzm.18 3k09.37 3k0a.17 3k0c.20 3k0e.30 3k0f.16 3k5h.10 3ke5.4 3keu.7
3kmw.3 3lcb.10 3ldl.1 3lev.7 3lf0.3 3lfz.2 3lki.5 3lkk.1 3ll3.6 3ll5.6 3lmi.10
3lqq.12 3lqr.12 3lrr.1 3lss.7 3m0e.18 3m1f.2 3m6g.3 3mey.1 3mhy.2 3mmv.2 3mn5.2
3mn6.7 3mn7.2 3mn9.2 3na3.3 3ncq.1 3nem.6 3o8l.17 3os3.1 3ovb.3 3pgk.6 3pp1.1

3q53.1 3q60.3 3qal.6 3qam.4 3qo7.3 3qun.2 3quo.3 3qxc.2 3r1r.9 3r5f.3 3r5x.4 3rep.3 3reu.4 3rgl.3 3rk1.1 3rrf.1 3rtc.1 3rte.1 3rtg.2 3s1a.23 3s3t.7 3se7.5 3sez.26 3si7.7 3sjh.2 3sl2.1 3t0z.1 3t54.1 3t8o.4 3ta0.3 3ta2.3 3tlx.7 3tpq.6 3tut.2 3tux.3 3tw3.4 3u4l.2 3u8x.3 3u9d.3 3ub5.2 3ufg.3 3v01.1 3v04.1 3v2u.17 3vh4.3 3zs7.1 4a2a.6 4aff.1 4at1.2 4dh1.1 4dh3.3 4din.7 4dug.32 4dw1.5 4dxl.1 4ed4.1 4en4.7 4erp.23

NAD

1gy8.13 1hku.1 1hl3.2 1iy8.4 1li4.4 1llu.8 1lsj.5 1lso.1 1m75.2 1m76.3 1m8f.1 1m8g.1 1m8j.1 1m8k.1 1m9h.1 1mew.2 1mfp.2 1mg5.2 1mi3.8 1nbo.1 1nfb.8 1nm5.5 1nr5.5 1nvm.17 1nzz.27 1o02.30 1oc2.5 1oc4.5 1og3.1 1ojz.1 1omo.2 1p44.9 1p45.4 1p9l.3 1pj3.17 1pjl.14 1pjs.2 1pl6.12 1pl8.11 1qv6.5 1qv7.5 1r37.5 1r66.2 1r6d.1 1rkx.7 1rlz.3 1roz.1 1rqd.4 1rwb.1 1s20.12 1sb8.3 1sb9.1 1sby.3 1sg6.2 1sm9.7 1sow.4 1t24.3 1t2d.3 1t2f.5 1t90.7 1u1i.11 1u3t.4 1u3u.4 1u3v.4 1u3w.2 1u5c.2 1u7h.1 1u7t.7 1u8x.4 1uh5.5 1up6.22 1up7.17 1ur5.1 1uwk.5 1uwl.5 1uxg.3 1uxh.2 1uxi.5 1uxj.3 1uxk.3 1uxt.2 1v9l.2 1vbi.2 1vc2.1 1vi2.3 1vko.6 1vm6.4 1vrw.5 1w1u.5 1wdk.17 1wnb.14 1wpq.5 1wwk.3 1x0x.1 1x15.5 1x1t.1 1x7d.1 1x87.5 1xag.1 1xah.4 1xaj.3 1xal.3 1xcb.1 1xlt.18 1ye4.7 1ye6.8 1yl7.16 1z0z.5 1z2i.8 1z45.5 1z9a.6 1zbq.17 1zjz.1 1zk1.1 1zsn.3 1zw1.5 1zxb.6 1zxl.2 2a5f.4 2a9k.1 2ag5.5 2aq8.2 2aqh.2 2aqi.2 2aqk.2 2b35.10 2b36.10 2b37.7 2b69.1 2bhp.5 2bi4.3 2bjk.4 2bl4.2 2c20.4 2c54.4 2c59.4 2c5a.3 2c5e.3 2c8f.4 2cfc.5 2cnb.16 2czc.9 2d1y.10 2d37.1 2d4v.13 2d8a.1 2dc1.5 2dfv.6 2dph.6 2dt5.1 2dvm.7 2eer.2 2ehu.6 2eii.6 2eit.6 2ejv.3 2ekl.1 2ekp.1 2ep7.4 2ewm.2 2fkn.14 2fm3.7 2fn7.7 2fnz.7 2foi.5 2g5c.3 2g76.4 2g82.14 2gdz.2 2gsd.5 2gwl.1 2h4f.1 2h4h.1 2h5l.21 2h7i.2 2h7l.2 2h7m.2 2h7n.2 2h7p.2 2hae.3 2hu2.2 2hun.3 2i29.2 2i2f.2 2i9p.4 2ixa.1 2ixb.1 2izz.7 2j40.6 2j5n.6 2jhf.6 2jhg.4 2jjy.6 2nsd.1 2o23.4 2o2s.7 2o2z.4 2o4c.2 2ol4.2 2onm.44 2oos.4 2op0.3 2op1.2 2p5u.6 2p5y.1 2pd3.3 2pd4.3 2pla.4 2pzj.2 2pzk.2 2pzl.2 2pzm.2 2q1t.1 2q1u.1 2q1w.8 2q2q.12 2q2v.7 2qg4.17 2qio.5 2v7g.6 2v7p.4 2vut.24 2w2l.5 2wdz.6 2wn7.2 2wsb.5 2wyv.4 2wyw.1 2x0i.2 2x0n.10 2x0r.4 2x22.2 2xaa.1 2xxj.5 2y42.10 2y42.7 2y99.5 2yvf.1 2yvg.2 2yy7.2 2zit.25 2ziz.4 2zj1.5 2ztl.4 2ztm.6 2ztu.5 2ztv.1 3a1n.2 3a28.13 3a4v.1 3a9w.3 3abi.1 3ajr.4 3am3.4 3am4.5 3am5.4 3aw9.3 3b1j.2 3b20.6 3b4w.1 3b6j.1 3b78.26 3b82.26 3b8h.27 3bts.7 3c7c.3 3c7d.2 3cea.4 3cin.4 3clh.5 3cps.3 3czm.4 3d4p.2 3e18.7 3ec7.1 3ehe.3 3enk.3 3f3s.7 3f4b.13 3fg0.6 3flk.6 3fmx.2 3fne.5 3fnf.6 3fng.2 3fnh.2 3geg.3 3gfb.2 3ggg.6 3ggp.2 3h3j.2 3h9e.1 3h9u.16 3i0p.2 3i9k.2 3icp.1 3iqd.2 3iwk.37 3jsa.2 3ju8.9 3jv7.1 3jyo.1 3jyp.1 3jyq.2 3jzd.3 3k2b.12 3kb6.3 3keq.2

3ko8.1 3l0d.1 3ln3.2 3lqf.3 3lsy.6 3lt0.4 3lt1.5 3lt2.4 3lt4.4 3lu1.6 3m2t.5 3m6i.12 3n7u.32 3nj4.10 3nj8.4 3nrc.4 3nt2.6 3nt5.3 3ntq.3 3ntr.3 3oet.14 3oew.2 3oey.2 3of2.2 3oig.1 3om9.13 3orf.2 3ox4.4 3p2o.3 3pfw.1 3pgx.4 3pjd.4 3pje.3 3pjf.4 3pqd.20 3pqf.15 3pxx.7 3q2i.4 3q2k.8 3q3c.1 3q6i.11 3q9o.7 3qj5.5 3qv1.12 3qvs.3 3qvw.3 3qvx.3 3qw2.8 3rf7.1 3riy.6 3riy.7 3rj5.2 3rj9.6 3ru7.5 3ru9.4 3rua.3 3ruc.7 3rud.4 3rue.7 3ruf.4 3ruh.5 3rvd.12 3s2e.11 3s2f.16 3s2g.12 3s2i.7 3sx2.10 3sxp.22 3syt.14 3t4e.4 3tnl.6 3toz.5 3tsc.10 3tsc.8 3u31.1 3uic.14 3uq8.2 3uwr.9 3uxy.6 3v9l.8 3vdq.7 3vdr.6 3vps.1 3zv5.3 3zv6.2 4a0m.6 4e5k.8 4e5n.8 4e5p.11 4ebf.13 4ef7.3 4egb.16 4f3x.12 4f7i.6

A.0.3 Dataset of 429 Pockets Binding FAD Non-covalently and 114 Pockets Hosting FAD Covalently

FAD non-covalently bound

1bqe.1 1c0i.1 1c0k.2 1c0l.2 1c0p.2 1d7y.2 1e62.1 1e63.1 1e64.1 1ep2.6 1f20.3
1f3p.4 1fl2.1 1gaq.3 1gaw.1 1gg5.3 1gjr.1 1go2.1 1gr1.1 1gv4.8 1gvh.2 1h66.4
1h69.6 1h81.12 1h82.11 1h83.13 1hsk.1 1i7p.4 1i8t.3 1ib0.4 1ijh.2 1iqr.2 1iqu.2
1jb9.1 1jeh.7 1jnr.12 1jqi.7 1jr8.1 1ju2.4 1k0i.2 1k0j.2 1k0l.1 1k87.8 1kbo.3
1kbq.3 1knp.5 1knr.5 1krh.6 1mo9.11 1mok.21 1n1p.1 1n4v.2 1n4w.1 1ng3.4 1ng4.3
1np7.8 1ogi.1 1ogj.1 1oqc.2 1owm.3 1own.3 1pn0.15 1qgy.1 1qgz.1 1qh0.1 1r2j.1
1rm6.18 1rp4.4 1rq1.4 1rx0.14 1ryi.5 1sb3.15 1sg0.1 1siq.3 1sir.3 1sm4.2 1tez.10
1tiw.3 1tt0.7 1u3c.5 1u3d.6 1u8v.9 1udy.9 1ukw.7 1umk.2 1v5e.5 1v5f.5 1v5g.4
1v93.3 1ve9.8 1w34.1 1w35.1 1xdi.6 1xhc.2 1xi2.1 1ybh.5 1yhy.5 1yhz.5 1yi0.5
1yi1.5 1ykj.5 1yoa.2 1z8n.5 1zmc.13 1zmd.9 1zp3.5 1zp4.6 1zx1.1 2a87.3 2apg.8
2aqj.6 2ar8.6 2b3d.2 2b5o.4 2b9w.2 2b9x.2 2b9y.4 2ba9.2 2bab.4 2bac.2 2bgi.1
2bmw.1 2bsa.1 2bzs.1 2c12.13 2c3c.11 2c7g.2 2cul.1 2cvj.1 2cz8.3 2d29.6 2dji.4
2du8.16 2dvl.4 2e0i.6 2eba.19 2eq6.6 2eq8.14 2eq9.24 2f1o.8 2f5z.22 2fg9.1 2fja.14
2fje.13 2fmn.3 2fmo.2 2g37.3 2gew.2 2gj3.2 2gmj.10 2gpj.1 2gqf.2 2gqt.1 2gqu.1
2gqw.3 2gr0.3 2gr1.3 2gr2.2 2gr3.2 2h94.6 2hj3.1 2hq9.1 2hti.1 2i0k.4 2i0z.2
2ijg.2 2iw5.7 2ix5.7 2ix6.10 2j09.2 2j4d.6 2jkc.7 2oal.7 2ok8.6 2oln.3 2pan.22
2pd7.2 2pd8.3 2pg0.5 2pgn.13 2qa1.5 2qcu.6 2qdx.2 2qmy.2 2qmz.2 2qwx.2 2qx4.2
2qx6.2 2qx8.2 2qx9.1 2r0m.3 2r0n.3 2r45.8 2r46.8 2r4e.9 2r4j.7 2r6h.4 2rc5.10
2rc6.12 2rgh.8 2rgj.2 2rgo.9 2uxw.4 2v1d.6 2v5z.11 2vig.16 2vnh.2 2vni.1 2vnj.1
2vnk.2 2vou.5 2vtb.22 2vvl.26 2vvm.8 2vyq.1 2vzl.1 2wb2.5 2wes.12 2wet.11 2wp5.6
2wpc.8 2wpe.9 2wpf.8 2wq6.4 2wq7.3 2ws3.29 2wsi.1 2x0l.7 2x3n.2 2x3u.1 2xnc.4
2xnj.3 2xry.6 2xrz.8 2y48.6 2y6r.5 2yau.3 2yg3.10 2yg4.9 2yg5.4 2yg6.11 2yg7.5
2yvj.9 2ywl.2 2yyi.3 2yyj.2 2yyl.1 2yym.2 2z5y.6 2zbw.2 3ab1.3 3ah5.3 3ahq.1
3ahr.1 3alj.2 3alk.2 3all.6 3awi.27 3axb.4 3b96.5 3c4a.1 3c96.2 3cir.24 3cnj.2
3crz.1 3cvu.3 3cvv.3 3cvw.6 3cvx.2 3cvy.3 3d1c.1 3d72.4 3da1.4 3djd.2 3djl.6
3e1t.1 3e2q.3 3e2r.4 3e2s.5 3ef6.4 3es9.9 3ewk.1 3f8d.3 3fbs.3 3fim.1 3fjo.5
3fpk.6 3fst.5 3fsu.2 3fy4.9 3g5m.1 3g5q.5 3g5r.4 3g5s.4 3g6k.11 3gam.1 3gdn.3
3gdp.2 3gwd.4 3gwl.2 3gwn.2 3gyi.2 3gyj.2 3gyx.39 3hdq.10 3hdy.12 3he3.13 3hji.2
3hjk.1 3i99.2 3if9.8 3ihg.10 3is2.3 3jqr.1 3jsx.5 3k7m.3 3k7q.3 3k7t.5 3ka7.4
3kpf.6 3l1r.7 3lcm.6 3lli.1 3lo8.1 3lvb.1 3lzw.2 3lzx.3 3m0o.1 3m12.1 3m13.12
3m31.3 3mbg.3 3mhp.4 3mj4.16 3mpi.10 3mpj.13 3nf4.9 3nfr.1 3ng7.3 3ngc.4 3nh3.3
3nhf.1 3nhj.2 3nhk.1 3nhl.2 3nho.3 3nhr.2 3nhs.1 3nhu.1 3nhw.1 3nhy.2 3nix.9

