Computational Methods for the Modulation of Protein-Protein Interactions

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von

Dipl.-Inform. Philipp Thiel

aus Tettnang

Tübingen
2014

Tag der mündlichen Qualifikation:	22. Januar 2015
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Oliver Kohlbacher
2. Berichterstatter:	Prof. Dr. Christian Ottmann

Zum Optimismus gibt es ja keine Alternative.

Harald Lesch

Abstract

During the last decades, drug discovery development has made considerable progress. However, annual numbers of released drugs for novel targets have been decreasing concomitantly. Limited success rates of combinatorial chemistry and high-throughput screening, as well as availability of feasible targets are some reasons for this problem. A strategy to overcome it is exploration of novel target classes in order to expand the druggable space. An example are protein-protein interactions (PPIs) that can be inhibited or stabilized. Inhibition aims at developing binders for one protein to prevent complex formation. However, known PPI inhibitors differ significantly from conventional drugs and current active site-biased compound libraries are probably inappropriate to discover them. The design of novel screening libraries is thus very important. PPI stabilization aims at developing molecules that bind to a protein complex to increase its stability like a *molecular glue*. In contrast to inhibition, it is rather unexplored but ground-breaking examples from nature inspire research efforts.

This work presents novel theoretical and experimental drug discovery approaches for these challenges. In the first part, we introduce novel chemoinformatics approaches for clustering of large chemical libraries. The development of a fast algorithm for pairwise similarity calculations forms the basis for an exact and deterministic clustering method, which is able to process the available chemical space in a short time. We complement our chemoinformatics work by a novel approach for fast classification of small molecules according to the similarity of their frameworks, the so-called *scaffolds*. The method generates families of molecules that share geometry conserving scaffolds and we show that family members possess similar activity on identical targets.

The second part introduces computational methods for PPI modulation. First, we present structure-based analysis of known stabilized PPIs, which enables the development of novel *in silico* approaches to screen for small molecule PPI stabilizers. We demonstrate their applicability by an experimentally tested virtual screening for 14-3-3 protein interaction stabilizers. Finally, we present a virtual screening approach dedicated to identify small molecule inhibitors of 14-3-3 protein interactions. Predicted inhibitors are experimentally verified and characterized by *in vitro* assays and X-ray crystallography. Structure-activity relationship studies yielded PPI inhibitors in the low micromolar range, which are also active in cell-based experiments.

Zusammenfassung

Die technologische Entwicklung der niedermolekularen Wirkstoffforschung hat in den vergangenen Jahrzenten große Fortschritte gemacht. Dennoch geht die jährliche Anzahl neuer Wirkstoffe zurück. Die niedrige Erfolgsquote von kombinatorischer Chemie und Hochdurchsatz-Screening sowie die mangelnde Verfügbarkeit handhabbarer molekularer Ziele sind Gründe hierfür. Eine Strategie dieses Problem zu lösen ist die Erforschung neuer molekularer Zielklassen. Ein Beispiel sind Protein-Protein Interaktionen (PPIs), deren Modulation zwei Möglichkeiten umfasst: (1) Inhibition und (2) Stabilisierung. Das Ziel der Inhibition ist die Entwicklung von Bindern, die sich an ein Protein anlagern und Komplexbildung verhindern. Bekannte Inhibitoren unterscheiden sich aber deutlich von herkömmlichen, wirkstoffähnlichen Molekülen und vorhandene Bibliotheken sind deshalb möglicherweise zur Suche ungeeignet. Die Entwicklung neuer Molekülbibliotheken ist deshalb von großer Bedeutung. PPI Stabilisierung hat zum Ziel, Moleküle zu entwickeln, die an Proteinkomplexe binden und deren Stabilität wie ein *Molekularkleber* erhöhen. Im Gegensatz zur Inhibition ist dieser Ansatz kaum erforscht, aber die Natur liefert wegweisende Beispiele die Forschung auf diesem Gebiet anregen.

Diese Arbeit stellt neue theoretische und experimentelle Methoden für diese Herausforderungen vor. Der erste Teil beschreibt Methoden der Chemoinformatik zum Clustering großer Substanzbibliotheken. Die Entwicklung eines schnellen Algorithmus zur paarweisen Ähnlichkeitsberechnung bildet die Grundlage eines exakten und deterministischen Clusteringverfahrens, das den verfügbaren chemischen Raum in kurzer Zeit verarbeiten kann. Ergänzend stellen wir eine Methode zur Klassifizierung von niedermolekularen Substanzen nach der Ähnlichkeit ihrer Grundgerüste vor. Die Methode erzeugt Molekülfamilien mit konservierter Grundgerüstgeometrie und wir zeigen, dass Mitglieder einer Familie ähnliche Aktivität auf einem Zielprotein haben. Der zweite Teil behandelt computergestützte Methoden zur PPI Modulation. Auf Basis strukturbioinformatischer Analysen stabilisierter PPIs entwickeln wir *in silico* Methoden zur PPI Stabilisatorsuche. Ein Beispiel für eine ausgewählte 14-3-3 PPI und die experimentelle Überprüfung der Ergebnisse zeigen deren Anwendung. Schließlich stellen wir die Ergebnisse eines virtuellen Screenings nach Inhibitoren von 14-3-3 PPIs vor. Vorhergesagte Inhibitorkandidaten werden *in vitro*, durch Röntgenkristallographie und in einem zellulären Assay experimentell validiert.

Acknowledgments

First and foremost, I thank my advisors Prof. Oliver Kohlbacher and Prof. Christian Ottmann for giving me the opportunity to conduct exciting interdisciplinary work that finally led to this thesis. Christian, who unhesitatingly started the adventure of integrating a bioinformatician into his group, gave me the chance to explore the fascinating field of non-virtual drug design and crystallography. Thank you for the patience and your believe in my theoretical work. Oliver, who always had time to listen to my concerns, ever managed to find time for a meeting with me, thereby considerably reducing the distance between Dortmund and Tübingen. Thank you for your immediate interest in the Dortmund project and especially for taking me over whenever necessary.

The Ottmann lab et al.: Benjamin Schumacher, Manuela Molzan, Michael Weyand, Sven Hennig, Lars Röglin, Malgorzata Skwarczynska, Maria Bartel, David Bier, Rolf und Michelin Rose, Adrijana Kubicek-Pejic, Stefanie Bovens, Svenja Schäfers, Nicole Meißner, Gernot Hahne, Marco Bürger, Stefan Baumeister, Tom Grossmann, and Oliver Koch. You all supported me and my work in various manners. I am glad of not only having found nice colleagues, but also good friends at the Chemical Genomics Centre and the Max Planck Institute in Dortmund.

Especially, I would like to thank Michael Weyand for sharing his exhaustless knowledge on X-ray crystallography with me. You awakened my fascination for this subject and sharpened my senses for looking deeply behind the atomic coordinates. Thanks to my 'cellar office' colleagues Ingrid Vetter, Claude Ostermann, and Georg Holtermann for their help and fruitful discussions. I'm grateful to Nina Fischer and Karin Boczek, who proof read parts of the manuscript.

Finally, I'm most deeply grateful to Caroline. You always motivate me to keep moving, you always encourage me in my work, and you always believe in me. Thank you!

General Remarks

- In accordance with the standard scientific protocol, I will use the personal pronoun ***we*** to indicate the reader and the writer, or my scientific collaborators and myself.
- Unless stated otherwise, all figures of protein structures were generated using the freely available visualization software BALLView.[1,2]
- We use the symbol '∘' to indicate the interaction of molecules in a complex. Thus, the notation Ras∘Raf would indicate a binary complex between the proteins Ras and Raf.

Contents

Chapter 1

Introduction

Motivation

The history of drug discovery can be traced back to the famous Egyptian *Papyrus Ebers*, a medical record which is dated before 1600 BC.[3] Despite this long history, the past 200 years were probably the period with most rapid and important progress leading to modern drug discovery. Starting with the identification of single organic molecules as active ingredient, their synthesis and modification in the 19th century, organic chemistry was the first discipline contributing to this development. In the 20th century, advancements in molecular biology and biochemistry led to the availability of purified proteins, which formed the basis for the development of *in vitro* assay systems and the elucidation of the three-dimensional structure of biological macromolecules by X-ray crystallography, nuclear magnetic sesonance (NMR) or cryo-electron microscopy. Finally, the initial sequencing of the human genome in 2001 by the groups of Lander and Venter formed a further milestone for various disciplines.[4,5] In the area of drug discovery, the human genome in the first place provided the material to estimate the number of druggable targets in humans, which is also known as *druggable genome*.[6] Various studies have been performed and the size of the druggable genome has been estimated to range in the order of 10^3, whereof a few hundred are already addressed by marketed drugs.[7,8]

However, despite this steady progress in drug discovery and the availability of not yet addressed targets, the number of new molecular entities (NME) released per year has been decreasing continuously. As shown in Fig. 1.1, this holds true especially for small molecule based drugs. A possible reason for this trend could be an increased intractability of the non-addressed fraction of currently considered drug targets using the existing drug discovery strategies. To overcome this problem, novel approaches are being explored and developed aiming at expanding the druggable genome. These efforts are of utmost importance for prospective drug discovery. They will form the basis for novel therapies and probably pave the way for the treatment of yet incurable diseases.

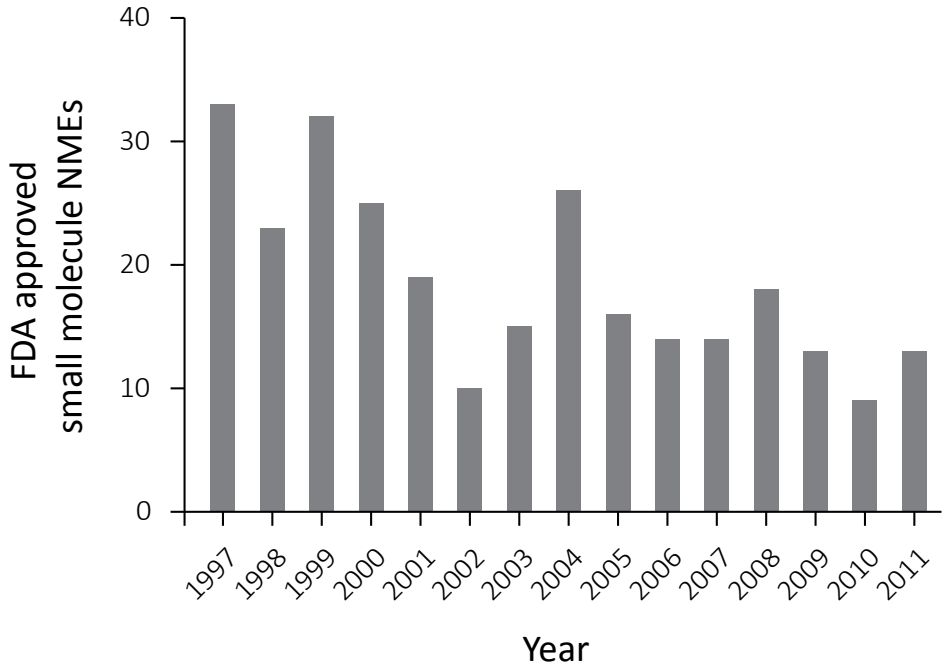


Figure 1.1: Number of yearly approved small molecule based NMEs by the U.S. Food and Drug Administration (FDA).

Expanding the Druggable Genome

To expand the space of druggable targets various strategies were developed in the recent past. One promising approach is to move from classic enzyme inhibition to enzyme activation.[9,10] Here, allosteric sites or regulatory subunits are modulated by small molecules. An advantage of this strategy is the possible reutilization of existing, established, and over the past decades optimized technologies like high-throughput screening (HTS) or protein-ligand docking. Another interesting approach is targeted therapy, where for example highly cytotoxic compounds are attached to monoclonal antibodies for site-specific application in malignant tissue.[11,12] Here, the high selectivity and specificity of monoclonal antibodies is a major advantage. An actual expansion of the druggable genome is the targeting of cellular processes on the DNA and RNA level. Here, small non-coding RNAs like siRNAs or shRNAs are used to modulate gene expression by RNA interference.[13,14] Even addressing RNA by small molecules is subject of current research.[15] However, these strategies are quite challenging because of major obstacles like application and transport of RNA.

A further target class which has the potency to tremendously expand the druggable genome are protein-protein interactions (PPI). The latter are of fundamental importance for all living organisms and they form a huge and complex network termed as *interactome*, which substantially contributes to the regulation and execution of the majority of biological processes. The

size of the binary human interactome has recently been re-estimated to comprise over 300,000 binary PPIs, of which only a fraction has yet been identified.[16]

In the recent past, the inhibition of PPIs by small molecules or by modified peptides has been accepted as a viable way to interfere with disease-related signaling pathways.[17–19] Successful examples have been presented for various anticancer, antiviral, antibacterial, and anti-inflammatory applications. However, it has also been recognized that the discovery of PPI inhibitors is a quite challenging task and standard HTS attempts often fail to yield validated hits.[20] As a possible reason for this observation, physicochemical property differences of small molecule PPI inhibitors and classical active site binders are discussed.[20,21] Existing screening libraries are historically grown and probably lack the chemotypes appropriate for binding to the surfaces of PPIs.

For a long time, the scientific community has overlooked that PPI modulation does not exclusively mean complex inhibition. Evidently, the complementary side of disrupting a transient biological system is its stabilization.[22] This fascinating mechanism is demonstrated by impressive examples from nature, where PPIs are stabilized by a small molecule. It has already been shown that protein complexes exhibit surface-exposed pockets at their binding interfaces with structural and physicochemical characteristics that are comparable to typical enzymatic active sites.[23,24] Additionally, the proof of concept for rational PPI stabilizer discovery has recently been provided by Ottmann and co-workers.[25] At all times, drug discovery has learned, copied, and adapted from nature and in contrast to the increasing number of examples for PPI stabilization, these examples for PPI inhibition are yet missing. Thus, PPI stabilization is a novel drug design concept with the potency to expand the druggable genome.

Challenges in Computer-Assisted PPI Modulation

Computational methods made their way into modern drug discovery and today some of them are an integral part of various steps in the drug discovery pipeline.[26] In general, these methods can be classified into three major fields, namely computational chemistry, chemoinformatics and structural bioinformatics.

A major challenge in chemoinformatics is the increasing number of – also virtually – available compounds. As described previously, contemporary HTS libraries seem to be unsuited for PPI inhibitor screening and the design and assembly of new compound collections will be necessary. Taking into account that virtually curated compound libraries nearly comprise a billion druglike molecules, sophisticated algorithms and tools are needed to handle and to analyze such huge amounts of data. Some of the existing chemoinformatics tools for similarity-based tasks like clustering are too slow and improved algorithms and approaches are needed. In addition, abstractions and representations of molecules reflecting medicinal chemistry needs would also be helpful to meaningfully reduce the complexity of these data sets.

In contrast to inhibition, the stabilization of PPIs is a rather unexplored field, especially from a computational perspective. Currently, no comprehensive analysis of the known examples has been performed. The availability of their crystal structures enables a detailed analysis of their mode of action. Learning from these examples allows us to define rules to augment existing virtual screening (VS) techniques and protocols for the identification of PPI stabilizers. An interesting task is also to analyze the stabilizing ligands themselves. If these compounds possess classic druglike chemotypes, this would preset a reasonable chance to identify candidates in existing screening libraries.

Thesis outline

This work is located at the interface of theoretical computer-assisted and experimental drug discovery. Novel theoretical approaches in chemoinformatics and structure-based drug design are presented and transferred into *in vitro* experiments to discover and characterize novel modulators of PPIs. Chapter 2 gives the theoretical and biological background necessary to understand the following chapters. First, selected topics from computer science, chemoinformatics and structure-based drug design are introduced. Second, fundamentals of X-ray crystallography are given, followed by the biological background on PPIs and their modulation. Finally, the protein family 14-3-3 is introduced, which forms the model system for our studies.

Part I: Chemoinformatics

The first part of this work introduces approaches for rapid analysis, clustering, and classification of large chemical spaces. Chapter 3 presents a novel chemoinformatics method for fast clustering of large chemical spaces. It is based on a sophisticated data structure in combination with an efficient algorithm that enables the calculation of all pairwise similarity coefficients – that is the similarity matrix – for libraries comprising tens of millions of compounds using standard hardware. The design goals of our algorithm are architecture independence and optimization for usage on modern multi-core machines. We demonstrate that this method is competitive to state-of-the-art methods for high-throughput similarity calculations. At its peak performance it calculates almost 400 million Tanimoto similarities per second. Analsyis of the method's runtime behavior and hardware demands allows us to infer important guidelines for parallel application. This method forms the core for our clustering approach, which yields an exact and deterministic clustering of a given compound library. We compare the clustering method to implementations of *Jarvis-Patrick* and *Ward* and show that its runtime is competitive or significantly better. As a final application example, we demonstrate that clustering of the available chemical space with over 17 million compounds takes only 64 hours to complete.

In Chapter 4, we introduce a method for the classification of druglike molecules based on their molecular frameworks, that is their *scaffolds*. The goals of this method are to generate a compound grouping in a way that is meaningful to medicinal chemists and a short computation time even for large compound libraries. Using the fast similarity calculation method introduced before, we show that similarity network generation at high Tanimoto thresholds yields homogeneous *scaffold families* and leads to a significant reduction of the input data set size. We show examples of *scaffold families*, which comprise molecules that are related by their structure-activity on two different targets.

Part II: Structure-based PPI Modulator Discovery

The second part of this work introduces computer-assisted drug discovery methods applied to the promising and challenging target class of PPIs and presents approaches for the development of stabilizers as well as inhibitors. In Chapter 5 we focus on PPI stabilization. We analyze the crystal structures of the currently described protein complexes that are stabilized by a small molecule in order to obtain quantitative knowledge on this mode of action. We use this knowledge to develop approaches to explicitely screen for PPI stabilizers *in silico*. As a first application, we use the approach to search for stabilized PPIs within the Protein Data Bank and indeed uncover six stabilized PPIs that were not part of the analyzed input data set. As a second application, we set up, perform, and experimentally test a VS for stabilizers of the interaction between 14-3-3 protein and its target Task3. Our virtual screening identified 258 stabilizer candidates from which 89 were tested in an *in vitro* stabilization assay. Finally, one compound shows stabilizing activity in the low micromolar range.

In the final Chapter 6, we present a VS study dedicated to the identification of 14-3-3 protein interaction inhibitors. We analyze the binding modes of representative crystal structures of 14-3-3 bound to different of its partner proteins. The information is used to formulate a binding mode hypothesis and use it to set up a virtual screening for inhibitors of 14-3-3 protein interactions. From a compound library comprising more than 8 million small molecules we retrieve candidates from virtual screening and manually selected 14 candidates for *in vitro* testing. Two candidates have been validated and show inhibitory activity in the low micromolar range. Structure-activity relationship studies of these hits led to more potent inhibitors, from which two are active in a cellular assay. We finally present an inhibitory compound that covalently binds to 14-3-3 .

Chapter 2

Background

The content of Section 2.4 is an extended version of the review article:

Small-molecule stabilization of protein-protein interactions: an underestimated concept in drug discovery?[22]

This chapter introduces the theoretical and biological background of this thesis. The first two parts provide the theoretical background of selected topics from computer science and chemoinformatics. Part three gives experimental and theoretical basics of X-ray crystallography. The biological background of PPIs and their modulation are content of part four. The family of 14-3-3 proteins, which serves as an example to study PPIs, is introduced in part five. Each of these topics forms a wide area of research by themselves. Thus, the single parts only focus on the topics that are necessary to understand the following chapters.

2.1 Graph Theory

Graph theory is a major research area in theoretical computer science and also in mathematics. Thus, it is a tremendously wide field and we only introduce the fundamentals necessary to understand the work presented in Chapter 3. The following content is primarily based on the textbook by Cormen *et al.*.[27]

2.1.1 Definition of a Graph

A *graph* G is a tuple (V, E), where V is a set of *vertices* with $n = |V|$ members. E is a set of $m = |E|$ edges. An *edge* is simply a pair (u, v) with $u, v \in V$ and $u \neq v$, and thus connects a

pair of vertices. Edges can have a direction or they can be undirected. In the latter case, the incident vertices of an edge form an unordered pair $\{u, v\}$. The resulting graph is then termed *undirected*, otherwise it is a *directed graph*. The number of incident edges of a single vertex is its *degree*. A *path* from vertex u to vertex v is a sequence of edges

$$\{u, w_1\}, \{w_1, w_2\}, \ldots, \{w_{k-1}, w_k\}, \{w_k, v\} \tag{2.1}$$

and the connected vertices u and v are called *reachable* from each other. The length of a path is the number of edges it spans. If every vertex in a graph is reachable from every other vertex, the graph is connected. Otherwise the graph is disconnected. A connected and undirected graph without cycles is a tree. A simple example of an undirected and disconnected graph is shown in Fig. 2.1.

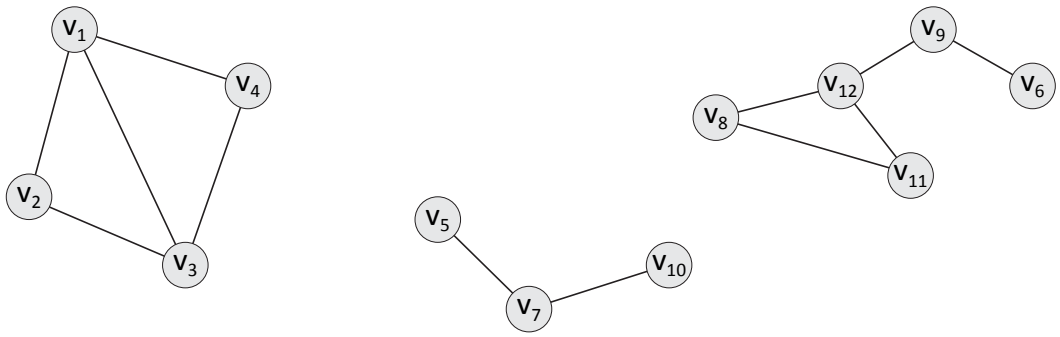


Figure 2.1: Example of an undirected and disconnected graph with $n = 12$ vertices and $m = 12$ edges. The graph is split up into three connected components.

2.1.2 Connected Components

Connected components (CC) are disjoint subsets of V where all vertices within a CC are reachable from each other. In contrast, there is no path that connects any two vertices of different CCs. The size of a CC is its cardinality, that is the number of its member vertices. The graph shown in Fig. 2.1 consists of three connected components. For the calculation of CCs in-memory and external memory algorithms have been described.[27,28] For our tasks, we concentrated on in-memory algorithms to avoid hard disk write access. The most basic in-memory algorithm to solve this problem is a depth-first search and its time complexity is linear in $|V| + |E|$. For huge graphs with millions of vertices and a high average vertex degree the algorithm can become infeasible due to memory limitations.

If the actual structure of a graph is not important, the CCs can be calculated incrementally without storing the edges. Such algorithms use *union-find* data structures, which are disjoint sets supporting the operations *union* and *find*. Initially, every vertex is a CC and $E = \{\emptyset\}$. An

edge $\{u, v\}$ is processed by a *find* operation, which identifies the representative vertices of the CCs to which u and v belong. A subsequent *union* operation concatenates the identified subtrees, if necessary. By the application of path compression strategies, which try to reduce the path length of vertices to their CC representative, these algorithms have near-linear runtimes.

2.2 Computational Methods in Drug Discovery

This section introduces selected fundamental concepts of computational methods in modern drug discovery that is commonly referred to as computer-aided drug design (CADD). We start with a brief overview on chemoinformatics in general and then describe the relevant theory in more detail. The second part of this chapter shortly introduces essential theory from structural bioinformatics.

2.2.1 Chemoinformatics

Chemoinformatics is a research field covering all aspects of information processing, storage and analysis of druglike molecules. Fundamental tasks comprise the digital representation of molecules, their comparison and classification, as well as database searching.[29,30]

Small Molecules

Small molecules – also frequently referred to as compounds – is a rather loose definition, which comprises stable organic molecules with a molecular weight (MW) roughly lower than 1,000 Da. The basic chemical element of these organic molecules is carbon. Together with hydrogen, nitrogen, oxygen, phosphorus, and sulfur it builds connected scaffolds that can contain further elements as substitutes including fluorine, chlorine, bromine and iodine. Depending on the elements and their hybridization, two atoms are connected by covalent single, double, aromatic or triple bonds.

In contrast to biologics, small molecules are with more than 90 % still the most important active substance of currently marketed drugs.[31] It has been observed that approved drugs form a subgroup of the entire set of small molecules with respect to selected physicochemical properties. This was first described by Lipinski *et al.*, who analyzed the MW, the number of hydrogen bond donors and acceptors, and the calculated logP of marketed drugs.[32] These properties are commonly referred to as *Lipinski's Rule-of-Five* (Ro5) and they are frequently used to estimate the oral bioavailability of small molecules.

In Silico Representations

Atoms and the connecting bonds describe the topology of a molecule structure, that is its two-dimensional (2D) appearance. This information can be described explicitly, for example, by

the *Simplified Molecular Input Line Entry System* (SMILES).[33] The latter is a chemical language to store the information in a single string, which can be translated back into a 2D topology. If the topology of a small molecule is extended by three-dimensional (3D) coordinates for every atom, the representation is a 3D geometry. The latter enables further representations like molecular surfaces. Examples for these representations are shown in Fig. 2.2A-C. An appropriate *in silico* data structure to map small molecules in 2D and 3D are graphs, where atoms are stored as vertices and bonds as edges.[34] Various properties like atom type, charge or the bond order can be assigned to vertices and edges.

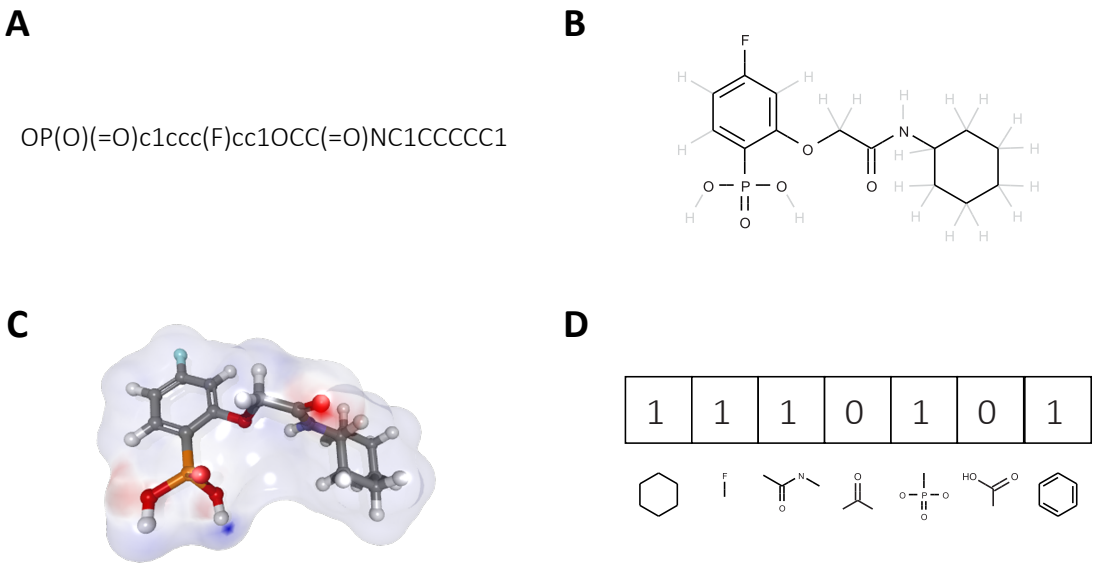


Figure 2.2: Different representations of a small molecule. **(A)** SMILES notation. **(B)** 2D topology. **(C)** 3D geometry and solvent-excluded surface (SES). **(D)** Abstract *structural key* fingerprint.

Molecular Fingerprints

In contrast to these explicit representations, molecule structures can be stored in abstract ways, for example as molecular fingerprints. The most basic fingerprint type is the *structural key* and an example is shown in Fig. 2.2D.[35] Here, a set of substructural patterns is defined and small molecules are represented by a boolean array, where every array position indicates the presence or absence of one of these patterns in the molecule it represents. If a substructure is present in a molecule the corresponding array position is *true* (1-bit), otherwise it is *false* (0-bit). A drawback of *structural keys* is their limitation to the set of predefined patterns. Molecules containing other substructural patterns cannot be described comprehensively by such fingerprints.

An advancement of *structural keys* are hashed fingerprints, which dynamically generate the set of substructural patterns from a small molecule itself. An example are path-based fingerprints (PBFP).[35] Here, all linear paths up to a predefined length are enumerated and obtain a unique ID. These IDs are then hashed into fixed-length bit arrays and all hashes for a single small molecule are combined by bitwise OR. A similar strategy is the extraction of radial substructures instead of linear paths.[36] In contrast to PBFPs, branching patterns can be captured by this fingerprint type. An example for this type are extended-connectivity fingerprints (ECFP).[37]

The density of fingerprints describes the fraction of 1-bits.[35] The fingerprint types tend to yield different densities. *Structural keys* and PBFPs tend to result in a higher density, whereas ECFPs are of low density.

A further level of abstraction can be achieved by introducing different atom descriptions. Frequently used descriptions are force-field-based atom types or the generalization of explicit atom types by pharmacophoric features like hydrogen bond donor and acceptor, hydrophobic, aromatic or charged.

Chemical Similarity

A commonly used technique to express the similarity between a pair of small molecules is to compare their substructural compositions. As described in the previous subsection the latter can be encoded by 2D fingerprints, where a distinct position indicates the absence (0-bit) or presence (1-bit) of a substructural element or a higher order feature of the corresponding molecule. To compare the substructural compositions of a molecule pair, two important parameters can be calculated from their fingerprints. First, the number of 1-bits of every fingerprint itself is calculated, which is computationally easy and has to be done only once. Second, the number of shared 1-bits of a fingerprint pair is calculated, which is used as an estimate for their substructural overlap. This shared feature count is individual for every fingerprint pair and is a computationally quite demanding task. Based on these simple parameters various related similarity and dissimilarity coefficients have been defined. One of the most intensively studied and frequently used similarity measures in chemoinformatics is the Jaccard or Tanimoto coefficient S_{Tan}:[38,39]

$$S_{Tan} = \frac{c}{a + b - c} \tag{2.2}$$

Here, a is the number of 1-bits in fingerprint **A**, b is the number of 1-bits in fingerprint **B** and c is the number of shared 1-bits of **A** and **B**. The Tanimoto similarity is a metric and has a codomain of $[0.0, 1.0]$. $S_{Tan} = 0.0$ indicates two maximal dissimilar fingerprints and $S_{Tan} = 1.0$ two identical fingerprints. It is important to mention that $S_{Tan} = 1.0$ does not imply identity

of the underlying molecules in general. As already mentioned, the computational challenge of all similarity and dissimilarity coefficients based on these parameters is the calculation of the number of shared 1-bits, c. In set notation, the number of shared 1-bits between two fingerprints **A** and **B** can be expressed as

$$c = |\mathbf{A} \cap \mathbf{B}| \tag{2.3}$$

and, for this reason, the calculation can be split up into a binary AND operation of **A** and **B** followed by counting the 1-bits in the resulting array. Different strategies and algorithms have been developed to calculate c efficiently. In Chapter 3, we present an algorithm we developed as the core part of a clustering method and we give an overview of existing solutions and applications.

Clustering

Clustering – or cluster analysis – in general is an unsupervised data mining technique aiming at splitting up a set of input objects into multiple groups, thereby maximizing inter-object similarities within the generated groups. Various clustering approaches have been described and only a subset of them is frequently used in chemoinformatics, whereat all have different strengths and weaknesses.[40] The methods can further be classified as shown in Fig. 2.3

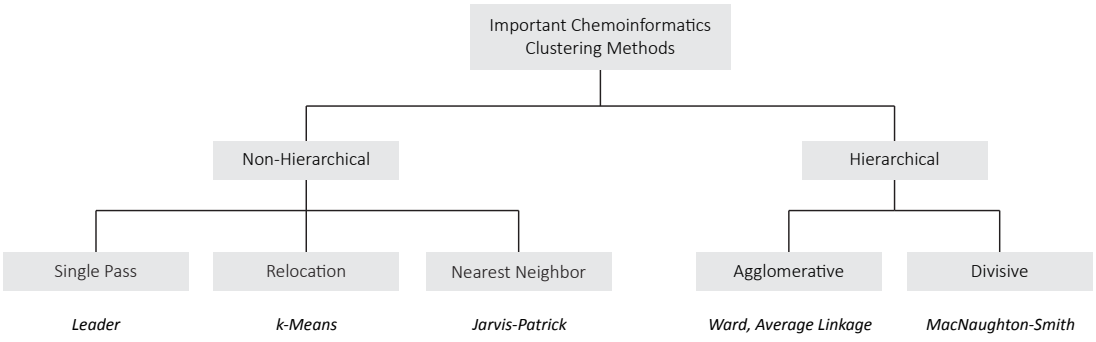


Figure 2.3: Classification schema of clustering methods. For every clustering strategy, examples are given that are frequently used in chemoinformatics.

Non-hierarchical methods yield a flat clustering, which is the result of only grouping input molecules without giving information on the relationship of the clusters. The relocation method *k-Means* starts by randomly selecting c molecules as initial cluster centers.[41,42] In an iterative procedure the remaining compounds are added to the cluster with maximum similarity to the center and subsequently the cluster centers are reassigned to the member with highest mean intra-cluster similarity. These steps are repeated until the cluster assignment

is stable. An advantage of this method is the ability to process large compound libraries due to its speed. The major disadvantages are the need to initially select the number of final clusters and, especially, that this selection process is generally random.

An example for a single-pass clustering is the *Leader* algorithm.[43] A similarity threshold and an initial starting molecule has to be chosen. Now, all molecules of which the similarity to the selected molecule exceeds the threshold are added to this cluster. If unassigned molecules are left, a new reference molecule is chosen in a subsequent assignment step. This is repeated until all molecules are assigned to a cluster. Here, the computational efficiency is again the advantage of this method. However, its major drawback is the outcome's dependency on the ordering of the input data.

The nearest neighbor method *Jarvis-Patrick* is also frequently used in chemoinformatics.[44] This method takes the two parameters j and k as input. For every molecule, the j nearest neighbors are calculated. Two molecules are clustered together if they are *vice versa* members of their j nearest neighbors and if they have k of their nearest neighbors in common. Again, the advantage is the computational speed. However, choosing appropriate values for j and k is difficult. Furthermore, *Jarvis-Patrick* clustering often produces few large clusters and has a high singleton rate.[45]

Hierarchical clustering methods can be subdivided into agglomerative and divisive techniques. Agglomerative ones iteratively cluster the initially unassigned input data by binary merge steps of next nearest neighbors. In contrast, divisive techniques start with one single cluster comprising all input molecules and divide it by iteratively splitting up clusters. As the agglomerative methods are mostly used in chemoinformatics we will briefly introduce their main representatives, namely the *Ward* method and *Average Linkage*.[40,46] Both methods are based on the same algorithm, but differ in their definitions of similarity between objects. Starting with an unclustered input data set, the most similar molecule pair is merged into a new cluster. Recalculation of the similarity of the newly formed cluster to all other objects and again merging the most similar pair is repeated until only one cluster is left. In the case of *Average Linkage*, the similarity between two clusters is the average similarity of all inter-cluster molecule pairs. The *Ward* method tries to minimize the increase in variance when merging two clusters. To create a final cluster assignment the hierarchical clustering tree has to be cut at a certain level, which is a disadvantage of these methods. Due to a space complexity of $O(N^2)$ and a time complexity of $O(N^3)$, naive implementations of hierarchical methods are infeasible for large data sets. This fact is their major disadvantage. However, in their performance to cluster related molecules together the hierarchical methods have been shown to perform better than *Jarvis-Patrick*, which is routinely used for large data sets.[47,48]

Molecular Scaffolds

The scaffold of a small molecule can be described as the core moiety, which confers the molecule's 3D geometry. However, various scaffold definitions have been proposed and even medicinal chemists have diverging opinions on how to define a molecule's scaffold. The most important formal scaffold definitions were introduced by Bemis and Murcko as well as by Schuffenhauer *et al.*.[49–51]

Both methods distinguish between atoms of the molecular framework – the scaffold – and side-chain atoms forming the decoration. The basic definition treats all atoms within cycles as part of the framework as well as all atoms lying on a direct path between cycles. Additionally, direct neighbors of these atoms belong also to the framework if the connecting bond order is greater than one, because bonds with these hybridization states contribute to the molecule's rigidity. The remaining atoms form side-chains. An example for a small molecule and its *Murcko scaffold* is shown in Fig. 2.4. The method of Schuffenhauer is called *scaffold tree*. It fragments a small molecule in a hierarchical way according to a set of meaningful chemical rules. The first step also removes decorations and thus yields a *Murcko scaffold*. Every further step removes one ring system and finally leaves a single ring at level 0. A *scaffold tree* decomposition is also shown in Fig. 2.4.