3nk0.3 3nk1.3 3nk2.3 3nlc.2 3nn0.3 3nvj.5 3nyc.2 3nye.2 3nyf.2 3o2n.2 3o55.1 3o73.2 3ovm.2 3owh.1 3owx.2 3ox1.1 3ox2.2 3ox3.2 3p0k.3 3p4p.22 3p4q.23 3p4r.21 3p4s.19 3pl8.6 3pnd.2 3ps9.7 3pvc.7 3q6j.8 3qfs.3 3qft.2 3qj4.5 3qse.7 3qvp.1 3qvr.1 3qzy.2 3r7c.5 3r9u.2 3red.24 3rnm.7 3rp6.1 3rp7.1 3rp8.1 3sgl.7 3szc.4 3t2z.7 3t58.22 3t59.17 3te7.1 3tem.1 3tx1.1 3tzb.3 3u2l.1 3u2m.1 3u33.41 3ukk.17 3ukl.32 3ukp.32 3ukq.5 3umv.9 3uxe.1 3uxh.2 3zxs.12 4af7.1 4at0.1 4at2.3 4dna.4 4dqk.4 4dql.2 4dsg.4 4eh1.2 4f8y.5 4fdn.3 4fdo.2 4fdp.6 4feh.2 4ff6.4 4fk8.5

FAD covalently bound

1e7p.49 1el5.2 1el7.1 1el8.1 1el9.2 1eli.2 1i19.7 1l9d.4 1l9e.1 1nek.10 1nen.11 1oj9.10 1oja.12 1ojd.53 1w1j.7 1w1k.8 1w1l.8 1w1m.12 1w1r.3 1w1s.3 1yq3.7 1yq4.8 1zov.6 1zr6.2 2acz.8 2axr.2 2bk5.11 2c64.13 2c67.12 2c70.11 2ekg.1 2f5v.7 2f6c.7 2gb0.1 2gf3.1 2gf3.2 2h88.16 2q6u.3 2qkn.4 2qpm.4 2v60.10 2v61.11 2vfr.2 2vfs.2 2vft.2 2vfu.3 2wdq.22 2wdr.22 2wdv.26 2wp9.29 2wqy.14 2wu2.25 2wu5.24 2xfn.13 2xfp.10 2yvf.1 2yvg.2 2z5x.7 3abv.11 3ae1.9 3ae2.8 3ae3.10 3ae4.8 3ae5.10 3ae6.7 3ae7.10 3ae8.10 3ae9.9 3aea.9 3aeb.9 3aec.7 3aed.9 3aee.9 3aef.9 3aeg.9 3bhf.6 3bhf.7 3bhk.2 3d2d.2 3d2j.3 3dje.4 3dq0.5 3fdy.3 3fw7.3 3fw8.3 3fw9.3 3fwa.3 3gsy.3 3hsu.2 3i3l.3 3js8.3 3k4b.5 3k4j.4 3kjm.2 3lsk.9 3nne.24 3po7.8 3qsm.6 3qss.6 3rh8.1 3rj8.2 3rja.3 3s1c.4 3s1d.3 3s1e.4 3s1f.4 3sfd.9 3sfe.10 3vr8.18 3vrb.19 3zyx.10 4a79.10 4a7a.10 4ec3.3

A.0.4 Culled Dataset of ATP-, NAD- and FAD-hosting Structures

ATP

1a0i 1kvk 1yfr 2e89 2qui 3a8t 3gah 3ll3 3rgl 1bcp 1mau 1yp3 2fxu 2r7l 3amt 3gni
3ll5 3rk1 1csn 1mb9 1z0s 2gbl 2r9v 3att 3gqk 3lmi 3rtg 1dy3 1nsf 1zao 2hix 2vhq
3c16 3h39 3lrr 3sez 1e2q 1obd 1zfn 2hmu 2vt3 3c9r 3h8v 3lss 3sl2 1e8x 1pk8 2a5y
2hs0 2xcw 3cqd 3hav 3mhy 3t54 1ee1 1qhx 2a84 2ivp 2y27 3crc 3hy2 3na3 3tlx 1gn8
1r0x 2aqx 2iyw 2ych 3dkc 3i7v 3nem 3tut 1gtr 1rdq 2aru 2o0h 2yj4 3dnt 3ibq 3os3
3v2u 1gz4 1s9j 2bek 2ogx 2yw2 3e7e 3ie7 3ovb 3vh4 1j09 1svm 2bup 2olr 2yww 3ehg
3ikh 3q60 3zs7 1j7k 1uf9 2c8v 2q0d 2z08 3eps 3inn 3quo 4a2a 1kj9 1vjd 2c96 2q7g
2z1u 3eth 3iq0 3qxc 4aff 1kmn 1wkl 2dto 2qkm 2zan 3fkq 3lev 3r1r 4dw1 1kp2 1xdn
2e5y 2qrd 2zsf 3g59 3lfz 3r5x 4ed4

NAD

1m8g 1sby 1wpq 2bjk 2ixa 3ajr 3i9k 3p2o 3vdq 1mew 1sg6 1x15 2d37 2izz 3b6j 3jsa
3pjf 1nvm 1t2d 1x7d 2d4v 2jhf 3bts 3jyo 3q3c 1og3 1up7 1z0z 2dt5 2nsd 3c7d 3ln3
3q9o 1pjs 1uwk 1z45 2dvm 2o2s 3cea 3m2t 3qvx 1pl8 1uxg 1zjz 2ekp 2o2z 3cin 3nrc
3rf7 1rkx 1v9l 2a5f 2g76 2qg4 3f3s 3nt2 3riy 1rlz 1vbi 2a9k 2g82 2vut 3ggg 3orf
3syt 1s20 1vm6 2b69 2i2f 2yvf 3h9u 3ox4 3uq8

FAD

1c0p 1rm6 2c12 2gpj 2qa1 3ah5 3g5s 3nlc 3tem 1ep2 1ryi 2cul 2gqt 2qcu 3axb 3g6k
3nyc 3zxs 1gvh 1tez 2cz8 2gqw 2qdx 3d1c 3gwl 3p0k 4dna 1jr8 1u8v 2dji 2hq9 2uxw
3djl 3gwn 3pnd 4feh 1ju2 1umk 2ed4 2ijg 2v5z 3e2q 3lli 3qj4 1n4w 2aqj 2fg9 2oln
2xry 3f8d 3lo8 3qvp 1r2j 2b9w 2gj3 2pgn 2yyj 3fst 3m31 3rp8

A.0.5 EC Dataset Containing 502 Pockets which Spreads Over All 6 Main EC Classes

1.1.1.21

1us0.1 2pfh.1 2iq0.2 1el3.2 1ads.2 1x96.1 1x97.1 2i17.1 2hv5.2 2inz.2 2ine.2 1t41.2 1x98.2 2ikg.1 1z8a.1 2f2k.1 2nvc.2 2i16.1 1pwl.2 2acq.2 2duz.1 1pwm.2 2ikj.2 2fzd.1 2nvd.2 2hvo.2 2iki.1 2acr.2 2agt.1 2dux.1 1t40.2 2iqd.2 1z3n.1 2ikh.1 2fzb.2 2hvn.1 2ipw.2 2is7.2 2pf8.1 2fz8.1 2fz9.2 2isf.2 2acs.2 2pev.1 1az1.2 2j8t.1 1z89.1 2dv0.2 2acu.2

1.1.1.62

1a27.1 3hb5.2 3klm.2 1i5r.1 1qyx.1 3dhe.1 3hb4.1 1qyw.1 1jtv.1 1fdt.2 3dey.2 1bhs.3 1fds.1 1dht.1 1iol.1 1qyv.1

1.14.13.2

1pbe.2 1pdh.2 1cj3.2 1pbb.2 1iuu.2 1iut.2 1cc6.2 1ykj.5 1iuv.3 1pbd.2 1iux.3 1cc4.3 1bkw.2 1dod.3 1doe.3 1k0l.1 1k0i.2 1doc.4 1dob.3 1k0j.2 1iuw.1 1ius.2 1pxc.3 1pxb.3 1bf3.2 1bgn.2 1cj4.2 1d7l.1 1phh.1 1pxa.4

2.7.1.37

1urw.1 2clx.3 2b53.1 1h1r.8 1oir.2 1oiu.6 1h08.2 2b52.2 2cch.6 2btr.2 2exm.1 2b54.2 2bts.4 1ke6.1 1pye.4 2c68.1 2b55.1 1ke9.1 1pxi.1 2c6i.2 2c6k.1 1ke8.1 1e1x.2 1e1v.2 1y8y.2 1ke7.1 1aq1.1 2fvd.1 1hcl.2 1pw2.2 1h0v.2 2bhe.1 1h00.3 1r78.2 1ykr.3 1gz8.1 1pxo.3 1hck.3 1h07.1 1b38.1 1jvp.2 1h01.1 1oit.2 2c6m.1 1h1s.6 2a0c.4

2.7.1.112

2src.3 1yi6.2 2ptk.1 1y57.2 1yol.1 2bdj.1 1fmk.1 2bdf.4

2.7.4.9

1nn5.1 1e2q.1 1e9e.1 1e2f.1 1e9d.1 1e99.1 1e2g.1 1nmz.1 1e98.1 1nn0.2 1e9f.1 1e9c.1 1e2d.1 1e9b.1 1nmx.1 1e2e.1 1nn3.1 1nmy.1 1nn1.1 1e9a.1

3.4.21.5

2cn0.2 1c5l.3 1oyt.2 1xmn.16 1ype.2 1ypl.2 1doj.1 1a4w.3 1ypm.2 1z71.4 2jh0.2

1qbv.2 1c1u.4 1gj4.3 2bvr.2 2bvs.2 2cf8.2 1o5g.2 1nm6.2 1shh.6 1ypg.2 1ghx.3 1ghy.2 1ba8.2 1vzq.3 1h8i.4 2cf9.2 1a3b.2 1tom.2 1c5n.3 1sl3.3 1w7g.2 2bdy.2 1ghv.3 1o2g.2 1aht.2 1ai8.2 1eb1.3 1ghw.3 1gj5.2 1ypj.2