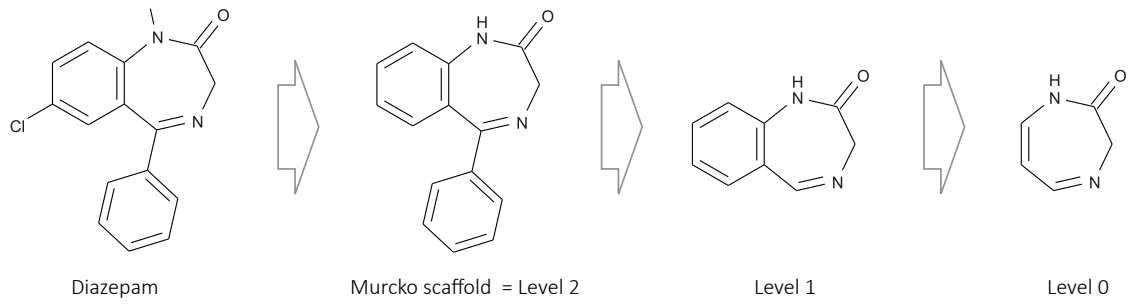


Figure 2.4: Scaffold decomposition of Diazepam. This example follows the *scaffold tree* rules and produces the corresponding hierarchy levels. The first level above the native molecule is its *Murcko scaffold*.

2.2.2 Structural Bioinformatics

In comparison to chemoinformatics, structural bioinformatics also deals with the macromolecular structure of drug targets. These structures are on the one hand used to predict binding pockets and their druggability. On the other hand they are used to predict the binding of small molecules to a target's binding pockets or to simulate and study their flexibility. If no structure is available, structural bioinformatics also tries to generate suitable models, which can be used to perform the mentioned tasks.

Protein-Ligand Docking

In this work we make extensive use of protein-ligand docking and thus give a brief introduction into it. Protein-ligand docking in general is the prediction of ligand-binding to a target receptor. In theory, this prediction comprises two interlocked steps, namely *docking* and *scoring*. The docking task is a sampling problem, which has to generate meaningful ligand conformations in a binding pocket. Various algorithms have been proposed and they differ in several key aspects like the treatment of ligand and/or receptor flexibility and the algorithms for conformational searches.[52]

The scoring has two major tasks: (1) identification of the native protein-ligand conformation(s) for a single ligand by discriminating 'true' and 'false' binding modes. (2) Prioritization of the 'true' protein-ligand conformations from different ligands in order to rank them by their experimentally derived binding affinities. These tasks are accomplished by a so-called scoring function, which is used to predict the binding affinity of a given protein-ligand complex (e.g. a docking pose). It evaluates intermolecular interactions and intramolecular terms of a complex and calculates a score.

Three types of scoring functions can be distinguished. First, empirical scoring functions, which describe the binding free energy ΔG_{bind} as the additive contribution of physically motivated energetic interaction terms. The coefficients of such an additive function are empirically adjusted to fit experimentally determined data sets (complex structures + corresponding binding affinities) by regression methods. Frequently used energetic terms describe electrostatic, ionic, aromatic, hydrophobic contacts, and entropic contributions. Examples are the scoring function developed by Böhm or Friesner.[53,54] The second group comprises knowledge-based scoring functions like Potential of Mean Force or DrugScore.[55,56] Here, the probabilities for the interaction of defined atom type pairs are extracted from structural databases and turned into pair potentials to score protein-ligand complexes. These scores do not estimate experimental ΔG_{bind} values but they usually correlate with them. The third type of scoring function is based on force fields from molecular mechanics. However, this type is not frequently used.

2.3 X-ray Crystallography

Structure-based approaches for the development or improvement of potential therapeutic agents are heavily based on 3D structures of their target proteins. The predominant techniques to determine protein structures are NMR and X-ray crystallography. As of April 2013, X-ray data account for 79,104 (88 %) and NMR data account for 9,892 (11 %) structures in the Protein Data Bank (PDB) as the central repository for macromolecular 3D structures.

The structures presented in this thesis were all determined by X-ray crystallography. This technique is based on the observation that X-rays are scattered by atom electrons and that pe-

riodic 3D arrangements of a macromolecule can lead to an amplification of this signal through constructive and destructive interference of scattered X-rays. In the following subsections, X-ray structure determination is described with a focus on protein crystallization, diffraction experiments and structure determination. The theoretical background is mainly based on the textbooks of Rupp and Drenth.[57,58]

2.3.1 Protein Crystallization

A periodic 3D arrangement of a molecule suitable for a diffraction experiment can occur in crystals. These can be grown in crystallization experiments. Its prerequisite is a sufficient amount of purified protein, which in general is produced by heterologous expression. The aim of crystallization experiments is to increase the concentration of a protein in solution up to a point of supersaturation, where it can undergo transition into solid phase. In case of a successful experiment, the resulting solid phase is a crystal and not amorphous precipitate.

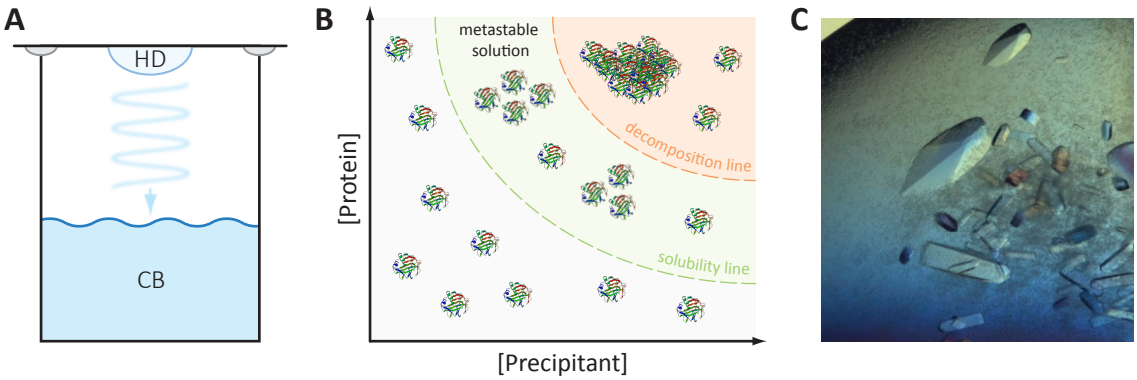


Figure 2.5: Crystallization experiment. **(A)** Hanging-drop crystallization. Crystallization buffer (CB) in a reservoir. The hanging-drop (HD) is a mixture of protein and crystallization buffer. Water diffuses from the lower concentrated drop into the reservoir. **(B)** Crystallization diagram relating protein and precipitant concentration to solution phases. In the metastable phase, crystal germination and growing can occur. Decomposition is the process of unstructured phase separation leading to protein precipitate. **(C)** Protein crystals of 14-3-3 protein in complex with PPI inhibitor.

The experimental technique we use to grow protein crystals is the hanging-drop vapour diffusion shown in Fig. 2.5A. Here, a small volume of a crystallization buffer from a reservoir is mixed to equal amounts with the protein solution. This mixture is positioned on a silicone-coated glass slide, which hermetically closes the greased aperture of the reservoir. Mixing reservoir and protein solution yields a lower concentrated protein solution in the hanging-drop. As a consequence, water diffuses from the drop back into the reservoir, leading to a volume loss of the drop and thus to an increasing concentration of drop ingredients. The crystallization buffer is usually a mixture of a buffering component to adjust the pH value, a precipitant and optional additives. The solution's pH value influences the charge distribution

on the protein surface. In a crystal, intermolecular contacts between neighboring proteins are usually weak and sparsely distributed, making protein crystals unstable. Thus, an appropriate charge distribution on the protein surface is necessary to enable crystal formation. Precipitants are chemicals that lower protein solubility with increasing concentration. Commonly used precipitants are salts, organic molecules like polyethylene glycols (PEG) or polyalcohols. Additives can be in principle all kinds of substances. They are used for optimization in cases where already identified crystallization solutions lead to crystals with insufficient quality for a successful diffraction experiment.

The crystallization experiment leads to an increased concentration of drop ingredients, which at the same time slowly decreases the protein solubility in the drop. These process can be outlined in a crystallization diagram as shown in Fig. 2.5B. This simplified diagram shows three protein phases. With increasing concentrations the protein enters a metastable phase, which separates stable solubility from an unstable phase at high concentrations where protein spontaneously precipitates. In the metastable phase, the system is not at equilibrium and supervening kinetic processes can induce spontaneous nucleation events. If such seed crystals exceed a critical size, further crystal growth can lead to macroscopic protein crystals and can bring the system back to equilibrium. Successfully grown crystals are shown in Fig. 2.5C. A successful crystallization condition for a specific protein has to be identified by trial and error.

X-ray irradiation damages proteins by free radical formation leading to defective crystals and finally can wipe out its diffraction power. This is prevented by cryo-conservation, where crystals are shock-frozen in liquid nitrogen and diffraction experiments are performed at a temperature of 100 K.

Protein-Ligand Crystallization

In structure-based drug design, crystal structures of proteins with bound ligands are of high value. These protein-ligand complexes can be generated by two different techniques, which are variants of the crystallization protocol described above: (1) by *co-crystallization*, which works by adding the ligand to the crystallization buffer. In a successful experiment the ligand binds to the protein and this complex undergoes crystal formation. (2) By *soaking*, where ligand solution is added to drops with existing protein crystals. Under successful conditions ligand molecules diffuse into the crystals and bind to surface-exposed pockets.

2.3.2 X-ray Diffraction

In a diffraction experiment, a frozen crystal is placed in the path of a monochromatic X-ray beam in front of a detector sensitive to X-ray photons. After a certain exposure time and rotation of the crystal a unique diffraction pattern emerges on the detector as shown in Fig. 2.6C. Depending on the crystal quality, sharp spots of varying intensity become visible.

They contain information on the crystal's geometry, its protein content and the resolution of the 3D structure. X-rays are electromagnetic waves, which can be expressed as complex wave vectors of the form:

$$\mathbf{F} = Fe^{i\varphi} \tag{2.4}$$

with φ being the phase angle and $F = |\mathbf{F}|$. The electric field vector of an X-ray wave interacts with electrons of an atom and the wave is scattered with a certain probability into a certain direction. This process is specific for the number of electrons and thus for the atom type. In fact, scattering by atoms is described by the atomic scattering factor f_s and can be calculated by integration over the atoms electron density $\rho(r)$

$$f_s = \int_{\mathbf{r}}^{V} \rho(r) e^{i(2\pi \mathbf{Sr})} dr = \mathscr{F}[\rho(r)] \tag{2.5}$$

where $\mathbf{S}$ is the vector difference of the incoming and scattered wave vector and $2\pi\mathbf{Sr}$ is the relative phase of a partial wave scattered by a certain part of $\rho(r)$. The right part of Eq. 2.5 indicates that f_s corresponds to the Fourier transformation ($\mathscr{F}$) of the atoms electron density. This is of great importance for reconstruction of electron density from diffraction data. A diffraction image is an interference pattern resulting from summing up the scattered X-ray waves of all atoms in a crystal. The relation of crystal geometry and content to the occurrence, position and intensity of spots will be described briefly.

A lattice can be imposed on the periodic protein arrangement in a crystal. This crystal lattice is an infinite stacking of identical points and belongs to one out of six primitive 3D lattices. Fig. 2.6A shows a 2D projection of a crystal with four unit cells (black lattice). It also shows a set of imaginary parallel lattice planes (grey), which slice the crystal lattice into periodic slabs. In principle, an infinite number of such planes can be constructed and systematically enumerated by so-called Miller indices h, k, and l. These mathematical concepts enable the interpretation of spots in a diffraction image as reflections of an X-ray wave by lattice planes. Accordingly, these spots are termed *reflexes*. A single reflex can be assigned to a single set of lattice planes. Thus, reflexes are also identified by the Miller indices of their corresponding lattice planes. The most important equation to relate reflecting positions to their corresponding set of lattice planes is given by Bragg's law:

$$n\lambda = 2d_{hkl} \sin\theta \tag{2.6}$$

Here, n is an integer value, λ is the wavelength of the incoming X-ray beam, d_{hkl} is the distance of lattice planes (hkl) and θ is half the reflecting angle. A graphical illustration of Bragg's law is shown in Fig. 2.6B. Additionally, the equation explains that maximum constructive

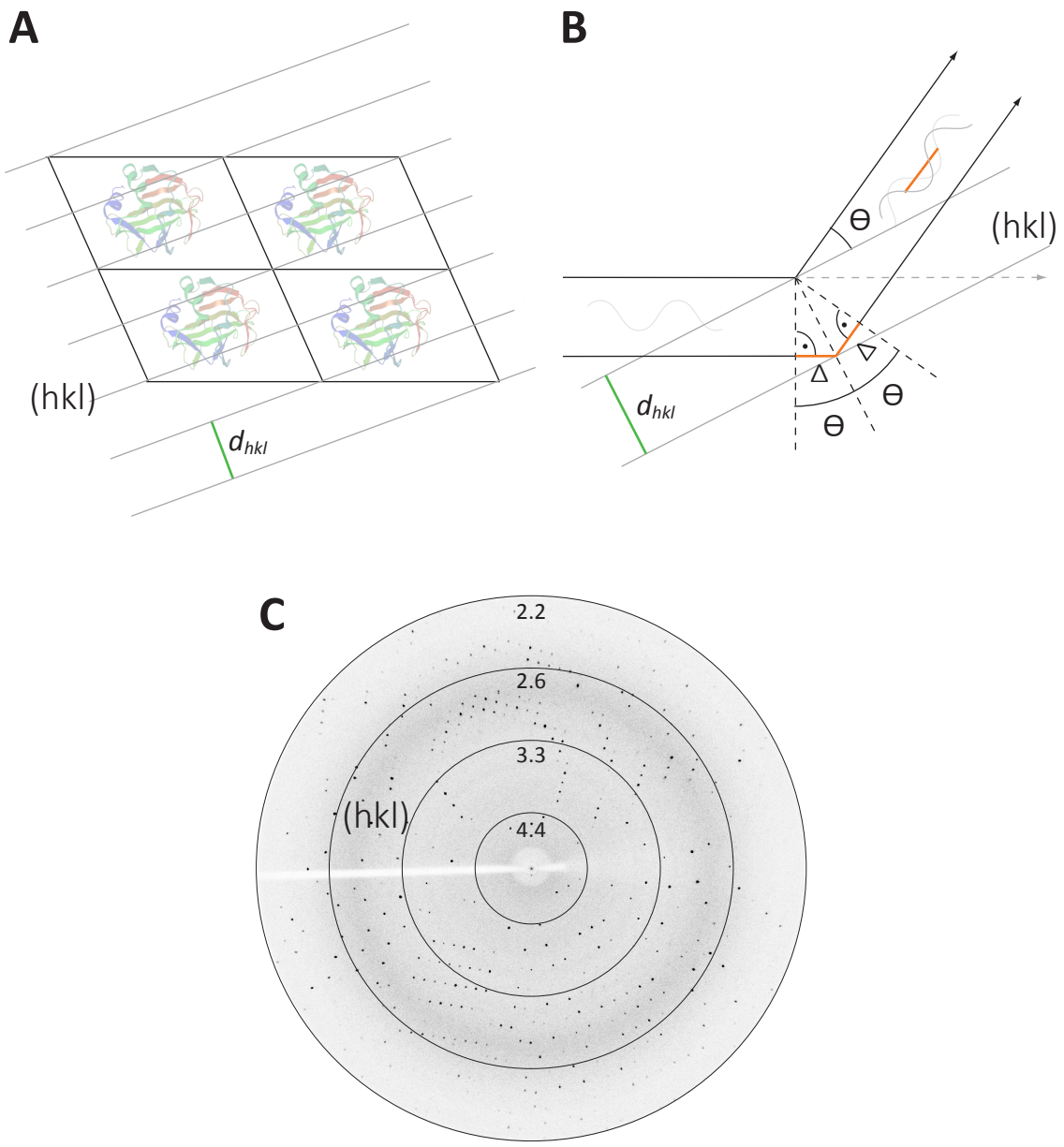


Figure 2.6: X-ray diffraction by a crystal. **(A)** 2D projection of a crystal consisting of four unit cells (black lattice). One set of parallel lattice planes (grey) identified by Miller indices (hkl) with plane distance d_{hkl}. **(B)** Reflection of an X-ray wave by lattice planes (hkl) with reflection angle θ illustrating Bragg's law. The orange lines equal the path difference Δ of a wave reflected by two adjacent lattice planes. **(C)** A diffraction image with sharp reflexes resulting from maximum constructive interference

interference occurs when the path difference Δ (orange line) is an integer multiple of the X-ray wavelength. This is a necessary condition for origination of reflexes and integer values fulfilling these conditions in all three dimensions can again be interpreted as the Miller indices.

As already mentioned, the total scattering of a crystal is the sum of all scattering contributions of atoms in the crystal. This so-called complex structure $\mathbf{F_h}$ is calculated as follows:

$$\mathbf{F_h} = \sum_{j=1}^{|Atoms|} f_{s,j} e^{i(2\pi \mathbf{h} \mathbf{x}_j)} \propto \sqrt{I_\mathbf{h}} \tag{2.7}$$

where $f_{s,j}$ is the atomic scattering factor of atom j, $\mathbf{h}$ the scattering direction as vector of Miller indices and $\mathbf{x}_j$ the fractional coordinate vector of atom j in the unit cell. The right half of Eq. 2.7 indicates that $\mathbf{F_h}$ is proportional to the observed intensity $I_\mathbf{h}$ of the reflex with Miller index $\mathbf{h} = (h,k,l)$. These fundamental equations can be used to recalculate the electron density at discrete points in the unit cell from experimentally measured intensities:

$$\rho(\mathbf{x}) = \frac{1}{V} \sum_\mathbf{h} \mathbf{F_h} e^{-i(2\pi \mathbf{h} \mathbf{x}) + i\alpha_h} \tag{2.8}$$

Here, $\mathbf{x}$ is a grid point (x,y,z), $\mathbf{h}$ the reflex (h,k,l), $\mathbf{F_h}$ its structure factor amplitude and $\alpha_\mathbf{h}$ its associated phase. This equation points out that the phase angle $\alpha_\mathbf{h}$ of reflex $\mathbf{h}$ in addition to its amplitude $\mathbf{F_h}$ is necessary to reconstruct electron density. However, $\alpha_\mathbf{h}$ cannot be measured by X-ray detectors and this information is lost during a diffraction experiment. This fact is known as the *phase problem* of X-ray crystallography and different techniques have been developed to solve it. The next subsection describes the molecular replacement (MR) method, which was used to solve all crystal structures presented in this thesis.

2.3.3 Structure Determination by Molecular Replacement

The aim of MR is to solve the crystallographic phase problem by using a model structure similar to the protein of interest to retrieve initial phases. This strategy is based on the observation that proteins with a sequence homology above ~ 30 % usually possess a conserved fold. Thus, if a structural homolog of the crystallized protein is known, its 3D model can be used to search for suitable arrangements in the unit cell. In a brute force approach, a six-dimensional search, that is rotation and translation in parallel, could be performed to find an optimal model placement. The placement quality is measured by correlating observed structure factors (F_{obs}) and calculated structure factors (F_{calc}) derived from the model.

Due to the combinatorial explosion of brute force MR more efficient strategies are used, which split the six-dimensional search up into consecutive rotational and translational searches. Although no structure factors can be calculated directly in the rotational search to score possible solutions, so-called *Patterson functions* can be used to identify correct spatial orientations of the search model. Patterson maps are calculated without phase information and have peak values at the tips of inter-atomic distance vectors. Additionally, Patterson functions are sym-

metric to their origin. The latter corresponds to the zero distances of all atom self-bijections. Rotational search works by calculating the Patterson maps for intramolecular distances from observed data and calculated model structure factors and to search for optimal peak alignments. The best solutions are forwarded to translational searches, which use intermolecular Patterson functions to find optimal peak alignments. The top solutions from rotation-translation searches are optimized by rigid-body refinement and finally yields the MR solutions.

Refinement and Model Building

Initial models from phasing procedures have to be corrected and optimized in an iterative working cycle. The latter consists of manual model building followed by automatic refinement of model parameters. Manual model building is performed by adjusting the current model to electron density maps in cartesian space. This comprises amongst others placement of correct residues, adapting side-chain conformations, adding solvents or even the rebuilding of whole domains. Subsequent automatic refinement procedures try to optimize model parameters by minimizing the target residual (R) functions:

$$R = \frac{\sum |F_{obs} - F_{calc}|}{\sum F_{obs}} \tag{2.9}$$

where F_{obs} are the measured structure factor amplitudes and F_{calc} are the amplitudes of the calculated model structure factors. To detect and avoid over-fitting, a subset of usually 5 % of the measured reflexes is excluded from refinement and used to evaluate the model quality using Eq. 2.9. The corresponding residual function is termed R_{free} and is related to the mean phase error. In contrast, the reflexes used for refinement yield the crystallographic R_{work}.

Atomic Displacement and Occupancy

The arrangement of protein atoms in a real crystal lattice is not strictly periodic because of two main reasons. First, thermal vibration causes atoms to oscillate around their mean position. Second, crystal lattices are not perfectly periodic due to protein disorder and imperfect crystal growth. An important parameter related to this atomic displacement used during refinement is the *isotropic B-factor* or *temperature factor*:

$$B_{iso} = 8\pi^2 \langle u_{iso}^2 \rangle \tag{2.10}$$

B_{iso} is the isotropic B-factor and $\langle u_{iso}^2 \rangle$ is the mean square isotropic displacement in Å^2 of an atom from its equilibrium position. Thus, B_{iso} is directly related to $\langle u_{iso}^2 \rangle$.

A further important parameter is atom *occupancy*, which is influenced by two major effects. First, solvent content like ligands or ions are not necessarily occurring in all possible positions

and unit cells. Second, protein parts can have varying positions. Especially amino acid side-chains can adopt different conformations and the conformation of symmetry related amino acids follow a typical distribution. This also leads to crystal atom positions that are only partially occupied. Both effects contribute to the occupancy of an atom, which is expressed as the fraction of actually occupied positions.

B-factor and occupancy are listed for every single atom in a PDB file and they provide important information on the reliability of atom occurrence and position.

2.4 Modulation of Protein-Protein Interactions

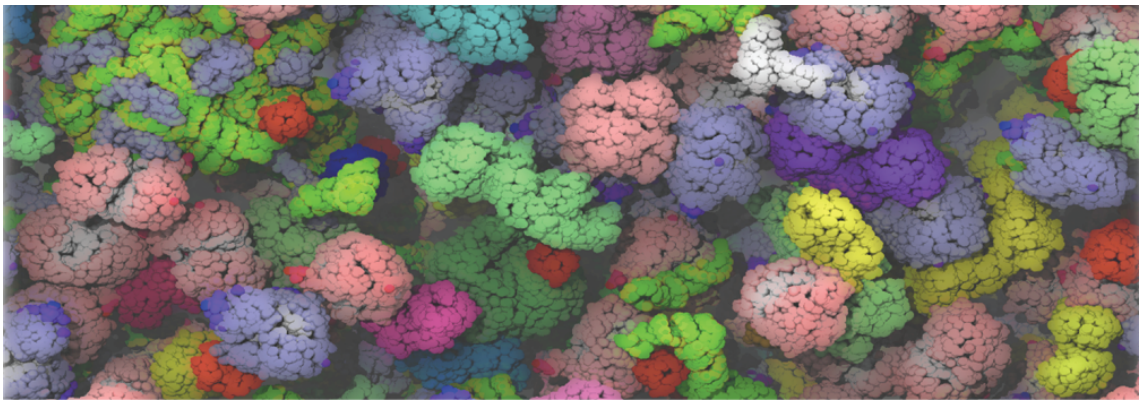


Figure 2.7: Snapshot of cytosolic protein crowding from a molecular simulation. The image is a detail taken from McGuffee and Elcock (*PLOS Comp. Biol.* (2010), **6**, e1000694).[59]

The human genome has been approximated to contain about 20,500 protein-coding genes.[60] About two-thirds of these proteins are constitutively expressed in human cells and the fraction of macromolecules in living cells is about 20-30 %.[61] The concentration of protein together with RNA was experimentally determined to be up to 400 grams per liter where protein makes up 80 % of it.[61,62] Thus, the interior of a cell is highly crowded as illustrated in Fig. 2.7 and as a simple consequence proteins are constantly involved in PPIs. In the recent past, the modulation of PPIs has been accepted as a possible strategy in drug discovery.[22,63] In the following subsections, we will describe the most important characteristics of PPIs relevant for their modulation. Stabilization and inhibition of PPIs are discussed with a focus on computational approaches.

2.4.1 Protein-Protein Interactions

PPIs are of fundamental importance for all living organisms. The underlying association of proteins into functional complexes as well as their dissociation is a highly dynamic process, which is regulated by various cellular mechanisms as shown in Fig. 2.8A.

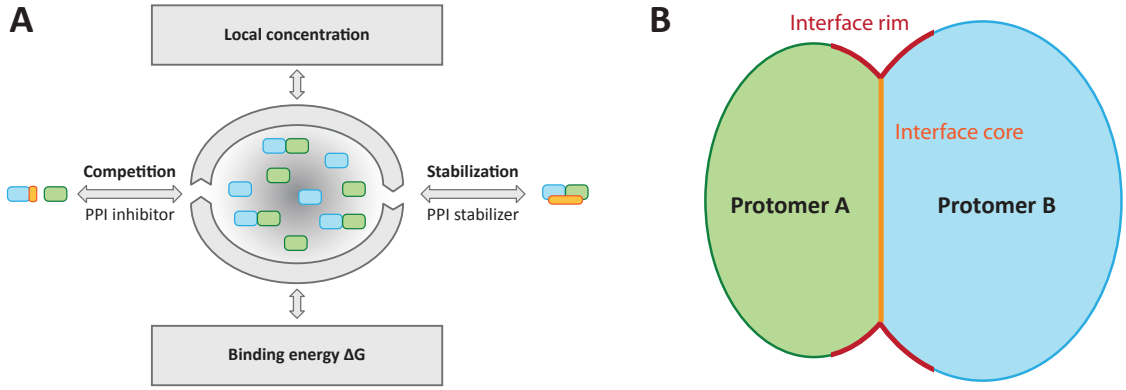


Figure 2.8: (A) Important regulatory mechanisms for the association state of interacting proteins. This equilibrium is regulated by the local concentration of the partners and their mutual binding affinity. External factors can compete for one partner or stabilize the complex. This figure is taken from Thiel *et al.* (*Angew. Chem. Int. Ed.* (2012), **51**, 2012-8).[22] Reproduction is granted under license number 3274110722038 of John Wiley and Sons. **(B)** Schematic representation of a binary heterooligomer.

Classification of PPIs

PPIs are reversible assemblies of protein subunits – so-called protomers –, which form the quaternary structure of proteins. Depending on their composition, PPIs can be classified into homo- and heterooligomers. Homooligomers are built up of identical subunits whereas heterooligomers consist of at least two different protein subunits. Furthermore, protein complexes are classified by their lifetime into permanent and transient PPIs. In permanent PPIs, protomers have a high binding affinity to each other and these complexes are often obligate.[64]

Depending on the protomers' binding affinity, transient comlexes are sub-divided into strong and weak transient PPIs. The oligomerization state of proteins is mainly determined by the mutual binding affinity and the local concentration of the protomers. The binding affinity is influenced by the physicochemical environment like the pH or the ionic strength on the one hand and by posttranslational modifications of the protomers like phosphorylation or methylation on the other hand as illustrated in Fig. 2.8A. The local concentration of protomers is influenced by a variety of factors like co-expression, compartmentalization or degradation.

Structure of PPIs

According to the *O-ring* model proposed by Bogan *et al.*, we will further refer to the *interface core* as the surface residues losing their solvent accessibility upon complex formation and to the *interface rim* as the surface residues in close proximity to the interface core.[65,66] Fig. 2.8B schematically illustrates this classification. A lot of work has been spent on analyzing PPI interfaces and the results vary. A comprehensive overview can be found in the review of Janin *et al.*.[67] The size of interface cores ranges from patches $< 1,000$ Å^2 buried surface area (BSA) to large cores with $> 10,000$ Å^2 BSA with a mean value of $\sim 1,910$ Å^2. The interface cores have been described to be rather flat.[68] Conformational changes in the protomers upon complex formation are more frequent for larger interfaces. These changes range from small variations to whole domain movements. In interface cores, the contribution of single residues to the binding free energy ΔG_{bind} is not equally distributed. So-called *hotspot* residues contribute significantly to complex stability.[65] Furthermore, interface cores show enrichment in aromatic and aliphatic residues and especially of arginine, which makes up 10 % of core and rim residues.

2.4.2 Inhibition of Protein-Protein Interactions

The inhibition of PPIs is considered a promising strategy for the development of drugs for various kinds of diseases.[20,63] However, it turned out that the experimental discovery of PPI inhibitors is quite challenging. Classical techniques like HTS yield low hit rates probably because current screening libraries are biased towards enzymatic active site inhibitors.[21] It has been observed that the latter markedly differ from known PPI inhibitors.[69] Thus, *in silico* approaches can offer beneficial alternatives to identify interaction inhibitors. As described previously, PPI epitopes are quite large, often feature multiple shallow pockets with distinct hotspot residues and they can undergoe conformational changes to different extends. These characteristics make the direct application of standard *in silico* protocols and tools difficult and requires the development of new tools or the adaption of standard protocols.[70] Due to the currently limited number of known inhibitors for a single PPI target, most approaches developed so far are based on the analysis of PPI crystal structures.[71]

A successful example for the computer-aided development of novel inhibitors of the transcription factor and tumor-suppressor p53 and mouse double minute 2 homologue (MDM2), which increases the degradation of p53, was reported by Czarna *et al.*.[72] They identified a tryptophan residue in p53, which is located in the center of the PPI and intrudes deeply into MDM2. Using this residue as an anchor, they constructed a virtual compound library based on an indole scaffold and its bioisosters. Protein-ligand docking led to the identification of MDM2 binders that are able to disrupt the p53∘MDM2 complex.

A similar example was published by Koch *et al.* who identified inhibitors of the thioredoxin reductase interaction with its substrate thioredoxin from *Mycobacterium tuberculosis*.[73] Complex structure analysis led to the identification of a distinctive sequence patch with prominent interactions as constraints for docking a large compound library with low accuracy. Subsequent filter steps and more accurate redocking yielded active PPI inhibitors.

In principle, these examples make use of a single crystal structure, which can be regarded as one representative conformational snapshot of a flexible and dynamic system. Thus it is possible that conformations of non-complexed proteins deviate significantly from their conformation in complex. This would invalidate any binding or pharmacophore hypothesis. Nevertheless, small molecules can bind to a protein by trapping a conformation which is different from that observed in the complexed form as shown for a small molecule inhibitor of Bcl-X_L, which normally binds and inactivates the pro-apoptotic BAD protein.[74] To tackle this problem *in silico*, Eyrisch and Helms developed a molecular dynamics (MD) simulation protocol to analyze for the occurrence of transient surface pockets, which can be addressed with standard tools.[75] In the case of Bcl-X_L, they also demonstrated accurate redocking of a known ligand into a transient pocket derived from MD simulation.

2.4.3 Stabilization of Protein-Protein Interactions

The stabilization of PPIs by small molecules is still in its infancy and not as well established as PPI inhibition. The basic principle of PPI stabilization is to enhance the binding affinity of interacting proteins. Fig. 2.9 illustrates this mechanism. Here, ligand L serves as *molecular glue* and binding to the PPI leads to a decrease in apparent binding energy ΔG_{AB}.

Two general modes of action can be observed for PPI stabilizers. First, a stabilizer can bind to a single protein partner, thereby increasing the mutual binding affinity of the protein partners in an allosteric fashion. Second, the stabilizing molecule binds to the interface rim of a protein complex, making contacts to both protein partners thereby decreasing the mutual binding energy as well. Correspondingly, we termed the different types allosteric (one protein partner) and direct (at least two protein partners) PPI stabilizers, respectively.

Allosteric Stabilization

Allosteric PPI stabilization has so far been described only for the interaction of the α- and β-tubulin heterodimer. The $\alpha\circ\beta$-heterodimers assemble to linear protofilaments and form cylindrical polymers – the Microtubules (MT). MTs have important functions in non-dividing and dividing cells. To fulfil the different tasks, MTs have to be permanently rearranged by continuous polymerization and depolymerization.[76] Disturbing this process has severe consequences for a cell, particularly during division, where MTs form the mitotic spindle apparatus that segregates the chromosomes.[77] Several natural products and derivatives induce

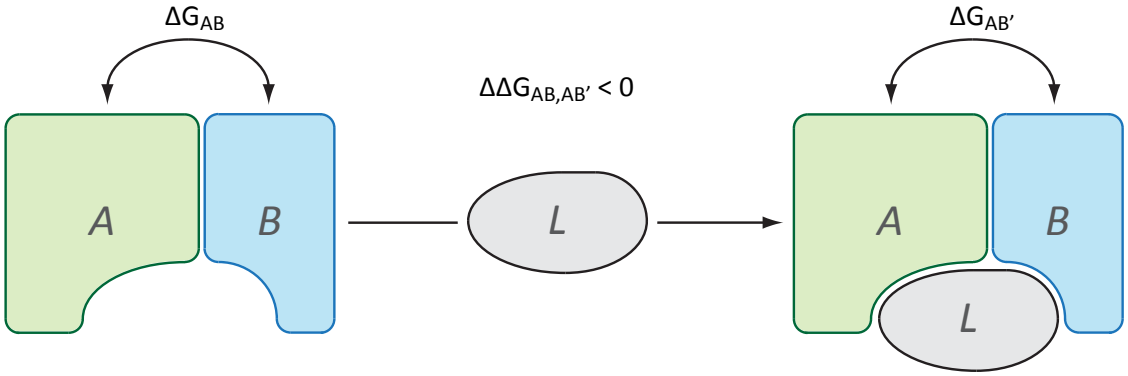


Figure 2.9: Schematic illustration of direct PPI stabilization.

cell cycle arrest by modulating MT polymerization and depolymerization leading to severe cellular disturbances and even to apoptosis. Thus, some of these molecules are used as antimitotic agents and belong to the most important drugs in the treatment of cancer.[78] One of the most intensely studied MT stabilizers is paclitaxel, which is isolated from the bark of *Taxus brevifolia*.[79] It binds with high affinity to a hydrophobic pocket of MTs, which is exclusively located on the β-subunit and thereby allosterically stabilizes the MT.[80]

Direct Stabilization

If a stabilizing small molecule simultaneously contacts multiple chains of a PPI, the mechanism is termed direct stabilization.[22] This can be further split up in two different modes of action. First, the stabilizing molecule binds to one of the proteins and creates an interaction surface for the second protein. This stabilizing effect can be so strong that dimerization of two proteins can be induced, even though they do not bind to each other in the absence of the ligand. This extreme case has been observed for the FKBP binding molecules FK506 and rapamycin and will not be further discussed.[81] The second mode of action is characterized by binding of a ligand to the interface rim of a PPI, thereby increasing the apparent PPI binding affinity. Three impressive examples are the natural products fusicoccin A (FSC), brefeldin A (AFB) and forskolin (FOK), which are illustrated in Fig. 2.10A-F.

FSC is a metabolite from a wilt-inducing fungus. Studies on the molecular target resulted in the identification of a complex between the regulatory domain of the plasma membrane H^+-ATPase 2 (PMA2) and 14-3-3 adapter proteins.[82] FSC binds to the interface rim of this complex and enhances the apparent affinity of the proteins about 90-fold.[83] It fills a hydrophobic gap in the interface of the two proteins. A central terpene ring is deeply buried in a funnel-like pocket formed by 14-3-3 and the absolute C-terminus of PMA2. Two chemically diverse compounds, unrelated to FSC, were identified as stabilizers of 14-3-3∘PMA2 in an HTS.[25]

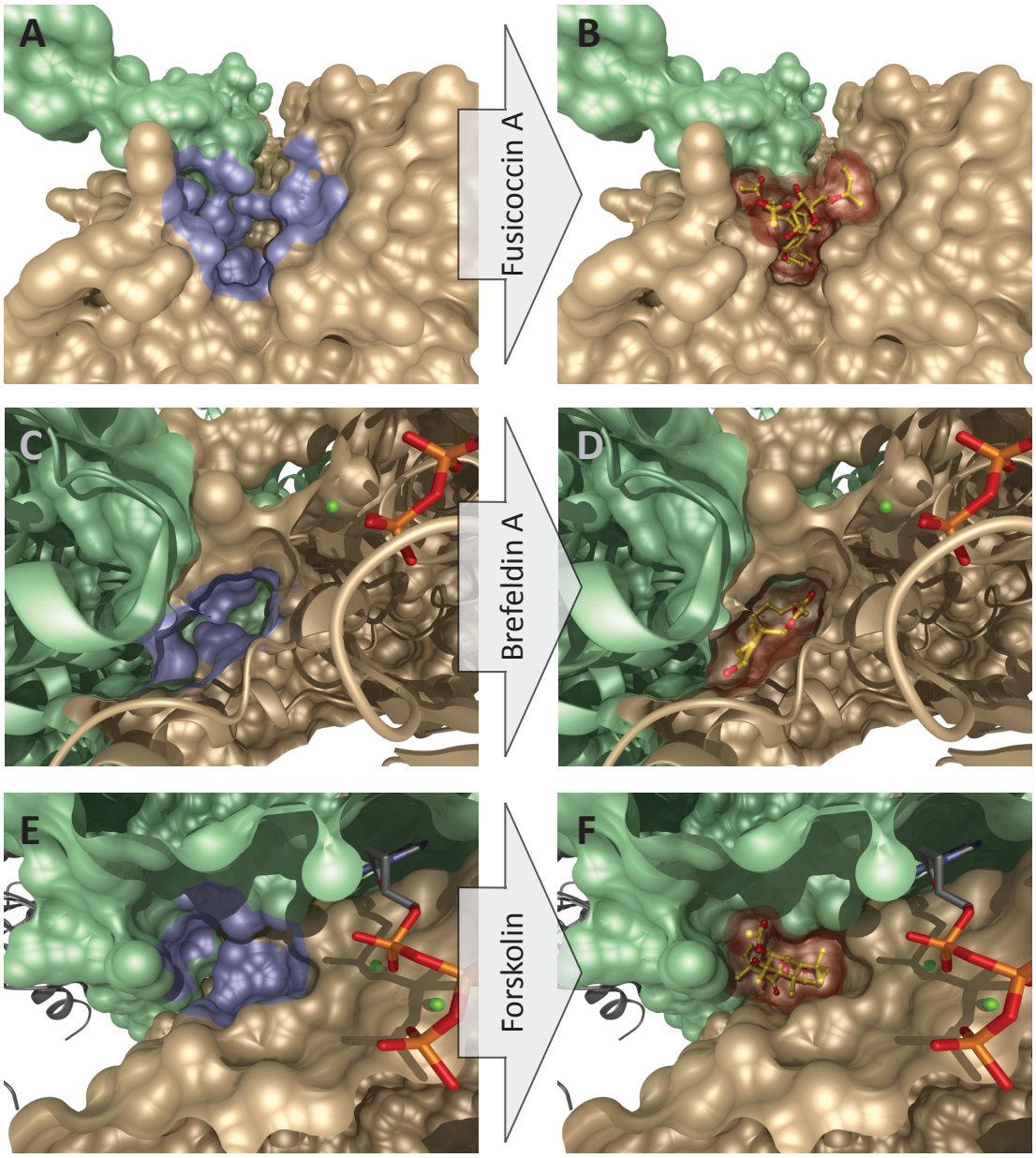


Figure 2.10: Examples of stabilized PPIs. Ligands are represented as ball-and-stick models and a semitransparent SES. **(A)** Binary complex of 14-3-3 protein (gold SES) and C-terminal domain of PMA2 (green SES). The FSC pocket is highlighted in blue. (PDB ID: 2o98) **(B)** FSC bound to the binary 14-3-3∘PMA2 complex. **(C)** Section through the complex of ARF1 (gold SES and cartoon) with a bound GTP analogue (stick model) and a Sec7 domain (green SES and cartoon). (PDB ID: 1r8q) **(D)** AFB deeply buried in the interface rim pocket. **(E)** Catalytic subunit of AC with bound ATP analogue (stick model). The C_{1a} domain (green SES) and C_{2a} domain (gold SES) are shown together with the FOK. (PDB ID: 1cju) **(F)** FOK bound to the interface rim pocket. This figure is taken from Thiel *et al.* (*Angew. Chem. Int. Ed.* (2012), **51**, 2012-8).[22] Reproduction is granted under license number 3274110722038 of John Wiley and Sons.

Both compounds bind to distinct pockets in the interface rim. The dipeptide epibestatin binds to a narrow surface cleft and is tightly sandwiched between the two proteins. The molecule interacts to equal parts with 14-3-3 and PMA2. A trisubstituted pyrrolidone occupies a more solvent-accessible site which substantially overlaps with the pocket of FSC.

The fungal metabolite AFB potently inhibits protein secretion by stabilizing the complexes of the small guanine nucleotide-binding protein ADP ribosylation factor 1 (ARF1) and several guanine nucleotide exchange factors like Sec7.[84] This leads to blockage of the GDP-GTP exchange activity of Sec7 and ultimately results in impairment of Golgi function.[85] The crystal structure of the ARF1-GDP∘Sec7∘AFB complex shows that AFB is deeply buried between the two proteins.[86,87] AFB binding is mostly hydrophobic in nature with few additional polar contacts. Ligand binding leads to 10-fold stabilization of ARF1-GDP∘Sec7. AFB exclusively binds to the ternary complex and no binding to ARF1-GDP or Sec7 alone has been observed.[88]

FOK is a cardioactive and blood-pressure lowering plant metabolite. Its molecular mechanism is a reversible increase of adenylyl cyclase (AC) activity, resulting in significant increase of cAMP levels in various tissues.[89] AC is a transmembrane protein with the cytoplasmic domains C_{1a} and C_{2a} forming the catalytic core.[90] FOK increases the apparent affinity of these subunits from a $K_D > 10$ μM down to 1 μM and results in a 60-fold enhanced catalytic activity of AC.[91] FOK binds to a deep and primarily hydrophobic pocket terminating a long cleft in the interface rim of the C_{1a}∘C_{2a} dimer. It shares equivalent contacts to both protomers, buries ∼ 90 % of its accessible surface and closes a hydrophobic pocket between the subunits.[92,93]

Computational campaigns to discover PPI stabilizers are rare and the published examples are described in Chapter 5, where we present our approaches to identify stabilizers *in silico*.

2.5 The Family of 14-3-3 Proteins

For our attempts to identify PPI modulating small molecules *in silico*, we used human 14-3-3 proteins as a model system. These proteins are versatile molecular adapters and their main function is to bind other protein partners. Several 14-3-3 PPIs have been shown to be related to diseased cellular states where PPI inhibition or stabilization with small molecules could be a valuable concept for therapeutic intervention. The following subsections briefly describe the physiological functions, the structure and the mode of action of 14-3-3 proteins.

2.5.1 Physiological Functions

Discovered in 1967, 14-3-3 proteins owe their name to the chromatographic fraction number (14) and the 2D electrophoretic coordinates (3,3) they were isolated from.[94] The homologous family of 14-3-3 proteins is conserved in all eukaryotes. In mammals, the seven homologs β (α), γ, ε, η, σ, τ (Φ) and ζ (δ) are ubiquitously expressed. However, different tissues show

varying expression patterns of these homologs.[95] With an average sequence identity of 46 %, the human homologs are highly conserved as shown in Fig. 2.11.

The basic task of 14-3-3 proteins is binding to other proteins. Up to date, more than 140 human binding partners have been identified.[96] Almost all of these partners bind phosphorylation dependent to 14-3-3 . The phosphorylated residue is either a serine or a threonine.[97] Additionally, a few target proteins have been identified where binding to 14-3-3 is phosphorylation-independent.[98–100] 14-3-3 proteins do not exhibit an intrinsic enzymatic activity. 14-3-3's activity becomes manifest in influencing the subcellular localization or the co-localization with other proteins, or the conformation of their partner proteins.[101]

Due to these versatile roles and number of targets, 14-3-3 proteins directly or indirectly take part in the regulation of all major cellular processes including cell-cycle, transcription, protein biosynthesis, signal transduction, and apoptosis.[97,105,106] Thus, various diseases are linked to the 14-3-3 interactome and PPIs within this network are considered potential tar-

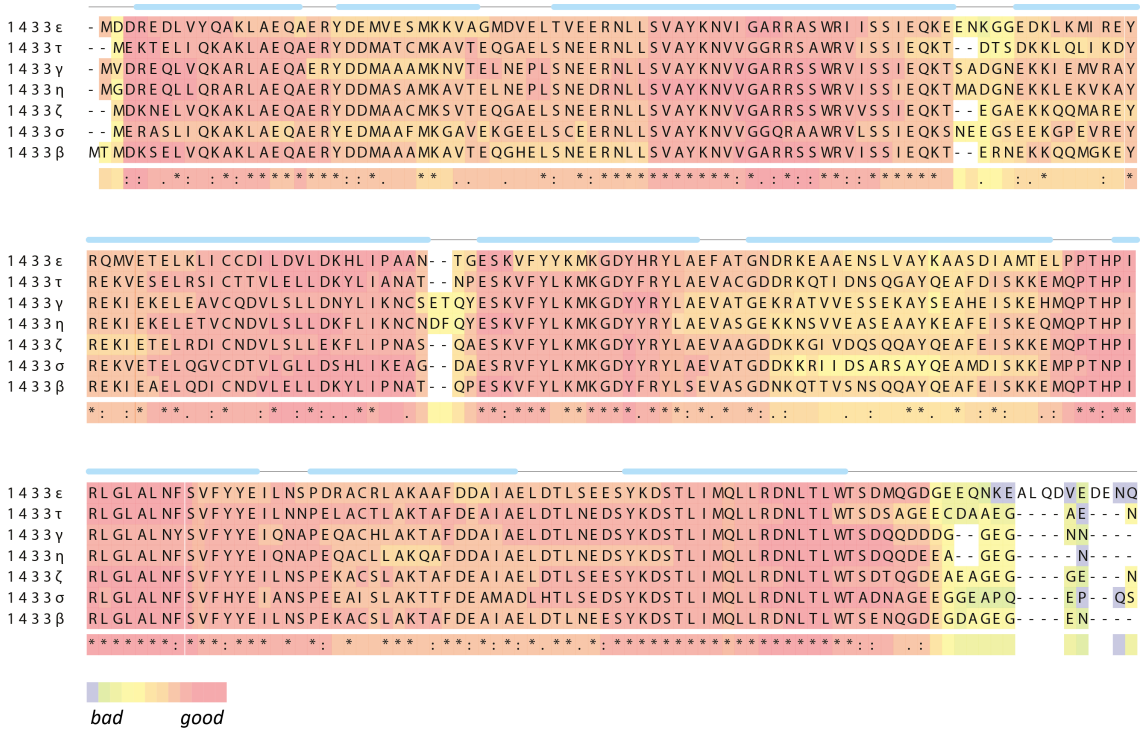


Figure 2.11: Multiple sequence alignment (MSA) of the human 14-3-3 homologs. The top line marks the secondary structure elements (coils and turns: grey line, α-helices: blue bars). The bottom line assigns conservation information (entirely conserved: '*', amino acids with same size and hydropathy: ':', amino acids with comparable size or evolutionary preserved hydropathy: '.'). With the exception of the C-terminal coil 14-3-3 proteins are highly conserved. The MSA was generated with T-Coffee (version 8.14) accessed through the Bioinformatics Toolkit of MPI Tübingen.[102,103] Secondary structure annotation was added manually based on the UniProt entry of 14-3-3ϵ (Accession: P62258).[104]

gets.[107] In the following, three 14-3-3 PPIs are shortly described to exemplify the function of 14-3-3 proteins and their potential medical relevance.

Influencing the subcellular localization by 14-3-3 has been shown for many binding partners. For example the human protein kinase C-Raf is a central element of the Ras-Raf-MAPK pathway, which couples extracellular signals to transcriptional regulation.[108] Phosphorylation of C-Raf at Ser259 enables binding to 14-3-3 and inhibits the recruitment of C-Raf to the plasma membrane, thereby suppressing MAPK downstream signaling.[109] Interestingly, C-Raf mutations clustering around Ser259 lower its affinity to 14-3-3, which results in constitutively increased downstream signaling.[110] These mutations are frequently found in patients suffering from *Noonan syndrome*, a severe developmental disorder. Thus, the PPI of 14-3-3 and mutated C-Raf forms a promising stabilizer target for this disease.

Promoting the co-localization of two proteins which do not interact directly is another function of 14-3-3. This mechanism has been observed for several kinases and their protein substrates where 14-3-3 brings enzyme and substrate in spatial proximity. An interesting example is the phosphorylation of the tau protein. Tau is an MT-associated protein in neural tissues with various functions and multiple phosphorylation sites.[111] Glycogen synthase kinase 3β (GSK3β) is one of the kinases phosphorylating tau. However, phosphorylation is mediated by binding of both proteins to 14-3-3 .[112] Tau itself has been found to accumulate in various neurodegenerative disorders like Alzheimer's disease and is discussed as a possible target for treatment of these disorders. As phosphorylated tau also binds directly to 14-3-3 , various starting points for beneficial PPI modulations are possible.[113]

The change of a protein's conformation upon 14-3-3 binding can lead to a modulated enzymatic activity. An example for this mechanism forms the arylalkylamine N-acetyltransferase (AANAT). This enzyme plays a central role in the circadian rhythm and its activity positively influences the production of melatonin.[114] Binding of phosphorylated AANAT to 14-3-3 stabilizes a protein conformation which increases the enzymes affinity to its substrate.[115] Imbalanced melatonin production has been found to be involved in various diseased states like metabolic, sleep or mood disorders.[116,117] Thus, stabilization or inhibition of the 14-3-3∘AANAT PPI could possibly serve as a strategy for drug intervention.

2.5.2 Structure and Mode of Action

The size of 14-3-3 monomers ranges between 25-30 kDa and their physiological assemblies are homo- as well as heterodimers. An important consequence of 14-3-3's dimeric assembly is the possibility to bind two phosphorylation sites of a single protein simultaneously. This has been described for multiple 14-3-3 target proteins including C-Raf.[118,119] On the one hand, synergistic binding increases the overall binding affinity of 14-3-3 to its target. On the other hand, dual binding sites enables fine-grained pathway regulation.

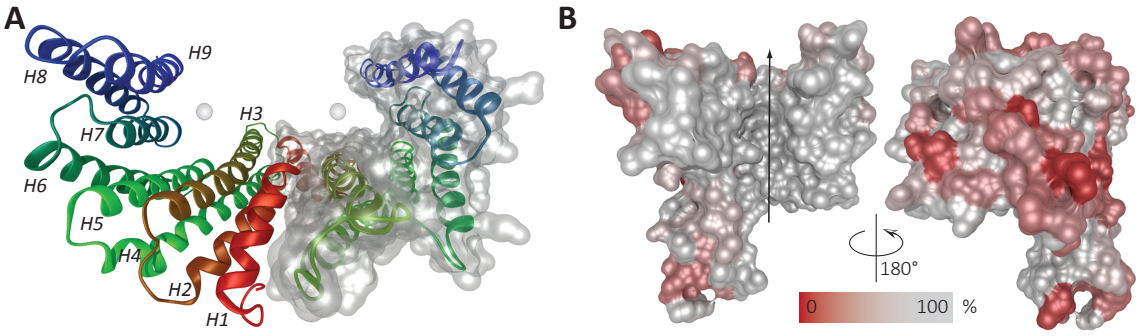


Figure 2.12: Structure and conservation of mammalian 14-3-3 proteins. **(A)** One 14-3-3 dimer is shown in cartoon representation colored by residue index (red: N-terminus, blue: C-terminus, H: helix). The spheres indicate the amphipathic groove, which is enclosed by helices 3, 5, 7, and 9. **(B)** One 14-3-3 monomer represented by it's SES and colored by residue conservation of all mammalian homologs. (grey SES: 0 % conservation, red SES: 100 % conservation) The arrow direction indicates the course of partner proteins from N- to C-terminus.

The crystal structures of all human homologs have been solved and possess the same fold. The average pairwise root-mean-square deviation (RMSD) of the C_α backbone is 0.74 Å. Fig. 2.12 shows different representations of 14-3-3σ. Fig. 2.12A represents a typical W-like shaped 14-3-3 dimer and the monomer in cartoon representation shows 9 constitutive anti-parallel α-helices (H1-H9). The most important structural feature is a longitudinal amphipathic groove enclosed by H3, H5, H7, and H9, which spans the entire protein. The amino acid variability on the 14-3-3 surface is not equally distributed (Fig 2.12B). A mapping of the amino acid conservation onto the protein surface shows no variability in the amphipathic groove. The highest variability can be observed on the outer surface. The flexible C-terminus is not resolved in the crystal structure.

The amphipathic groove forms the functional unit of 14-3-3 proteins because the partner proteins bind into it. Due to its narrow shape, no secondary structure elements like α-helices or β-sheets fit into the groove and partner proteins bind via elongated sequence stretches. As mentioned before, the most important characteristic of almost all binding motifs is a phosphorylated serine or threonine residue. Typically, the binding sequences of known binding partners are classified into one out of three different binding motifs listed in Table 2.1. The mode I and II motifs are mainly characterized by a proline at position +2 relative to the phosphorylated residue.[120,121] In most binding partners, these prolines induce a bend, forcing the sequence out of the amphipathic groove. Mode III motifs are characterized by their location at the very C-terminus of the partner proteins with only one residue following the phosphorylation site.[122]

It is noteworthy to mention that these motifs were defined at an early point in 14-3-3 research on the basis of only a fraction of the binding partners known today.[96] Thus, a lot of

Table 2.1: 14-3-3 consensus motifs. X indicates a variable position, π indicates aromatic amino acids and $COOH$ marks the C-terminus.

Motif	*Consensus sequence*						
	-4	-3	-2	-1		$+1$	$+2$
Mode I		R	S	X	[pS|pT]	X	P
Mode II	R	S	π	X	[pS|pT]	X	P
Mode III					[pS|pT]	X $_{COOH}$	

binding sequences discovered later on do not fit into one of these historical classes like the yes-associated protein (YAP), which spans the entire binding groove or the cyclin-dependent kinase inhibitor 1B where the phosphorylated residue itself forms the C-terminus.[123,124] Structural details of the binding geometry surrounding the phosphorylation site are discussed more detailed in Chapter 6, where it forms the basis for development of a VS approach in order to identify 14-3-3 PPI inhibitors.

Chapter 3

Deterministic Clustering of Large Chemical Spaces

The content of this chapter is an extended version of the article:

Blocked Inverted Indices for Exact Clustering of Large Chemical Spaces. [125]

3.1 Introduction

As described in Section 2.2.1, clustering is one of the most important tasks in chemoinformatics. It has manifold applications in theoretical and in experimental setups. Various methods have been developed to address this problem and the most frequently used are agglomerative hierarchical variants and the non-hierarchical methods *k-Means* and *Jarvis-Patrick*. A major difference between these hierarchical and most non-hierarchical methods is that non-hierarchical ones take all pairwise molecule similarities in a data set into account. The non-hierarchical methods prevent the calculation of the entire similarity matrix at the cost of also not being deterministic. This is especially true for *k-Means* but not for naïve implementations of *Jarvis-Patrick*. However, to gain efficiency, the latter approach often uses non-deterministic heuristic algorithms for nearest neighbor calculation to explicitly avoid all pairwise similarities calculations. [126]

In contrast, the advantage of being deterministic due to complete similarity matrix calculation makes the hierarchical methods quite inefficient and computationally demanding with respect to time and memory requirements. Even when using sophisticated algorithms, the upper limit of input data sets for hierarchical methods lies in the order of 10^6. As a conse-

quence, the methods of choice for clustering larger compound libraries based on 2D fingerprint similarities are still *k-Means* and *Jarvis-Patrick*.

Bearing in mind that the size of chemical libraries comprising commercially available compounds already exceeds 30×10^6 compounds and that systematically generated chemical space with more than 10^9 compounds even is three orders of magnitude larger, it is highly desirable to develop novel deterministic and efficient clustering methods, which are capable of processing data sets of such dimensions.[127,128] A key step to such efficient clustering methods is the availability of fast algorithms to calculate 2D fingerprint similarities like the already introduced Tanimoto coefficient. A naive implementation to calculate the similarity matrix leads to an algorithmic complexity of $O(n^2)$, where n is the number of input molecules and thus is not well suited for huge data sets. After a long period of stagnated algorithmic progress in this field, novel and interesting approaches were presented in the recent past (see also subsection on related work below).

Goals of the Project

In the present work we introduce a deterministic approach to cluster very large chemical spaces comprising tens of millions of compounds. To handle such huge data sets we developed a fast and flexible algorithm for high-throughput calculation of all pairwise similarities of an input compound library based on the concept of inverted index data structures (iiDS). An overview of our clustering workflow is shown in Fig. 3.1 and it is described in the following.

The first step shown in Fig. 3.1A makes use of a non-uniqueness property of binary fingerprints. This is a consequence of the feature generation from substructural patterns, which normally do not encompass an entire molecule. Based on this intrinsic property duplicates of fingerprints are identified and removed in order to reduce the total number of similarities, which have to be calculated in subsequent steps.

Fig. 3.1B shows the second step, which performs a preliminary grouping of the remaining and unique fingerprint set. The basic concept used for this purpose is CC decomposition as described in Section 2.1.2. In our case, every molecule or its corresponding fingerprint represents a graph vertex and an edge between two vertices represents the similarity between the fingerprints of its incident vertices. Due to the symmetry of the Tanimoto coefficient, all edges and thus the graph, are undirected. Initially, the set of edges is empty and edges are gradually inserted into the graph during this step. For this purpose, all pairwise similarities are calculated using our efficient algorithm and all vertex – or molecule – pairs with $S_{Tan} \geq S_{Tan}^{Cut}$ are connected by a newly added edge. Here, the adjustable parameter S_{Tan}^{Cut} is a lower similarity cutoff to discard molecule pairs with too low similarity. Up to this point, this step yields a similarity network, which contains only a subset of all possible edges. Depending on the choice of S_{Tan}^{Cut}, the generated network is disconnected and CCs can directly be calculated. The

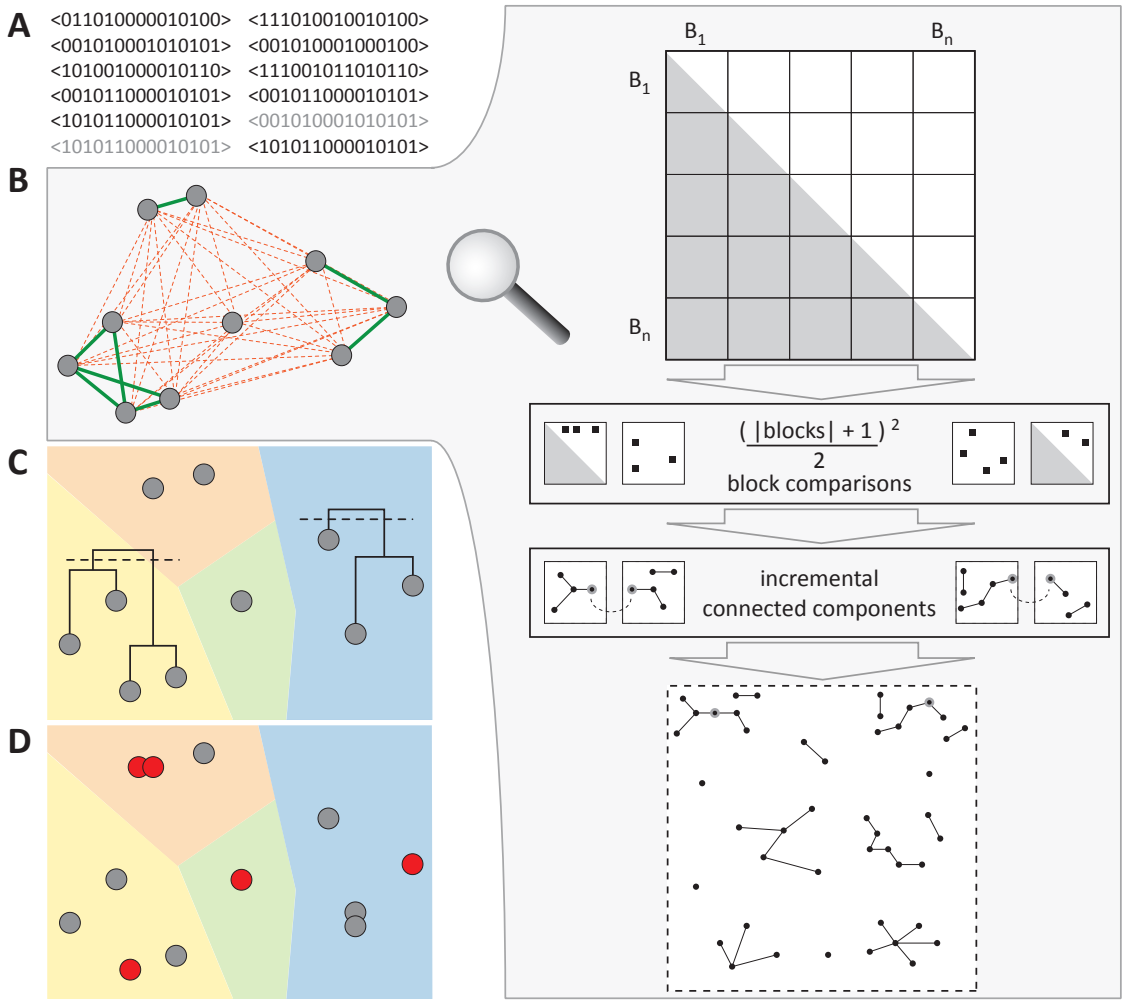


Figure 3.1: Clustering method schema. **(A)** Removal of duplicates and creation of a unique set of input fingerprints. **(B)** Calculation of all pairwise Tanimoto similarities and construction of a similarity network by applying a similarity cutoff to retrieve the induced CCs. The right half sketches the CCs decomposition using blocked inverted indices via construction of the complete similarity matrix. **(C)** Application of hierarchical clustering on large CCs. **(D)** Remapping of fingerprint duplicates onto clusters containing their representative fingerprint.

applicability of CC decomposition as a feasible strategy to cluster chemical libraries has been shown by Zahoránszky *et al.*.[129]

The third step shown in Fig. 3.1C performs hierarchical clustering of every CC, which exceeds a predefined size. As hierarchical clustering we have implemented the agglomerative average linkage method for three reasons. First, the method was shown to perform well for clustering of chemical databases.[47] Second, the average linkage method is compatible with our algorithm for fast similarity calculation. Third, efficient algorithms have been described

for average linkage clustering. Finally, a cluster assignment is generated using automatic level selection for cutting the clustering hierarchy.

In the fourth step shown in Fig. 3.1D, a representative molecule for every cluster is calculated. Again, we use our fast similarity calculation method for this purpose. Finally, the duplicate fingerprints that were removed in step one are merged into the clusters containing their representative fingerprint.

To demonstrate the performance of our clustering method, we process the available chemical space on a current compute server. We compare our fast similarity calculation algorithm as well as the clustering method to state-of-the-art implementations and tools. Furthermore, we present a detailed analysis of the advantages and disadvantages of inverted index methods for 2D fingerprint similarity applications, especially its hardware demands, and compare it to hardware-accelerated approaches.

Related Work

The calculation of molecular similarity on the basis of 2D fingerprints has been in the focus of several studies while this thesis was in progress. Recently published solutions for chemoinformatics applications tackle this problem in two different ways. On the one hand by using specialized machine-dependent instructions on hardware level and on the other hand by algorithms using iiDS.[30]

Hardware-Accelerated Methods. The problem of counting the number of 1-bits in an array is well known and often referred to as the population count (popcount) of an array. Its efficient solution is of such high importance that recent instruction set extensions of modern CPUs (for example SSE4.2 of Intel) and modern GPUs introduced this operation as a single-cycle instruction on hardware level. Haque *et al.* described a thoroughly implemented popcount-based method and a combination with cache-efficient strategies, which show impressive performances.[130] Nevertheless, the described cache-efficient version only provides advantages when calculating highly regular data structures like similarity matrices. Furthermore, the performance of popcount methods drop with increasing fingerprint sizes as will be shown in the results section of this chapter.

Inverted Index Methods. The iiDS arose from the field of information retrieval and is used as a technique to encode text documents for subsequent similarity calculations.[131] Its relationship to tasks in the field of chemoinformatics has recently been described by Nasr *et al.*.[132] They used iiDS based similarity calculations as one key element to speed up similarity searching. An inverted index algorithm to accelerate LINGO similarity calculations has been described by Kristensen *et al.*.[133] However, this work also focused on speeding up similarity searching. Our work is based on an initial inverted index implementation developed by Lisa Peltason within the scope of her Studienarbeit in the group of Prof. Dr. Oliver Kohlbacher.[134]

3.2 Materials and Methods

This section first describes implementation details of our iiDS algorithm for fast 2D fingerprint similarity calculation. It is followed by implementation details of the single steps of the outlined clustering workflow. Finally, we give detailed information on employed software, hardware and data sets for software development, evaluation, and benchmarking.

3.2.1 Implementation Details

The software we developed is implemented in the object-oriented programming language C++ and is incorporated into the structural bioinformatics framework BALL (Biochemical Algorithms Library).[2]

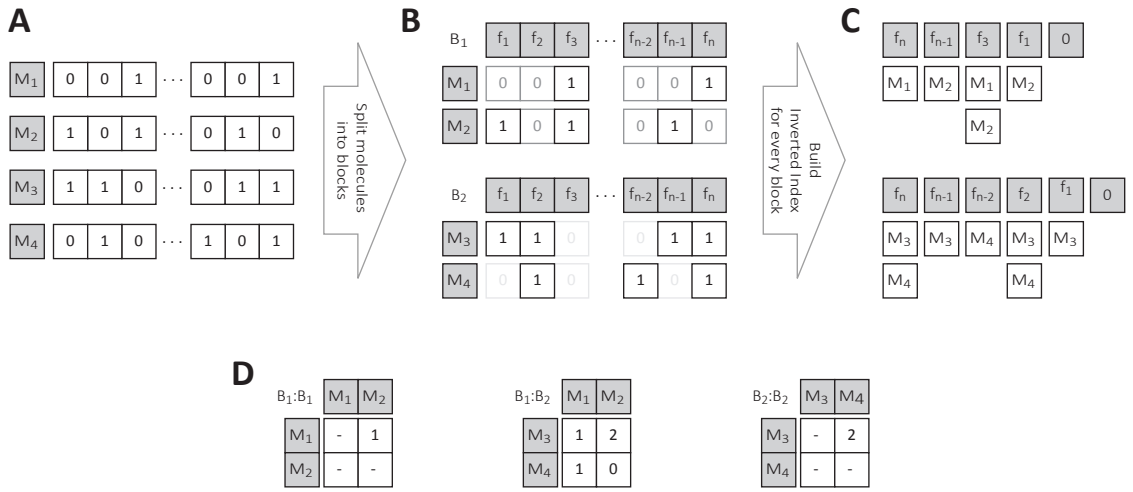


Figure 3.2: Blocked inverted index algorithm. **(A)** The input data set is split up into equally sized blocks. **(B)** A block is a matrix where every row is a single fingerprint and columns are fingerprint features. **(C)** For every block a single iiDS is generated. **(D)** Pairwise comparison of all blockwise iiDS enables efficient shared feature count calculation.

Blocked Inverted Index Algorithm

A major aim was to develop a pure algorithmic solution for the problem of 2D binary fingerprint similarity calculation with no need for specialized hardware. We thus used the concept of iiDS for our similarity calculation and developed an improved and flexible algorithm for pairwise similarity calculations. A simplified overview is shown in Fig. 3.2. As shown in Fig. 3.2A, the algorithm takes as input a set of molecules encoded as 2D fingerprints. Internally, fingerprints are represented as feature lists, which is a list of the 1-bit array indices or higher order substructure hashes. Fig. 3.2B illustrates the key idea of our algorithm, which is to split

the input molecules up into blocks of equal size. The block size is a critical parameter to the algorithm's performance as shown in the results section. We thus call our algorithm the blocked inverted index (bII) method. If the input library contains n fingerprints and a block size of s_b is chosen, we generate a set of blocks $\mathfrak{B}$ with

$$m = |\mathfrak{B}| = \lceil \frac{n}{s_b} \rceil \tag{3.1}$$

members. A single block b_i can be seen as a matrix where each row contains a fingerprint and the columns are the fingerprint features f_i.

In Fig. 3.2C, the generation of an iiDS for every molecule block is shown. In principle, iiDS generation is a reordering of the block matrix. The UML diagram of our iiDS implementation for a block b_i is shown in Fig. 3.3. As can be seen, an InvertedIndex struct has four attributes. The attribute *n_molecules* stores the number of molecules contained in b_i. In general, the number of molecules per block is s_b. However, the size of the last block is the remainder $a \equiv n \bmod s_b$ and could be less than s_b. The attribute *n_features* is an array, which stores the total feature count of every corresponding fingerprint. The optional attribute *parent_clusters* is only used during hierarchical clustering and indicates the cluster a molecule belongs to. Finally, *feature_skip_list* is a pointer to the first instance of multiple FeatureList structs, which are connected as a skip list. A FeatureList struct has two attributes. First, *feature_id* is a unique identifier of a distinct fingerprint feature. Second, *block_position* is an array storing the positional indices of all molecules in b_i. These indices are used during shared feature count calculation to address cells in a matrix storing the shared feature counts for all pairs of molecules between two blocks b_i and b_j. The FeatureLists, which are stored in an InvertedIndex, are decreasingly sorted according to their *feature_id* and terminated by a default FeatureList with *feature_id* $= 0$.

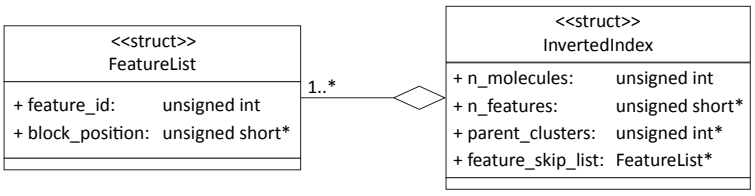


Figure 3.3: UML diagram of inverted index core data structure.

Fig. 3.2D finally shows, how these data structures enable the efficient calculation of shared feature counts c between all pairs of molecules of a processed block pair $\{b_i, b_j\}$. As an important consequence, the time complexity of our algorithm is $O(m^2)$ as opposed to $O(n^2)$ for a naïve pairwise Tanimoto calculation.