3.4.23.16

2hs1.1 3b7v.1 2aoc.1 3a2o.1 1sdu.1 2ien.1 1sdt.1 2hs2.1 3djk.1 3cyx.1 2f8g.1 2avq.1 2idw.1 3d1y.1 2nnp.1 2nmy.1 3dk1.1 2avs.1 2a1e.1 2f80.1 1sdv.1 2qd7.1 2qd6.1 2aod.1 3d20.1 3bva.1 3b80.1 3i6o.1 1k1t.1 2avm.1 2aoh.1 2nmz.1 2aof.1 3bvb.1 2aog.1 2qci.1 3d1x.1 2pk6.1 3cyw.1 2nnk.1 2avo.1 2nmw.1 3fx5.1 2f81.1 2aoi.1 3k4v.3 2qd8.1 3d1z.1

3.4.24.86

2ddf.3 2oi0.2 3ewj.2 3edz.3 2a8h.1 3l0t.4 2i47.9 1bkc.10 2ddf.1 2fv5.4 3l0v.2 3lgp.1 2fv9.3 3g42.12 1zxc.2 3b92.2

4.2.1.1

2fou.1 1zfq.2 1zge.1 2cbd.1 2hl4.1 2ili.1 2nno.1 1fr7.1 1zgf.1 2cbb.1 2ax2.2 2hd6.1 1cam.1 1ray.1 2pou.1 2nns.1 2nnv.1 2cba.1 2fov.2 1yo1.2 1zh9.1 2gd8.2 2aw1.1 2foq.1 1lg5.1 1fr4.1 1tg3.1 2f14.1 1g52.1 1xq0.1 2hnc.1 2pow.1 2nng.1 1cil.1 1mua.1 1rzb.1 1thk.1 1th9.1 1moo.1 1g0e.1 2eu2.1 2eu3.1 2fos.1 2fmz.1 1g0f.1 1oq5.1 1zfk.1 2fmg.1 1z9y.1

5.3.1.5

1xim.12 9xim.12 3xin.12 4xim.10 6xim.12 1xim.13 3xim.12 8xim.13 7xim.12 2xim.12 2xin.11 5xin.11 1xin.11

5.4.2.1

1bq4.12 5pgm.22 4pgm.10 1bq4.15 1qhf.5

1.1.1.42

1ai3.3 8icd.3 1p8f.3 1idf.3 9icd.3 1ide.2 1cw4.3 1cw1.4 1idc.3 1bl5.2 1gro.2 4icd.1 7icd.1 5icd.1 1ai2.4 3icd.2 1grp.2 1pb1.3 1hj6.2 1iso.3 1pb3.4

1.1.1.21

1ah4.2 1eko.2 1ah3.2 1ah0.2

2.7.1.112

1iep.5 2g1t.9 2g2h.6 1fpu.4 2hzi.8 2f4j.2 2hyy.5 2gqg.3 1opk.5 1opj.4 2fo0.3 2hiw.4

2.7.4.9

1n5l.3 1mrn.1 1n5j.1 1w2h.4 1n5k.3 1gsi.1 1g3u.1 1w2g.4 1gtv.1 1mrs.1 1n5i.1 3tmk.17 1tmk.3 1tmk.2 2tmk.4

1.1.1.21

1ye6.11 1jez.7 1mi3.8 1mi3.9 1z9a.6 1ye4.9 1r38.9 1sm9.10 1k8c.7

4.2.1.1

1bzm.2 2foy.1 1hcb.2 2it4.4 2nn7.1 2fw4.4 3lxe.4 1crm.2 1hug.1 1azm.2 1jv0.5 1czm.2 2nn1.1 2cab.1 2nmx.1 1huh.2 3fw3.3 3f7u.3 3f7b.3 1z93.1 1z97.2

4.1.1.23

3gdr.5 3gdk.5 1dqx.3 3gdl.2 3gdm.3 1dqw.3 3gdt.6 3ewz.5 3bgg.2 3ex4.1 3ewu.6 3ex5.1 3ex0.1 3ex1.3 2qcl.2 3ex6.4 3g3d.3 2qcm.1 3dbp.2 3ex2.3 3ex3.3 2qcg.5 2qcf.1 3eww.4 3l0n.3 3g3m.1 2qce.1 3bk0.2 3ewy.1 3ewx.1 2qch.4 3mi2.3 3l0k.2 2p1f.2 2qcn.2 3bvj.3

6.3.2.1

3ivg.4 1n2e.4 3ivc.5 3coz.3 1n2b.5 2a84.2 3cow.3 3cov.3 3ioe.5 3img.4 1n2h.5 3isj.5 2a86.5 1mop.5 3coy.2 2a7x.2 3iob.4 1n2j.4 2a88.2 1n2o.2 3le8.3 3iue.4 1n2g.4 3imc.2 1n2i.4 3iub.4 3iod.3

6.3.4.4

2gcq.1 1hoo.7 1ade.7 1cg1.1 1hon.8 1adi.8 1ch8.1 1cg4.1 1hop.7 1qf4.2 1cg3.1 1nht.1 1qf5.2 1cib.1 1cg0.1 1gim.1

A.0.6 Dataset Containing 1028 Serine Protease Pockets of 40 Subclasses of the EC Class 3.4.21

3.4.21.0

1cmv.4 1dua.3 1eax.1 1euf.1 1exf.2 1fi8.3 1hyl.1 1iav.1 1m9u.3 1mza.1 1o5e.3
1o5f.3 1op0.1 1op2.1 1p57.2 1q3x.9 1s2n.3 1sh7.2 1sot.14 1soz.8 1te0.3 1umu.1
1v6c.2 1wmd.1 1wme.2 1wmf.2 1wpo.3 1wsd.1 1wvm.2 1y9z.1 1ym0.2 1z8g.3 2b6n.1
2ea3.1 2gv6.1 2gv7.1 2hd4.1 2hlc.1 2oq5.1 2psx.2 2psy.2 2qf0.16 2qf3.4 2r0l.1
2r3y.8 2w2m.2 2w2n.5 2w2q.3 2z9i.5 2zec.1 2zgc.1 2zgh.2 2zgj.2 2zle.19 3bps.4
3dfj.1 3dfl.1 3e0p.1 3e16.2 3e1x.1 3f7m.1 3fvf.2 3gdv.5 3gyl.1 3h42.7

3.4.21.1

1ab9.1 1acb.3 1afq.1 1ca0.1 1dlk.2 1gg6.1 1ggd.1 1gl1.3 1hja.1 1k2i.1 1oxg.1
1p2m.6 1p2n.6 1p2o.8 1p2q.6 1t7c.6 1t8l.6 1t8m.6 1t8n.6 1t8o.6 1vgc.1 2jet.5
2p8o.1 2vgc.1 3vgc.1 4cha.2 4vgc.1 8gch.1

3.4.21.11

1hax.2 1hay.2 1haz.3 1hb0.2 1qix.2

3.4.21.12

1boq.1 1gba.1 1gbb.1 1gbc.1 1gbd.1 1gbe.1 1gbf.1 1gbh.1 1gbi.1 1gbj.1 1gbk.1
1gbl.1 1gbm.1 1p01.1 1p02.1 1p03.1 1p05.1 1p06.1 1p09.1 1p10.1 1p11.1 1p12.1
1qq4.2 1qrw.2 1qrx.1 1ssx.1 2h5c.1 2h5d.1 2lpr.1 3lpr.1 5lpr.2 6lpr.1 7lpr.1
8lpr.1 9lpr.1

3.4.21.14

1mee.1 2sic.1 2sni.1

3.4.21.20

1au8.1 1cgh.1 1t32.1

3.4.21.21

1cvw.3 1dan.6 1kli.3 1klj.3 1w0y.6 1w7x.2 1wqv.8 1wtg.4 1wun.5 1ygc.3 1z6j.7
2a2q.4 2aer.2 2b7d.5 2bz6.3 2c4f.4 2ec9.4 2fir.4 2flr.7 2zwl.5 2zzu.6

3.4.21.22

2wph.2 2wpi.2 2wpj.1 2wpk.2 2wpl.3 2wpm.2

3.4.21.27

1zhm.2 1zhr.2 1zlr.2 1zmj.2 1zml.2 1zmn.2 1zom.2 1zpb.2 1zpz.2 1zrk.2 1zsj.2
1zsk.2 1zsl.2 1ztj.2 1ztk.2 2fda.2 3bg8.2

3.4.21.32

1azz.1

3.4.21.34

2anw.2 2any.2

3.4.21.35

1gvz.1 1spj.1 2pka.2

3.4.21.36

1b0e.2 1bma.3 1c1m.3 1e34.2 1e35.2 1e36.2 1e37.3 1e38.2 1ela.2 1elb.2 1elc.3
1eld.2 1ele.3 1fzz.3 1gvk.3 1gwa.3 1h9l.3 1hv7.2 1l0z.3 1l1g.2 1lka.3 1lkb.3
1mmj.4 1nes.3 1qgf.2 1qnj.2 1qr3.3 1uo6.3 1uvo.2 1uvp.2 2a7c.3 2a7j.1 2bb4.3
2bd2.2 2bd3.3 2bd4.3 2bd5.2 2bd7.3 2bd8.3 2bd9.3 2bda.3 2bdb.2 2bdc.3 2blo.2
2blq.2 2cv3.3 2de8.2 2de9.3 2est.2 2fo9.2 2foa.2 2fob.2 2foc.2 2fod.2 2foe.2
2fof.2 2fog.2 2foh.2 2g4t.3 2g4u.3 2h1u.2 2iot.2 2oqu.3 2v0b.2 2v35.3 3hgn.2
3hgp.2 4est.3 5est.2 6est.2 7est.1 8est.3 9est.2

3.4.21.37

1h1b.4 1ppf.3 2rg3.1 2z7f.2

3.4.21.39

1nn6.1 1pjp.1 1t31.1

3.4.21.4

1aks.1 1amh.1 1avx.2 1brb.1 1brc.2 1btw.2 1btx.1 1btz.1 1bzx.3 1c1o.1 1c1p.1
1c1q.1 1c1r.1 1c1s.1 1c1t.1 1c2d.1 1c2e.1 1c2f.1 1c2g.1 1c2h.1 1c2i.1 1c2j.1
1c5p.1 1c5q.2 1c5r.1 1c5s.2 1c5t.1 1c5u.1 1c5v.1 1ce5.1 1co7.1 1eb2.1 1ejm.10

1ept.1 1ezs.3 1f0t.2 1f0u.1 1f5r.2 1f7z.1 1fmg.1 1fn6.1 1fn8.1 1fni.1 1fy4.1
1fy5.1 1fy8.1 1g36.1 1g3c.1 1g3d.1 1g3e.1 1gdn.1 1gdq.1 1gdu.1 1ghz.1 1gi0.1
1gi1.1 1gi2.1 1gi3.2 1gi5.1 1gi6.1 1gj6.1 1h4w.1 1hj9.1 1j14.1 1j15.1 1j16.1
1j17.1 1j8a.1 1jir.2 1jrs.1 1jrt.1 1k1i.2 1k1j.1 1k1l.2 1k1m.2 1k1n.2 1k1o.2
1k1p.1 1k9o.4 1lqe.1 1mbq.2 1n6x.1 1n6y.1 1nc6.2 1o2h.2 1o2i.2 1o2j.2 1o2k.2
1o2l.2 1o2m.1 1o2n.1 1o2o.2 1o2p.1 1o2q.1 1o2r.1 1o2s.2 1o2t.2 1o2u.1 1o2v.1
1o2w.1 1o2x.2 1o2y.1 1o2z.2 1o30.1 1o31.2 1o32.1 1o33.2 1o34.2 1o35.2 1o36.2
1o37.2 1o38.1 1o39.2 1o3b.2 1o3c.2 1o3d.2 1o3e.2 1o3f.2 1o3g.2 1o3i.2 1o3j.2
1o3k.2 1o3l.1 1o3m.2 1o3n.2 1o3o.2 1oph.3 1os8.1 1oss.1 1ox1.1 1oyq.1 1p2i.1
1p2j.1 1p2k.2 1ppc.1 1pph.2 1ppz.1 1pq5.1 1pq7.1 1pq8.1 1pqa.1 1qa0.1 1qb1.2
1qb6.1 1qb9.1 1qbn.1 1qbo.1 1ql7.1 1ql9.1 1qqu.1 1rxp.1 1s0q.1 1s0r.2 1s5s.1
1s6f.1 1s6h.1 1s82.1 1s83.1 1s84.2 1s85.1 1sbw.1 1sfi.1 1smf.2 1taw.2 1tio.1
1trn.3 1tx7.1 1tx8.1 1uhb.1 1utj.2 1utk.1 1utl.1 1utm.1 1utn.2 1uto.1 1utp.1
1utq.1 1v2j.1 1v2k.1 1v2l.1 1v2m.1 1v2n.1 1v2o.1 1v2p.1 1v2q.1 1v2r.1 1v2s.1
1v2t.1 1v2u.2 1v2v.1 1v2w.1 1v6d.2 1xvm.1 1xvo.1 1y3u.2 1y3v.2 1y3w.2 1y3x.2
1y3y.1 1y59.1 1y5a.1 1y5b.1 1y5u.1 1yf4.1 1ykt.1 1ylc.1 1yld.1 1yp9.2 1yyy.1
1z7k.3 2a31.1 2a32.1 2a7h.2 2age.2 2agg.2 2agi.2 2ah4.1 2ayw.1 2blv.2 2blw.2
2by5.1 2by6.1 2by7.1 2by8.1 2by9.1 2bya.1 2bza.2 2cmy.2 2d8w.2 2f91.1 2fi4.2
2fi5.2 2fmj.1 2ftm.2 2fx4.1 2fx6.1 2g51.1 2g52.1 2g55.1 2g5n.1 2g5v.1 2g8t.1
2j9n.2 2o9q.2 2otv.1 2oxs.1 2plx.1 2ptc.2 2ra3.5 2sta.1 2stb.1 2tio.1 2uuy.2
2vu8.1 2zdk.1 2zdl.1 2zdm.1 2zdn.1 2zfs.1 2zft.1 2zhd.1 2zq1.1 2zq2.2 3btd.3
3btf.3 3btm.2 3btq.1 3e8l.3 3fp6.1 3fp7.1 3fp8.2 3tgi.1 3tgj.1 3tgk.2 5ptp.1

3.4.21.42

1elv.3

3.4.21.43

2odp.9 2odq.9

3.4.21.46

1dfp.3 1dic.1

3.4.21.47

1rrk.2 1rtk.5 2ok5.5 3hrz.12

3.4.21.5

1a2c.2 1a3b.2 1a3e.3 1a46.4 1a4w.3 1a5g.2 1a61.2 1abi.2 1abj.2 1ad8.2 1ae8.3
1afe.3 1aht.2 1ai8.2 1aix.3 1awf.2 1b5g.3 1b7x.3 1ba8.2 1bb0.2 1bbr.9 1bcu.3
1bhx.2 1bth.7 1c1u.4 1c1v.2 1c1w.2 1c4u.3 1c4v.1 1c5l.3 1c5n.3 1c5o.4 1d3d.2
1d3p.3 1d4p.2 1d6w.1 1d9i.1 1de7.4 1dit.3 1doj.1 1dx5.14 1eb1.3 1eoj.3 1eol.3
1etr.2 1ets.3 1ett.2 1fpc.3 1fph.4 1g30.2 1g32.2 1g37.3 1ghv.3 1ghw.3 1ghx.3
1ghy.2 1gj4.3 1gj5.2 1h8i.4 1hag.4 1hah.4 1hai.2 1hgt.2 1hxe.2 1hxf.3 1id5.5
1iht.2 1jwt.3 1k21.3 1k22.3 1kts.2 1ktt.3 1lhc.3 1lhd.2 1lhe.2 1lhf.2 1lhg.1
1mu6.2 1mu8.2 1mue.3 1nm6.2 1no9.2 1nrr.2 1nrs.2 1nt1.2 1nu7.8 1nu9.16 1ny2.3
1nzq.2 1o0d.3 1o2g.2 1o5g.2 1oyt.2 1ppb.2 1qbv.2 1qhr.3 1qj1.3 1qj6.3 1qj7.3
1rd3.4 1riw.2 1sb1.3 1sfq.4 1sg8.4 1sgi.3 1shh.6 1sl3.3 1t4u.3 1t4v.3 1ta2.3
1ta6.3 1tbz.3 1thp.3 1thr.2 1ths.3 1tmb.3 1tmt.2 1tmu.2 1tom.2 1tq7.3 1twx.3
1ucy.9 1uma.3 1uvt.2 1vr1.3 1vzq.3 1w7g.2 1way.3 1wbg.4 1xm1.1 1xmn.13 1ycp.5
1ype.2 1ypg.2 1ypj.2 1ypl.2 1ypm.2 1z71.4 1z8i.2 1z8j.2 1zgi.3 1zgv.3 1zrb.3
2a2x.4 2ank.4 2anm.3 2b5t.7 2bdy.2 2bvr.2 2bvs.2 2c8w.4 2c8x.3 2c8y.2 2c8z.3
2c90.3 2c93.5 2cf8.2 2cf9.2 2cn0.2 2feq.3 2fes.3 2gde.2 2gp9.4 2h9t.1 2hgt.3
2jh0.2 2jh5.2 2jh6.2 2pgb.1 2pgq.2 2r2m.3 2thf.2 2uuf.3 2uuj.3 2uuk.2 2v3h.2
2v3o.2 2zc9.2 2zda.2 2zdv.2 2zf0.2 2zff.2 2zfp.2 2zfq.2 2zfr.2 2zg0.2 2zgb.2
2zhe.2 2zhf.2 2zhq.2 2zhw.2 2zi2.2 2ziq.2 2znk.2 2zo3.2 3bef.3 3bei.3 3bf6.2
3biu.2 3biv.3 3bv9.3 3c1k.3 3c27.3 3d49.3 3da9.2 3dd2.2 3dhk.2 3dt0.2 3dux.3
3e6p.7 3egk.3 3eq0.2 3f68.2 3gic.3 3gis.6 3hat.3 3hk3.2 3hki.7 3jz1.2 3jz2.1
4htc.3 4thn.3 5gds.2 7kme.2 8kme.3

3.4.21.59

2bm2.2 2f9n.6 2fpz.8 2fs9.5 2fww.1

3.4.21.6

1c5m.1 1ezq.1 1f0r.2 1f0s.2 1fjs.2 1g2l.1 1ksn.1 1lpg.1 1lpk.1 1lpz.1 1mq5.1
1mq6.1 1nfu.1 1nfw.2 1nfx.2 1nfy.2 1v3x.1 1wu1.1 1xka.2 1xkb.3 1z6e.2 2boh.2
2bok.3 2bq7.2 2cji.1 2d1j.2 2ei6.2 2ei7.2 2ei8.1 2fzz.1 2g00.1 2h9e.1 2j2u.1
2j34.2 2j38.2 2j4i.1 2j94.1 2j95.2 2jkh.1 2p3t.1 2p3u.2 2p93.1 2p94.2 2p95.2
2phb.2 2pr3.2 2q1j.2 2ra0.1 2uwl.1 2uwo.1 2uwp.1 2vh0.1 2vh6.1 2vvc.4 2vvu.3
2vvv.2 2vwl.1 2vwm.3 2vwn.1 2vwo.1 2w3i.2 2w3k.2 3cen.2 3cs7.2 3ffg.2

3.4.21.61

1ot5.4 1r64.2

3.4.21.62

1bh6.1 1c3l.2 1c9j.1 1c9m.1 1cse.1 1dui.2 1gnv.3 1lw6.1 1ndq.1 1ndu.1 1oyv.2
1q5p.1 1r0r.2 1sbn.1 1scn.2 1sib.1 1sua.2 1tk2.1 1tm1.1 1tm7.1 1tmg.1 1y1k.1
1y33.1 1y34.1 1y3c.1 1y3d.1 1y4a.2 1y4d.1 1yu6.2 2gko.1 2sec.1 2z2x.2 3sic.2
5sic.2

3.4.21.64

1bjr.1 1cnm.1 1egq.1 1ht3.1 1ic6.1 1oyo.1 1p7v.1 1p7w.1 1pek.1 1pfg.1 1pj8.1
2dqk.1 2duj.1 2g4v.1 2hpz.1 2id8.1 2pq2.1 2pwa.1 2pwb.1 2pyz.1 2v8b.1 3d9q.1
3ddz.1 3de0.1 3de1.1 3de2.1 3de3.1 3de4.1 3de5.1 3de6.1 3de7.1 3dvq.2 3dvr.1
3dvs.1 3dw1.1 3dw3.1 3dwe.1 3dyb.1 3gt3.1 3gt4.1 3prk.1

3.4.21.66

1tec.1 2tec.1 3tec.1

3.4.21.7

1ddj.5 1rjx.1

3.4.21.71

1bru.2

3.4.21.73

1c5w.1 1c5x.1 1c5y.1 1c5z.1 1ejn.1 1f5k.1 1f5l.1 1gi7.1 1gi8.1 1gi9.1 1gj7.1
1gj8.1 1gj9.1 1gja.1 1gjb.1 1gjc.1 1gjd.1 1o3p.1 1o5a.1 1o5b.1 1o5c.1 1owd.1
1owe.2 1owh.2 1sc8.1 1sqa.2 1sqo.2 1sqt.2 1u6q.2 1vj9.1 1vja.1 1w0z.1 1w10.1
1w11.1 1w12.1 1w13.1 1w14.1 2o8t.2 2o8u.1 2o8w.1 2vin.2 2vio.1 2vip.2 2viq.1
2viv.1 2viw.1 3ig6.3

3.4.21.74

2aip.1 2aiq.1

3.4.21.78

1op8.6 1orf.1

3.4.21.79

1iau.1

3.4.21.81

1cso.2 1ct0.2 1ct2.1 1ct4.2 1ds2.2 1sgd.2 1sge.2 1sgn.2 1sgp.2 1sgq.2 1sgr.2 1sgy.2 2gkv.3 2nu0.1 2nu1.2 2nu2.1 2nu3.2 2nu4.1 2qa9.1 2sgd.2 2sge.2 2sgf.2 2sgp.2 2sgq.2 3sgq.2

3.4.21.82

1hpg.1

3.4.21.88

1jhf.3 1jhh.3

3.4.21.89

1kn9.2 3iiq.3

3.4.21.9

1ekb.1

3.4.21.92

1tg6.27 1tyf.3 1yg6.8 2f6i.14 2fzs.4 2zl2.28 2zl4.22

3.4.21.97

1id4.5 1iec.5 1ied.3 1ief.6 1ieg.8 1njt.19 1o6e.7

A.0.7 Specific Binding to the Rat NR1/NR2B Receptor

Incubation mixtures used for inhibition experiments contain 5 nmol/l [^{3}H]-Ifenprodil, an optimized amount of rat brain membrane preparation (male Wistar rats), 5 mM Tris / 1 mM EDTA (pH 7.4, 100 µM R(+)-3-PPP, 1 µM GBR-12909, 1 µM GBR-12935) and inhibitor in 1 % DMSO within a total amount of 200 µl. Nonspecific binding was estimated in the presence of 10 µM CP101.606.

Samples were incubated for 60 minutes at 4° C. Binding was terminated by filtration of the incubated membrane preparations using Filtermat B (Pharmacia, Uppsala Sweden) and a Micro Cell Harvester (Skatron, Lier, Norway) and carefully washed with 50 mM Tris / HCl-buffer pH = 7.7 to separate free and bound radioactivity. The binding of [^{3}H]-Ifenprodil (KD = 9 nM, specific binding ca. 80 %) was determined by counting the remaining activity with a scintillation counter (Betaplate 1205, Berthold, Wildbad, Germany). Based on these raw data, IC$_{50}$ were calculated using the Hill model given sufficient inhibition. Otherwise, percent inhibition data were recorded.

The assay was performed by Ute Egerland at BioCrea GmbH.