Listing 3.1: C++ implementation of shared feature count calculation between two sets of fingerprints represented as iiDS. For clarity, braces were skipped.

```
1  void sharedFeatureCounts(InvertedIndex *ii1, InvertedIndex *ii2,
2                           unsigned short **cc_matrix)
3
4       FeatureList *f1 = ii1->feature_skip_list;
5       FeatureList *f2 = ii2->feature_skip_list;
6
7       unsigned short *ii1_position, *ii2_position, *cc_matrix_f1;
8
9       while (f1->feature_id != f2->feature_id)
10            if (f1->feature_id > f2->feature_id)
11                 ++f1;
12            else
13                 ++f2;
14
15      while (f1->feature_id && f2->feature_id)
16            ii1_position = f1->block_positions;
17
18            while (*ii1_position)
19                 cc_matrix_f1 = cc_matrix[*ii1_position];
20                 ii2_position = f2->block_positions;
21
22                 while (*ii2_position)
23                      ++*(cc_matrix_f1 + *ii2_position);
24                      ++ii2_position;
25
26                ++ii1_position;
27
28           ++f1;
29           ++f2;
30
31           while (f1->feature_id != f2->feature_id)
32                if (f1->feature_id > f2->feature_id)
33                     ++f1;
34                else
35                     ++f2;
```

The core part of the shared feature count calculation is shown in Listing 3.1 and described in the following. The function takes as input pointers to both InvertedIndex structs to be processed and a pointer to a zero-initialized shared feature counts matrix, First, pointers to the first element (FeatureList) of the *feature_skip_lists* are initialized (lines: 4-5). Furthermore, pointers for later use to iterate the *block_position* arrays and to point to a certain matrix row are declared. The next step iterates over both *feature_skip_lists* until the underlying FeatureLists have the same or the default *feature_id*. The latter case indicates that no molecules between the processed blocks possess common features (lines: 9-13). The outer loop spanning lines 15 to 35 iterates both *feature_skip_lists* as long as there are FeatureLists with same *feature_id*. Two

FeatureLists with same *feature_id* contain the *block_positions* of all molecules in the processed blocks, which possess this feature. Now, the previously initialized *block_position* pointers are used to iterate over the corresponding arrays (lines: 16, 20). Thereby, all pairs of molecules between the processed blocks sharing this fingerprint feature are visited in the first inner loop (lines: 18-26). The dereferenced *block_positions* are used to address the matrix cell of the visited molecule pair, which is then incremented (line: 23). When the end of one or both *block_position* arrays is reached the function steps ahead to the next FeatureList (lines: 28, 29). The second inner loop again tries to find the next FeatureList pair with same *feature_id*. If one *feature_skip_list* reaches the default FeatureList, the outer loop exits and the function returns.

Using the calculated shared feature counts matrix and the total feature counts of every fingerprint, which can also be accessed via the iiDS, immediate calculation of S_{Tan} is possible using Eq. 2.2. To exploit the full power of modern shared memory multi-core architectures, we implemented this algorithm in a thread-parallel fashion. For this purpose we used the BOOST thread library.

Clustering Step 1: Duplicate Fingerprint Detection

For the unique fingerprints filter, feature lists are hashed using the collate class of the C++ standard library. Duplicates are detected on the basis of the calculated hash values. The first occurrence of every molecule is used for further processing and the remaining occurrences are temporarily stored.

Clustering Step 2: Connected Components Decomposition

Using the described bII method, all pairwise similarities (S_{Tan}) are calculated for the unique fingerprint set. As we are solely interested in the final CC decomposition and do not need the complete similarity network topology, we do not store the edges exceeding the similarity cutoff. This proceeding enables the application of the *union-find* based CC algorithm introduced in Chapter 2.1. This method starts with an empty set of edges and dynamically generates the CCs by inserting new edges into the graph. For this purpose we use the BOOST graph library, which provides incremental CC functionality.[135] This implementation uses union-find with path-compression and has a time complexity of O($V + \alpha E$), where α is the extremely slow growing inverse Ackerman function.

In the following clustering step we need the nearest neighbor of every molecule. Instead of recalculating these similarities, we immediately store the nearest neighbor information while calculating all pairwise similarities.

Clustering Step 3: Hierarchical Clustering

The agglomerative average linkage method is used to cluster CCs exceeding a predefined size. As mentioned in this chapter's introduction, efficient methods have been described for this clustering methodology on the basis of so-called reciprocal nearest neighbors (RNN).[136,137] RNNs are cluster pairs which are mutual nearest neighbors and have the property that they can immediately be merged to create a new cluster. Murtagh described two RNN-based algorithms. First, an iterative parallel approach where a nearest neighbor is calculated for every element with subsequent determination and merging of all RNN pairs at once. The second algorithm constructs a chain of subsequent nearest neighbors. At a certain point the chain ends up with an RNN pair, which has to be merged and the algorithm continues from the last but two chain link. As already mentioned, we calculated the nearest neighbor information for every molecule during similarity graph construction. Thus we decided to start the hierarchical clustering using the parallel RNN version until the entire similarity matrix for the remaining clusters fits into the main memory. At this point we calculate the similarity matrix for the remaining clusters and switch to the nearest neighbor chain algorithm to finish the clustering.

Clustering Step 4: Cluster-level Selection

For cluster-level selection we implemented the method developed by Kelley *et al.*.[138] The method is often termed as the Kelley criterion. In their work, the authors also used average linkage to generate a clustering on their data set. A penalty value is calculated for every internal node of the dendrogram. The node with minimum penalty value is chosen to cut the dendrogram yielding a final cluster selection with maximally populated clusters and minimal internal spread. The spread of a cluster is an information about the similarity of its members. In our case, small spread values indicate high similarities of the molecules in a cluster. As the calculation of a clusters spread has to sum up the pairwise similarities of its members, we calculate this sum during clustering and store it for every internal node when it is created.

Clustering Step 5: Calculation of Cluster Representative

To select a representative molecule for every cluster C_i we calculate its medoid. The latter is the molecule $m_i \in C_i$ with the highest mean Tanimoto similarity to all other molecules in C_i. If multiple molecules share the maximum mean similarity, all of them are marked as representatives. Additionally, if the fingerprint of a medoid is shared by multiple input molecules all of them are marked as representatives in the output, too. This step also uses our bII algorithm to efficiently calculate all pairwise similarities within all clusters.

3.2.2 Data Sets

As a virtual compound library we used the *all purchasable* subset of the ZINC database (version 12, accessed 20/6/2012) comprising 17,833,934 compounds at that time.[127] The library was prepared using Pipeline Pilot and OpenEye.[139,140] Salts were removed and canonical SMILES were calculated. Two commonly used types of binary fingerprints were calculated using tools from ChemAxon's JChem Base (GenerateMD).[141] The first type was extended connectivity fingerprints (ECFP), which generates higher order features based on radial substructures of increasing radii.[37] We used a maximum radius of four bonds. The second type was ChemAxon's default path-based chemical fingerprints (PBFP), which also generates higher order features on the basis of linear paths of a predefined length. Here we used the default maximum path length of seven bonds. ECFPs are used as an example for sparse fingerprint types and PBFP as an example for fingerprints with high density. To benchmark the similarity calculation methods, we used a random subset selected with Pipeline Pilot.

3.2.3 Hardware and Software

The popcount versions for similarity matrix construction of Haque *et al.* were compiled as provided in their supporting information using CUDA toolkit version 4.2.9. For runtime comparison of our clustering method to *Ward* and *Jarvis-Patrick*, we used the implementations provided by ChemAxon JChem Base (Ward, Jarp).[141] For performance evaluation of these methods, we used a current desktop computer equipped with an Intel Core™ i7-2600 processor (4 cores, 3.8 GHz) and 8 GB main memory. For clustering of the entire ZINC all purchasable subset we used a compute server equipped with 4 AMD Opteron™ 6274 processors (64 cores, 2.2 GHz) and 512 GB main memory.

We used the C++ compiler from the GNU Compiler Collection (versions: Intel system 4.6.2, AMD system 4.4.7). Cache analysis were performed using Cachegrind from the Valgrind Tool Suite (version 3.7.0). Additionally, we used the thread and graph libraries from BOOST (versions: Intel system 1.49.0, AMD system 1.46.1).

3.3 Results

In this section, we first show the benchmarking results of our bII similarity calculation method in comparison to state-of-the-art tools to solve this task. We then show the results of block size and parallel execution analysis. Next, we show the clustering method comparison results. Finally, we demonstrate the performance of the method to process the currently available chemical space.

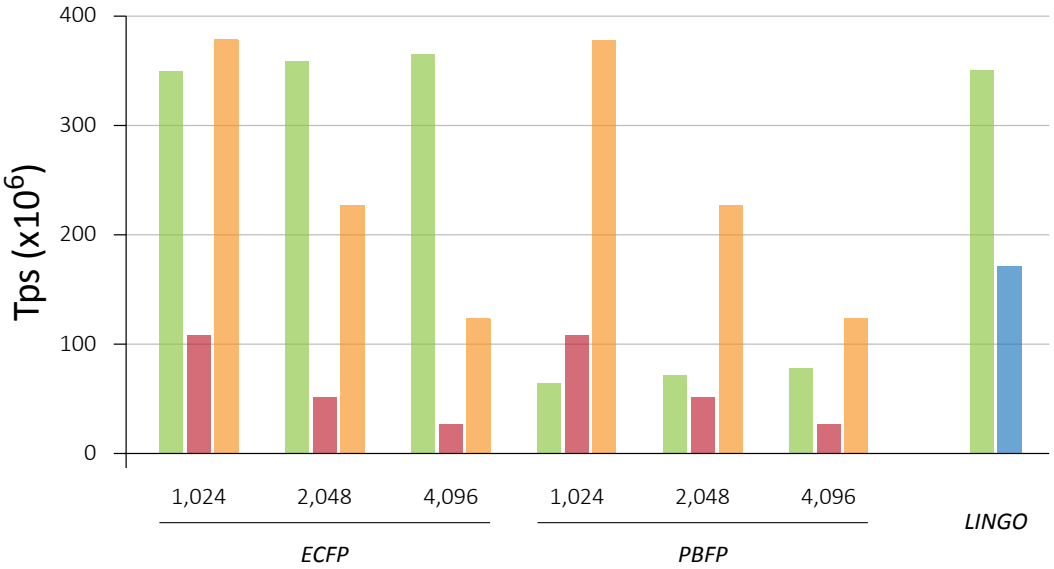


Figure 3.4: Performance of similarity calculation. The bars show maximum Tps values for our bII method (green bars) in comparison to naïve SSE4 implementation (red bars) and to the cache-efficient SSE4 implementation (orange bars) for ECFP and PBFP fingerprint types folded to different lengths. Additionally, we compared our bII implementation to the inverted index implementation for LINGO similarity calculations (blue bar).

3.3.1 Blocked Inverted Index Performance

To evaluate the maximum performance of our blocked inverted index algorithm we conducted threshold searches for ECFP and PBFB fingerprints folded to different lengths (1,024, 2,048, and 4,096 bits). We used a scenario as described in the study of Haque and compared the results to their popcount-based implementations.[130] In brief, we calculated all pairwise similarities of an input library and reported all molecule pairs with a similarity exceeding a similarity of $S_{Tan} > 0.8$. The results are displayed as a bar chart in Fig. 3.4. For these benchmark runs, we set the block size parameter for our bII algorithm to $s_b = 480$ molecules per block. For the best performing settings with ECFPs folded to 4,096 bit, our bII algorithm achieved the equivalent of 365 million Tanimoto calculations per second (Tps). These results demonstrate that our bII algorithm performs as well as the popcount implementations and for several cases it is even better. This is particularly remarkable because our method is a purely algorithmic solution for this problem. These results point out two important characteristics of the compared methods. First, the performance of inverted index algorithms is rather independent of the fingerprint size whereas the performance of popcount-based algorithms decreases linearly with fingerprint length. Second, the performance of inverted index algorithms is output-sensitive with respect to the final shared feature counts whereas the performance of popcount algorithms is independent thereof. This is shown by the throughput of the bII for the dense PBFPs, which is approximately an order of magnitude lower than for the ECFPs.

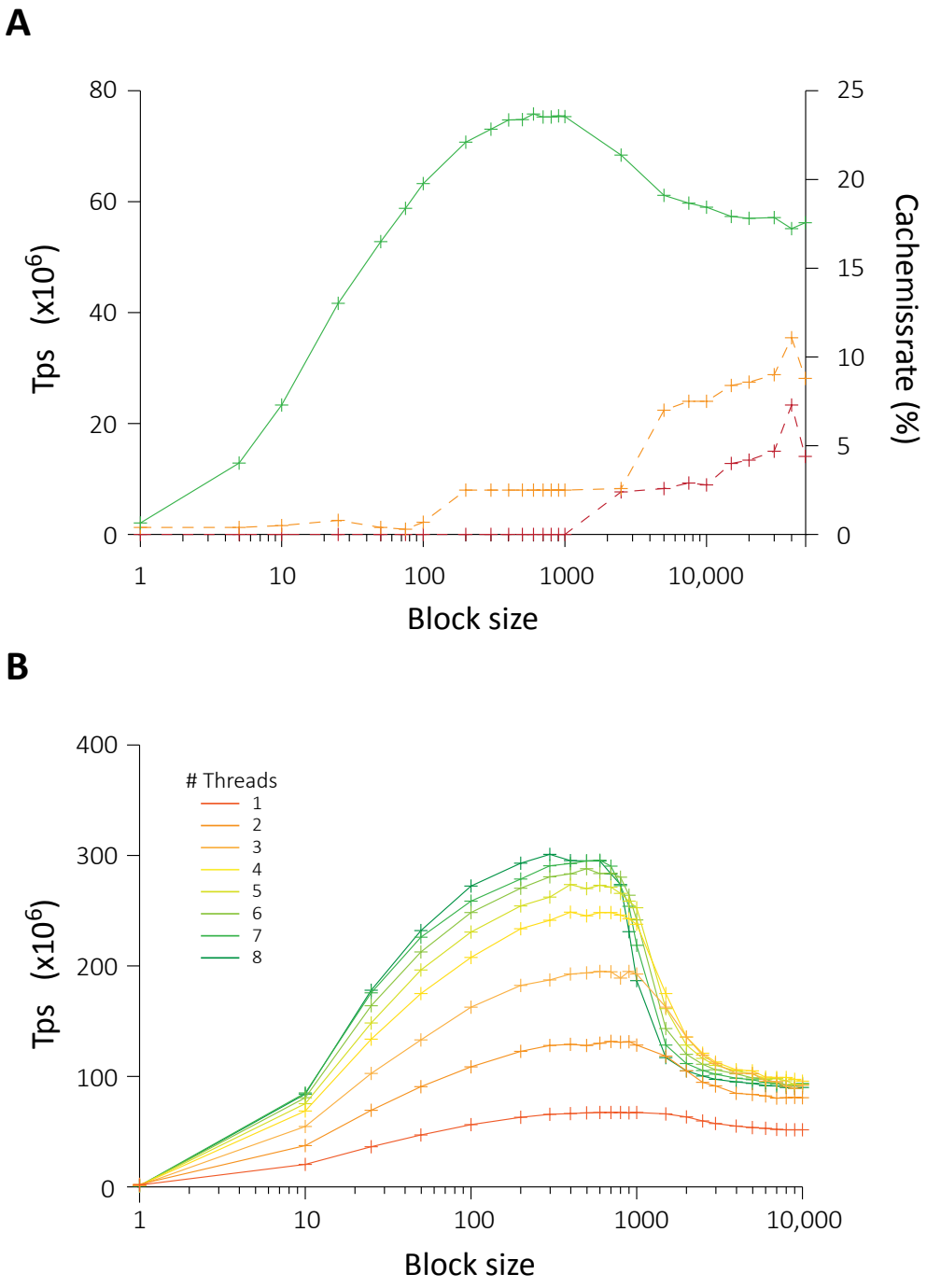


Figure 3.5: Block size dependency of the bII algorithm. **(A)** The primary vertical axis shows the performance in million Tps (green line). Cache write misses in percent are shown on the secondary vertical axis for the L1 data cache (dashed blue line) and the L3 data cache (dashed red line). **(B)** The chart shows the bII performance for 1-8 parallel threads and increasing block sizes.

3.3.2 Block Size and Hardware Scalability

As we were interested to gain knowledge on the influence of the block size parameter s_b, we evaluated the performance of the bII algorithm for varying block sizes. This parameter affects three important aspects. First, the number of pairwise block comparisons decreases. Second, the size of a single iiDS increases and third, the size of the matrix to store the shared feature counts grows quadratic with the block size. Fig. 3.5A shows the Tanimoto throughput depending on the block size using a single threaded version of our algorithm. Starting with a block size of one, which corresponds to an iiDS for every input molecule, the minimal throughput of our algorithm is about 2×10^6 Tps. This can be regarded as an efficient implementation of naïve Tanimoto calculation. Increasing the block size up to 500 molecules leads to a hyperbolic performance gain and reaches a maximum throughput of about 75×10^6 Tps. This throughput remains nearly constant up to a block size of about 900. Further increase leads to an asymptotic performance loss tending to a throughput of $\sim56\times10^6$ Tps for a block size of 50,048 molecules. In this extreme case, a single iiDS is created including all input molecules. To explain this effect, we analyzed the processor cache behavior. Fig. 3.5A thus also shows the relative cachemissrates for L1 and L3 data cache writes. The initial performance gain seems to decelerate due to an increasing miss rate of the L1 data cache. The performance decrease following the plateau phase seems to be an additive effect of increasing miss rates of both, the L1 and the L3 data caches. However, the dominating effect is supposed to be the L3 behavior because cache misses of the latter are typically an order of magnitude more expensive than L1 cache misses.

Fig. 3.5B shows the hardware scalability of the bII algorithm with increasing number of parallel threads. The results demonstrate that the implementation scales almost linearly up to a number of four threads. Interestingly, the use of more than four threads does not scale linearly any more. The increase in Tps is less than 25 % for every single thread compared to the first four threads. The reason therefor is most probably also the cache limitations because the employed Intel processor provides four physical and eight virtual cores. Thus, using more than four threads has the consequence that physical cores and their dedicated cache have to be shared by multiple threads.

3.3.3 Comparison to Standard Clustering Methods

The chart in Fig. 3.6 shows the runtimes of our clustering method and ChemAxon's *Ward* and *Jarvis-Patrick* implementation evaluated on a desktop computer. Our method was tested with the ECFP and PBFP fingerprints folded to 2,048 bits. The ChemAxon methods were run with ChemAxons standard fingerprint settings (PBFP, 512 bit). The similarity cutoffs for CC decomposition were separately chosen for the different fingerprint types. To yield reasonable and comparable sizes of the largest CCs, we set the similarity cutoffs to $S_{Tan}^{Cut} = 0.6$ for ECFPs

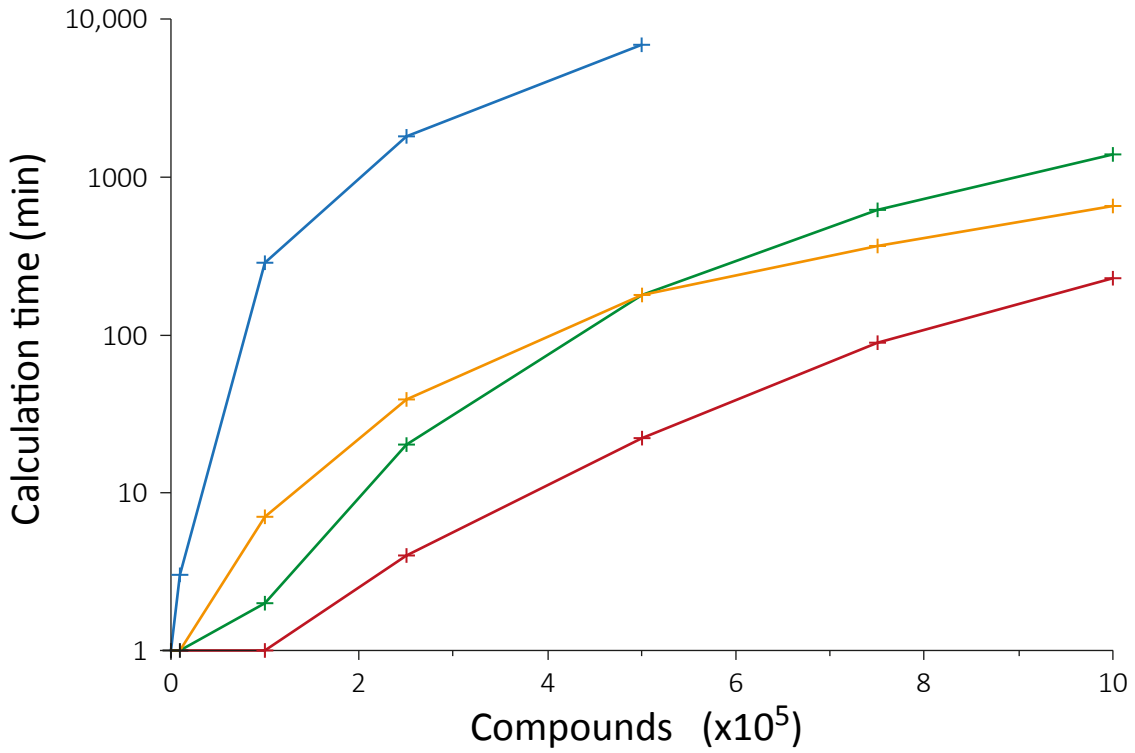


Figure 3.6: Clustering runtime comparison. We compared *Ward* (blue marks), *Jarvis-Patrick* (orange marks) and our clustering method. We evaluated the latter on PBFPs (green marks) and ECFPs (red marks).

and to $S_{Tan}^{Cut} = 0.7$ for PBFPs. The lower cutoff for hierarchical clustering of CCs was set to 1,000 molecules per CC for all library sizes. Our method clearly outperforms the runtime of the *Ward* method. As expected, the *Jarvis-Patrick* implementation is much faster than the *Ward* method due to the use of heuristic algorithms.[141] The runtimes of our clustering method lie in the range of the performance of *Jarvis-Patrick*. For the ECFP fingerprint type, our method is even faster than *Jarvis-Patrick*. These observations show that the runtimes of chemoinformatics clustering methods are also highly dominated by the speed of the similarity calculation. Thus, the latter forms a bottleneck and optimizations can lead to significant runtime improvements of clustering applications.

3.3.4 Clustering the Available Chemical Space

The input database contained 17,833,934 compounds, which were represented as ECFPs folded to 2,048 bit length. The duplicate fingerprints filter step reduced this input library down to 12,976,486 unique fingerprints. To choose appropriate values for S_{Tan}^{Cut} to generate the similarity network and the minimum CC size for hierarchical clustering, we performed a pre-analysis step. For this purpose, we calculated all CC decompositions in a Tanimoto

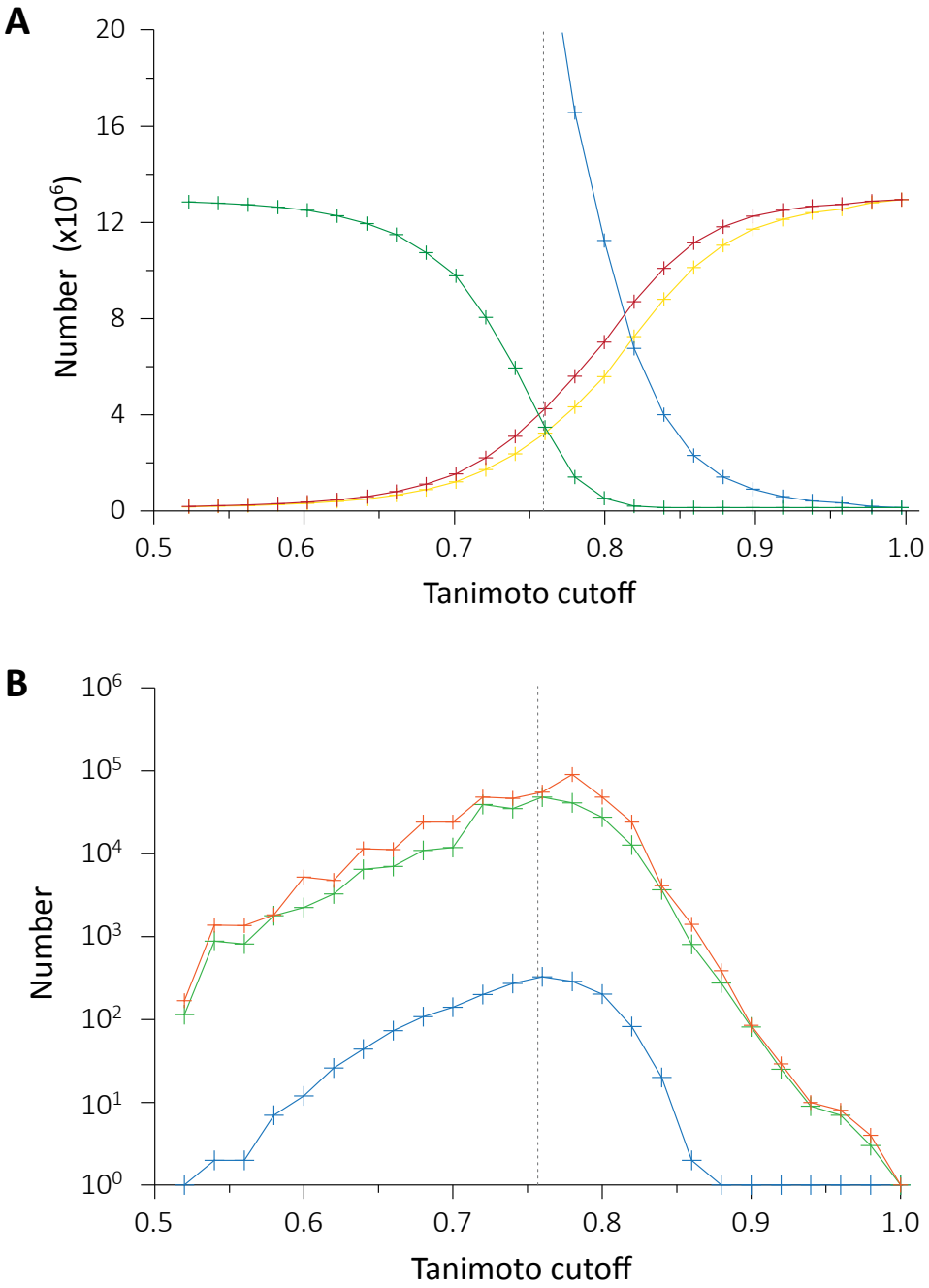


Figure 3.7: Similarity network analysis of the ZINC *all purchasable* subset. The finally selected similarity cutoff $S_{Tan}^{Cut} = 0.76$ is marked by dotted vertical lines. **(A)** The plot shows the number of edges exceeding S_{Tan}^{Cut} (blue line), the size of the largest CC (green line), the total number of CCs (red line) and the number of singletons (yellow line). **(B)** The second and third largest CCs are shown (red and green line, respectively) as well as the number of CCs with more than 1,000 members (blue line).

range from 0.52 to 1.0 using a step width of 0.02. The resulting graph is shown in Fig. 3.7A. It reveals similar results as previously calculated for three different and small compound datasets reported by Lepp *et al.*.[142] They also recognized a single dominating CC, which gradually decomposes in a Tanimoto range of 0.5 to 0.8. The second and third largest CCs are an order of magnitude smaller than the largest CC, which is shown in Fig. 3.7B. Additionally, this graph shows the number of CCs with more than 1,000 members.

Based on these results, we decided to set the cutoff for similarity network generation to $S_{Tan}^{Cut} = 0.76$ and the CC size exclusion cutoff for hierarchical clustering to 1,000 members. Thus, CCs with less than 1,000 members were not directed to hierarchical clustering. Using these parameters we executed our method on the described Opteron compute server using all cores and 100 GB of main memory, which took 64 hours to complete.

3.4 Discussion

In the present study we describe the development of a deterministic clustering method, which is capable of processing extremely large chemical spaces represented as 2D fingerprints within days on current standard hardware. As described in this chapter, one critical bottleneck of chemoinformatics clustering methods is the speed and efficiency of the underlying similarity calculation method. Thus, we developed and implemented an extremely fast and flexible algorithm for the calculation of binary 2D fingerprint similarities especially for this purpose. This algorithm is highly optimized to compute all pairwise similarities within a given dataset. It is based on inverted index data structures, which is an efficient representation of binary fingerprints and enables extremely fast calculation of shared feature counts between fingerprints. We have modified this data structure by dividing the input molecules into equally sized blocks, which enables the choice of an optimal block size to maximize the number of similarity calculations per second. Recently introduced methods to calculate fingerprint similarities use specialized instruction set extensions on hardware level or specialized hardware itself, like GPUs. In contrast, the presented bII algorithm is purely algorithmic and can be run on every standard CPU.

We compared our bII algorithm to popcount-based methods and the results demonstrate that our algorithm can perform equally well and for several cases even better. Furthermore, these results nicely highlight the advantages and disadvantages of the compared methods for different fingerprint types. The bII method performs better for sparse fingerprints and is insensitive to fingerprint length, whereas popcount-based methods perform better on short fingerprints and are insensitive to fingerprint density. Additionally, we have analyzed the block size dependence of our algorithm which points out that splitting of molecules influences the cache performance and an optimal choice of this parameter is critical to the overall performance.

Using this algorithm, we implemented our deterministic clustering method to process very large compound libraries. The method combines different techniques to decompose an input library in three consecutive steps. These steps comprise a unique fingerprint filtering, the graph theoretic concept of CC decomposition and standard hierarchical clustering. By runtime comparison of our clustering method to the frequently used methods *Ward* and *Jarvis-Patrick*, we could demonstrate that our method is faster or at least competitive for commonly used 2D fingerprint types. To finally show the power of our clustering method, we processed the commercially available chemical space comprising more than 17 million compounds on a compute server which took 2.5 days of computation time. For the clustering of such large chemical spaces, normally the non-deterministic *k-Means* or *Jarvis-Patrick* approaches are used. Deterministic hierarchical methods like *Ward* cannot be used for data sets of these dimensions because their runtime complexity is too bad.

Our blocked inverted index method and the exact clustering are currently integrated into the free chemical structure database ChemSpider, which is provided by the Royal Society of Chemistry (`http://www.chemspider.com/`).[143] The methods are used to generate similarity networks for their newly developed webservices in order to speed up similarity searching.

Chapter 4

Scaffold Families

4.1 Introduction

The concept of small molecule scaffolds was briefly introduced in Section 2.2.1. Scaffolds are frequently used for various chemoinformatics tasks, which comprises amongst others the design and the analysis of compound libraries,[144] scaffold hopping,[145] or rational and scaffold-driven drug design approaches.[146] As already described, different definitions of molecular scaffolds exist and the opinions on how to define a molecule's scaffold diverge. However, in general it is reasonable to use data set-independent scaffold definitions like *Murcko scaffolds* or *Scaffold Trees*.[49,51,147] This is for example not true if maximum common substructure approaches are employed to dynamically calculate the minimal scaffold within a data set.[148]

In the case of compound library analysis the distribution and occurrence of scaffolds can yield useful information on the composition of a library. These can possibly be used to specifically expand the library by adding novel scaffolds. However, depending on the employed scaffold definition the results of such analysis differ vastly.[150] Using scaffolds reduces compound complexity and can help to rationalize chemical space by grouping structurally related compounds. This procedure is static in a sense that it does not allow for structural variations in the scaffolds and the created groups. Nevertheless, allowing limited structural variations is common practice when medicinal chemists explore structure-activity of identified hits. An example is illustrated in Fig. 4.1. It shows two dihydrofolate reductase (DHFR) inhibitors as well as their corresponding *Murcko scaffolds*.[149] Despite slightly different *Murcko scaffolds*, their receptor-bound 3D conformations are highly similar. As both compounds are thus closely related by their biological activity, it would makes sense to group them together from a drug design perspective.

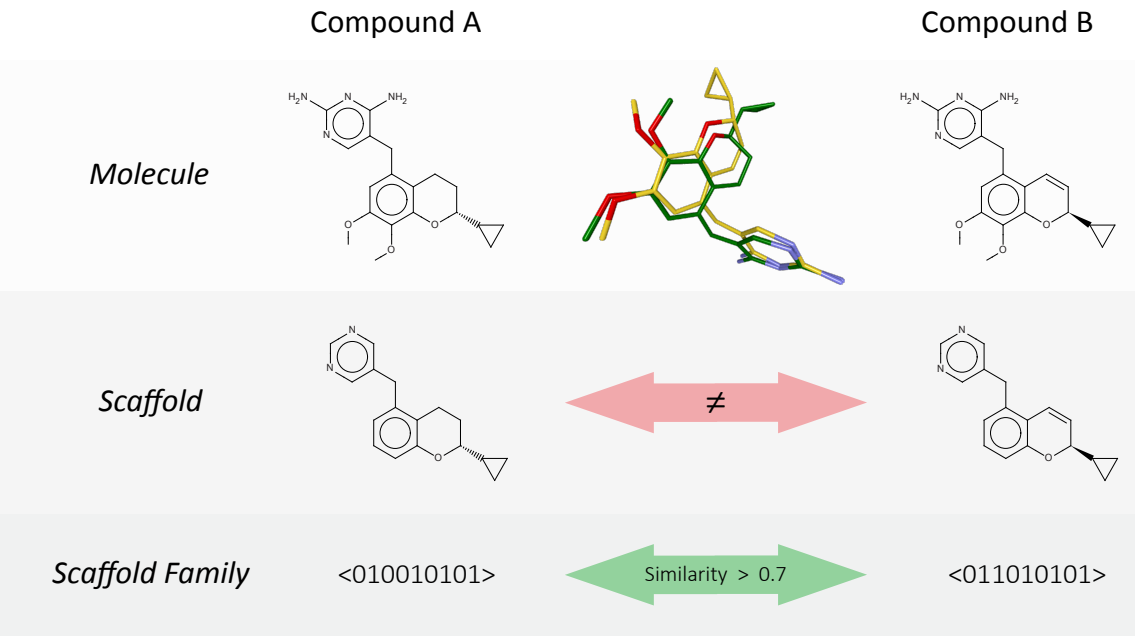


Figure 4.1: *Scaffold family* example. Compound A and B are chromene-derived antibiotics, which inhibit the enzyme DHFR. Compound A is the drug candidate Iclaprim and its conformation is taken from PDB entry 3fra.[149] Compound B is ligand XCF taken from PDB entry 3frf.[149] The binding mode of both inhibitors is conserved as can be seen from their superposed conformations. However, their *Murcko scaffolds* are not equal but the fingerprint-based 2D similarity between A and B indicates a highly related scaffold. Thus, A and B can be classified into the same *scaffold family*.

Goals of the Project

In this project, we introduce a fast method to abstract molecule scaffolds in order to create so-called *scaffold families* with improved information content when referring to medicinal chemistry aspects. As described in the introductory DHFR inhibitor example, *scaffold families* are intended to group compounds together that show closely related activity on a single target but do not share the same scaffold. In structure-activity relationship (SAR) studies, the exploration of related scaffolds is part of a medicinal chemist's core competencies, but for these experts it is not feasible to manually classify larger compound data sets or even libraries according to related scaffolds. Thus, an automatic method to generate *scaffold families* is desirable. We quantitatively evaluate the method on selected target specific data sets and a large virtual compound library. Furthermore, we show selected *scaffold families* generated by our method to highlight the advantages of this approach.

Related Work

Previous research projects to analyze the scaffold diversity of compound libraries mainly focused on assessing the number of unique scaffolds and their population sizes.[144] Langdon *et al.* performed a similar study.[150] However, they were especially interested in the comparison of compound libraries' scaffold compositions for different scaffold definitions. As a result, they recommend level 1 *scaffold tree* scaffolds as a useful abstraction method to analyze compound libraries. Additionally, they used 2D fingerprints to hierarchically cluster level 1 scaffolds to enable data visualization using tree maps for diversity and population analysis.

4.2 Materials and Methods

This section describes our method to calculate *scaffold families* for a given compound collection, the analyzed data sets and their preparation. The latter was performed using standard components from Pipeline Pilot using tools from its Chemistry package.[151] To calculate the scaffold similarity network of all data sets we used the software that we presented in Chapter 3.

4.2.1 Scaffold Family Calculation

Fig. 4.2 gives an overview of our method to calculate *scaffold families* from a compound library. First, we generate the *Murcko scaffold* for every molecule. We decided to use the *Murcko scaffold* definition because resulting scaffolds are closest to the original molecule. Additionally, these scaffolds are represented by at least one molecule in the library, which is not necessarily given for lower level *scaffold tree* scaffolds. We group the library according to their scaffolds, yielding a set of unique scaffolds.

Second, we calculate 2D binary fingerprints from the unique *Murcko scaffold*s. As a binary fingerprint type we employ ECFPs with a maximum radius of four bonds and pharmacophore atom abstraction (Pipeline Pilot fingerprint abbreviation: FCFP_4).

In the third step, we calculate the *scaffold families* on the basis of all pairwise fingerprint similarities using the Tanimoto coefficient as similarity measure. Using our presented methods for fast similarity calculation (Chapter 3), we generate the scaffold similarity network at a predefined Tanimoto cutoff. From this scaffold network we extract the CCs, which finally represent the *scaffold families*.

4.2.2 Data Sets

We used 13 target specific data sets extracted from the ChEMBL database (release 17) to analyze the classification characteristics of our *scaffold families*.[152] ChEMBL is one of the most comprehensive and freely availably databases containing bioactivity data for small molecule

ligands on various targets. Compound structures of the data sets were downloaded from the ChEMBL website in SMILES format. All data sets are protein targets of varying sizes and comprise four organisms. Additionally, we analyzed a manually curated compound database, which contains all immediately available compounds from selected and representative chemical suppliers.

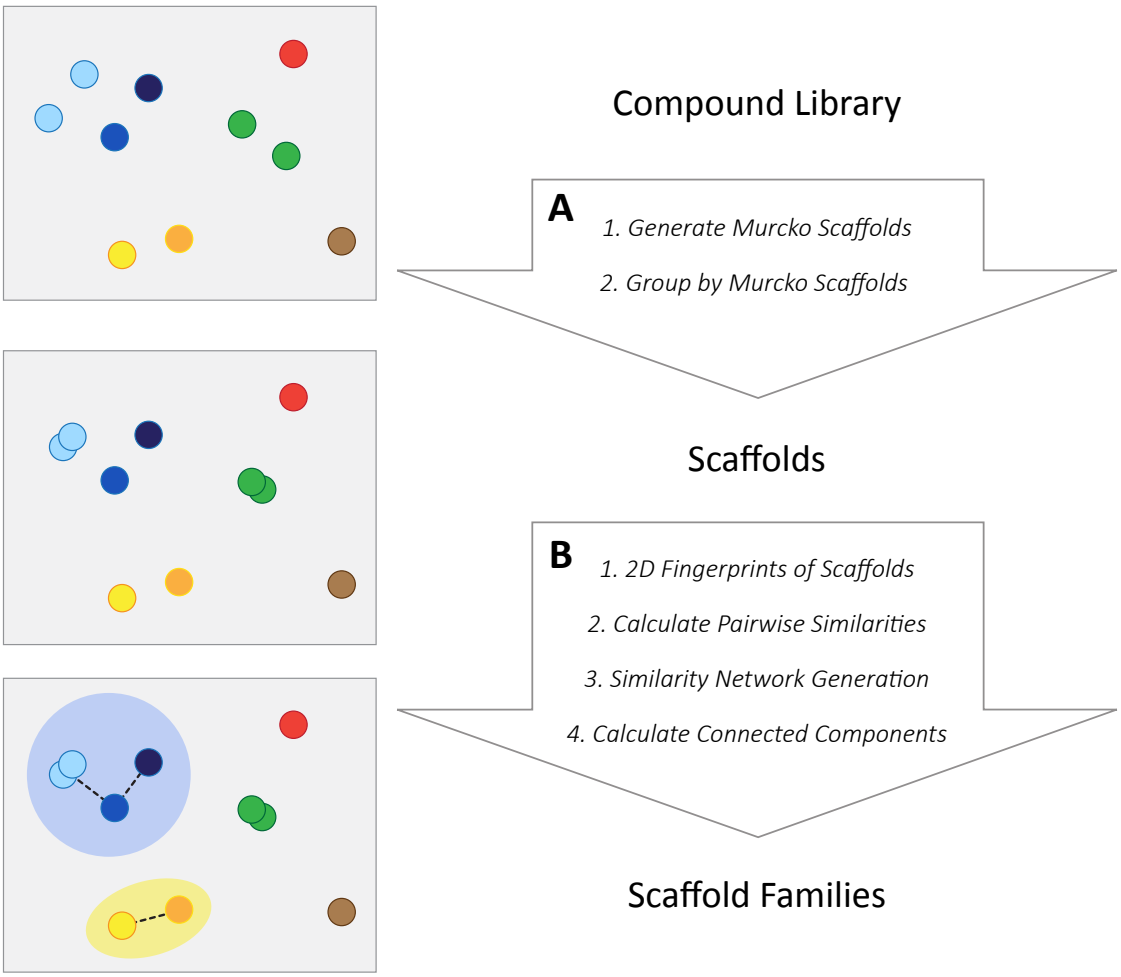


Figure 4.2: *Scaffold family* generation schema. **(A)** First, *Murcko scaffolds* are generated for every molecule in a virtual library. Second, duplicate removal yields the virtual scaffold library. **(B)** *Scaffold families* are obtained by calculating the 2D fingerprint-based similarity network at a defined similarity cutoff and subsequent CCs decomposition. The resulting CCs form the *scaffold families*.

Data Set Preparation

The data sets we selected and the resulting compound numbers are listed in Table 4.1. Duplicate compounds in the single subsets were removed. Subsequently, we analyzed the compounds' scaffold fractions and discarded all compounds with less than 50 % non-hydrogen scaffold atoms as compared to the decoration. The remaining compounds were used for the generation and analysis of *Murcko scaffold*s and *scaffold families*.

The custom compound library was compiled from the supplier catalogues using the following procedure. First, salts and counter ions were removed, charges and stereo information was standardized and canonical SMILES were generated. Based on the latter we started with the largest supplier data set to select compounds for our library. The other suppliers were sequentially added by decreasing size and duplicates were skipped. In total, this data set finally comprised 6,494,794 compounds. The sizes of the single supplier subsets are also listed in the appendix (Table E.1).

Table 4.1: Target-specific compound data sets extracted from ChEMBL database. M: redundancy-reduced (unique) compounds. N: number of compounds with sufficient scaffold fraction used for further analysis.

Target	*Abbr.*	*Organism*	*UniProt ID*	*M*	*N*
Neuraminidase		*Influenza A*	P03468	405	160
Beta lactamase		*P. aeruginosa*	Q932Y6	658	628
Angiotensin-converting enzyme	ACE	*H. sapiens*	P12821	668	431
Aldose reductase	AR	*H. sapiens*	P15121	748	680
Dihydrofolate reductase	DHFR	*H. sapiens*	P00374	1,147	1,091
Urokinase		*H. sapiens*	P00749	1,230	1,110
Heat shock protein 90-α	HSP90	*H. sapiens*	P07900	1,316	1,247
Reverse transcriptase	HIV-RT	*HIV 1*	Q72547	3,284	2,781
Histamin H3 receptor	H3R	*H. sapiens*	Q9Y5N1	3,350	3,238
Cylcooxygenase 1	COX-1	*H. sapiens*	P23219	3,836	3,417
Acetylcholinesterase	AChE	*H. sapiens*	P22303	4,134	3,554
SRC kinase		*H. sapiens*	P12931	5,311	5,039
Thrombin		*H. sapiens*	P00734	7,322	6,497

4.3 Results

This section first shows the quantitative results for the analysis of *Murcko scaffolds* and *scaffold families* from the ChEMBL data sets. Additionally, the *scaffold family* decomposition of our custom compound library is presented. Second, we highlight two *scaffold families* generated from the ChEMBL data sets to demonstrate the beneficial properties of this method.

4.3.1 Quantitative Analysis of Scaffold Fingerprints

The fractions of unique *Murcko scaffolds* and *scaffold families* are shown in the bar chart of Fig. 4.3. The *scaffold families* were calculated using a Tanimoto cutoff $S_{Tan} = 0.7$. Various cutoffs were tested and the latter was chosen because it leads to a significant size reduction of all analyzed data sets. Light-colored base parts of the bars indicate the fraction of resulting singletons for every type. Here, singletons are scaffolds or *scaffold families* that cover only one original compound in the analyzed data set.

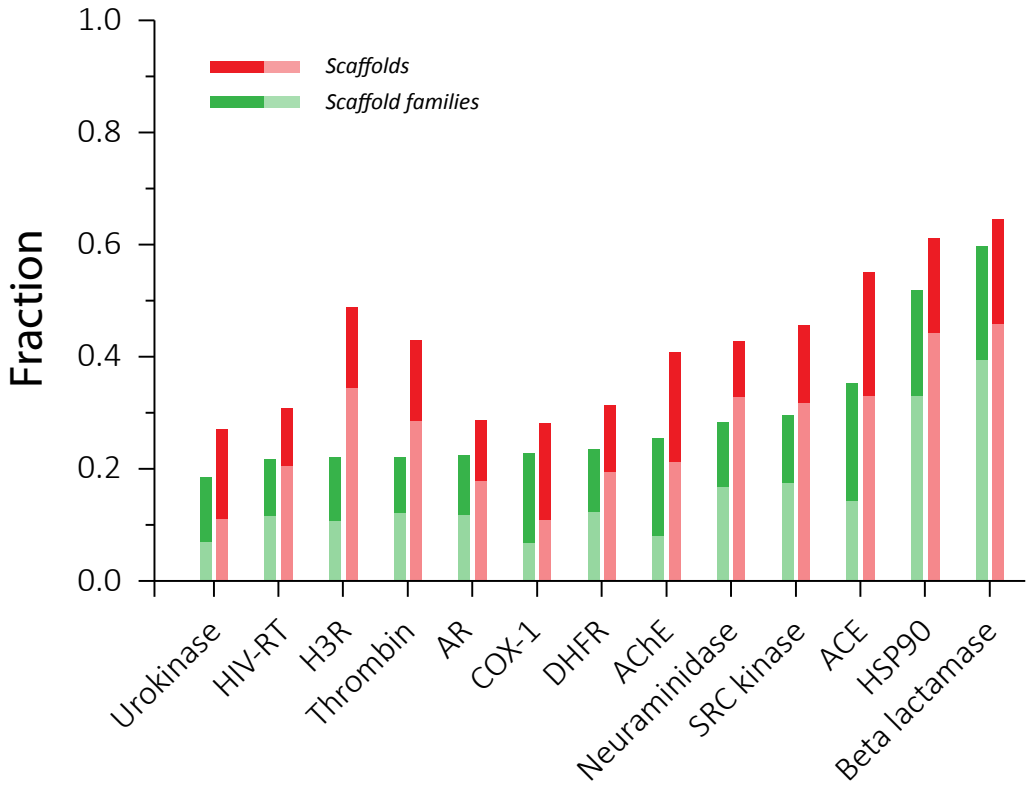


Figure 4.3: Fractions of *Murcko scaffolds* (red bars) and *scaffold families* (green bars) for the ChEMBL data sets. *Scaffold families* were calculated using a Tanimoto cutoff of 0.7. The light coloured base parts of the bars show the singleton percentages of the corresponding types. The data sets are sorted by increasing *scaffold family* fraction.

The calculation of *scaffold families* from *Murcko scaffolds* reduces the fraction of representative groups to various extents. The least difference can be observed for the COX-1 inhibitor data set. Here, the number of *scaffold families* is only $\sim$ 5 % smaller than the number of unique *Murcko scaffolds*. The maximum difference can be observed for the H3R inhibitor data set. Here, the fraction of *scaffold families* is less than half the fraction of unique *Murcko scaffolds*. This large reduction points out that the actually explored scaffold space is much smaller than the number of unique *Murcko scaffolds* present in this data set. In contrast, a small reduction as in case of the COX-1 data set indicates that the explored scaffold space indeed equals the number of *Murcko scaffolds*.

The comparison of singleton fractions also yields interesting insights. In all analyzed data sets, the magnitude of *Murcko scaffold* reduction by *scaffold family* generation is correlated to the reduction of singletons.

Custom Compound Library

From the 6,494,794 unique members in our custom compound library 6,480,058 compounds possess a valid *Murcko scaffold*. Application of the minimum scaffold fraction filter to skip compounds with $>$ 50 % decoration leaves 6,085,875 compounds. The number of *Murcko scaffolds* and *scaffold families* are listed in Table 4.2. The library comprises $1,259,875$ unique *Murcko scaffolds* whereof $\sim$ 62 % are singleton scaffolds.

The *scaffold families* were calculated using a Tanimoto cutoff of 0.85. We have evaluated various cutoffs in the range of 0.6 to 0.9 in order to choose this final cutoff because it yields a homogeneous library decomposition where the single huge CC is broken up and no single CC is dominating. The number of *scaffold families* using this similarity cutoff yields 626,257 *scaffold families* from which only $\sim$ 51 % are singletons.

Table 4.2: Fraction of *Murcko scaffolds* and *scaffold families* in the custom library.

Type	*N*	*Singletons*
Murcko scaffolds	1,259,875	789,122
Scaffold families	626,257	324,322

4.3.2 Qualitative Analysis of Scaffold Fingerprints

To demonstrate properties of *scaffold family* classification we selected two interesting examples from the analyzed ChEMBL data sets. They illustrate how *scaffold families* can be utilized to generate a classification of chemical spaces with improved medicinal chemistry information content.

Aldose Reductase Inhibitors

Fig. 4.4 shows a group of phtalazinone-based inhibitors of the human aldose reductase (AR). This enzyme catalyzes the reduction of glucose into sorbitol. It is one of the most important targets for the treatment of hyperglycemic states in patients suffering from type 2 diabetes to prevent long-term consequences.[153] The selected compounds were optimized in a typical SAR study by Mylari *et al.*.[154] Starting from the lead structure zopolrestat, the authors explored various oxodiazole and thiadiazole derivatives with AR inhibitory activity. The visualized example shows one *scaffold family* that merges four different *Murcko scaffolds* representing both chemotypes. The latter demonstrates the benefits of using pharmacophore feature atom encoding for 2D fingerprint generation.

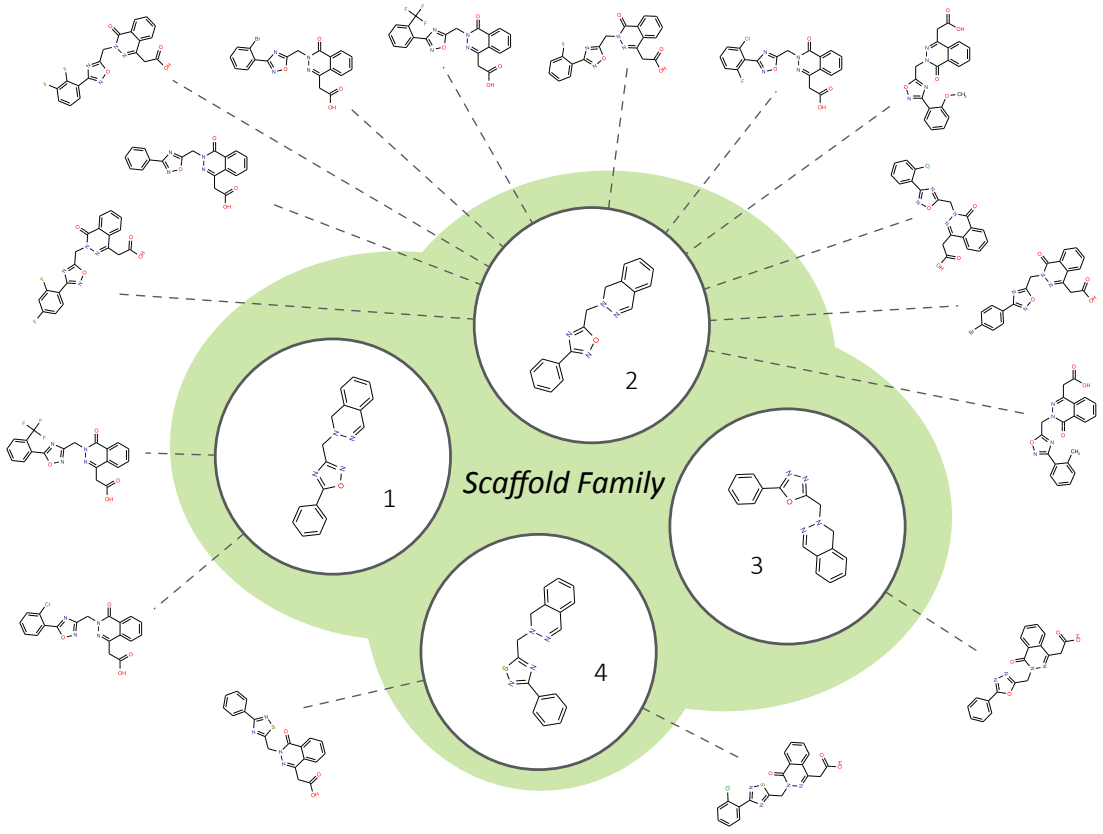


Figure 4.4: A *scaffold family* from AR inhibitors data set. This *scaffold family* comprises four different *Murcko scaffolds*, which represent 16 inhibitors in total. The grouping of oxodiazole and thiadiazole derivatives exemplifies the benefits of pharmacophore feature atom type encoding.

Serine Protease Inhibitors

The second example is shown in Fig. 4.5 and highlights serine protease inhibitors, which were synthesized in an effort to identify selective Factor Xa (FXa) modulators. FXa is one of the key enzymes in the early blood clotting cascade and an important anticoagulant target.[155] The selected compounds were all identified in a SAR study exploring amidine-substituted (bis)benzylidene-cycloketone olefines as candidates for non-peptidic FXa inhibitors.[156] The presented *scaffold family* is formed by five different *Murcko scaffolds*. Here, the variant key feature is the central cycloketone. This central aliphatic ring spans all sizes from cyclohexanone up to cyclodecanone.

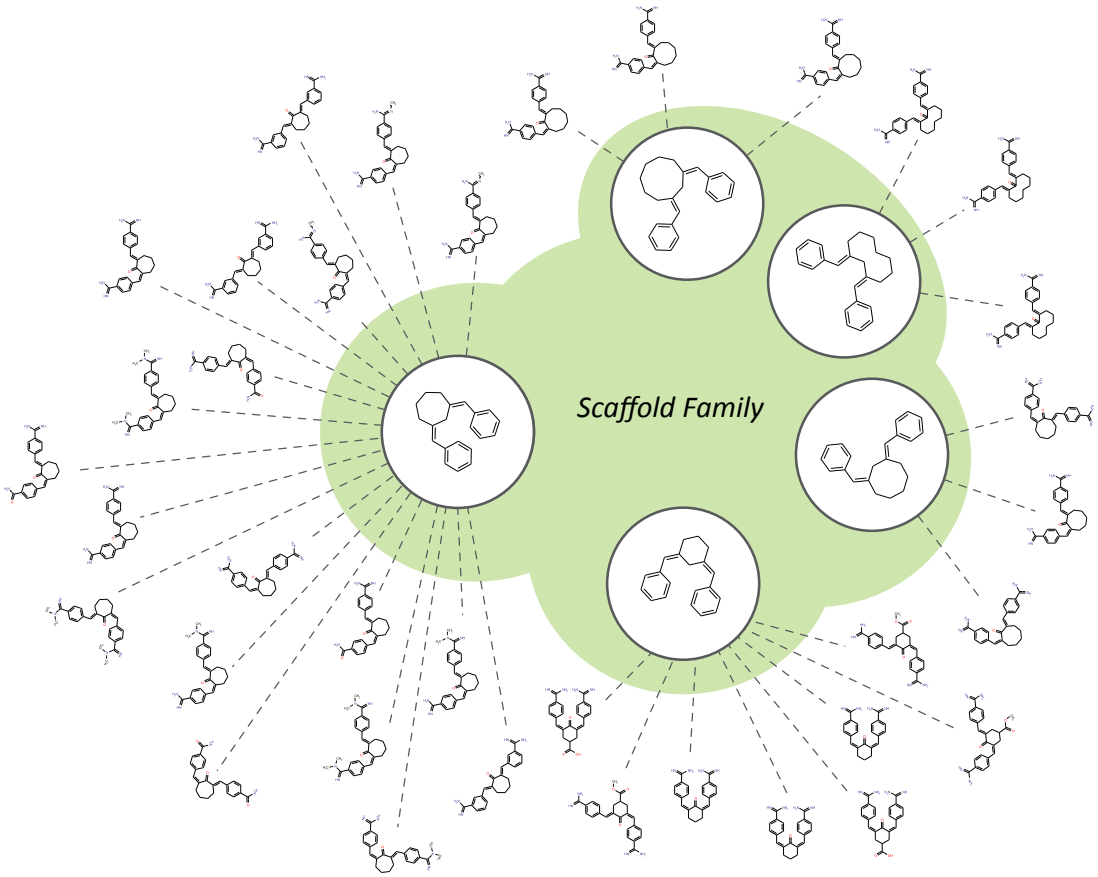


Figure 4.5: A *scaffold family* from thrombin serine protease inhibitors data set. This *scaffold family* comprises five different *Murcko scaffolds*, which represent 36 inhibitors in total.

4.4 Discussion

A meaningful classification of compounds in a way that reflects medicinal chemistry notions is of great interest and chemoinformatics tools to perform this task are of great value. Possibly the most important approach is the calculation and distribution of molecule scaffolds, which splits up a molecule into a functional framework and its decoration. Different scaffolding definitions have been proposed and evaluated and lead to quite different outcomes. However, a common substructure is a property of all scaffold definitions. This implies that a scaffold shared by a set of molecules is an invariant common substructure of them. On the one hand, this invariant representation is very useful to analyze the precise scaffold diversity within compound libraries. On the other hand, this view does not embrace the multifaceted kinds of structural relationships from a medicinal chemistry and structure-activity point of view.

In this chapter, we introduced *scaffold families* as a computational approach to push the rather static scaffold definitions towards a more medicinal chemistry meaningful way of representing compound framework relationships. *scaffold families* are generated by CC decomposition of a similarity network calculated using 2D fingerprints from *Murcko scaffolds*. In combination with the tools for fast fingerprint similarity calculations presented in the previous chapter, *scaffold families* can efficiently be calculated, even for huge virtual compound libraries. Our results demonstrate, that the use of *scaffold families* instead of *Murcko scaffolds* can reduce the number of representative groups and especially the number of singletons. The examples of AR and FXa nicely point out the benefits of our approach. Both cases show the successful co-classification of compounds from typical SAR studies. In summary, *scaffold family* oriented compound classification is extremely useful to rationalize scaffold and chemical space. Additionally, it is only marginally more computationally demanding than the generation of scaffolds alone.

Chapter 5

In Silico Analysis of Protein-Protein Interaction Stabilization

The content of this chapter is in parts published in the review article:

Small-molecule stabilization of protein-protein interactions: an underestimated concept in drug discovery?[22]

5.1 Introduction

The number of publications on the inhibition of PPIs by small molecules has been continuously growing over the last two decades.[17,18,20] This has also inspired theoretical investigation of PPI inhibition and a lot of effort has already been spent on gaining knowledge for the development of tools for *in silico* discovery of small molecule PPI inhibitors. A recent summary is given in the review of Villoutreix *et al.*.[70]

In contrast, only one publication is available yet reviewing the stabilization of PPIs by small molecules.[22] The work describes the currently known and structurally characterized examples for this fascinating molecular mechanism. Interestingly, without being aware of the underlying mechanism, the stabilization of PPIs has been one of the most successful industrial strategies for herbicide development since the early 1940s.[157] Today, assays to monitor PPI stability *in vitro* are available and feasible for HTS as demonstrated by Rose *et al.*.[25] Nevertheless, costs for HTS campaigns remain high and get even more expensive due to increasingly elaborate technology. Computational techniques are nowadays routinely used to support drug discovery campaigns because they are more cost effective.

Thus, it is of great importance to promote *in silico* research in the field of small molecule stabilization of PPIs. As a consequence of the novelty of this research field, no work has as yet been spent on analyzing the properties of small molecule binding to such complexes and to turn this knowledge into the development of *in silico* tools to screen specifically for PPI stabilizers. However, a couple of straightforward screens for PPI stabilization have been reported and will be discussed briefly in the following.

Goals of the Project

The first goal of this work is to analyze the ligand-bound stabilized PPIs, which are listed in Thiel *et al.* to infer structural principles of this mode of action.[22] Based on this knowledge we want to develop tools to identify PPIs which are stabilized by small molecule ligands.

The second goal is to use these tools for screening the PDB in order to identify as yet overlooked candidate PPIs that are possibly stabilized by a bound ligand. The latter is particularly promising because PPI stabilization by small molecules is not yet a well established mechanism and the description of it in the literature is heterogeneous and inconsistent. Additionally, the identification of further stabilized complexes will serve as a positive control for the developed tools and the conclusions we will draw from known cases.

The third goal is to use the resulting set of stabilized PPI complexes to evaluate the ability of current protein-ligand docking tools to correctly predict the binding pose of stabilizing ligands into their rim-exposed pockets.

Finally, we will use the obtained knowledge from the above and the developed tools to set up a VS approach to predict small molecule stabilizers of a selected PPI target. A set of final candidates will be experimentally evaluated for their potency to stabilize the target PPI and – in case of success – we will try to elucidate the binding mode of validated hits by means of X-ray crystallography.

Related Work

Efforts to find stabilizers of PPIs by means of *in silico* techniques are sparse but three projects have already been described in the literature. The first one published was performed by Ray *et al.* who tried to find small molecule ligands that stabilize the homodimer of superoxide dismutase 1 (SOD1).[158] SOD1 is a key player in familial amyotrophic lateral sclerosis (fALS), a severe degenerative motor neuron disease.[159] Here, point mutations in SOD1 lead to its aggregation, which is suspected to play an important role in disease progression. A prerequisite of aggregation is the dissociation of the SOD1 homodimer. The latter served as a model system for a standard protein-ligand docking approach that yielded stabilizers active *in vitro*.

The second VS for PPI stabilizers was reported by Block *et al.*.[23] As a target they used the interaction of plant 14-3-3 like protein C and PMA2, which is stabilized by the natural product

FSC (see also Section 2.4.3). Using a combination of ligand-based and structure-based VS, the authors' aim was to demonstrate the feasibility of PPI stabilizer screening *in silico*. However, the study revealed no active stabilizers for this interaction. A possible reason for this could have been the strong emphasis the authors put on the reference ligand FSC for the ligand-based step because the scaffold of this natural product is highly singular and similar compounds in public databases are limited to structurally close derivatives.[160]

The third and most recent report in VS for PPI stabilizers was presented by Jiang *et al.*.[161] In this study, the PPI of interest was the homodimer of myc-associated factor X (Max) as a promising target in various human cancer types. In brief, the transcription factor c-Myc has been shown to be overexpressed in many tumors and many of its biological activity depends on the heterodimerization with Max.[162] Thus, the c-Myc∘Max complex has been subjected to various PPI inhibitor discovery campaigns. As an interesting alternative, the authors proposed Max∘Max homodimer stabilization as a way to reduce available Max for heterodimerization with c-Myc. By means of standard protein-ligand docking they selected 68 compounds for experimental testing and reported 13 active stabilizers.

The described studies applied standard protein-ligand docking tools for structure-based VS, which have all been developed for the prediction of small molecule binders to enzymatic active sites. None of these works tried to include knowledge of already known PPI complexes, which are stabilized by small molecule ligands. However, Block *et al.* also performed a quantitative and qualitative analysis of rim-exposed interface PPI pockets to assess their frequency and potential druggability.[23] To get an estimate on the number of such pockets they analyzed 198 transient protein complex structures for the occurrence of rim-exposed interface pockets. In total they identified 380 pockets. Thus, a transient complex features on average two pockets. To assess the druggability of the identified rim-exposed interface pockets they calculated typical descriptors for druggability estimation and compared the results to 636 ligand-bound pockets of 243 enzymes. The results revealed considerable similarities with enzymatic binding pockets in terms of hydrophilicity, cavity volume, and burial suggesting a reasonable chance to successfully address PPI pockets with small molecules.

Recently, Gao *et al.* performed a more comprehensive analysis of PPIs in the PDB on the occurrence of rim-exposed interface pockets.[24] They analyzed 1,611 transient PPIs for pockets at the rim of the corresponding PPI interface. They identified 3,045 pockets at the PPI interface whereof 30 % exclusively exist in the formed protein complexes. Additionally, they reported 782 ligands with more than five heavy atoms that bind in proximity of these pockets. However, the authors drew no further conclusions from their work regarding the possibilities to stabilize PPIs with small molecules.

5.2 Materials and Methods

All materials, which were used for wet lab experiments and which are not directly described in the following sections are listed in the appendix (Table E.6). We refer to small molecule ligands of crystal structures by their chemical component identifier from the PDB. In the analysis of crystal structures only parts relevant to this work on the concept of PPI stabilization are introduced and discussed.

5.2.1 Principles Underlying known PPI Stabilizers

To gain information on structural principles of PPI stabilization we analyzed nine protein complexes, which have been found to be directly stabilized by a small molecule and for which the trimeric crystal structures are available. The PPIs stabilizes by AFB (PDB ID: 1s9d), FOK (PDB ID: 1cs4), and FSC (PDB ID: 2o98) were introduced in Section 2.4.3 and are visualized in Fig. 2.10. The remaining structures are shown in Fig. 5.1 and described in the following.

The PDB entry 3p1o contains a trimeric complex of 14-3-3σ, a mode III phosphopeptide from the human TWIK-related acid-sensitive potassium (K^+) channel 3 (Task3) and the fungal toxin FSC.[163] The C-terminal end of Task3 can be phosphorylated, which enables its binding to 14-3-3σ with an K_D of 1.5 μM. The corresponding structure is analog to the stabilized complex of 14-3-3 like protein C and PMA2, which is shown in Fig. 2.10A-B. The binary complex of 14-3-3 and Task3 reveals a large cavity at the interface rim. This cavity is the binding pocket for FSC, which contacts both chains and thereby decreases the affinity to an apparent K_D of 50 nM. The protein-ligand interaction is mainly hydrophobic in nature. Only one deeply buried hydrogen bond is formed between the conserved Lys129 and FSC's methoxy group.

Three complex structures are related to 14-3-3∘PMA2 (PDB ID: 2o98). They all contain a stabilized protein complex consisting of a plant 14-3-3-like protein C or E and a C-terminal construct of phosphorylated PMA2. Again, the binary complexes reveal large cavities at the interface rim, which are occupied by stabilizing ligands. In crystal structure 3m51 the ligand is a pyrrolidone derivative (YR1) and in 4dx0 a pyrazole derivative (OMT).[25,164] Both compounds occupy a similar pocket as FSC and increase the affinity of the proteins. The structure of YR1 is shown in Fig. 5.1F. The binding mode of OMT is comparable. Epibestatin (EBT) is the stabilizing ligand in PDB entry 3m50 but binds to a different pocket in the interface rim (Fig. 5.1D).[25]

Crystal structure 2p1q, shown in Fig. 5.1A-C, contains a trimeric complex of the F-box protein TIR1, a peptide from the auxin-responsive protein IAA (Aux/IAA) and the phytohormone auxin (IAC).[165] IAC binds into a deep cavity formed at the interface rim of TIR1∘Aux/IAA. The indole part occupies a mainly hydrophobic pocket inside a leucine-rich repeat of TIR1 and is stabilized by a hydrogen bond to its nitrogen. The carboxy group contacts polar residues of TIR1. The Aux/IAA peptide π-stacks with a tryptophan side-chain onto the aromatic indole.

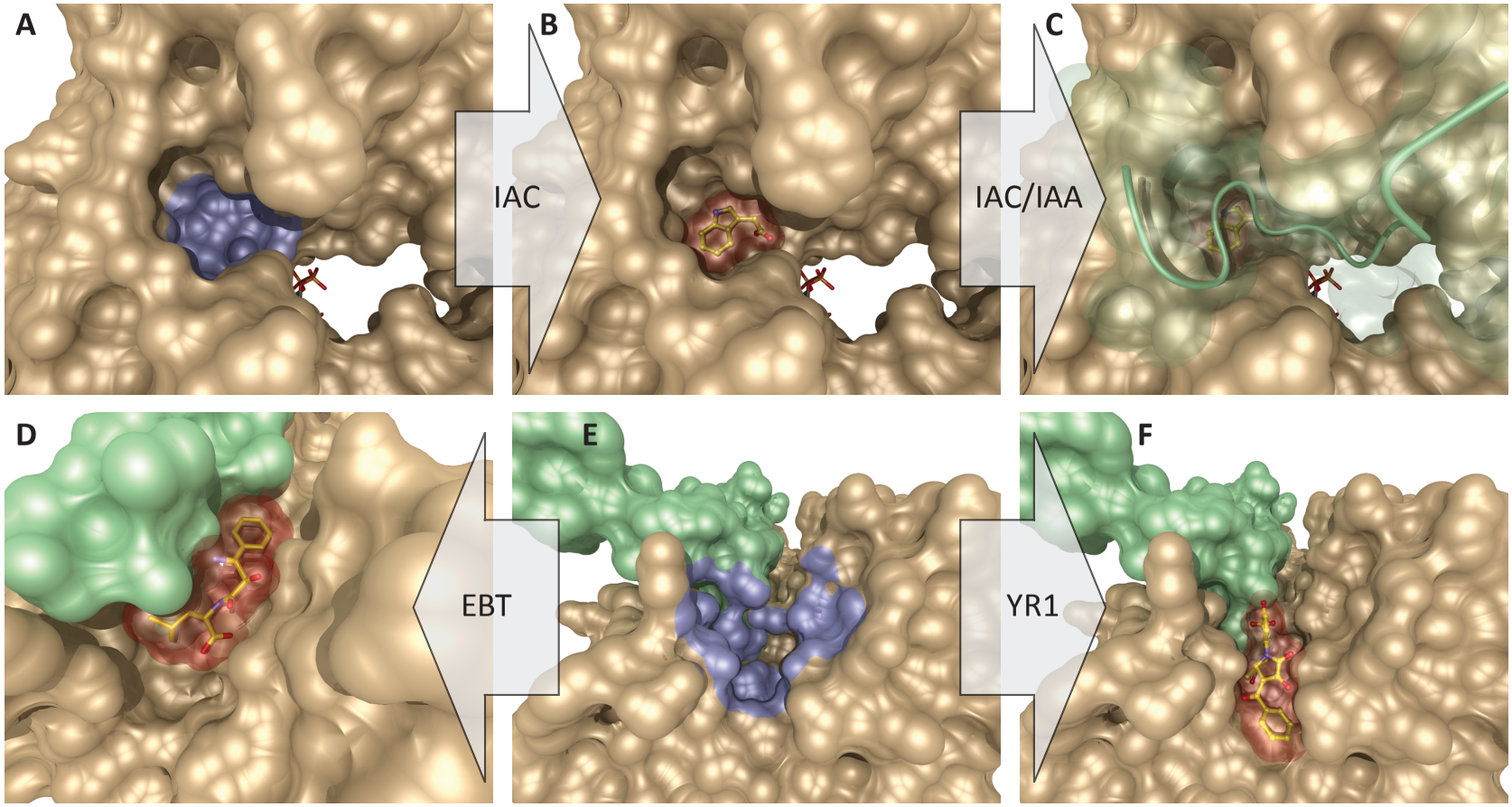


Figure 5.1: Stabilized PPIs. Ligands are represented as ball-and-stick models and a semitransparent SES. **(A)** TIR1 as gold SES. Binding pocket of IAC is highlighted in blue. (PDB ID: 2p1q) **(B)** AUX bound to the pocket in TIR1. **(C)** Aux/IAA stacked onto the complex of TIR1 and IAC. **(D)** EBT bound to the distant pocket in the interface of 14-3-3∘PMA2. (PDB ID: 3m51) **(E)** 14-3-3 as gold SES in complex with PMA2 shown as green SES. The binding pocket of FSC, YR1 and OMT is highlighted in blue (PDB ID: 3m50) **(F)** YR1 bound to the rim-exposed pocket in the interface of 14-3-3∘PMA2. (PDB ID: 3m51) This figure is partially taken from Thiel *et al.* (*Angew. Chem. Int. Ed.* (2012), **51**, 2012-8). Reproduction is granted under license number 3274110722038 of John Wiley and Sons.[22]

A comparable mechanism to TIR1∘Aux/IAA stabilization by IAC is found in structure 3ogk.[166] The binary receptor for ligand OGK is formed by the F-box protein coronatine insensitive 1 (COI1) and the transcriptional repressor JAZ1. The ligand OGK is also a phytohormone from the family of jasmonates. OGK is deeply buried in a hydrophobic pocket in the inner side of the COI1 leucine-rich repeat. Additionally, OGK is coordinated via four hydrogen bonds to COI1 side-chains. The contacts to JAZ1 are rather small but the ligand increases the affinity of this complex.