A.0.8 Overview of the Dataset Containing 90 Protease Pockets Including their Merops Information

Clan	Subclan	Family	Pocket
MP	MA	M02	1o8a.1
MP	MA	M02	1r42.1
MP	MA	M10	2tcl.1
MP	MA	M10	1zs0.1
MP	MA	M10	1gkc.3
MP	MA	M10	1ciz.1
MP	MA	M10	1mmq.1
MP	MA	M10	1rm8.1
MP	MA	M12	3b92.2
MP	MA	M12	2rjp.5
MP	MA	M12	2v4b.3
MP	MA	M12	3b8z.1
MP	MA	M13	1r1h.5
MP	MA	M13	3dwb.3
MP	MC	M14	1aye.4
MP	MC	M14	1uwy.3
MP	MC	M14	3d68.7
MP	MC	M14	2pcu.1
MP	MS	M19	1itu.5
MP	MG	M24	1b6a.2
MP	MG	M24	2okn.8
MP	MP	M67	2znr.1
MP	MA	M10	1qib.1
MP	MA	M10	1q3a.3
MP	MA	M10	3f19.1
MP	MA	M10	1you.1

Clan	Subclan	Family	Pocket
SP	PA	S01	2zft.1
SP	PA	S01	1iau.1
SP	PA	S01	2bm2.3
SP	PA	S01	2psx.2
SP	PA	S01	2oq5.1
SP	PA	S01	1cgh.1
SP	PA	S01	1orf.1
SP	PA	S01	1klt.1
SP	PA	S01	2f9n.8
SP	PA	S01	3e0p.1
SP	PA	S01	2zch.8
SP	PA	S01	1h4w.1
SP	PA	S01	1bio.1
SP	PA	S01	1md8.1
SP	PA	S01	1elv.3
SP	PA	S01	2odp.9
SP	PA	S01	2any.2
SP	PA	S01	1zsk.2
SP	PA	S01	1kli.3
SP	PA	S01	2jkh.1
SP	PA	S01	1vzq.3
SP	PA	S01	3f6u.1
SP	PA	S01	1o5e.3
SP	PA	S01	1q3x.9
SP	PA	S01	1gj7.1
SP	PA	S01	1a5h.2
SP	PA	S01	1lo6.2
SP	PA	S01	1spj.1
SP	PA	S01	2qxj.1
SP	PA	S01	1eax.1
SP	SC	S09	2g63.19
SP	SC	S09	1z68.12
SP	SC	S09	1p0i.5
SP	SC	S09	1f6w.5
SP	SC	S10	1ivy.7
SP	SC	S33	2ocg.1
SP	SC	S33	3c5v.1
SP	SB	S53	3edy.5
SP	PA	S01	2ok5.5

Clan	**Subclan**	**Family**	**Pocket**
AP	AA	A01	1qrp.3
AP	AA	A01	2vij.2
AP	AA	A01	2g24.2
AP	AA	A01	1lya.1
AP	AA	A01	1tzs.3

Clan	**Subclan**	**Family**	**Pocket**
CS	CA	C01	1fh0.2
CS	CA	C01	1m6d.1
CS	CA	C01	1nqc.1
CS	CA	C01	1mem.2
CS	CA	C01	2ipp.1
CS	CA	C01	1cb5.13
CS	CA	C02	1zcm.1
CS	CA	C02	1kfu.6
CS	CA	C12	2etl.3
CS	CD	C14	2dko.3
CS	CD	C14	2qlb.3
CS	CD	C14	2c2z.2
CS	CD	C14	1rwn.2
CS	CD	C14	1pyo.3
CS	CA	C64	3dkb.14
CS	CA	C65	1tff.1
CS	CA	C01	1mhw.2
CS	CA	C01	2djg.5
CS	CA	C02	1ziv.4

Clan	**Subclan**	**Family**	**Pocket**
TP	PB	T02	2a8j.4

A.0.9 Sufficiently Populated Interaction Pairs for DSX H-Bond Potentials

PDB (79)

O.3oh_O.co2 (83 527), O.co2_O.3oh (83 526), O.2p_N.ams (80 035), N.ams_O.2p (80 001), O.3oh_O.am (68 475), O.am_O.3oh (68 473), O.co2_N.guh (44 012), N.guh_O.co2 (44 003), O.3oh_N.ams (42 287), N.ams_O.3oh (42 216), O.2p_N.guh (38 926), N.guh_O.2p (38 917), O.am_N.ams (38 057), N.ams_O.am (38 013), O.co2_N.ams (30 834), N.ams_O.co2 (30 799), O.3oh_N.guh (27 542), N.guh_O.3oh (27 537), O.co2_N.4h (19 466), N.amp_O.3oh (18 614), O.3oh_N.amp (18 614), N.4h_O.co2 (18 595), O.3oh_N.4h (18 032), N.4h_O.3oh (17 969), O.2p_N.4h (15 598), N.4h_O.2p (15 546), O.am_N.amp (13 689), N.amp_O.am (13 688), O.3oh_N.ar3h (12 979), N.ar3h_O.3oh (12 972), O.am_N.guh (12 410), N.guh_O.am (12 406), N.aap_O.am (12 268), O.am_N.aap (12 268), O.am_N.4h (11 951), N.4h_O.am (11 393), O.co2_N.ar3h (9371), N.ar3h_O.co2 (9364), N.amp_O.co2 (8948), O.co2_N.amp (8946), N.ams_O.3po (7167), O.3po_N.ams (7162), N.amp_O.2p (6711), O.2p_N.amp (6711), O.am_N.ims (5682), N.ims_O.am (5676), O.3oh_N.ar2 (5470), N.ar2_O.3oh (5467), O.2p_N.ar3h (4959), N.ar3h_O.2p (4955), N.aap_O.co2 (4844), O.co2_N.aap (4844), O.am_N.mih (4828), N.mih_O.am (4826), N.ar3h_O.am (4548), O.am_N.ar3h (4548), N.ams_O.3et (4454), O.3et_N.ams (4429), N.guh_O.3et (4157), O.3et_N.guh (4151), N.4h_O.3et (4131), O.3et_N.4h (4129), O.2s_N.ams (4015), N.ams_O.2s (4006), N.guh_O.3po (3922), O.3po_N.guh (3922), N.amp_O.3et (3098), O.3et_N.amp (3090), O.3po_N.4h (2868), N.4h_O.3po (2865), N.ar2_N.ams (2515), N.ams_N.ar2 (2504), N.mih_O.co2 (2428), O.co2_N.mih (2428), N.guh_O.2s (2272), O.2s_N.guh (2272), N.amp_N.ar2 (2127), N.ar2_N.amp (2127), N.ams_O.carb (2010)

CSD (144)

O.co2_O.h2o (75 538), O.co2_N.4h (64 786), N.4h_O.co2 (44 462), O.3oh_N.4h (37 926), N.4h_O.3oh (30 817), O.am_N.ams (29 162), N.ams_O.am (29 128), O.co2_O.3oh (22 816), O.3oh_O.co2 (22 815), O.3oh_O.carb (17 275), O.2s_N.4h (15 652), N.ar2_N.ar3h (14 163), N.ar3h_N.ar2 (14 151), O.am_O.h2o (14 104), O.3oh_O.3et (13 660), O.2s_O.h2o (13 524), O.3et_O.3oh (13 378), N.4h_O.2s (11 624), N.4h_O.h2o (11 304), O.3et_N.4h (11 105), O.am_O.3oh (10 516), O.3oh_O.am (10 514), O.carb_O.3oh (10 234), O.co2_N.ar3h (10 191), N.ar3h_O.co2 (10 177), N.ar2_O.3oh (10 090), O.3oh_N.ar2 (10 086), O.2p_N.4h (9614), N.amp_O.am (9126), O.am_N.amp (9126), N.4h_O.carb (9050), O.co2_O.ph (8977), O.ph_O.co2 (8977), N.4h_O.3et (8881), O.am_N.4h (8788), N.ar2_O.h2o (8679), O.carb_N.4h (8403), O.noh_N.oh (8390), O.ph_O.carb (8362), O.3et_O.h2o (8209), O.co2_N.guh

(8121), N.guh_O.co2 (8116), N.4h_O.2p (8049), S.3_N.4h (7804), O.carb_O.h2o (7482), O.2p_O.h2o (7126), N.oh_O.noh (7076), N.4h_S.3 (6898), N.4h_N.1 (6796), O.n_N.4h (6752), N.2n_N.4h (6686), O.2s_N.guh (6147), N.guh_O.2s (6146), O.co2_N.ams (6096), N.ams_O.co2 (6079), O.3oh_N.ams (5796), N.ams_O.3oh (5786), N.4h_O.am (5616), N.ar3h_O.carb (5528), O.ph_N.4h (5221), O.am_N.ims (5123), N.ims_O.am (5107), N.4h_N.2n (5031), N.4h_O.n (4895), O.2es_O.3oh (4638), O.3oh_O.2es (4638), O.co2_N.amp (4614), N.amp_O.co2 (4612), N.3p_N.4h (4497), N.ar3h_O.h2o (4374), O.r3_N.4h (4229), N.ar2_N.4h (4111), N.ams_O.carb (4098), N.ams_O.h2o (3900), N.4h_N.3p (3825), N.4h_P.3 (3797), O.ph_O.3et (3763), O.3et_O.ph (3710), O.carb_O.ph (3655), P.3_N.4h (3654), N.4h_O.r3 (3591), N.ar3h_O.3oh (3561), O.3oh_N.ar3h (3561), O.3oh_N.2n (3512), O.am_O.ph (3472), O.ph_O.am (3472), O.2es_N.ams (3460), O.co2_N.mih (3459), N.ams_O.2es (3456), N.mih_O.co2 (3449), N.2n_O.3oh (3422), N.4h_O.ph (3348), N.amp_O.3oh (3161), O.3oh_N.amp (3161), O.3oh_S.3 (3117), O.n_O.h2o (3103), N.aap_O.co2 (3086), O.co2_N.aap (3086), O.3eta_O.3oh (2987), O.3oh_O.3eta (2987), S.3_O.3oh (2927), O.noh_O.carb (2832), O.3oh_N.guh (2800), N.guh_O.3oh (2797), N.ams_S.3 (2755), S.3_N.ams (2749), O.3oh_O.r3 (2676), N.4h_N.ar2 (2671), O.r3_O.3oh (2633), N.ar3h_O.3et (2619), N.guh_O.h2o (2593), N.ar3h_P.3 (2585), O.3et_N.ar3h (2578), O.3oh_O.n (2514), O.n_O.3oh (2514), O.carb_N.ar3h (2507), O.3oh_N.mih (2481), N.mih_O.3oh (2480), O.3oh_N.1 (2440), O.o_O.h2o (2360), N.3s_O.co2 (2347), O.co2_N.3s (2347), O.co2_O.noh (2315), O.noh_O.co2 (2315), N.amp_S.3 (2312), S.3_N.amp (2308), N.oh_N.4h (2283), O.ph_N.2n (2247), O.o_N.4h (2214), O.am_N.mih (2201), N.mih_O.am (2198), O.ph_P.3 (2173), N.2n_O.ph (2168), O.3oh_O.o (2143), O.3oh_O.2po (2121), O.2po_O.3oh (2114), O.ph_N.ar2 (2098), N.ar2_O.ph (2096), N.guh_O.carb (2065), O.n_N.guh (2026), N.guh_O.n (2025), S.3_O.h2o (2015), O.am_N.guh (2011), N.guh_O.am (2008)

Bibliography

[1] Frank H. Allen. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallographica Section B*, 58(3 Part 1):380–388, Jun 2002. doi: 10.1107/S0108768102003890. URL http://dx.doi.org/10.1107/S0108768102003890.

[2] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–10, 1990.

[3] R. C. Amorim and B. Mirkin. Minkowski Metric, Feature Weighting and Anomalous Cluster Initialisation in K-Means Clustering. *Pattern Recognition*, 45:1061–1075, 2012.

[4] J. An, M. Totrov, and R. Abagyan. Pocketome via Comprehensive Identification and Classification of Ligand Binding Envelopes. *Molecular & Cellular Proteomics*, 4: 752–761, 2005.

[5] E. Arunan, G. R. Desiraju, R. A. Klein, J. Sadlej, S. Scheiner, I. Alkorta, D. C. Clary, R. H. Crabtree, J. J. Dannenberg, P. Hobza, H. G. Kjaergaard, A. C. Legon, B. Mennucci, and D. J. Nesbitt. Definition of the Hydrogen Bond. *Pure and Applied Chemistry*, 83(8):1637–1641, 2011. doi: 10.1351/PAC-REC-10-01-02.

[6] O. Bachar, D. Fischer, R. Nussinov, and H. Wolfson. A Computer Vision Based Technique for 3-D Sequence-independent Structural Comparison of Proteins. *Protein Engineering*, 6:279–287, 1993.

[7] Pedro J. Ballester and W. Graham Richards. Ultrafast shape recognition to search compound databases for similar molecular shapes. *Journal of Computational Chemistry*, 28(10):1711–1723, 2007. ISSN 1096-987X. doi: 10.1002/jcc.20681. URL http://dx.doi.org/10.1002/jcc.20681.

[8] J. A. Barker and J. M. Thornton. An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinformatics*, 19(13):1644–1649, 2003.

[9] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1): 235–242, 2000.

[10] 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.

[11] T. A. Binkowski, L. Adamian, and J. Liang. Inferring functional relationships of proteins from local sequence and spatial surface patterns. *Journal of Molecular Biology*, 332(2):505–526, 2003.

[12] C. Bissantz, B. Kuhn, and M. Stahl. A Medicinal Chemist's Guide to Molecular Interactions. *Journal of Medicinal Chemistry*, 53(14):5061–5084, 2010. doi: 10.1021/jm100112j. URL `http://dx.doi.org/10.1021/jm100112j`.

[13] I. Boukhris, Z. Elouedi, T. Fober, M. Mernberger, and E. Hüllermeier. Similarity analysis of protein binding sites: a generalization of the maximum common subgraph measure based on quasi-clique detection. In *ISDA'09: International Conference on Intelligent Systems Design and Applications. Proceedings*, pages 1245–1250, Pisa, Italy, November 2009.

[14] Andrew P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30:1145–1159, 1997.

[15] G. P. Brady and P. F. W. Stouten. Fast Prediction and Visualization of Protein Binding Pockets with PASS. *Journal of Computer-Aided Molecular Design*, 14:383–401, 2000.

[16] 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*, 56:250–260, 2004.

[17] C. Bron and J. Kerbosch. Finding all cliques of an undirected graph. *Communications of the ACM*, 16(9):575–577, 1973.

[18] I. J. Bruno, J. C. Cole, J. P. Lommerse, R. S. Rowland, R. Taylor, and M. L. Verdonk. IsoStar: a library of information about nonbonded interactions. *Journal of Computer-Aided Molecular Design*, 11(6):525–537, 1997.

[19] I. J. Bruno, J. C. Cole, P. R. Edgington, M. Kessler, C. F. Macrae, P. McCabe, J. Pearson, and R. Taylor. New software for searching the Cambridge Structural Database and visualising crystal structures. *Acta Crystallographica Section B*, 58: 389–397, 2002. doi: 10.1107/S0108768102003324.

[20] H. Bunke and K. Shearer. A Graph Distance Metric Based on the Maximal Common Subgraph. *Pattern Recognition Letters*, 19(3-4):255–259, 1998.

[21] C. Carter, J. Benavides, P. Legendre, J. D. Vincent, F. Noel, F. Thuret, K. G. Lloyd, S. Arbilla, B. Zivkovic, and E. T. MacKenzie. Ifenprodil and SL 82.0715 as cerebral anti-ischemic agents. II. Evidence for N-methyl-D-aspartate receptor antagonist properties. *Journal of Pharmacology and Experimental Therapeutics*, 247 (3):1222–1232, 1988.

[22] A. J. Chalk, C. L. Worth, J. P. Overington, and A. W. Chan. PDBLIG: classification of small molecular protein binding in the Protein Data Bank. *Journal of Medical Chemistry*, 47(15):3807–16, 2004.

[23] Y. Chen, T. Kortemme, T. Robertson, D. Baker, and G. Varani. A new hydrogen-bonding potential for the design of protein-RNA interactions predicts specific contacts and discriminates decoys. *Nucleic Acids Research*, 32(17):5147–5162, 2004.

[24] T. Cheng, X. Li, Y. Li, Z. Liu, and R. Wang. Comparative assessment of scoring functions on a diverse test set. *Journal of Chemical Information and Modeling*, 49(4): 1079–1093, 2009. doi: 10.1021/ci9000053.

[25] R. Chikhi, L. Sael, and D. Kihara. Real-time Ligand Binding Pocket Database Search Using Local Surface Descriptors. *Proteins*, 78:2007–2028, 2010.

[26] H. Choi, H. Kang, and H. Park. New angle-dependent potential energy function for backbone-backbone hydrogen bond in protein-protein interactions. *Journal of Computational Chemistry*, 31(5):897–903, 2010.

[27] I. R. Craig, C. Pfleger, H. Gohlke, J. W. Essex, and K. Spiegel. Pocket-Space Maps To Identify Novel Binding-Site Conformations in Proteins. *Journal of Chemical Information and Modeling*, 51(10):2666–2679, 2011.

[28] Ido Dagan, Lillian Lee, and Fernando Pereira. Similarity-Based Methods For Word Sense Disambiguation. In *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 56–63, 1997.

[29] S. Dambhare, S. A. Soman, and M. C. Chandorkar. Current Differential Protection of Transmission Line Using the Moving Window Averaging Technique. *IEEE Transactions on Power Delivery*, 25(2):610–620, 2010.

[30] 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*, 49:2863–2872, 2009.

[31] M. Debela, P. Hess, V. Magdolen, N. M. Schechter, T. Steiner, R. Huber, W. Bode, and P. Goettig. Chymotryptic specificity determinants in the 1.0 A structure of the zinc-inhibited human tissue kallikrein 7. *Proceedings of the National Academy of Sciences*, 104(41):16086–91, 2007.

[32] 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.

[33] G. R. Desiraju. A Bond by Any Other Name. *Angewandte Chemie International Edition*, 50(1):52–59, 2010. doi: 10.1002/anie.201002960.

[34] M. M. Deza and E. Deza. *Encyclopedia of Distances*. Springer, Heidelberg, Germany, 2009.

[35] P. D. Dobson and A. J. Doig. Predicting Enzyme Class From Protein Structure Without Alignments. *Journal of Molecular Biology*, 345(1):187–199, 2005. ISSN 0022-2836. doi: 10.1016/j.jmb.2004.10.02. URL `http://www.sciencedirect.com/science/article/pii/S0022283604013166`.

[36] O. Dym and D. Eisenberg. Sequence-structure analysis of FAD-containing proteins. *Protein Science*, 10(9):1712–1728, September 2001.

[37] S. R. Eddy. Hidden Markov models. *Current Opinion in Structural Biology*, 6(3): 361–365, 1996.

[38] Ö. D. Ekici, M. Paetzel, and R. E. Dalbey. Unconventional serine proteases: Variations on the catalytic Ser/His/Asp triad configuration. *Protein Science*, 17:2023–2037, 2008.

[39] T. J. Ewing, S. Makino, A. G. Skillman, and I. D. Kuntz. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *Journal of Computer-Aided Molecular Design*, 15(5):411–428, 2001.

[40] U. Feige, S. Goldwasser, L. Lovasz, S. Safra, and M. Szegedy. Approximating clique is almost NP-complete. In *Proc. 32nd IEEE Symp. on Foundations of Computer Science*, pages 2–12, 1991.

[41] H. J. Feldman and P. Labute. Pocket Similarity: Are α Carbons Enough? *Journal of Chemical Information and Modeling*, 50:1466–1475, 2010.

[42] Alan R. Fersht, Jian-Ping Shi, Jack Knill-Jones, Denise M. Lowe, Anthony J. Wilkinson, David M. Blow, Peter Brick, Paul Carter, Mary M. Y. Waye, and Greg Winter. Hydrogen bonding and biological specificity analysed by protein engineering. *Nature*, 314:235–238, 1985. doi: 10.1038/314235a0.

[43] T. Fober and E. Hüllermeier. Similarity Measures for Protein Structures based on Fuzzy Histogram Comparison. In *World Congress on Computational Intelligence*, pages 2808–2814, Barcelona, Spain, 2010.

[44] T. Fober, M. Mernberger, G. Klebe, and E. Hüllermeier. Evolutionary construction of multiple graph alignments for the structural analysis of biomolecules. *Oxford Bioinformatics*, 25(16):2110–2117, August 2009.

[45] T. Fober, S. Glinca, G. Klebe, and E. Hüllermeier. Superposition and Alignment of Labeled Point Clouds. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(6):1653–1666, November/December 2011.

[46] Thomas Fober, Marco Mernberger, Gerhard Klebe, and Eyke Hüllermeier. Efficient Similarity Retrieval for Protein Binding Sites based on Histogram Comparison. In *German Conference on Bioinformatics*, pages 51–60, Braunschweig, Germany, 2010.

[47] Thomas Fober, Marco Mernberger, Gerhard Klebe, and Eyke Hüllermeier. Fingerprint Kernels for Protein Structure Comparison. *Molecular Informatics*, 31(6-7):443–452, 2012.

[48] 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.

[49] D. Ghersi and R. Sanchez. Beyond structural genomics: computational approaches for the identification of ligand binding sites in protein structures. *Journal of Structural and Functional Genomics*, 12(2):109–17, 2011.

[50] R. J. Gillespie and R. S. Nyholm. Inorganic stereochemistry. *Q. Rev. Chem. Soc.*, 11:339–380, 1957. doi: 10.1039/QR9571100339. URL `http://dx.doi.org/10.1039/QR9571100339`.

[51] S. Glinca and G. Klebe. Cavities Tell More than Sequences: Exploring Functional Relationships of Proteases via Binding Pockets. *Journal of Chemical Information and Modeling*, 53(8):2082–92, 2013. doi: http://dx.doi.org/10.1021/ci300550a.

[52] H. Gohlke, M. Hendlich, and G. Klebe. Knowledge-based scoring function to predict protein-ligand interactions. *Journal of Molecular Biology*, 2(295):337–356, 2000.

[53] Holger Gohlke and Gerhard Klebe. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angewandte Chemie International Edition Engl.*, 41(15):2644–2676, 2002.

[54] D. S. Goodsell and A. J. Olson. Automated docking of substrates to proteins by simulated annealing. *Proteins*, 8(3):195–202, 1990.

[55] A. Grishaev and A. Bax. An empirical backbone-backbone hydrogen-bonding potential in proteins and its applications to NMR structure refinement and validation. *Journal of the American Chemical Society*, 126(23):7281–7292, 2004.

[56] C. Guda, S. Lu, E. D. Scheeff, P. E. Bourne, and I. N. Shindyalov. CE-MC: a multiple protein structure alignment server. *Nucleic Acids Research*, 32(Web Server Issue): W100–W103, 2004.

[57] J. Günther, A. Bergner, M. Hendlich, and G. Klebe. Utilising structural knowledge in drug design strategies: applications using Relibase. *Journal of Molecular Biology*, 326(2):621–636, 2003.

[58] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.

[59] E. Hellinger. Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. *Journal für die reine und angewandte Mathematik*, 136:210–271, 1909.

[60] M. Hendlich, F. Rippmann, and G. Barnickel. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *Journal of Molecular Graphics and Modelling*, 15(6):359–363, 1997.

[61] M. Hendlich, A. Bergner, J. Günther, and G. Klebe. Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions. *Journal of Molecular Biology*, 326(2):607–620, 2003.

[62] B. Hoffmann, M. Zaslavskiy, J. Vert, and V. Stoven. A New Protein Binding Pocket Similarity Measure Based on Comparison of Clouds of Atoms in 3D: Application to Ligand Prediction. *BMC Bioinformatics*, 11:99, 2010.

[63] 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.

[64] L. Holm and C. Sander. Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Research*, 25(1):231–234, 1997.

[65] N. Howard, C. Abell, W. Blakemore, G. Chessari, M. Congreve, S. Howard, H. Jhoti, C. W. Murray, L. C. A. Seavers, and R. L. M. van Montfort. Application for Fragment Screening and Fragment Linking to the Discovery of Novel Thrombin Inhibitors. *Journal of Medicinal Chemistry*, 49:1346–1355, 2006.

[66] P. Jaccard. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

[67] X. Jalencas and J. Mestres. Identification of Similar Binding Sites to Detect Distant Polypharmacology. *Molecular Informatics*, 32:976–990, 2013.

[68] Mikael P. Johansson and Marcel Swart. Intramolecular halogen-halogen bonds? *Physical Chemistry Chemical Physics*, 15:11543–11553, 2013. doi: 10.1039/C3CP50962A. URL `http://dx.doi.org/10.1039/C3CP50962A`.

[69] W. Kabsch. A solution of the best rotation to relate two sets of vectors. *Acta Crystallographica*, 32:922–923, 1976.