5.2.2 Analysis of PPI Stabilization

To describe the structural characteristics of PPI stabilization by small molecule ligands in a quantitative manner, we analyzed the selected ternary complex structures in two different ways: First, we analyzed the contact sizes of all interacting molecules. Second, we analyzed selected physicochemical properties of the stabilizing ligands.

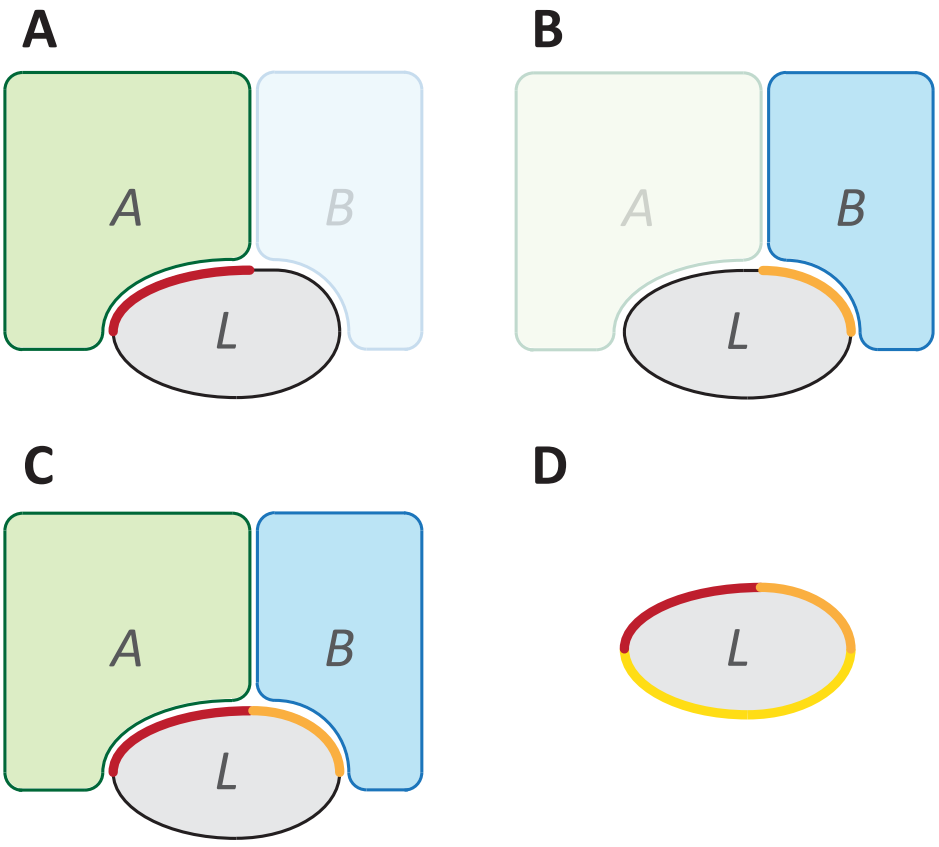


Figure 5.2: Evaluated surface areas of stabilized PPI complexes. Calculated BSAs are differently colored. **(A)** Contact surface between ligand and protein A (BSA_{AL}, red). **(B)** Contact surface between ligand and protein B (BSA_{BL}, orange). **(C)** Contact surface between ligand and the entire protein complex (BSA_{ABL}, red + orange). **(D)** Surface area of unbound ligand (red + orange + yellow), which is used to calculate $FracBSA_L$.

Ligand Analysis

To analyze the eight stabilizing small molecules we used the isomeric SMILES provided by the Chemical Component Dictionary of the PDB. These were used as input for physicochemical property calculation with Pipeline Pilot.[151] Molecules were generated from SMILES using standard parameters. Property calculators from the Chemistry package were used to calculate the number of hydrogen bond donors and acceptors, the MW, and the Alog$P$.[167–169] The latter is an estimate for the experimentally determined partition coefficient of octanol in water, log$P$.

Contact Size Analysis

In the following, we use A and B to denote the participating proteins and L to denote the stabilizing ligand of an analyzed PPI. We describe the interaction properties of a stabilized PPI $A{\circ}B{\circ}L$ by three parameters. To calculate them, we implemented a tool called *MultivalentInteractionAnalyzer*. It is built on top of the biochemical algorithms library BALL.[2]

Fig. 5.2 gives a schematic overview of the following parameters. The first parameter – BSA_{AL} – is the BSA size between protein A and ligand L in Å^2. The corresponding contact surface is highlighted in Fig. 5.2A. The second parameter – BSA_{BL} – is the BSA size between protein B and ligand L in Å^2. The corresponding contact surface is highlighted in Fig. 5.2B. The third parameter – $FracBSA_L$ – is the fraction of ligand surface area, which is buried upon binding to the rim-exposed surface pocket. To calculate the latter, we need BSA_{ABL}, which is the contact size between protein complex $A{\circ}B$ and ligand L as well as the total ligand surface area. Fig. 5.2C and 5.2D highlight these surfaces.

The *MultivalentInteractionAnalyzer* uses the Numerical Solvent-Accessible Surface (NumericalSAS) class of BALL to calculate these parameters. However, this class does not calculate BSAs directly but it provides functionality to calculate the solvent-accessible surface areas (SASA) of individual atoms in a structural assembly. The class uses a probe sphere of predefined size to calculate accessible surface points. We set the probe radius to 1.4 Å, which is commonly used for water molecules as solvent.[170] To retrieve the described BSA parameters, we calculate SASAs for L and for the subcomplexes $A{\circ}L$, $B{\circ}L$ and $A{\circ}B{\circ}L$. Using these values, the BSA parameters are calculated as follows:

$$BSA_{AL} = \sum_{i=1}^{|LigandAtoms|} SASA_L(atom_{L,i}) - SASA_{AL}(atom_{L,i}) \tag{5.1}$$

$$BSA_{BL} = \sum_{i=1}^{|LigandAtoms|} SASA_L(atom_{L,i}) - SASA_{BL}(atom_{L,i}) \tag{5.2}$$

$$BSA_{ABL} = \sum_{i=1}^{|LigandAtoms|} SASA_L(atom_{L,i}) - SASA_{ABL}(atom_{L,i}) \tag{5.3}$$

The four functions $SASA_{ABL}$, $SASA_{AL}$, $SASA_{BL}$ and $SASA_{L}$ return the SASA of ligand atom i using the complexes $A{\circ}B{\circ}L$, $A{\circ}L$, $B{\circ}L$ and unbound ligand as input, respectively. Parameter $FracBSA_{L}$ is calculated by the following equation:

$$FracBSA_L = \frac{BSA_{ABL}}{\sum_{i=1}^{|LigandAtoms|} SASA_L(atom_{L,i})} \tag{5.4}$$

5.2.3 Screening the Protein Data Bank for Stabilizer Candidates

Based on the properties which we inferred from analyzing currently described stabilized PPI complexes, we set up a screening for potentially ovelooked stabilized complexes within the PDB. Therefore, we implemented a workflow to filter the PDB to select possible candidate PPIs, which are possibly stabilized by a small molecule ligand. This workflow is shown in Fig. 5.3.

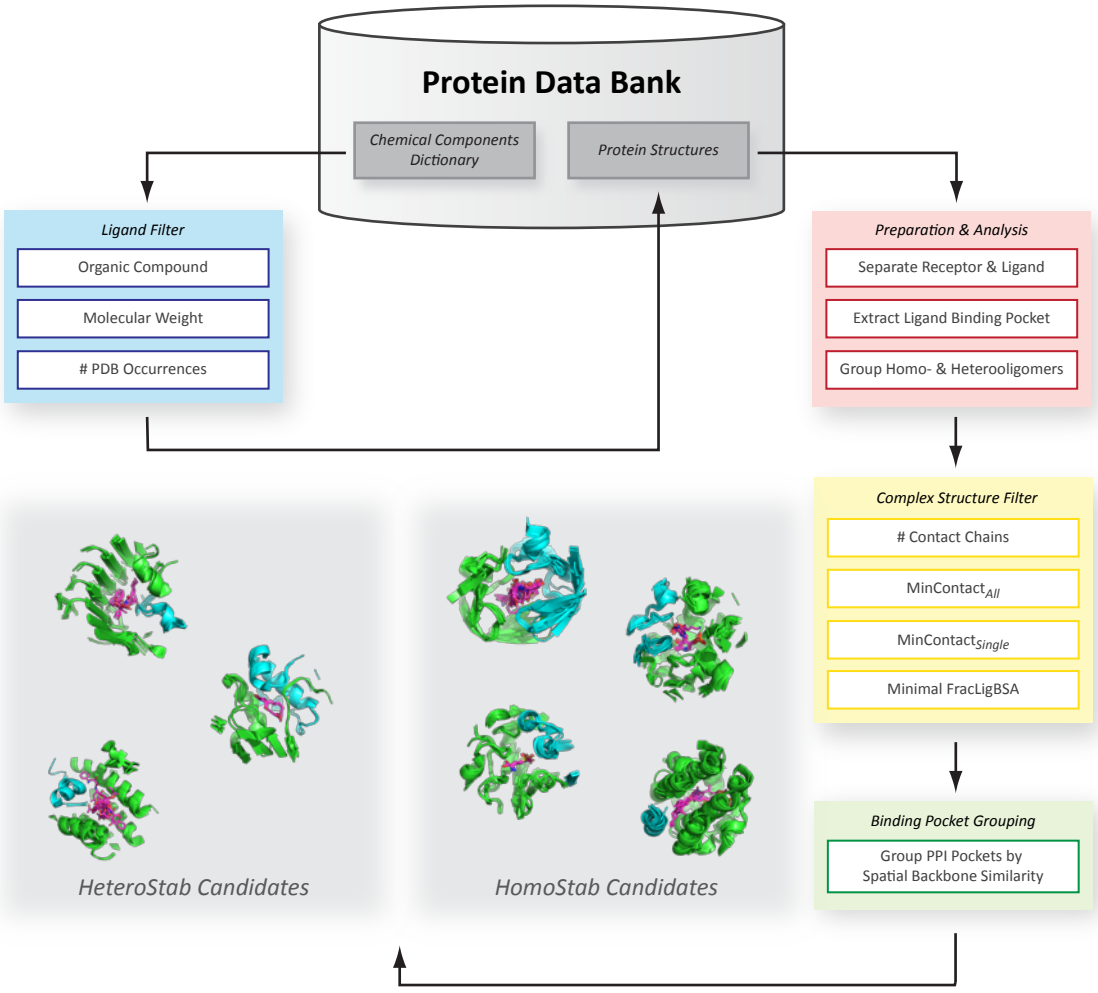


Figure 5.3: Filtering the PDB for candidates of small molecule stabilized PPIs.

Ligand-guided Preselection

The first workflow step is a filtering of the PDB Chemical Component Dictionary. This dictionary contains all non-peptidic ligands occurring in the PDB. The workflow was set up in Pipeline Pilot.[151] We obtained the dictionary in SD format and started by removing all inorganic compounds. This filter rejects compounds containing other atom types than H, C, N, O, P, S, F, Cl, Br, or I. Thereby compounds like metal clusters were removed, in which we were not interested. The second step was an MW filter. Here, we used the lowest and highest MW from the previously discussed stabilizer analysis as a lower and an upper threshold. Finally, we rejected all compounds that were present in more than 15 PDB entries. This value has been arbitrarily chosen in order to include the most frequent PPI stabilizer FOK. We included this constraint because various chemical components are frequently found in crystallization buffers and can be regarded as promiscuous binders. For example, glycerol is present in more than 8,000 PDB entries. For the remaining chemical components we obtained the corresponding PDB entries for subsequent structure-based analysis.

Preparation of Preselected Candidate Structures

For further analysis, we split every PDB file into a receptor and a ligand file. We exclusively extracted the ligand binding pocket from the receptor part. As a simple criterion to define the ligand binding pocket, we selected all residues containing at least one atom with a maximum distance of 17.0 Å to the ligand's geometric center. This value has been arbitrarily chosen based on the observation that for the known stabilizer-bound structures a sufficiently large pocket was extracted. We only kept receptors and corresponding ligands if the receptor consisted of at least two protein chains to ensure that only PPI complexes were selected.

Additionally, we extracted information of the receptor chains to group the candidates into homo- and heterooligomeric PPIs. The latter is not a straightforward procedure because the necessary information in PDB file headers are rather inconsistent. However, a substantial part can be captured by checking for the occurrence of a single COMPND entry listing multiple chain identifiers. This uniquely identifies homooligomers. In the case of multiple COMPND entries, we checked if they contain the same molecules by comparing their names, which can also be used to distinguish between homo- and heterooligomers.

PPI Analysis of Preselected Candidate Structures

We used the *MultivalentInteractionAnalyzer* to calculate the contact properties of the prepared structures in order to filter for stabilized PPI candidates. For this purpose, we used the BSA parameter values we retrieved from the known stabilized PPIs to define three filter criteria.

$FracBSA_L^{Min}$: Criteria to reject candidate structures with too few contacts between ligand and complex. As lower cutoff we used the minimum observed $FracBSA_L$ within the set of known PPI stabilizers.

BSA_{CL}^{Minor}: Criteria to reject candidate structures where at least one protein-ligand contact is too small. As lower cutoff we applied the smallest observed BSA_{CL} between ligand and any single protein contact chain C within the set of known PPI stabilizers.

BSA_{CL}^{Major}: Criteria to reject candidate structures where no protein-ligand contact is large enough. This is also a lower cutoff and was chosen as the smallest BSA_{CL}, where L is again the stabilizing ligand and C are those contact chains from every analyzed PPI that form the larger contact area with the ligand.

Pocket-guided Grouping of PPI Stabilizing Candidates

Finally, we grouped the remaining binding pockets by their 3D structural conservation. For this purpose we implemented the BALL tool *PPIReceptorMapper*. This tool takes as input two extracted PPI receptors and calculates an optimal mapping of their C_α backbone using BALL's StructureMapper class with default parameters. Our tool provides the optional parameters CA_{min} and $RMSD_{max}$, which can be used to control if the mapping of a receptor pair is accepted, that is if they are structurally conserved. Parameter CA_{min} is a lower cutoff for the number of C_α atoms that must be aligned to accept a receptor pair. $RMSD_{max}$ is an upper cutoff that receptor pairs mustn't exceed in order to be an accepted match. Using a straightforward procedure, we iterated over all PPI receptor candidates and grouped successfully superposed receptor pairs and the corresponding ligands. For every evaluated receptor, we first searched for matching pocket in already existing groups. In case of a group match, the receptor was added to it. Otherwise, it was compared to all ungrouped receptors and in case of an accepted matching a new group was formed.

5.2.4 Redocking of Known PPI Stabilizers

To test the ability of protein-ligand docking software to correctly reproduce and rank the binding poses of small molecule PPI stabilizers, we performed redocking experiments of a subset of the known and newly identified stabilizing ligands. As an exemplary protein-ligand docking software we evaluated Glide from the Schrödinger Small-Molecule Drug Discovery Suite.[171]

Structure Preparation

Protein-ligand structures from the PDB were analyzed and prepared using Schrödinger Maestro.[172] First, we deleted all waters as well as other organic compounds. Additionally, all ions

without contact to the ligand were deleted. To prepare the PPI receptors we also removed the ligands from their binding pockets. Preparation was performed according to the *Protein Preparation Wizard* protocol with slightly modified settings: In case of alternative side-chain conformations, we kept the candidate with highest occupancy. We further checked if all side-chains in the binding pocket and adjacent to it were complete. We included sampling of water orientations in the hydrogen bond refinement step. Finally, we applied restrained energy minimization to the entire protein.

Suitable ligand conformations for docking were generated using LigPrep.[173] As input structures we used the crystallized ligands. Possible protonation states were generated in a pH range from 5.0 to 9.0. We disabled the option to generate tautomers and forced the software to determine ligand chiralities from the input 3D structures. A single low energy ring conformation was calculated for every ligand.

Receptor Grid Calculation and Protein-Ligand Docking

The rectangular grid boxes capturing the PPI receptor were centered at the ligand centers. The dimensions of the inner box were kept at the default size of 10 Å. The dimensions of the outer boxes were adjusted for the docking of ligands with similar size as the reference ligand. Calculation of receptor grids was performed without constraints using Glide.[171]

Flexible docking of the prepared ligands into the PPI receptor grid was also performed using the Glide single precision (SP) scoring function. Default parameters were used with two exceptions: The number of poses included for post-docking minimization was increased to 20 for every ligand and ten poses per ligand were written to the output.

Analysis of Docking Accuracy

To assess the accuracy of the employed docking software to reproduce and to rank PPI-ligand complexes, we calculated the RMSD between the top-ranked docking pose according to the glide docking score and the crystallized ligand conformation. Additionally, we calculated the RMSDs of the docking ranks 2-5 and the native ligand conformation to evaluate if better or possibly correct solutions are found on lower ranks.

5.2.5 Virtual Screening for 14-3-3∘Task3 Stabilizers

To translate the insights we gained from analyzing PPI stabilization by small molecules, we integrated this knowledge into a VS for ligands with the potency to stabilize the interaction of 14-3-3σ and the previously described potassium channel Task3 (Section 5.2.1). We decided to use this target as our model system because we had access to *in vitro* assays for validation as well as to materials and methods for crystallization of this target. Furthermore, Task3 has

been shown to be overexpressed in several types of cancer and it is linked to neuropathological disorders like ischemia or epilepsy.[174,175] Thus, small molecule stabilizers for this PPI would form useful tools in chemical biology research or even could serve as candidates for drug development. In a previous study, we evaluated the druggability of a comparable complex of 14-3-3 and PMA2 using the prediction method SCREEN developed by Nayal and Honig.[176] Here, this complex achieved a druggability index $DI > 0.8$ and is thus classified as highly druggable.[110]

Structure Analysis

Various high-resolution structures of 14-3-3σ in complex with mode III phosphopeptides are available from the PDB as well as currently unpublished data. We took a subset of these structures for further analysis and to select an appropriate model for structure-based VS. The analyzed structures are listed in the appendix (Tables E.2 and E.3).

All structures possess a single monomer of 14-3-3σ in the asymmetric unit. To perform further analysis we superposed these complexes using the macromolecular modeling tool Coot.[177,178] Superposition was performed using Coot's implementation of the Secondary Structure Matching (SSM) algorithm developed by Krissinel and Henrick.[179] As reference structure for superposition we used 14-3-3 chain A from PDB entry 3p1n without bound peptide.

Analysing Conserved Waters

Due to the availability of high-quality structures for 14-3-3σ in complex with naturally occurring and artificial mode III phosphopeptides, we performed an in-depth-analysis of the crystallographically resolved waters. We were especially interested in their spatial conservation throughout all analyzed structures to get hints which waters might to be kept for structure-based VS.

For this purpose, we implemented the tool *ConservedWaterFinder* using BALL. The tool requires a set of superposed PDB structures for water analysis. In the first step the distribution of water B-factors is analyzed for every input structure separately. Due to the dependence of the atomic B-factor on crystallographic resolution we performed a z-score normalization of the water B-factors using Eq. 5.5. As robust estimates for mean and standard deviation we used median and median absolute deviation (MAD), respectively.

$$z(w_i) = \frac{Median([b(w_1), b(w_n)]) - b(w_i)}{MAD([b(w_1), b(w_n)])} \tag{5.5}$$

Here, n is the number of waters and $b(w_i)$ is the B-factor of water i. $Median([b(w_1), b(w_n)])$ is the B-factor median of all waters. The MAD is calculated equivalently. As we were interested

in positive z for waters with low temperature factors, we substracted $b(w_i)$ from the median in the numerator. The calculated z-scores can be used to filter for spatially highly conserved crystallographic waters with respect to the considered crystal structure. The *ConservedWaterFinder* thus provides a parameter z_{cut}, which is used as a filter criterion to delete waters w_i with $z(b(w_i)) < z_{cut}$.

In the next step, the waters of all input structures are inserted into a global water map. This water map is then hierarchically clustered using *Ward's* minimum variance method with pairwise water distances as dissimilarity measure.[46] Again, the *ConservedWaterFinder* provides a cluster selection parameter *cs* to cut the cluster hierarchy. Additionally, an optional parameter s_{min} can be specified to delete clusters c_i with $|c_i| < s_{min}$. Finally, all waters that are part of a remaining conserved water cluster are written into a PDB file for visual inspection.

Compound Library for Virtual Screening

As compound library for virtual screening we used a subset of the custom vendor library introduced in Section 4.2. All compounds and their *Murcko scaffolds* were encoded as ECFPs using pharmacophore feature atom typing. To select this subset we clustered the vendor library using the previously introduced chemoinformatics tools with some modifications.

In brief, we used scaffold fingerprints as input and skipped CC decomposition. We directly used the parallel RNN clustering method to merge RNN *scaffold families* up to a lower similarity cutoff $S_{Tan} = 0.6$. The resulting clusters were then hierarchically clustered using the complete molecule fingerprints if their size exceeded 1,000 compounds. The medoid selection procedure was applied to select cluster representatives, which were included into the virtual screening library. The latter finally comprised 197,062 compounds.

Protein-Ligand Docking

The virtual screening library was prepared with *LigPrep* using default parameters except using Ionizer instead of Epik for ionization (option: –i 2).[173] This produced 513,900 conformers as input for protein-ligand docking.

The PPI receptor preparation was performed using the *Protein Preparation Wizard* of Schrödinger Maestro with slightly modified settings:[172] We kept the entire ternary complex and the waters surrounding reference ligand FSC for optimization of hydrogen bond networks. We restricted the energy minimization to hydrogens only. Finally, we removed all water molecules with the exception of those waters that were defined as conserved on the basis of the previously described conserved water analysis. The receptor grid generation was performed with Glide using default settings.[180] The grid center was defined by selecting FSC as the reference ligand. The cubic grid box had the outer dimensions of $x = y = z = 27.71$ Å.

The prepared ligand conformers were docked into the 14-3-3σ∘Task3 receptor grid using Glide in SP mode.[54,171,181] We set the maximum number of reported poses to 200,000 compounds and used default settings for all other options. For further processing, the docked compound poses were sorted according to decreasing docking score.

In order to choose an appropriate upper docking score cutoff to select solutions for further analysis, we also calculated the docking score of the reference ligand in place. Here, we used the FSC conformation from protein preparation as input ligand and calculated the docking score using the Glide SP scoring function.

PPI Contact Filter

To tailor the structure-based VS approach to the discovery of PPI stabilizing ligands, we filtered all generated docking poses using our *MultivalentInteractionAnalyzer*. Here, we again used the filtering criteria $FracBSA_L^{Min}$, BSA_{CL}^{Minor}, and BSA_{CL}^{Major}, which were introduced earlier in this section. In this way, we rejected all docking poses that did not match the PPI stabilizer criteria we derived from structure analysis described in Section 5.2.2.

Reference Ligand Filter

As a final step, we exploited the availability of the co-crystallized reference ligand FSC, which was not incorporated up to this point. However, none of the available studies yields information on SAR, which we could have used to define constraints on the docking poses. Thus, we decided to score the volume overlap of the filtered docking poses and the reference ligand FSC and to rank the remaining poses accordingly. For this step, we used the free software tool *Shape-it* from Silicos-it.[182] This software uses a method to align molecules described by Gaussian atom descriptions.[183] *Shape-it* calculates a Tanimoto-based volume overlap score which we used for final docking pose ranking.

Compound Selection

Based on this final ranking, we selected compounds for experimental validation. Selected candidates were acquired from MolPort (http://www.molport.com/). Compounds were delivered in solid form. We dissolved the compounds in dimethyl sulfoxide (DMSO) and stored them as 20 mM stock solutions at -20 °C.

In Vitro Validation of Selected Compounds

To assess the compounds' potential to stabilize the protein interaction of a Histidin-tagged 14-3-3σ and FAM-labeled phosphopeptide from Task3 *in vitro*, an fluorescence polarization (FP)-based assay was performed (Maria Bartel, TU Eindhoven).[184] An initial validation screen

was performed to identify potential candidates, which were analyzed in more detail to determine the compounds' EC_{50}.

Additionally, three control experiments were conducted. First, the assay was performed without peptide to test if the compounds themselves interfere with the assay. Second, the Task3 phosphopeptide was replaced by a peptide from C-Raf (residues 252-264) surrounding phosphorylated Ser259. This C-Raf peptide is no mode III motif and the sequence exceeds the +1 position following the phosphorylated residue. This markedly reduces the size of the binding pocket in the 14-3-3∘C-Raf complex and prevents FSC binding and complex stabilization. Finally, the compounds were tested on an entirely different PPI using the same assay. This target complex consisted of retinoid X receptor with a co-repressor and a known ligand as a negative control.

Protein Crystallization

To verify our structure-based VS approach we also tried to obtain ligand-bound crystal structures of the validated 14-3-3σ∘Task3 stabilizers. 14-3-3σ was used as a C-terminally shortened construct (14-3-3σΔC), which was truncated after Thr231. Cloning, expression and purification of 14-3-3σΔC was performed as described by Schumacher *et al.*.[185] Protein was provided by the Ottmann lab. Task3 peptides were synthesized by Biosyntan (Berlin) and re-suspended in Millipore water. Complexation solution was mixed from 14-3-3σ and peptide at a molar ratio of 1:2 and diluted using the complexation buffer listed in the appendix (Table E.7) to reach a final 14-3-3σΔC concentration of 14 mg/ml.

Crystallization buffers were created as a 2D grid variation consisting of 24 different conditions (C1-C24), which is based on a successful crystallization buffer for complexes of 14-3-3σ with a phosphopeptide from YAP. The original condition and the derived grid variations are listed in the appendix (Table E.8). The pH values of the Na-HEPES buffers were pre-adjusted and these solutions were used to set up the crystallization conditions with a final concentration of 95 mM Na-HEPES. This procedure has been performed in accordance to the production report of the original crystallization buffer (JCSG Core I, Qiagen®).

For soaking experiments, we performed hanging-drop crystallization experiments at 4 °C. For this purpose, reservoirs from a 24-well plate were filled with 500 µl of crystallization buffer. Hanging-drops were mixed from 2 µl complex solution with 2 µl crystallization buffer from the reservoir and placed on a cover glass, which was used to seal the well. Diffracting crystals grew within a week. We then soaked these crystals with candidate compounds. Therefore, we injected compound DMSO stock solution directly into the hanging-drops and incubated them four days. As the mother liquors in the hanging-drops were cryogenic, we directly flash-froze the crystals in liquid nitrogen for subsequent diffraction experiments.

Data Collection and Structure Elucidation

Diffraction experiments were performed in-house using a rotating copper anode from Rigaku (MicroMAX-007 HF) as an X-ray source. This beamline was equipped with a MAR345 image plate. Data processing was carried out using the software package XDS.[186] During this step, choice of the lowest resolution to trim the data sets was based on three parameter values of the outermost shells. First, we required a minimum signal to noise ratio of 4.0. Second, we required a minimum completeness of 90.0 %. Third, we required the redundancy independent R-factor (R_{meas}) to be lower than 40.0 %. Crystal parameters and data collection statistics are listed in the appendix (Table E.9).

The CCP4 software collection for macromolecular X-ray crystallography was used for phase determination and automatic refinement.[187] MR was carried out with PHASER (version 2.1.4).[188] PDB entry 3p1n without waters was used as a search model. This model contains an identical complex of 14-3-3σ and Task3 phosphopeptide. MR solutions were used as starting data for iterative cycles of manual and automatic refinement with REFMAC (version 5.5) and COOT (version 0.6), respectively.[178,189] Final refinement statistics are listed in the appendix (Table E.9). Corresponding Ramachandran plots were generated with RAMPAGE and are shown in the appendix (Fig. D.4).

5.3 Results

The first part of this section shows the characteristics for small molecule stabilizers of PPIs and the analyzed determinants that we have drawn from the interaction properties with their PPI receptors. Next, we present the results of screening the PDB for unrecognized small molecule stabilized PPIs and the outcome of our stabilized redocking experiments. Finally, we present the results of our VS approach for the discovery of small molecule PPI stabilizers for the 14-3-3$\sigma\circ$Task3 PPI.

5.3.1 Structural Characterization of PPI Stabilization

Table 5.1 lists the calculated structural and physicochemical properties of the eight evaluated small molecule PPI stabilizers. Additionally, it shows the Ro5 thresholds for these properties. With the exception of FSC, all analyzed small molecules lie within the Ro5 limits and thus are typical druglike compounds. However, the fungal metabolite FSC is a plant toxin and thus not expected to be Ro5 compliant.

Table 5.1: Structural and physicochemical properties of small molecule PPI stabilizers. These properties form the Ro5 and are commonly used to describe the oral bioavailability. The Ro5 limits are listed in the bottom row.

		Hydrogen bond		
Ligand ID	*MW*	*Donors*	*Acceptors*	*AlogP*
AFB	280.4	2	4	2.2
EBT	308.4	4	5	-1.3
FOK	410.5	3	7	0.8
FSC	**680.8**	4	**12**	1.9
IAC	175.1	2	2	1.8
0MT	440.4	2	6	4.4
OGK	321.4	2	4	2.4
YR1	460.4	3	8	3.2
Ro5	< 500.0	< 5	< 10	< 5.0

Contact Analysis of Stabilized PPIs

The initial properties we analyzed were the contact surface sizes, BSA_{AL} and BSA_{BL}, of the ligands to every single protein chain of the stabilized PPI, which is shown in Fig. 5.4. Subfigure A shows the measured BSAs in $Å^2$. Fig. 5.4B displays the ratios of the smaller and the larger BSAs. A contact ratio of 1.0 would indicate equally sized contact surface sizes of the ligand to both PPI receptor chains.

Interestingly, these contact ratios span an extremely wide range. The smallest ratio is < 0.1 and it is observed for the phytohormone OGK binding to its PPI receptor. This is due to a very small surface contact between OGK and the minor contact chain as can be seen in Fig. 5.4A. In fact, it is the smallest contact size observed for all analyzed PPI complexes. In contrast, the ratios for EBT and FOK are 0.96 and 0.99, respectively. Both ligands share nearly equally sized contacts with both of the PPI receptor chains. In general, the apparent trend seems to be either the occurrence of one chain with large ligand contacts and one chain with small ligand contacts or complexes with equally distributed protein-ligand contacts.

Fig. 5.4C shows the results of evaluating the fraction of surface area that a stabilizing ligand looses upon binding to a rim-exposed PPI pocket ($FracBSA_L$). The result values span from 0.66 for FSC binding to 14-3-3σ∘Task3 up to 0.98 for IAC binding to AuxIAA∘TIR1. Interestingly, the lowest five $FracBSA_L$ values are observed for the complexes involving 14-3-3 proteins.

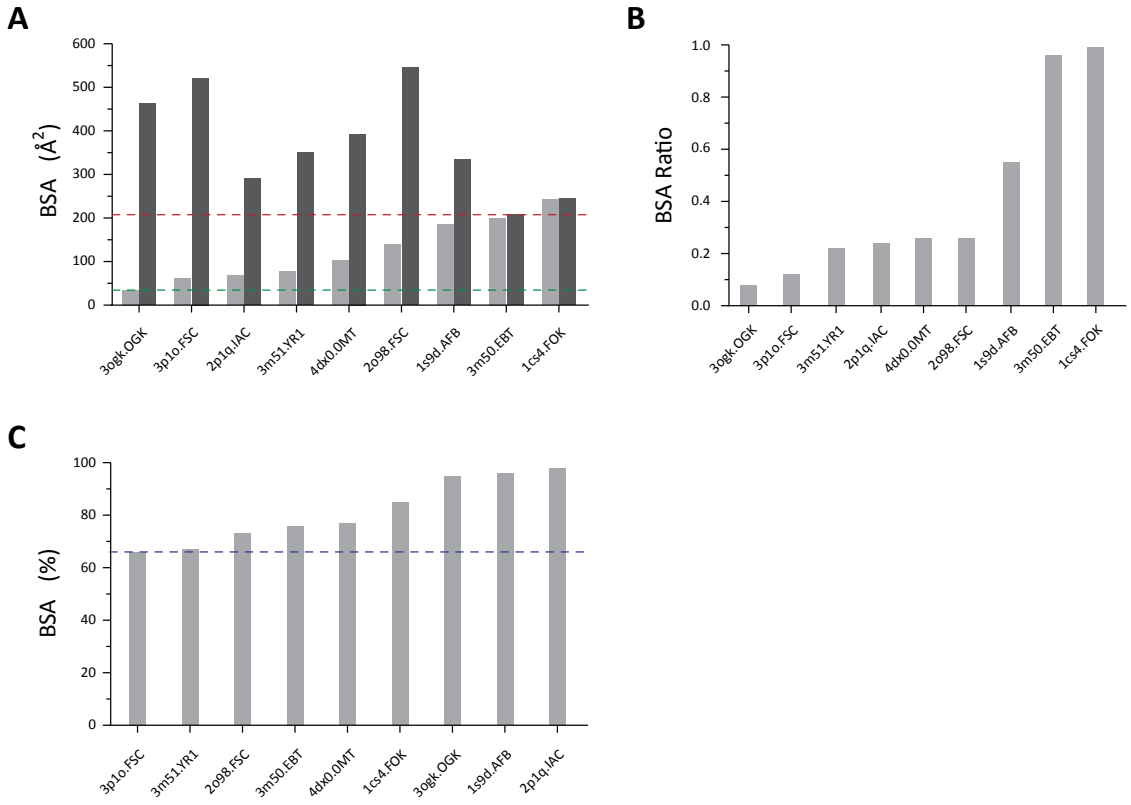


Figure 5.4: Surface contact analysis. The analyzed PPI complexes are identified by their PDB ID and the Ligand ID in the format <PDB_ID.Ligand_ID>. **(A)** Contact sizes between ligand and receptor chains (BSA_{AL} and BSA_{BL}). Light grey bars represent smaller ligand-chain contact sizes and dark grey bars the larger ones, respectively. Inferred lower cutoffs for filter criteria are indicated: BSA_{CL}^{Minor} (dashed green line) and BSA_{CL}^{Major} (dashed red line). **(B)** Contact ratios of BSA_{AL} and BSA_{BL}. **(C)** $FracBSA_L$ of the analyzed stabilizing ligands. The inferred lower cutoff for our filter criteria $FracBSA_L^{Min}$ is indicated by a dashed line.

5.3.2 Stabilizer Candidates in the Protein Data Bank

The workflow for our PDB screening is shown in Fig. 5.3. The Chemical Components Dictionary of the PDB comprised 16,727 compounds at that time (accessed 8/8/2013). The results of the initial ligand-based filtering are listed in Table 5.2. As upper and lower MW cutoffs we used 700 Da and 170 Da, respectively. These values were chosen on the basis of the MW range of known PPI stabilizers listed in Table 5.1. Besides these MW filters we applied no further Ro5 filter criteria because they rather reflect pharmacokinetic and -dynamic properties of compounds, which is of minor interest to this study. The applied filters reduced the number of candidate ligands down to 12,032 compounds.