[70] 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.

[71] E. Karakas and H. Furukawa. Crystal structure of a heterotetrameric NMDA receptor ion channel. *Science*, 344(6187):992–997, 2014.

[72] E. Karakas, N. Simorowski, and H. Furukawa. Subunit arrangement and phenylethanolamine binding in GluN1/GluN2B NMDA receptors. *Nature*, 475(7355): 249–253, 2011.

[73] E. Kellenberger, C. Schalon, and D. Rognan. How to Measure the Similarity Between Protein Ligand-binding Sites. *Current Computer-Aided Drug Design*, 4:209, 2008.

[74] E. A. Kennewell, P. Willett, P. Ducrot, and C. Luttmann. Identification of Target-specific Bioisosteric Fragments from Ligand-protein Crystallographic Data. *Journal of Computer-Aided Molecular Design*, 20:385–394, 2006.

[75] K. Kinoshita and H. Nakamura. Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. *Protein Science*, 12(8): 1589–1595, 2003.

[76] Andreas Klamt, Jens Reinisch, Frank Eckert, Arnim Hellweg, and Michael Diedenhofen. Polarization charge densities provide a predictive quantification of hydrogen bond energies. *Physical Chemistry Chemical Physics*, 14:955–963, 2012. doi: 10.1039/C1CP22640A. URL `http://dx.doi.org/10.1039/C1CP22640A`.

[77] Andreas Klamt, Jens Reinisch, Frank Eckert, Jerome Graton, and Jean-Yves Le Questel. Interpretation of experimental hydrogen-bond enthalpies and entropies from COSMO polarisation charge densities. *Physical Chemistry Chemical Physics*, 15: 7147–7154, 2013. doi: 10.1039/C3CP44611E. URL `http://dx.doi.org/10.1039/C3CP44611E`.

[78] Ina Koch. Enumerating all connected maximal common subgraphs in two graphs. *Theoretical Computer Science*, 250(1–2):1–30, 2001. doi: 10.1016/S0304-3975(00)00286-3. URL `http://www.sciencedirect.com/science/article/pii/S0304397500002863`.

[79] A. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4:83–91, 1933.

[80] T. Kortemme, A. V. Morozov, and D. Baker. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *Journal of Molecular Biology*, 326(4):1239–1259, 2003.

[81] T. Krotzky. *Analyse und algorithmische Erweiterung von Methoden zum strukturellen Vergleich von Proteinbindestellen.* Grin Verlag GmbH, München, Germany, 2010. ISBN 3640946200.

[82] T. Krotzky, T. Fober, E. Hüllermeier, and G. Klebe. Extended Graph-based Models for Enhanced Similarity Search in Cavbase. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 11(5):878–890, 2014. doi: 10.1109/TCBB.2014.2325020.

[83] Irina Kufareva, Andrey V. Ilatovskiy, and Ruben Abagyan. Pocketome: an encyclopedia of small-molecule binding sites in 4D. *Nucleic acids research*, 40(Database issue): D535–40, 2012. doi: 10.1093/nar/gkr825.

[84] D. Kuhn. *Beschreibung von Proteinbindetaschen für Funktionsstudien und de Novo-Design und die Entwicklung von Methoden zur funktionellen Klassifizierung von Proteinfamilien.* PhD thesis, Philipps-Universität Marburg, Marburg, Germany, 2004.

[85] D. Kuhn, N. Weskamp, E. Hüllermeier, and G. Klebe. Functional classification of protein kinase binding sites using Cavbase. *Journal of Medical Chemistry*, 2(10): 1432–47, 2007.

[86] S. Kullback and R. A. Leibler. On Information and Sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951. doi: 10.1214/aoms/1177729694.

[87] I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin. A geometric approach to macromolecule-ligand interactions. *Journal of Molecular Biology*, 161(2): 269–288, 1982.

[88] P. Labute and M. Santavy. Locating Binding Sites in Protein Structures. 2007. URL `https://www.chemcomp.com/journal/sitefind.htm`.

[89] G. N. Lance and W. T. Williams. Mixed-data classificatory programs, I.) Agglomerative Systems. *Australian Computer Journal*, 1:15–20, 1967.

[90] R. A. Laskowski. SURFNET: A Program for Visualizing Molecular Surfaces, Cavities, and Intermolecular Interactions. *Journal of Molecular Graphics*, 13:323–330, 1995.

[91] V. Le Guilloux, P. Schmidtke, and P. Tuffery. Fpocket: an Open Source Platform for Ligand Pocket Detection. *BMC Bioinformatics*, 10:168, 2009.

[92] C.-H. Lee, W. Lü, J. C. Michel, A. Goehring, X. Song J. Du, and E. Gouaux. NMDA receptor structures reveal subunit arrangement and pore architecture. *Nature*, 511 (7508):191–197, 2014.

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

[94] S. Leis, S. Schneider, and M. Zacharias. In silico prediction of binding sites on proteins. *Current Medicinal Chemistry*, 17:1550–1562, 2010.

[95] Christopher A. Lipinski, Franco Lombardo, Beryl W. Dominy, and Paul J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 23 (1–3):3–25, 1997.

[96] G. Liu and L. Wong. Effective pruning techniques for mining quasi-cliques. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, part II*, pages 33–49, Antwerp, Belgium, 2008.

[97] Y. Lu, T. Shi, Y. Wang, H. Yang, X. Yan, X. Luo, H. Jiang, and W. Zhu. Halogen Bonding - A Novel Interaction for Rational Drug Design? *Journal of Medicinal Chemistry*, 52(9):2854–2862, 2009. doi: 10.1021/jm9000133.

[98] I. K. McDonald and J. M. Thornton. Satisfying hydrogen bonding potential in proteins. *Journal of Molecular Biology*, 238(5):777–793, 1994.

[99] Mark McGann. FRED Pose Prediction and Virtual Screening Accuracy. *Journal of Chemical Information and Modeling*, 51(3):578–596, 2011. doi: 10.1021/ci100436p.

[100] M. Mernberger, G. Klebe, and E. Hüllermeier. SEGA - A Semi-Global Approach to Graph Alignment for Approximate Molecular Structure Comparison. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, PrePrints, 2011.

[101] J. Meslamani, D. Rognan, and E. Kellenberger. sc-PDB: a database for identifying variations and multiplicity of 'druggable' binding sites in proteins. *Bioinformatics*, 27 (9):1324–6, 2011. doi: 10.1093/bioinformatics/btr120.

[102] J. Mestres, E. Gregori-Puigjané, S. Valverde, and R. V. Solé. The topology of drug-target interaction networks: implicit dependence on drug properties and target families. *Molecular BioSystems*, 5(9):1051–7, 2009.

[103] J. E. Mills and P. M. Dean. Three-dimensional hydrogen-bond geometry and probability information from a crystal survey. *Journal of Computer-Aided Molecular Design*, 10(6):607–622, 1996.

[104] M. Moll and L.E. Kavraki. Matching of structural motifs using hashing on residue labels and geometric filtering for protein function prediction. In *CSB'08: 7th Conference on Computational Systems Bioinformatics. Proceedings*, pages 157–169, Palo Alto, USA, August 2008.

[105] L. Mony, J. N. Kew, M. J. Gunthorpe, and P. Paoletti. Allosteric modulators of NR2B-containing NMDA receptors: molecular mechanisms and therapeutic potential. *British Journal of Pharmacology*, 157(8):1301–1317, 2009.

[106] A. V. Morozov, T. Kortemme, K. Tsemekhman, and D. Baker. Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *PNAS, Proceedings of the National Academy of Sciences*, 101(18):6946–6951, 2004.

[107] A.G. Murzin, S.E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4):536–540, 1995.

[108] S. G. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453, 1970.

[109] G. Neudert and G. Klebe. DSX: a knowledge-based scoring function for the assessment of protein-ligand complexes. *Journal of Chemical Information and Modeling*, 51(10): 2731–2745, 2011.

[110] G. Neudert and G. Klebe. fconv: Format Conversion, Manipulation and Feature Computation of Molecular Data. *Bioinformatics*, 27:1021–1022, 2011.

[111] Gerd Neudert. *Development and Improvement of Tools and Algorithms for the Problem of Atom Type Perception and for the Assessment of Protein-Ligand-Complex Geometries.* PhD thesis, Philipps-Universität Marburg, Marburg, Germany, 2011.

[112] Britta Nisius, Fan Sha, and Holger Gohlke. Structure-based computational analysis of protein binding sites for function and druggability prediction. *Journal of Biotechnology*, 159(3):123–134, 2012. doi: http://dx.doi.org/10.1016/j.jbiotec.2011.12.005. URL `http://www.sciencedirect.com/science/article/pii/S0168165611006614`.

[113] 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*, 5(8): 1093–1108, 1997.

[114] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin. Shape Distributions. *ACM Transactions on Graphics*, 21(4):807–832, 2002.

[115] Linus Pauling. *The nature of the chemical bond and the structure of molecules and crystals*. Cornell University Press, Ithaca, NY, USA, 1938.

[116] X. Pennec and N. Ayache. A Geometric Algorithm to Find Small but Highly Similar 3D Substructures in Proteins. *Bioinformatics*, 14:516–522, 1998.

[117] S. Pérot, 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 Discov Today*, 15(15-16):656–67, 2010.

[118] C. L. Perrin and J. B. Nielson. “Strong” hydrogen bonds in chemistry and biology. *Annual Review of Physical Chemistry*, 48:511–544, 1997.

[119] K. P. Peters, J. Fauck, and C. Frömmel. The Automatic Search for Ligand Binding Sites in Proteins of Known Three-dimensional Structure Using only Geometric Criteria. *Journal of Molecular Biology*, 256:201–213, 1996.

[120] D. Plewczynski, M. Łaźniewski, R. Augustyniak, and K. Ginalski. Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. *Journal of Computational Chemistry*, 32(4):742–755, 2011.

[121] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe. A fast flexible docking method using an incremental construction algorithm. *Journal of Molecular Biology*, 261(3): 470–489, 1996.

[122] N. D. Rawlings, A. J. Barrett, and A. Bateman. MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Research*, 40:D343–50, 2012. doi: 10.1093/nar/gkr987.

[123] O. C. Redfern, A. Harrison, T. Dallman, F. M. Pearl, and C. A. Orengo. CATHEDRAL: a fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures. *PLoS Computational Biology*, 3(11):2334–2347, 2007.

[124] M. Rosen, S. L. Lin, H. Wolfson, and R. Nussinov. Molecular shape comparisons in searches for active sites and functional similarity. *Protein Engineering*, 11(4):263–77, 1998.

[125] B. Rost. Twilight zone of protein sequence alignments. *Protein Engineering*, 12(2): 85–94, 1999.

[126] R. B. Russell. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *Journal of Molecular Biology*, 279:1211–1227, 1998.

[127] L. Sael and D. J. Kihara. Detecting local ligand-binding site similarity in nonhomologous proteins by surface patch comparison. *Proteins*, 80(4):1177–1195, 2012.

[128] S. Sarkhel and G. R. Desiraju. N-H...O, O-H...O, and C-H...O hydrogen bonds in protein-ligand complexes: strong and weak interactions in molecular recognition. *Proteins*, 54(2):247–259, 2004.

[129] 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):1755–78, 2008. doi: 10.1002/prot.21858.

[130] P. Schmidtke, C. Souaille, F. Estienne, N. Baurin, and R. T. Kroemer. Large-scale comparison of four binding site detection algorithms. *Journal of Chemical Information and Modeling*, 50(12):2191–200, 2010.

[131] S. Schmitt, M. Hendlich, and G. Klebe. From structure to function: a new approach to detect functional similarity among proteins independent from sequence and fold homology. *Angewandte Chemie International Edition*, 40(17):3141–3146, 2001.

[132] 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.

[133] G. Schneider and K. H. Baringhaus. *Molecular Design*. Wiley-Vch, 2008.

[134] K. T. Schomburg and M. Rarey. Benchmark data sets for structure-based computational target prediction. *Journal of Chemical Information and Modeling*, 54:2261–2274, 2014.

[135] Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 1.3r1. August 2010.

[136] 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.