Next, we analyzed the entire PDB structure of every candidate ligand. Based on the observed PPI contact sizes described in the last subsection we set our PPI stabilizer filter criteria

Table 5.2: PDB ligand filtering results.

Filter	*Limit*	*Compounds*
Organic		15,977
MW lower	≥ 170	13,694
MW upper	≤ 700	12,946
PDB occurrence	≤ 15	12,032

as listed in Table 5.3a. The corresponding cutoffs are highlighted in Fig. 5.4A and 5.4C. We have not taken the contact ratio into account because we subsume that the real ligand-chain contact sizes are more important than their ratio. Additionally, the contact ratios span such a wide range that a filter criterion most probably would not be discriminative. As listed in Table 5.3b, these settings finally yielded 2,045 candidate PPIs, whereof 412 were classified as heterooligomers and 1,633 as homooligomers. To reduce these candidates to a non-redundant set, we grouped them by receptor conservation using the introduced *PPIReceptorMapper*. Mapping parameters were set to $CA_{min} \geq 35$ and $RMSD_{max} \leq 1.0$ Å. These settings finally yielded 86 unique heterooligomeric PPI candidates and 333 unique homooligomeric PPI candidates.

Table 5.3: Filter criteria and results of the structure-based screening for potentially overlooked stabilized PPIs in the PDB.

(a) Applied filtering criteria

Filter criteria	*Unit*	*Cutoff*
$FracBSA_L^{Min}$		0.66
BSA_{CL}^{Minor}	Å^2	34.0
BSA_{CL}^{Major}	Å^2	207.0

(b) Resulting stabilized PPI candidates

	Heterooligomer	*Homooligomer*
Total	412	1,633
Mapped	49	175
Unmapped	37	158

Manual Validation of True PPI Stabilizers

As we were interested to know if the selected candidates contained true stabilizers we manually inspected the 86 heterooligomeric PPI-ligand complexes. For this purpose, we searched in the corresponding PDB entries and the original literature for reports on a stabilizing effect of these ligands on the PPIs. Indeed, we found such information for six of these complexes, which are listed in Table 5.4 and briefly described in the following.

The structure of the enzyme histone deacetylase 3 (HDAC3) in complex with a co-repressor domain was recently reported by Watson *et al.*.[190] Binding of co-repressors to HDAC3 can regulate their enzymatic activity. In this work, the structure reveals an inositol tetraphosphate

molecule which acts as a molecular glue between the proteins. In the recent past, histone deacetylases moved into the focus as anticancer targets and this mechanism possibly hints at a strategy to modulate these enzymes.

N-methyl-D-aspartate (NMDA) receptors are transmembrane protein complexes, whose natural ligand is glutamate.[191] These receptors form heteromeric ion-channels and some family members contain the protein subunits GluN1 and GluN2B. The ion-channel activity of these subtypes can be modulated in an allosteric way by phenylethanolamine derivatives like the inhibitor *Ifenprodil*, which has neuroprotective activity.[192] The mechanism of these compounds has recently been identified by Karakas *et al.*.[193] They solved a crystal structure where the compound binds into a rim-exposed pocket in the interface of GluN1 and GluN2B. Binding leads to a stabilization of the heterodimer and establishes an interesting approach to modulate this important target class.

The actin-related protein (Arp) 2/3 protein complex is a heteromeric ATPase, which attaches to existing actin filaments and triggers the nucleation of new branches.[194] Small molecule inhibitors of this function have been described and their mode of action was revealed by crystal structures reported by Nolen *et al.*.[195] Some of these compounds bind into a rim-exposed pocket in the heterodimer interface of Arp2∘Arp3. The described mechanism is a blocked movement of these subunits into their active conformations, which can also be considered as a stabilization of a protein interaction.

We found several examples for related nuclear receptors whose regulatory effect on target-gene expression is physiologically modulated by binding to specific co-repressor proteins. An example is the retinoic acid receptor alpha (RARα), whose activity can be silenced by binding of nuclear receptor co-repressor 1. Le Maire *et al.* recently reported on a small molecule that strengthens this interaction and supports silencing.[196] Again, the crystal structure shows that the small molecule contacts both interacting proteins.

A well known example for an allosterically stabilized PPI is the $\alpha\circ\beta$ heterodimer in polymerized tubulin, which has been described in Section 2.4.3. In addition to the allosteric stabilizers we found direct stabilizers like the alkaloid colchicin. This molecule binds to the interface of the α and β subunits.[197]

Finally, we identified multiple PDB entries containing viral capsids, which are composed of different subunits. One example is a capsid from a human rhenovirus containing the coat proteins VP1 and VP2 as subunits. Here, a small molecule binds to an interface pocket of VP1∘VP2.[198] Binding of these compounds stabilizes the capsid and hinders uncoating of the virus, which forms an interesting starting point for antiviral drug development.

Table 5.4: Verified stabilized PPIs from PDB screening. The highlighted PPIs were already known and used for initial structure analysis. E: number of PDB entries; L: unique ligands; PDB: representative PDB entry with ligand ID in brackets; R: resolution in Å.

PPI	*E*	*L*	*PDB*	*Ligand name*	*R*
HDAC3 and co-repressor	1	1	4a69 (i0p)	D-myo-Inositol 1,4,5,6-tetrakis(phosphate)	2.06
NMDA receptor subunits	2	2	3qel (qel)	4-[(1R,2S)-2-(4-benzylpiperidin-1-yl)-1-hydroxypropyl]phenol	2.60
Arp2 and Arp3	2	2	3ukr (ckh)	2-fluoro-N-[2-(2-methyl-1H-indol-3-yl)ethyl]benzamide	2.48
Nuclear receptors and co-repressors	6	5	3kmz (eqo)	4-(E)-2-[5,5-dimethyl-8-(phenylethynyl)-5,6-dihydronaphthalen-2-yl]ethenylbenzoic acid	2.10
Tubulin α and β chain	8	4	3ut5 (loc)	Colchicine	2.73
Viral capsid protein subunits	21	17	1ncq (w11)	WIN63843	2.50
14-3-3 and phosphopeptides	8	7	3p1o (fsc)	Fusicoccin A	1.90
F-box protein and target	6	5	2p1q (iac)	Indole-3-acetic acid	1.91
ARF1 and Sec7	3	1	1s9d (afb)	Brefeldin A	1.80
C_{1a} and C_{2a} of AC	26	8	1cul (fok)	Forskolin	2.40

5.3.3 Redocking of Known PPI Stabilizers

For redocking experiments, we selected one representative PPI-ligand complex from every identified target group listed in Table 5.4. If a target group comprised more than one PPI-ligand complex, we have chosen the entry with best crystallographic resolution.

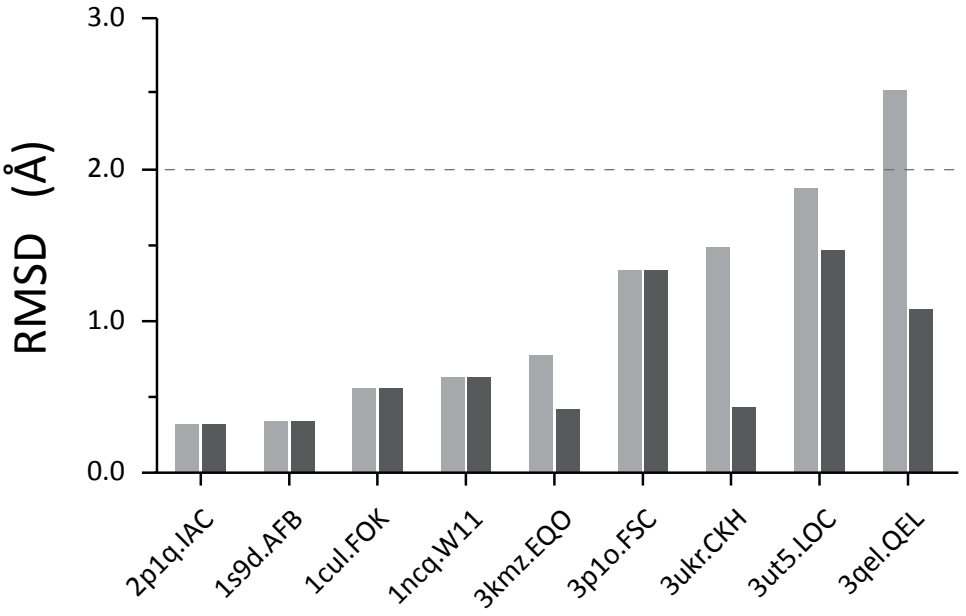


Figure 5.5: Redocking of a representative subset of PPI stabilizing small molecules (nomenclature: <PDB_ID.Ligand_ID>). The light grey bars indicate the RMSD between the stabilizers' experimental geometry to the top-ranked docking pose. The dark grey bars indicate the lowest RMSD observed between the stabilizers' experimental geometry and the five highest scoring docking poses.

For HDAC3 in complex with a co-repressor we only found one example in the PDB (4a69). The crystal structure shows HDAC3 in complex with a co-repressor, which is described to be stabilized by an inositol tetraphosphate (IP_4) molecule. In addition to IP_4, two glycerol molecules are located in the same binding pocket, which extensively contact the ligand. These contacts bridge the space between IP_4 and the HDAC3∘co-repressor complex. In principle, it is possible to treat these glycerol molecules as part of the receptor, but we have decided to skip this PPI because this arrangement appeared too artificial as to be included in this study.

All redocking experiments were performed as described in Section 5.2.4. For PDB entry 1ncq manual intervention was necessary because assigned bond orders of ligand W11 were wrong. Bond orders were corrected based on the chemical component description of W11.

Fig. 5.5 summarizes the redocking results. The bar chart shows RMSD values for non-hydrogen atoms between crystallized ligand conformation and docked ligand poses. The light-grey bars were calculated for the top-ranked docking pose according to the Glide SP

docking score. Dark-grey bars represent the pose within the five highest ranking docking poses that has the lowest RMSD to the crystal ligand conformation. Thus, if both bars have the same height, the top-ranked ligand is also closest to the native pose. The RMSD threshold of 2.0 Å is highlighted as a dashed grey line, which has frequently been used to evaluate the accuracy of docking algorithms and it is a commonly accepted cutoff to classify docking poses into correctly predicted ($\leq$ 2.0 Å) and false predicted ($>$ 2.0 Å).[199–201]

The docking software predicted correctly ranked docking poses for nine out of ten stabilizing ligands when docked into their native PPI complexes. For the stabilizers IAC, AFB, FOK, W11 and FSC the lowest RMSD pose is also the one with lowest Glide SP score. A more accurate ligand pose is found for the stabilizers EQO, CKH and LOC within the docking ranks 2-5. However, the difference between these conformations and the top-ranked pose is only 0.57 Å on average. For these cases, a pose clustering of the docking conformations would group these solutions together. The largest deviation is found for stabilizer CKH with an RMSD of 1.06 Å. This is due to the freely rotatable *para*-fluorobenzene ring of the ligands 2-fluorobenzamide moiety, which is flipped by 180° as shown in Fig. 5.6.

Only the docking of ligand QEL into the complex of GluN1 and GluN2B failed to rank a correct ligand conformation best. As indicated by the dark grey bar for this complex, the docking algorithm samples a correct ligand pose but the scoring function failed to correctly score this pose best.

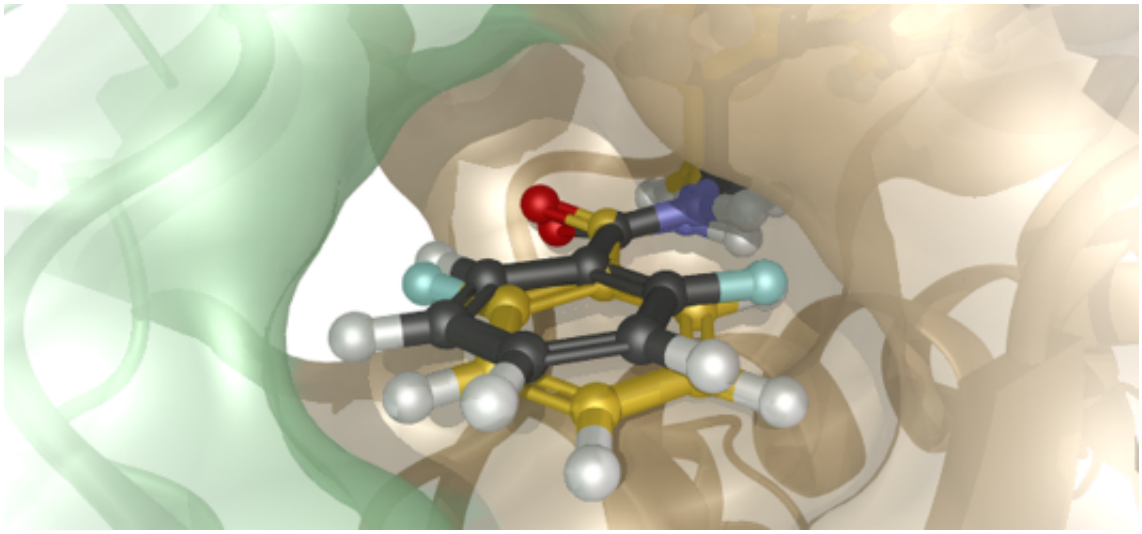


Figure 5.6: Binding pocket of PPI stabilizer CKH in the interface rim of the Arp2∘Arp3 complex (PDB ID: 3ukr). The protein is shown as cartoon and semi-transparent SES colored by chain (green: Arp3; gold: Arp2). The crystallized ligand conformation of CKH is shown as ball-and-stick model with grey carbons. The top-ranked docking pose of CKH is also shown as ball-and-stick model with yellow carbons. The major difference for CKH is the 180° flip of its *para*-fluorobenzene moiety.

5.3.4 Virtual Screening for 14-3-3∘Task3 Stabilizers

To select an appropriate structure of 14-3-3σ in complex with a mode III phosphopeptide for VS we analyzed 11 candidates from the PDB and unpublished in-house date sets, which are listed in the appendix (Tables E.2 and E.3). The collection comprises mammalian and plant 14-3-3 homologs and targets that were crystallized under different conditions.

Superposition of these structures yields an average RMSD of 0.87 Å for 14-3-3 monomers. This confirms the high structural conservation of this protein. However, visual inspection of the superposed proteins revealed the occurrence of two pronounced conformations of helix 9. Backbone models of two representative structures are shown in Fig. 5.7 where the two distinct conformations of helix 9 are highlighted in green and red, respectively. FSC (stick-model and yellow SES) is shown in the rim-exposed binding pocket formed by 14-3-3 and Task3 peptide. As helix 9 is part of this pocket, the selection of one conformer for protein-ligand docking will significantly affect its outcome.

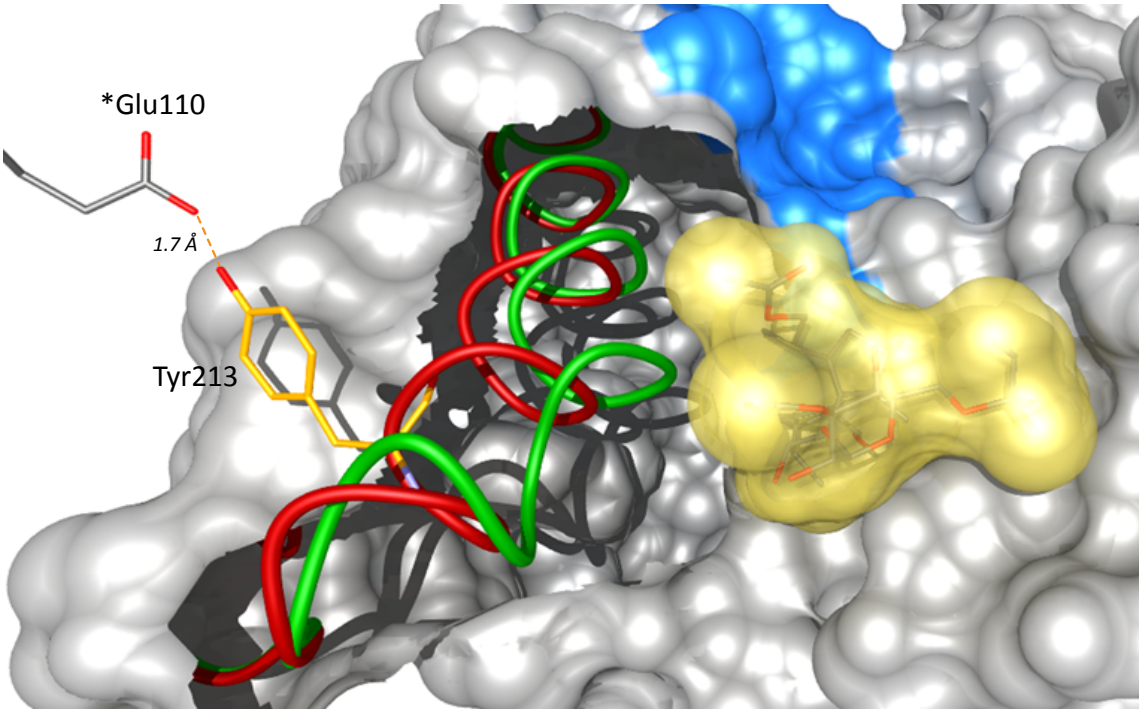


Figure 5.7: Superposition of the two major helix 9 conformations of the analyzed 14-3-3 crystal structures. One 14-3-3σ monomer is shown as grey SES and a bound Task3 peptide as blue SES. Reference ligand FSC is shown as ball-and-stick model and semi-transparent yellow SES. Observed helix 9 conformations are represented as green and red ribbons. The red conformation is only present in crystal structures of space group $C222_1$. Here, helix 9 is strongly curved most likely due to a charged-assisted hydrogen bond of Tyr213 to *Glu110 of a neighboring 14-3-3 symmetry mate. Therefore, the red conformation seems to be a crystallographic artifact leaving most likely the green conformation as the biological relevant.

We found that the red conformation is only present in the high-resolution structures crystallized in space group $C222_1$. In most of these structures the distorted conformation of helix 9 is most likely due to a charge-assisted hydrogen bond between Tyr213 and *Glu110 of a neighboring symmetry mate. 14-3-3 structures of other space groups lacking this spatial arrangement possess a straight helix 9. Hence, we subsume that the bent (red) conformation is a crystal artifact and the straight one (green) is its native conformation. Fortunately, a high-resolution crystal structure of space group $C222_1$ with a straight helix 9 conformation was available to us. This structure is not deposited with the PDB and was solved in-house at the MPI Dortmund. It is a complex of 14-3-3σ∘Task3 with bound FSC as the reference ligand and it was generated using the procedure described in Section 5.2.5. With a resolution of 1.65 Å this structure is well suited for modeling tasks and we chose it as template for VS.

Conserved Waters

The selection of waters to keep in our docking template was performed using the *ConservedWaterFinder* described in Section 5.2.5. As input structures we used the nine mammalian complex structures listed in the appendix (Table E.2). The superposed main chain atoms of the 14-3-3σ monomers yield an average RMSD of 0.48 Å. In total, the analyzed structures contain 2,619 waters. Fig. 5.8A shows the binding pocket with FSC and all resolved crystal waters. To select waters with a low B-factor, we applied the *ConservedWaterFinder* using $z_{cut} = 1.0$, which is exceeded by 891 waters in total. The B-factor filtered composition in the region of interest is shown in Fig. 5.8B and contains 87 waters.

The remaining 87 waters were clustered and using a level of 1.5 Å to cut the hierarchy yielded 29 water clusters. Furthermore, we required clusters to contain at least four members yielding 11 clusters. These are displayed in Fig. 5.8C and they comprise 60 low B-factor waters. From these candidates, we chose four clusters to select highly conserved waters to be preserved for docking.

These clusters were chosen for two reasons. First, they sit deeply buried in the binding pocket and are trapped between protein and FSC. Second, two of these clusters occupy space, wherein several alternative conformations of the highly flexible 14-3-3 residue Lys49 place their primary amine. Fig. 5.8D shows the selected water clusters and observed Lys49 conformations. Thus, keeping these waters additionally provides a reasonable alternative to the selection of an appropriate conformation of Lys49 in the docking template.

Virtual Screening Results

Using the selected 14-3-3σ∘Task3 complex in combination with the conserved waters chosen in the previous subsection, we performed the VS workflow as described in Section 5.2.5. The results are summarized in Table 5.5.

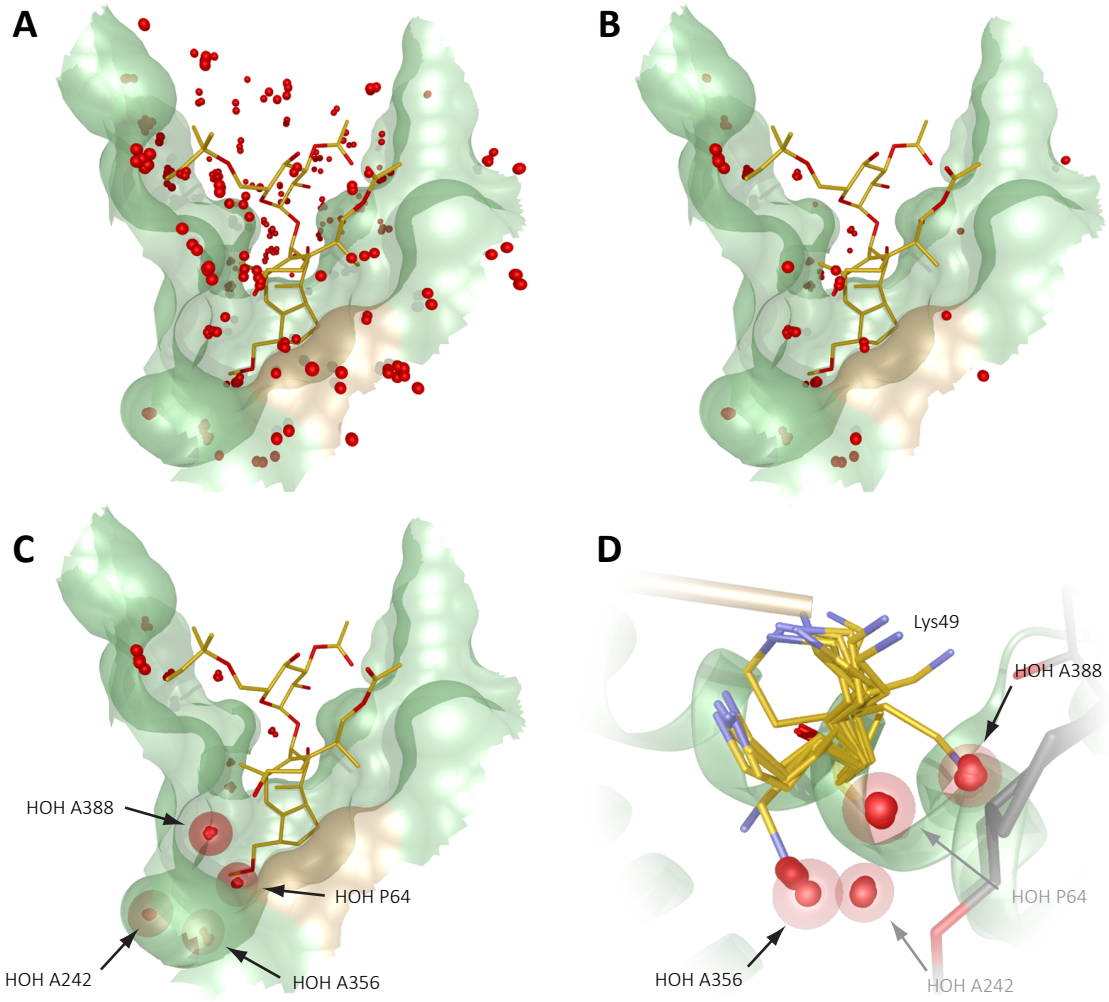


Figure 5.8: Conserved water analysis. Binding pocket, reference ligand FSC and waters within a 11.0 Å radius around the ligand center are shown. 14-3-3σ is shown as green SES, Task3 as gold SES. Reference ligand FSC is shown as stick model with yellow carbons. **(A)** Crystallographic waters from nine 14-3-3 structures. **(B)** Waters with a B-factor $z \geq 1.0$ **(C)** Low B-factor waters after cluster selection at a threshold of 1.5 Å. Finally selected waters are highlighted by red spheres centered at the corresponding water positions of our docking template. Numbering according to PDB entry 3spr. **(D)** Selected water clusters and observed conformations of 14-3-3 residue Lys49 (stick model, yellow carbons).

The first step in our VS workflow was a protein-ligand docking of the prepared library comprising 513,900 conformers, which was generated from 197,062 unique compounds. The described settings yielded 511,996 conformations, which passed the Glide SP energy funnel. To select an appropriate docking score cutoff for filtering these docked conformations, we calculated the docking score of reference ligand FSC in place. Thus, no conformational sampling was performed. Using the same receptor grid as for protein-ligand docking, FSC receives a docking score of -7.66 kcal/mol. Based on this value, we decided to apply an upper

Table 5.5: Filtering results from VS.

VS Step	*Candidates*
Protein-ligand docking	511,996
Reference ligand docking score cutoff	954
MultivalentInteractionAnalyzer	258
Purchased compounds	89

binding energy threshold of −7.0 kcal/mol. Interestingly, only 954 unique compounds fall below this energy value.

The last filtering step in our VS approach was the application of the *MultivalentInteraction-Analyzer* using the observed criteria for PPI stabilization by small molecule ligands. Here, we applied the same cutoffs as used for the PDB stabilizer screening from Section 5.3.2. These filtering criteria reduced the number of remaining stabilizer candidates down to 258 compounds. Additionally, our PPI stabilizer filter performed a similar task as a normalization function. These functions are frequently applied to balance MW and docking score because larger compounds tend to achieve better docking scores due to scoring functions additivity.[202,203] This balancing effect is visualized in Fig. 5.9. The histogram shows that the mean MW is shifted from 595 Da ($\sigma = 265$) down to 396 Da ($\sigma = 95$).

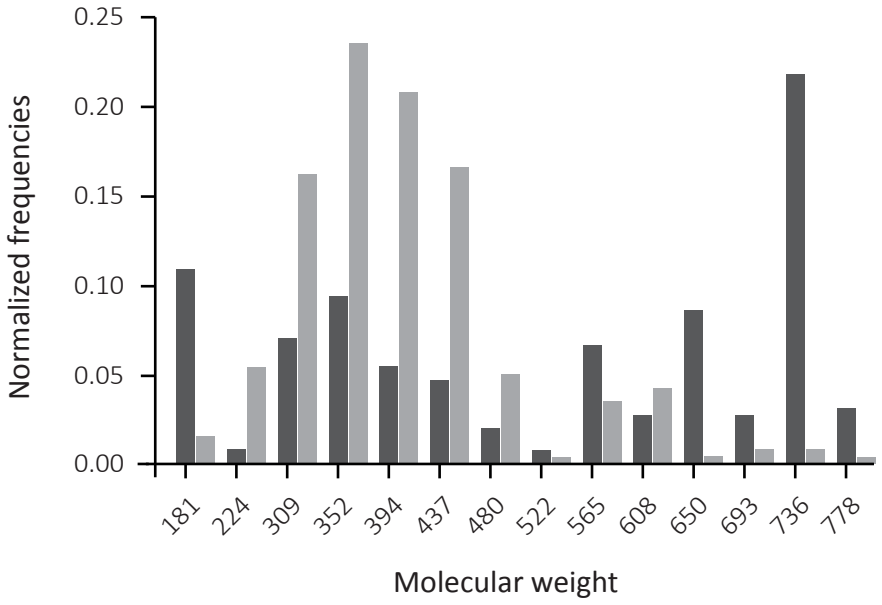


Figure 5.9: MW histogram. Dark grey bars represent a histogram for the 258 top-ranked docking poses from protein-ligand docking. Light grey bars represent a histogram for the 258 final candidates after application of the PPI stabilization filter.

Finally, this resulting candidate list was sorted by increasing volume overlap of ligand poses with reference ligand FSC as described in Section 5.2.5. From the final 100 top-ranked stabilizer candidates we were able to obtain 89 samples for *in vitro* testing, which are listed in the appendix (Table E.4).

In Vitro Validation Experiments

To test the 89 candidate compounds with respect to their ability to stabilize the interaction of 14-3-3σ and a Task3 phosphopeptide, we performed the described *in vitro* assay (Maria Bartel, TU Eindhoven). Results are shown in the appendix (Fig. D.2). Compounds S2, S37 and S43 showed preliminary stabilizing activity (Fig. 5.10D). Subsequent titration experiments are shown in Fig. 5.10A and revealed EC_{50} values of 115.9 μM for S2, 125.2 μM for S37 and 47.3 μM for S43. In comparison, the reference ligand FSC showed an EC_{50} of 1.0 μM.

The first control experiment was intended to test if candidates interfere with the FP assay (data not shown). Compound titration without Task3 peptide showed no significant signal increase. Thus, the measurements seem not to be the result of assay artifacts. The second control experiment was titration of the compounds against retinoid X receptor with a co-repressor as an entirely different target to test for unspecific binding to protein. The results, displayed in Fig. 5.10C, suggest presence of unspecific binding events for compounds S2 and S37, which classifies them as false positive hits. In contrast, compound S43 shows only a marginal signal increase at higher concentrations. However, this signal increase at higher concentrations is observed for most of the 89 tested compounds as can be seen in the appendix (Fig. D.2). Finally, all compounds were titrated against 14-3-3σ and a phosphorylated 10-mere from C-Raf (252–264; pSer259), which is shown in Fig. 5.10B. Here, S2 and S37 showed the same behavior as in the previous assays. In this assay, S43 shows also stabilizing activity, which could point to unspecific 14-3-3 complex stabilization because the binding pocket in this mode I complex is markedly smaller. Stabilizer candidate S43 is shown in Fig. 5.11.

Soaking and Crystallography

To test our VS hypothesis and to elucidate if the stabilizer candidate S43 binds as expected, we performed soaking experiments as described in the Section 5.2.5. Crystals tolerated up to 0.5 μl compound DMSO stock solution in the crystallization drops. The best data set from diffraction experiments was evaluated to a resolution of 1.85 Å. After several rounds of iterative model building, including building of water molecules, and refinement we carefully checked the difference electron density map for occurrence of compound S43. However, no extra density could be identified and up to now we were not successful in solving a ternary complex structure of 14-3-3σ, Task3, and S43.

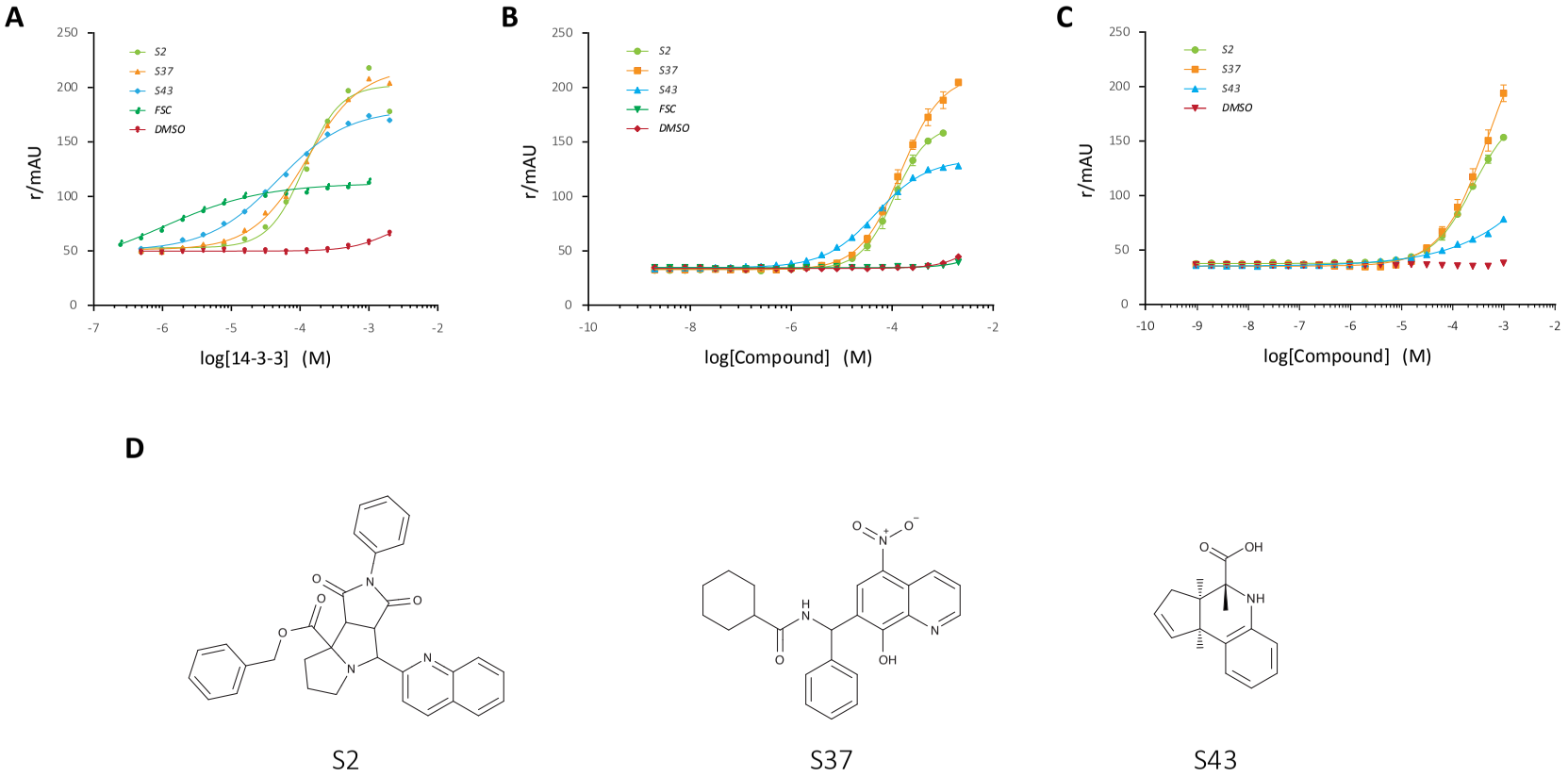


Figure 5.10: *In vitro* testing of stabilizer candidates (Maria Bartel, TU Eindhoven). **(A)** EC_{50} titration of S2, S37, S43, FSC and DMSO. **(B)** Titration of S2, S37, S43, FSC and DMSO against 14-3-3σ and C-Raf (252–264; pSer259). **(C)** Titration of S2, S37, S43 and DMSO against retinoid X receptor with a co-repressor. **(D)** 2D representation of stabilizer candidates.

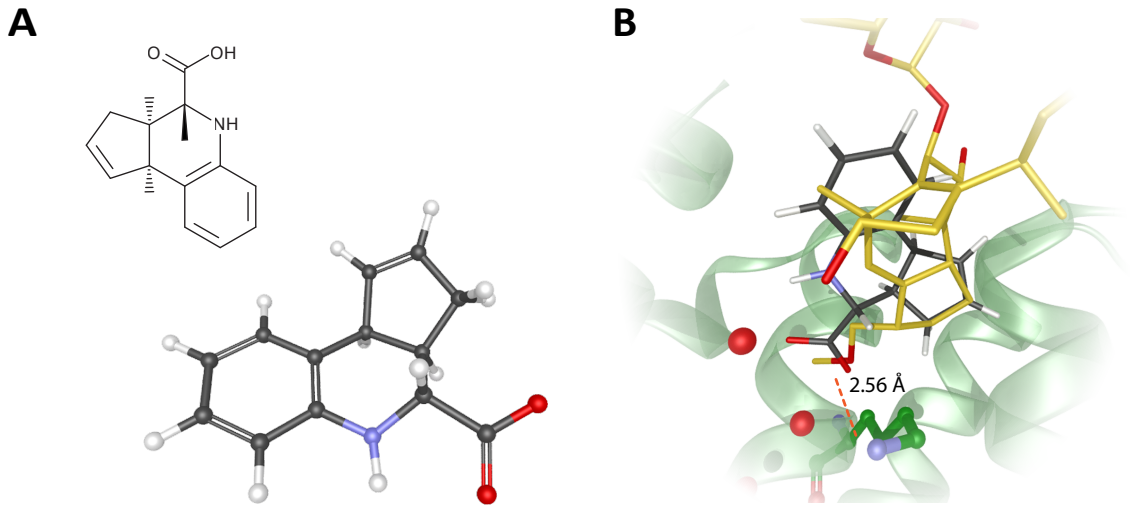


Figure 5.11: Stabilizer candidate S43. **(A)** 2D depiction of S43 and 3D conformation from docking. **(B)** Comparison of docked S43 (stick model; grey carbons) and crystallized reference ligand FSC (stick model; yellow carbons). 14-3-3σ is shown in green cartoon representation and the conserved waters as red spheres. The spatially conserved 14-3-3σ residue Lys122 is shown as ball-and-stick model.