[137] Sven Siggelkow and Hans Burkhardt. Improvement of Histogram-Based Image Retrieval and Classification. In *Proceedings of the 16 th International Conference on Pattern Recognition*, volume 3 of *ICPR '02*, page 30367, Washington, DC, USA, 2002. IEEE Computer Society. ISBN 0-7695-1695-X. URL `http://dl.acm.org/citation.cfm?id=839291.842844`.

[138] S. C. Sim and M. Ingelman-Sundberg. Pharmacogenomic biomarkers: new tools in current and future drug therapy. *Trends in Pharmacological Sciences*, 32(2):72–81, 2011.

[139] M. J. Sippl. Knowledge-based potentials for proteins. *Current Opinion in Structural Biology*, 5(2):229–235, 1995.

[140] N. Smirnov. Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics*, 19:279–281, 1948. doi: 10.1214/aoms/1177730256.

[141] Temple F. Smith and Michael S. Waterman. Identification of Common Molecular Subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.

[142] A. Stark and R. B. Russell. Annotation in three dimensions. PINTS: patterns in non-homologous tertiary structures. *Nucleic Acids Research*, 31(13):3341–3344, 2003.

[143] B. Stegemann and G. Klebe. Cofactor-binding sites in proteins of deviating sequence: Comparative analysis and clustering in torsion angle, cavity, and fold space. *Proteins*, 80(2):626–648, 2011.

[144] G. R. Stockwell and J. M. Thornton. Conformational Diversity of Ligands Bound to Proteins. *Journal of Molecular Biology*, 356:928–944, 2006.

[145] Robin Taylor. Which intermolecular interactions have a significant influence on crystal packing? *CrystEngComm*, 16(30):6852–6865, 2014. doi: 10.1039/C4CE00452C.

[146] F. Teichert, U. Bastolla, and M. Porto. SABERTOOTH: protein structural alignment based on a vectorial structure representation. *BMC Bioinformatics*, 8(1):425–442, 2007.

[147] J. M. Thornton. From genome to function. *Science*, 292(5524):2095–2097, 2001.

[148] A. E. Todd, C. A. Orengo, and J. M. Thornton. Evolution of function in protein superfamilies, from a structural perspective. *Journal of Molecular Biology*, 307(4): 1113–1143, 2001.

[149] H. F. Velec, H. Gohlke, and G. Klebe. DrugScore(CSD)-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *Journal of Medicinal Chemistry*, 48(20):6296–6303, 2005.

[150] M. L. Verdonk, J. C. Cole, and R. Taylor. SuperStar: A Knowledge-based Approach for Identifying Interaction Sites in Proteins. *Journal of Molecular Biology*, 289: 1093–1108, 1999.

[151] M. L. Verdonk, V. Berdini, M. J. Hartshorn, W. T. Mooij, C. W. Murray, R. D. Taylor, and P. Watson. Virtual screening using protein-ligand docking: avoiding artificial enrichment. *Journal of Chemical Information and Computer Sciences*, 44(3): 793–806, 2004.

[152] 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*, 50:2041–2052, 2010.

[153] 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*, 52:360–372, 2012.

[154] M. M. von Behren, A. Volkamer, A. M. Henzler, K. T. Schomburg, S. Urbaczek, and M. Rarey. Fast Protein Binding Site Comparison via an Index-Based Screening Technology. *Journal of Chemical Information and Modeling*, 53:411–422, 2013.

[155] A. Vulpetti, T. Kalliokoski, and F. Milletti. Chemogenomics in Drug Discovery: Computational Methods Based on the Comparison of Binding Sites. *Future Medicinal Chemistry*, 4(15):1971–1979, 2012. doi: 10.4155/fmc.12.147.

[156] M. Wagener and J. P. Lommerse. The Quest for Bioisosteric Replacements. *Journal of Chemical Information and Modeling*, 46:677–685, 2006.

[157] Guoli Wang and R. L. Dunbrack. PISCES: a protein sequence culling server. *Bioinformatics*, 19(12):1589–1591, 2003. doi: 10.1093/bioinformatics/btg224.

[158] R. Wang and S. Wang. How does consensus scoring work for virtual library screening? An idealized computer experiment. *Journal of Chemical Information and Computer Sciences*, 41(5):1422–6, 2001.

[159] 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.

[160] R. Wang, X. Fang, Y. Lu, C. Y. Yang, and S. Wang. The PDBbind database: methodologies and updates. *Journal of Medicinal Chemistry*, 48(12):4111–4119, 2005.

[161] 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):123–35, 2010. doi: 10.1021/ci900349y.

[162] M. Weisel, E. Proschak, and G. Schneider. PocketPicker: Analysis of Ligand Binding-Sites with Shape Descriptors. *Chemistry Central Journal*, 1:7, 2007.

[163] N. Weskamp, E. Hüllermeier, D. Kuhn, and G. Klebe. Multiple Graph Alignment for the Structural Analysis of Protein Active Sites. *IEEE Transactions on Computational Biology and Bioinformatics*, 4(2):310–320, 2007. ISSN 1545-5963. doi: 10.1109/TCBB.2007.358301.

[164] Nils Weskamp, Eyke Hüllermeier, and Gerhard Klebe. Merging chemical and biological space: Structural mapping of enzyme binding pocket space. *Proteins: Structure, Function, and Bioinformatics*, 76(2):317–330, 2009. ISSN 1097-0134. doi: 10.1002/prot.22345. URL `http://dx.doi.org/10.1002/prot.22345`.

[165] J. C. Whisstock and A. M. Lesk. Prediction of protein function from protein sequence and structure. *Quarterly Reviews of Biophysics*, 36(3):307–340, 2004.

[166] M. A. Williams and J. E. Ladbury. Hydrogen Bonds in Protein-Ligand Complexes. In H.-J. Böhm and G. Schneider, editors, *Protein-Ligand Interactions*, pages 137–161. Wiley-VCH, Weinheim, Germany, 2005. ISBN 978-3-52730521-6.

[167] D. J. Wood, J. de Vlieg, M. Wagener, and T. Ritschel. Pharmacophore Fingerprint-Based Approach to Binding Site Subpocket Similarity and its Application to Bioisostere Replacement. *Journal of Chemical Information and Modeling*, 52:2031–2043, 2012.

[168] 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. doi: 10.1186/1471-2105-11-47. URL `http://www.biomedcentral.com/1471-2105/11/47`.

[169] Y. Ye and A. Godzik. FATCAT: a web server for flexible structure comparison and structure similarity searching. *Nucleic Acids Research*, 32:582–5, 2004.

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

[171] R. J. Zauhar, G. Moyna, L. Tian, Z. Li, and W. J. Welsh. Shape signatures: a new approach to computer-aided ligand- and receptor-based drug design. *Journal of Medical Chemistry*, 46(26):5674–90, 2003.

[172] M. Zheng, B. Xiong, C. Luo, S. Li, X. Liu, Q. Shen, J. Li, W. Zhu, X. Luo, and H. Jiang. Knowledge-based scoring functions in drug design: 3. A two-dimensional knowledge-based hydrogen-bonding potential for the prediction of protein-ligand interactions. *Journal of Chemical Information and Modeling*, 51(11):2994–3004, Nov 2011. doi: 10.1021/ci2003939.

[173] Z. Zsoldos, D. Reid, A. Simon, S. B. Sadjad, and A. P. Johnson. eHiTS: a new fast, exhaustive flexible ligand docking system. *Journal of Molecular Graphics and Modelling*, 26(1):198–212, 2007.

Danksagung

Mein Dank gilt an erster Stelle *Herrn Prof. Dr. Gerhard Klebe* für die Möglichkeit in seiner Arbeitsgruppe Wirkstoffdesign promovieren zu dürfen und für das überaus interessante Promotionsthema. Das Arbeitsklima in seiner Gruppe war jederzeit ausgezeichnet und er verstand es immer, den optimalen Konsens zwischen richtungsweisender Betreuung und kreativem Freiraum herzustellen. So konnte ich zu jeder Zeit meine Ziele verfolgen und die in dieser Arbeit vorgestellten Methoden und Ergebnisse entwickeln.

Dr. Peter Kolb danke ich für seine Bereitschaft, diese Arbeit als Zweitgutachter zu beurteilen sowie für die gemeinsame Arbeit am DUPED-Projekt.

Ich danke darüber hinaus vielen Kollegen der Arbeitsgruppe Wirkstoffdesign, anderen kooperierenden Arbeitsgruppen sowie Studenten aus dem Fachbereich Pharmazie, ohne die ein großer Teil der präsentierten Studien nicht möglich gewesen wäre:

Dr. Thomas Fober danke ich vor allem für seine Unterstützung im LC-Projekt sowie bei der Untersuchung ligandbasierter Bindetaschen. *Thomas Rickmeyer* danke ich für die gemeinsame Arbeit an allen Cavbase-Projekten und die immer angenehme Büroatmosphäre. Im Rahmen des RAPMAD-Projekts danke ich *Dimitri Grigorev, Simon Müller, Bastian Weißkopf* und *Benjamin Wenzel*, die in ihren Praktika neue Testdatensätze zusammengestellt und die von mir entwickelten Methoden ausgiebig getestet haben. Hier ist zudem *Dr. Serghei Glinca* zu nennen, der mir freundlicherweise den EC- und Proteasedatensatz für die Bindetaschenvergleiche zur Verfügung stellte. *Dr. Marc Strickert* überließ mir darüber hinaus den umfangreichen Serinproteasedatensatz. *Dr. Christian Grunwald* danke ich für den angenehmen Kontakt zur BioCrea GmbH und die durchgeführten Bindungsassays am NMDA-Rezeptor im Rahmen des RAPMAD-Projekts. *Denis Schmidt* gilt mein Dank für die Verwendung der Bindetaschenvergleichsmethode RAPMAD im DUPED-Projekt.

Dr. Julian Fuchs danke ich für die gemeinsame Arbeit an den Proteasebindetaschen sowie die vielen hilfreichen Informationen zur Stadt Cambridge, die er mir via Skype mitgeteilt hat.

Dr. Gerd Neudert gebührt großer Dank für die hervorragenden Arbeiten an DSX und fconv, ohne die meine Weiterentwicklungen – aber auch viele Arbeiten an Bindetaschenprojekten – nicht möglich gewesen wären.

Dr. Johannes Schiebel danke ich für das Korrekturlesen von Teilen dieser Arbeit und zahlreiche Diskussionen und Hilfestellungen zu Wasserstoffbrückengeometrien.

Für diese Forschungsarbeit wurden Rechnungen auf dem MaRC2-Hochleistungsrechner der Philipps-Universität Marburg durchgeführt. Für die Programmierberatung danke ich *René Sitt* und *Christian Strack* vom Hessischen Kompetenzzentrum für Hochleistungsrechnen, gefördert durch das Hessische Ministerium für Wissenschaft und Kunst. Ferner waren *Dr. Thomas Gebhardt* und *Manuel Haim* Teil des Beratungsteams.

Allen Personen aus dem Administratorenteam der AG danke ich dafür, dass sie unser Linux-Netzwerk betreut haben. *Lydia Hartleben* danke ich dafür, dass sie mit ruhiger Art und organisatorischem Geschick die Arbeitsgruppe zusammenhielt.

Zudem danke ich auch vielen Kollegen für die zahlreichen Abende in Restaurants, Bars oder im Kino, an denen die Arbeit am Institut nicht im Vordergrund stand: *Michael Betz, Stefan Krimmer, Nicole Bertoletti, Thomas Rickmeyer, Chris Rechlin, Martin Stieler, Florent Chevillard* und *Dr. Nina Klee.*

Abschließend möchte ich meinen Eltern *Manfred* und *Doris Krotzky* danken, die mich jederzeit unterstützt haben. Ebenso danke ich *Tanja Faust* für ihre Unterstützung während meines Studiums und der Promotionszeit.

Erklärung

Ich versichere, dass ich meine Dissertation

"Methods for the Efficient Comparison of Protein Binding Sites and for the Assessment of Protein-Ligand Complexes"

selbstständig ohne unerlaubte Hilfe angefertigt und mich dabei keiner anderen als der von mir ausdrücklich bezeichneten Quellen bedient habe.

Die Dissertation wurde in der jetzigen oder einer ähnlichen Form noch bei keiner anderen Hochschule eingereicht und hat noch keinen sonstigen Prüfungszwecken gedient.

Marburg, den 8.6.2015

. . . .

(Timo Krotzky)