5.4 Discussion

Stabilization of PPIs is as yet an underrepresented mode of action of small molecules, but it is a very attractive alternative to active site and PPI inhibition. The majority of currently known and structurally resolved examples are natural products whose mechanism was coincidentally discovered. The first rational and successful approach to discover PPI stabilizers by means of HTS was presented by Rose *et al.* in 2010.[25] Additionally, three *in silico* trials to find stabilizers by VS have been published as mentioned in the Section 5.1. However, none of these works studied the structural properties of already known PPI stabilization examples and tried to integrate the obtained structural information into novel *in silico* tools to tailor standard VS approaches to the identification of PPI stabilizing small molecules. Thus, our aim was to advance this research area by acquiring quantitative knowledge on the principles of PPI stabilization and to use it to perform an exemplary stabilizer-tailored VS.

Based on a recent review on structurally characterized stabilized PPIs we selected nine described structures and analyzed the contact contributions shared between the protein partners and the stabilizing ligand.[22] A common property of all ligands is the large fraction of surface area that is buried upon binding into the rim-exposed PPI pockets (> 65 %). However, the partitioning of the contact areas to the protein chains turned out to be highly imbalanced and the ratio of contact portions ranges from 0.08 to 0.99. Thus, quite small contact surfaces (≥ 34 Å^2) between a ligand and one protein partner seem to be sufficient to cause a stabilizing effect. Additionally, interesting findings were obtained from ligand analysis itself. The distribution

of typical structural and physicochemical properties used to assess the druglikeness of a compound, as shown in Table 5.1, indicates that currently known stabilizing small molecules were perfect drug candidates with exception of FSC. However, the hosts of FSC producing fungi are plant species and thus other pharmacokinetic requirements are imposed on this molecule. The druglikeness of the other compounds is especially encouraging because it provides a reasonable chance to identify stabilizer candidates in currently available compound collections. In contrast, this is not necessarily given for PPI inhibitors because it has frequently been observed that these compounds tend to break these properties.[20,204]

We used the obtained quantitative knowledge on PPI stabilization to implement a tool to screen *in silico* for stabilizer candidates and applied it on ligand-based preselected candidate complexes from the PDB. Indeed, we were able to identify further PPIs whose stabilization by a ligand has been described in the literature. These examples serve as positive controls and a validation for the structural properties we used to describe PPI stabilization.

The availability of crystal structures of stabilized PPIs enabled us to assess the ability of state-of-the-art protein-ligand docking software to correctly predict the binding poses of co-crystallized ligands. This evaluation is of great interest because docking tools were trained on datasets that solely contain enzyme-inhibitor complexes. However, the results of our redocking experiments show that the used docking software Glide is well suited for docking stabilizing ligands into rim-exposed PPI pockets. Thus, the utilization of this tool to perform a VS for PPI stabilizers can be regarded a valid approach.

Finally, we used our information on PPI stabilization and the developed tools to design a VS for PPI-stabilizing small molecules. As physiological target we selected the interaction of 14-3-3σ and Task3 because various high-quality crystal structures of this complex and related 14-3-3 complexes are available. This fact led to the development of a tool to select water molecules that are spatially conserved across multiple crystal structures. Our method does not use a scoring function to evaluate energetics but is based only on the experimentally determined crystallographic water positions and their B-factors. Application on available 14-3-3 complexes significantly reduced the number of crystallographic waters and yielded a small subset of spatially conserved waters. In our opinion, this tool is of great value for CADD because it facilitates the selection of waters to be added to the receptor grid.

Docking of ∼200,000 compounds revealed a quite large number of potential candidates for binding into the PPI pocket, but application of our *MultivalentInteractionAnalyzer* reduced these poses to a manageable set of stabilizer candidates. Further docking score filtering using the FSC score as cutoff and volume overlap sorting of the remaining poses yielded a small set of final candidates. Thus, our PPI stabilizer-tailored VS successfully scaled the large number of predicted docking poses down to a set of highly promising stabilizer candidates.

In vitro testing of 89 compounds led to the identification of one final candidate (S43), which stabilized the 14-3-3σ∘Task3 PPI with an EC_{50} of $\sim$ 47 μM. Control experiments ex-

cluded assay artifacts and promiscuous binding to protein. However, until now crystallization trials of S43 in complex with 14-3-3σ∘Task3 were not successful. Interestingly, compound S43 also seems to stabilize the PPI of 14-3-3σ and a C-Raf peptide phosphorylated at Ser259.

This peptide is not a mode III motif and exceeds the +1 position. However, Molzan *et al.* could show that a reduced pocket is still accessible in this complex.[110] They also evaluated the druggability of this complex and report a $DI > 0.7$, which classifies this complex as druggable. In addition, the crystal structure of a FSC-based fragment is presented, which binds into the 14-3-3ζ∘C-Raf pocket. As can be seen in Fig. 5.11, compound S43 is rather small and therefore might also fit into the 14-3-3σ∘C-Raf pocket. This hypothesis is indeed supported when inspecting the superposition of docked S43 onto a 14-3-3∘C-Raf complex as shown in Fig. 5.12. Subfigure A shows the original ternary complex from VS and Fig. 5.12B shows a superposition onto a structure of 14-3-3σ and a C-Raf peptide phosphorylated at Ser259. Obviously, S43 perfectly fits into this pocket, which could explain its activity.

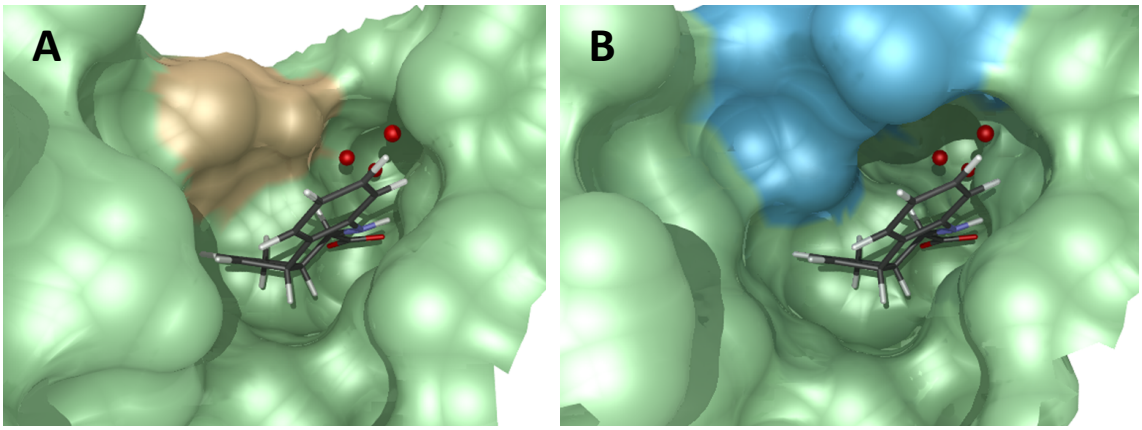


Figure 5.12: S43 docking poses and C-Raf superposition. **(A)** Docking pose of S43 bound into the interface pocket of 14-3-3σ (green SES fraction) in complex with Task3 (golden SES fraction). **(B)** S43 superposed onto binding pocket from 14-3-3σ (green SES fraction) in complex with C-Raf (blue SES fraction) from PDB entry 3iqj.

In summary, the presented VS for 14-3-3σ∘Task3 stabilizer candidates yielded one final hit. Further crystallization experiments, especially co-crystallization, are necessary in order to elucidate the compound's mode of action and to verify our hypothesis. On a first glance, a hit rate of about 1.1 % seems to be quite low for a VS approach in contrast to reports on VS yielding > 30 % hit rates.[205–207] However, the majority of reports yielding such high hit rates use enzyme targets, which are often well studied and the availability of known ligands with low complexity enable additional screening techniques. In contrast, PPIs are still no standard target for *in silico* drug discovery, especially in case of ligand-induced stabilization. Furthermore, the stabilization of 14-3-3σ and Task3 seems to be already an intrinsically difficult target. This is supported by three observations. First, a successful HTS campaign on a closely related target

complex (14-3-3 and PMA2) resulted in a hit rate < 0.01 %, which is also a low result for *in vitro* screening.[25] Second, a previous VS campaign on this closely related target yielded no hits.[23] Third, the stabilizer FSC is – from a chemical point of view – a highly complex molecule with only a handful closely related natural products with similar activity.

Chapter 6

Virtual Screening for 14-3-3 Protein-Protein Interaction Inhibitors

The content of this chapter is an extended version of the articles:

Covalent attachment of pyridoxal-phosphate derivatives to 14-3-3 proteins. [208]

Virtual Screening and Experimental Validation Reveal Novel Small-Molecule Inhibitors of 14-3-3 protein-protein interactions. [209]

6.1 Introduction

As described in Section 2.5.1, 14-3-3 proteins bind to numerous partner proteins in mammalian cells and are thus directly or indirectly connected to various diseases. Most intensely studied is 14-3-3's connection to several types of cancer. For example, 14-3-3ζ is found to be overexpressed in breast cancer, lung cancer, as well as neck cancer.[107] Thus, inhibition of this 14-3-3 homolog could be a beneficial treatment for these cancer types. Also in neurodegenerative diseases 14-3-3 proteins have been shown to play crucial roles.[210] In pathologic neural tissues various important binding partners of 14-3-3 have been identified like the microtubule-associated protein tau. 14-3-3 has frequently been found to influence aggregation of these partner proteins leading to the formation of neurofibrillary plaques, which is often correlated with cognitive impairment.[211] However, for diseases like Alzheimer's, the role of these plaques is still controversial and 14-3-3 inhibitors would be valuable research tools.

In addition, 14-3-3σ has been shown to be released into the extracellular matrix (EM) by keratinocytes where they act as signaling molecules in several types of fibroblasts.[212] Here, stimulation by 14-3-3 leads to an overexpression of matrix metalloproteinase (MMP) family

members, especially MMP-1.[213] MMPs are exported to the EM where they hydrolyze other EM components.[214] 14-3-3 thus influences tissue remodeling, which is a complex process where a sensitive balance between EM synthesis and its degradation is vital. The extracellular receptor for 14-3-3 in the plasma membrane has recently been identified as aminopeptidase N (APN) but the interacting sequence and its binding mode are yet unknown.[215] Beside APN's enzymatic activity, the protein has various roles on the plasma membrane including ligand sensing or signal transduction.[216] Thus, the interaction of 14-3-3σ and APN possibly forms an interesting target for diseases where EM remodeling is involved like non-healing wounds.[213] Additionally, such inhibitors could form beneficial tool compounds for chemical biology to study the various functions of APN on the outer membrane.

Goals of the Project

The goal of this project is to identify non-covalently binding and non-peptidic small molecule inhibitors of 14-3-3 PPIs by means of *in silico* screening. Numerous X-ray structures of 14-3-3 with various binding partners are available from the PDB. These structures represent different conformational snapshots, which we analyze to obtain detailed information on the binding mode of phosphopeptides and the flexibility of the protein partners. Based on these results, we try to define a minimal pharmacophore, which should subsequently be used to perform a ligand-based VS. The resulting compounds will subsequently be docked into a high-resolution structure of 14-3-3 to select final candidates for *in vitro* testing. Putative candidate compounds will be tested in crystallization trials to solve their structure in complex with 14-3-3. Finally, we plan to test the inhibitors activity in a cellular assay to block the interaction of 14-3-3 with APN.

Related Work

Several attempts have been done to develop inhibitors for 14-3-3 proteins, which comprise the search for high-affinity peptides or peptidomimetics and the search for small molecule ligands as antagonists of 14-3-3 PPIs. Using phage display, Wang *et al.* were the first to identify a peptide that binds with high affinity to 14-3-3 and inhibits its PPIs.[217] The identified peptide – known as R18 – is not selective for the different 14-3-3 homologs and has been shown to block the interaction with C-Raf *in vitro*. The R18 peptide does not feature a phosphorylated serine or threonine residue and binds primarily to a hydrophobic patch at the end of the amphipathic groove. A further development of R18 is its bifunctionalized version which can bind to both amphipathic grooves of a 14-3-3 dimer.[218]

Wu *et al.* employed a microarray-based strategy to screen a custom peptide-small molecule hybrid library.[219] The library they synthesized is based on a tri-peptide with a central phosphoserine flanked by two glycines. These peptides were modified using various members of

a building block library on both ends, thereby generating peptide-small molecule hybrids. In a microarray-based experiment, 14-3-3-binding peptidomimetics were identified and used for 14-3-3 pull-downs on cell lysate. In combination with immunoblotting, the pull-down experiment revealed disruption of the PPI of 14-3-3 with p53 as well as C-Raf by one peptide-small molecule hybrid.

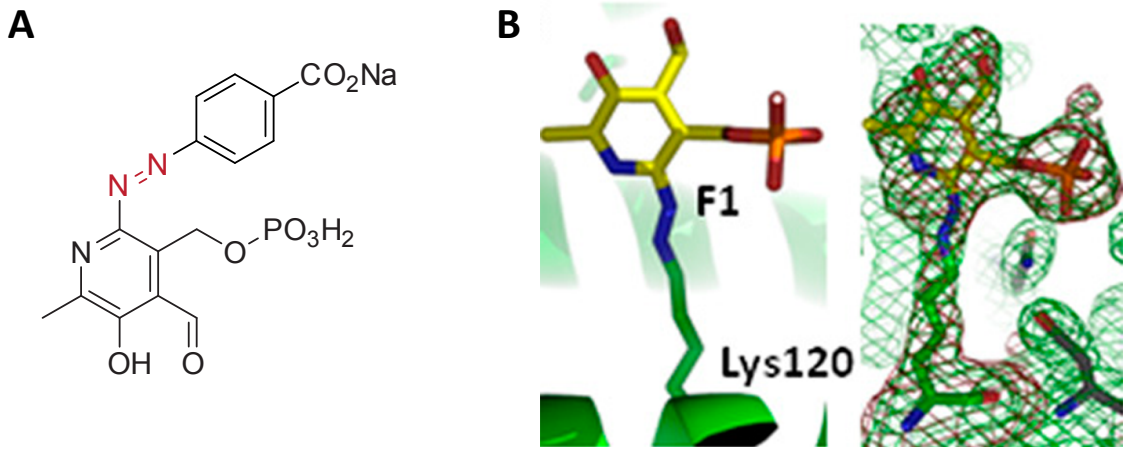


Figure 6.1: Covalent attachment of F101 to 14-3-3. **(A)** 2D representation of F101 with the proposed cleavage site highlighted in red. **(B)** Crystal structure of F101 covalently bound to Lys120 of 14-3-3ζ (PDB ID: 3rdh) at a resolution of 2.39 Å. (B) is taken from Zhao *et al.* (*Proc. Natl. Acad. Sci. USA* (2011), **108**, 16212-6).[107]

Fu *et al.* attempted to identify small molecule inhibitors of 14-3-3 via HTS.[107,184] They identified an active inhibitor – Fobisin101 (F101) – that covalently binds to a conserved lysine in the amphipathic groove of 14-3-3.[107] The compound's binding mode was characterized by X-ray crystallography and subsequent mass spectrometry (MS) from protein of the dissolved crystal, which was used for the diffraction experiment. Fig. 6.1 shows F101 and the published crystal structure. The proposed mechanism for covalent binding starts with an initial cleavage of the compound's N=N double bond caused by X-ray radiation with radical formation and subsequent attachment to Lys120 of 14-3-3ζ. This covalent modification is described to form a persistent inactivation of 14-3-3 proteins and therefore F101 is proposed as a lead compound for a unique class of radiation-triggered chemotherapeutics.

An *in silico* approach for the identification of 14-3-3 PPI inhibitors was described by Corradi *et al.*.[220,221] As an intracellular target they aimed to disrupt specifically the interaction of 14-3-3 with the oncogenic tyrosine kinase BCR-ABL. They employed a combination of pharmacophore-based virtual screening and molecular docking to select 14 compounds for biological testing. To generate a pharmacophore model, they used a complex of 14-3-3 with an artificial phosphopeptide to assign all interactions between phosphopeptide and 14-3-3. Due to a large number of interactions, they skipped several interactions. Most notably, they discarded all interactions established by phosphate coordinating residues of 14-3-3. However,

to account for the phosphate's negative charge in the pharmacophore model, they created a rather large negative ionizable feature at this site. The remaining interactions were monitored in a short MD simulation and revised if their distances or angles deviated from standard values. The final pharmacophore model was used to screen a library of 200,000 compounds, yielding 99 matches. The subset of 87 Ro5-compliant compounds were then docked into 14-3-3 and based on a consensus scoring, final compounds were selected for experimental testing. One compound – BV02 – was found to be active in a cell-proliferation assay monitoring cell viability depending on BCR-ABL activity. Additionally, the subcellular localization of BCR-ABL was analyzed using Western Blotting, which showed increased nuclear localization of this protein. However, the direct binding of BV02 to 14-3-3 has neither been shown in a biochemical assay nor by X-ray crystallography. Due to the complexity of the assay, which regulates the activity and localization of BCR-ABL, it is – in our opinion – not ultimately proven that 14-3-3 is indeed the target of BV02.

6.2 Materials and Methods

All materials that were used for wet lab experiments and which are not directly described in the following sections are listed in Table E.6.

6.2.1 Crystal Structure Analysis

We extracted complex structures from the PDB in order to analyze the binding geometries of mammalian 14-3-3 proteins to their phosphorylated targets. To identify the entries containing mammalian 14-3-3 proteins we used BLAST through the PDB website to search for protein sequences similar to 14-3-3σ as the query.[222] According to the MSA shown in Fig. 2.11, we truncated the C-terminal sequence following Glu237 due to low conservation. PDB default settings for BLAST were used with an e-value cutoff of 10.0 and masking of low complexity sequences. Using the advanced search interface of the PDB we additionally filtered the BLAST hits according to the following search criteria:

- *Source organism* Mammalia
- *Experimental Method* X-ray diffraction
- *Resolution* < 2.5 Å
- *Number of protein entities* > 1

Preparation and analysis of resulting 14-3-3 complex structures were performed using the *Molecular Operating Environment* (MOE, version 2010.10).[223] Only one 14-3-3 monomer with bound phosphopeptide per entry was kept. Furthermore, unresolved residues were

deleted before sequence alignment. Using MOE's *Protein superpose* functionality the remaining 14-3-3 monomers were mapped onto each other based on C_α atoms. Default settings were used with optional *Accent Secondary Structure Matches* enabled.

Structural flexibility of bound phosphopeptides was analyzed by superimposing them separately. Here, the corresponding sequences were manually aligned by centering them onto their phosphorylated residues using the sequence editor of MOE. Based on this alignment, the all-atom RMSD was calculated between all passively superimposed phosphopeptides.

6.2.2 Virtual Screening

Virtual Compound Library

To ensure maximum diversity in our virtual screening library, we used the ZINC *all now* subset (release 11).[127] This subset is a comprehensive collection of immediately available compounds and comprised 8,061,769 compounds at that time. The library was downloaded as a reference set generated at a pH value of 7 in SMILES format. Compound preparation was performed using Pipeline Pilot.[139] Components from the Chemistry Package were used to reconstruct molecules from SMILES (*Molecule from SMILES*), salts were removed (*Strip Salts*) and the largest fragment was kept for each entry (*Keep Largest Fragment*).

Ligand-based Virtual Screening

Ligand-based VS was also performed with Pipeline Pilot. Our filtering workflow started with a substructure filter (*Substructure Filter from File*) using default settings. The input was an SD file containing the query substructure. The second step was a *Lipinski Filter*, which forwards only Ro5-compliant compounds. In addition to the Ro5 filter we applied the *HTS Filter* of the Chemistry Package to reject poor HTS candidate compounds. The latter comprises compounds containing non-organic atom types, reactive substructures, or those with MW below 150 Da. To eliminate candidates with multiple query substructure occurrences and to gain internal rigidity, the last filter step forwarded only compounds with a single query match and possessing at least one ring. Finally, a diverse selection was performed on the remaining compounds using the *Diverse Molecules* tool with a desired number of 500 compounds to be selected. As a descriptor we used ECFPs with a maximum diameter of 4 bonds and plain atom types as atom abstraction.[37]

Structure-based Virtual Screening

Compounds from ligand-based VS were docked into a high-resolution structure of 14-3-3σ. The crystal structure we used is a complex of 14-3-3σ and a phosphorylated Task3 peptide (PDB ID: 3p1n).[163] Structure preparation was performed using the *Protein Preparation Wizard* of

Schrödinger Maestro with slightly modified settings:[224] We kept the peptide and surrounding waters for optimization of hydrogen bond networks and deleted them afterwards. The receptor grid was calculated using Glide and centered at the position of the phosphorylated residue.[180] The precise receptor grid dimensions are listed in Table 6.1. Default settings were used for all other parameter options. For protein-ligand docking, the selected compounds were prepared using *LigPrep*.[225] Parameter settings deviating from default values are listed in Table 6.1. Finally, the prepared compounds were docked into the generated 14-3-3 receptor grid using Glide in extra-precision (XP) mode.[54,180,181,226] Detailed parameter settings deviating from default settings are also listed in Table 6.1. Docked compound poses were sorted by increasing docking score and the top 200 docking poses were manually inspected to choose a subset for experimental testing.

Table 6.1: Parameter settings for receptor grid generation, ligand preparation, and protein-ligand docking using PDB entry 3p1n. Only options deviating from default settings are listed.

Glide: Receptor Grid Generation	
Grid center	$x = -17.3$ Å, $y = -14.7$ Å, $z = 9.6$ Å
Inner box dimensions	$x = y = z = 12.0$ Å
Outer box dimensions	$x = y = z = 23.0$ Å
LigPrep: Ligand Preparation	
Ionization	Neutralize and ionize
pH-Range	7.0 ± 3
Stereoisomers	Use chiralities from input geometry
Glide: Flexible Ligand Docking	
Precision	Extra precision (XP)
Post-docking minimization poses	20
Output poses per ligand	5

6.2.3 Experimental Validation and X-ray Crystallography

Selected compounds from VS and derivatives from SAR analysis were obtained from InterBioScreen. The pyridoxal-phosphate (PLP) derivative F101 was obtained from Sigma-Aldrich. 2D representations and supplier informations of all compounds are listed in the appendix (Table E.5). Compounds were delivered as solids, dissolved in DMSO as 20 mM stock solutions and stored at -20 ° C.

In Vitro Validation Experiments

To test the compounds' potential to compete with the binding of a phosphorylated peptide to 14-3-3 *in vitro*, a fluorescence polarization (FP)-based assay was performed (Dr. Lars Röglin, CGC Dortmund).[184] Active compounds were further tested in a cellular assay for their potency to inhibit upregulation of MMP-1 mRNA levels in response to extracellular application of 14-3-3 (Nicole Meissner and Svenja Schäfers, CGC Dortmund). Human lung fibroblast cells (IMR90) were purchased from ATCC.

The compounds' ability to permeate cell membranes was tested using the parallel artificial membrane permeation assay (PAMPA) by the LDC Dortmund.[227] PAMPA is an *in vitro* assay to model passive transmembrane permeation.

Protein Crystallization

For crystallization experiments of all inhibitor complexes we used a C-terminally shortened construct of 14-3-3σ (14-3-3σΔC), which was truncated after Thr231. Cloning, expression, and purification of 14-3-3σΔC was performed as described by Schumacher *et al.* and protein was provided by the Ottmann lab.[185]

For crystallization trials, 14-3-3∘inhibitor stock solutions were obtained by mixing the 14-3-3σΔC stock solution, compound DMSO stock solutions and complexation buffer to yield final concentrations of 12 mg/ml 14-3-3σΔC and 2 mM compound, respectively. The complexation buffer is listed in the appendix (Table E.7). Complex solutions were incubated overnight at 4 °C. Crystallization was performed at 4 °C using the hanging-drop or sitting-drop method, respectively. Crystallization buffers from the 2D grid described in Section 5.2.5 were used.

For the *in vitro* validated VS hits A1-A14, reservoirs from a 24-well plate were filled with 500 µl of crystallization buffer. Hanging-drops were mixed from 2 µl complex solution with 2 µl crystallization buffer from the reservoir and placed on a cover glass, which was used to seal the well. Since the final crystallization drops were already cryogenic, crystals were directly flash-frozen in liquid nitrogen. For the inhibitors from the second compound batch (B1–B31), a crystallization screening was performed using sitting-drops in 96-well plates. Plate reservoirs were filled with 75 µl of grid variation buffers C17-C24 column by column (A-H) using a 96-well pipetting system. Sitting-drops were mixed manually from 1 µl complex solution with 1 µl crystallization buffer from the reservoir. The 96-well plates were sealed with transparent adhesive foil. All successful crystallization conditions that yielded suitable crystals for diffraction experiments are listed in Table 6.2. 96-well plates were stored in an imaging system, which allowed automatic monitoring of the crystallization drops.

Table 6.2: Successful crystallization buffers of 14-3-3∘inhibitor complexes from the 2D grid. The constant ingredients Na-HEPES buffer (95 mM), $CaCl_2$ (190 mM) and Glycerol (5 % v/v) are not listed.

Compound	Condition	*pH*	*PEG 400* (% v/v)
A1	C24	7.7	28.0
A2	original	7.5	26.6
B1	C17	7.1	27.0
B2	C22	7.3	28.0
B3	C23	7.5	28.0
B4	C21	7.1	28.0
B5	C17	7.1	27.0
B6	C15	7.5	26.0
B7	C21	7.1	28.0
B8	C22	7.3	28.0
B9	C24	7.7	28.0
A3	original	7.5	26.6
F101	original	7.5	26.6

Data Collection

Diffraction experiments were performed in-house and at the Swiss Light Source (SLS) of the Paul Scherrer Institute Villigen (Switzerland), at beamline PXII. The in-house beamlines used rotating copper anodes from Rigaku (MicroMAX-007 HF) and from Bruker (AXS MICROSTAR) as X-ray source. Both beamlines were equipped with a MAR345 image plate. Data processing was carried out using the software package XDS.[186] During this step, choice of the lowest resolution to trim the data sets was based on three parameter values of the outermost shells. First, a signal to noise ratio of 4.0 was set as lower cutoff. Second, a completeness of 90.0 % was used as lower cutoff. Third, we set the upper limit for the redundancy independent R-factor (R_{meas}) to 40.0 %. Crystal parameters and data collection statistics of all complex structures are listed in the appendix (Table E.9).

Structure Elucidation and Model Building

The CCP4 software suite was used for phase determination and automatic refinement.[187] MR was carried out with PHASER (version 2.1.4) using a monomer of 14-3-3σ as a search model.[188] A unique solution was found for every 14-3-3∘inhibitor data set. These initial

models were used as starting data for iterative cycles of automatic and manual refinement with REFMAC (version 5.5) and COOT (version 0.6), respectively.[178,189] Refinement statistics of final models are given in the appendix (Table E.9). Corresponding Ramachandran plots were generated with RAMPAGE and shown in the appendix (Fig. D.3).[228] To visualize density with reduced model bias, simulated annealing composite-omit maps were calculated with PHENIX.[229] The final models of all 14-3-3∘inhibitor complexes were deposited with the PDB. All corresponding PDB IDs are listed in the crystallographic Table E.9.

6.3 Results

The first part of this section shows the binding geometry conservation of 14-3-3 in complex with target proteins. Based on the deduced pharmacophore hypothesis the results of ligand-based and structure-based VS are presented. Experimental hit validation and successfully solved complex structures are subsequently presented. Finally, we show the true binding mode of covalent 14-3-3 PLP inhibitors, which was elucidated by X-ray crystallography.

6.3.1 Structure Analysis

The search for complex structures of 14-3-3 and phosphopeptides as described in Section 6.2.1 yielded 14 entries within the PDB, which are listed in Table 6.3. The entries 3iqu and 3iqv were not used for binding analysis because bound phosphopeptides are truncated duplicates of the C-Raf construct already present in 3iqj. Entry 3cu8 is also a truncated duplicate of the construct contained in 3iqj. Although the 14-3-3 isoform in this entry is different from 3iqj, it was skipped because the surface-exposed residues of all mammalian 14-3-3 isoforms are entirely conserved in the amphipathic groove. Entry 1qjb was skipped because an artificial mode I phosphopeptide was already represented by entry 1ywt. Entry 2o02 was skipped because the binding partner is an unphosphorylated peptide from exoenzyme S, which binds to a distant surface patch at the end of the amphipathic groove. Thus, nine entries from the PDB remained for the analysis of phosphopeptide binding geometries.

Superposition of 14-3-3 chains revealed an average C_α RMSD of 0.4 Å for 14-3-3 monomers. Thus, 14-3-3 adopts identical conformations in all considered cases. In contrast, the passively superimposed phosphopeptides show highly variable C_α RMSD values as shown in Fig. 6.2. The bar chart in Fig. 6.2A shows the RMSD values for the phosphorylated residue at position 0, three N-terminal flanking residues (positions -3, -2, -1) and three C-terminal flanking residues (positions $+1$, $+2$, $+3$). The phosphorylated residue has the lowest C_α RMSD with 0.6 Å and only the -1 residue shows a comparably low RMSD. The RMSD values for the other flanking residues rapidly increase with the distance to the phosphorylation site and exceed 2.5 Å at the -3-position and at the $+2$-position. Weblogos representing the relative residue

Table 6.3: 14-3-3∘phosphopeptide complex structures extracted from the PDB.

PDB ID	*Residue*	*Resolution* (Å)	*Binding Partner*
1qja	pSer	2.0	Artificial mode II phosphopeptide
1ywt	pSer	2.4	Artificial mode I phosphopeptide
2br9	pSer	1.75	Consensus phosphopeptide
2c63	pSer	2.15	Consensus phosphopeptide
2c1n	pSer	2.0	Phosphopeptide from histone H3 (7-14)
2v7d	pThr	2.5	Phosphopeptide from integrin β-2 (755-746)
2wh0	pSer	2.25	Phosphopeptide from proteinkinase C (342-373)
3iqj	pSer	1.15	Phosphopeptide from C-Raf (255-264)
3lw1	pThr	1.28	Phosphopeptide from p53 (385-294)
1qjb			Duplicate of an artificial mode I phosphopeptide
2o02			Non-phosphorylated peptide from exoenzyme S
3cu8			Duplicate of phosphopeptide in 3iqj
3iqu			Shortened duplicate of 3iqj
3iqv			Shortened duplicate of 3iqj

frequencies for all positions are shown on top of the bars.[230] The low sequence conservation of flanking residues is a possible explanation for their low spatial conservation. Fig. 6.2B shows the superimposed structures and the diverging phosphopeptides. The high spatial conservation of the central phosphoserine or phosphothreonine residue, which is coordinated by 14-3-3 residues Arg56, Arg129, Tyr130, and partially by Lys49 is highlighted. Particularly, the phosphate group is the only side-chain moiety that is also spatially well conserved.

6.3.2 Virtual Screening and Experimental Validation

Based on the findings of the 14-3-3 complex structure analysis we concluded that the phosphate moiety of the phosphorylated amino acid contains the strongest pharmacophoric properties and combined a ligand-based VS with structure-based docking. The entire filtering workflow of ligand-based VS is shown in Fig. 6.3A. For ligand-based VS we used a phosphonate as an initial substructure filter, which had the following SMARTS pattern:

```
[C,O][P]([OH])(=[O])([OH])
```

The initial substructure filter reduced 8,061,769 compounds to 2,349 phosphonate derivatives. Filtering for Ro5 compliance in combination with the HTS filter further reduced the selection to 1,502 compounds. The last filter ensuring compound rigidity, requiring at least one ring and rejecting compounds with multiple phosphonates, yielded 1,012 compounds.

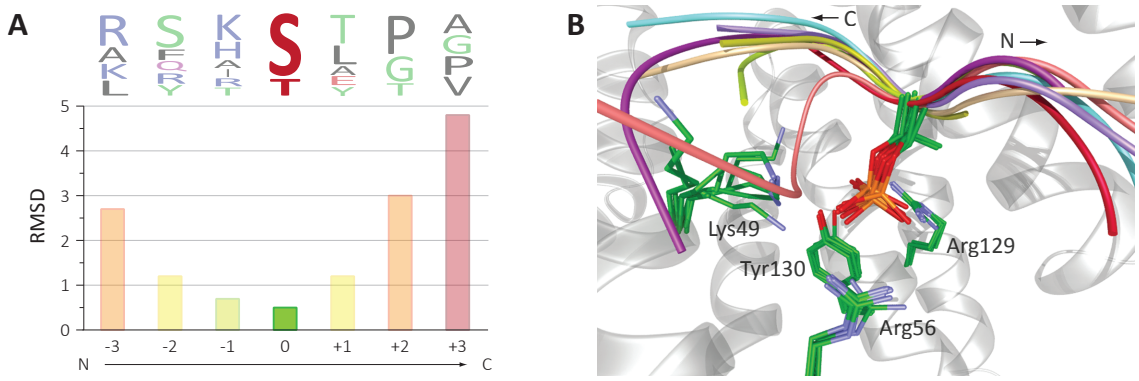


Figure 6.2: Superposition of 14-3-3∘phosphopeptide complexes. **(A)** Analysis of spatial and sequence conservation of phosphopeptide residues. Position 0 is the phosphorylation site. Negative positions indicate N-terminal residues, positive positions indicate C-terminal residues. The weblogo shows relative residue frequencies for the corresponding positions (green: polar residues, blue: basic residues, red: acidic residues, black: hydrophobic residues). **(B)** Superposed crystal structures. A single 14-3-3 monomer is represented as grey cartoon, phosphopeptides are represented as colored ribbons, phosphorylated residues are represented as sticks colored by element. The phospho-coordinating residues of 14-3-3 are also shown as green stick models coloured by element.

Visual inspection of this selection hinted at the presence of several compound clusters. Therefore, we calculated a diverse set of compounds based on ECFPs with a desired number of 500 compounds in the final selection. As a result, we ended up with 512 representative compounds for structure-based docking.

Ligand preparation of the selected compounds produced 698 input structures for protein-ligand docking, which was performed as described in Section 6.2.2. The resulting list of docking poses was sorted according to decreasing docking score. We visually inspected the 200 top-ranked docking poses, which spanned a docking score range from -10.2 to -7.7 and comprised 70 different compounds out of the 512 input candidates. From this set, we manually selected 14 compounds for experimental testing, which we refer to as A1–A14. Our primary selection criterion was the placement of the phosphonate moiety of the docked compounds in order to satisfy our pharmacophore hypothesis. The docking ranks and the corresponding scores for these compounds are listed in the appendix (Table E.5).

The compounds' potential to compete with binding of a phosphopeptide from C-Raf to 14-3-3 was assessed in an FP-based assay (Dr. Lars Röglin, CGC Dortmund). Fig. 6.3B shows the results of this validation screen at a compound concentration of 250 µM. Compounds A1 and A2 were validated as true actives because they exceeded our expected cutoff of 50 % normalized PPI inhibition (Table 6.4). Subsequent titration of A1 and A2 revealed IC_{50} values of 30 µM and 116 µM, respectively. The identified compounds were also active in a second orthogonal assay, which is based on a different physical phenomenon than the FP-based assay

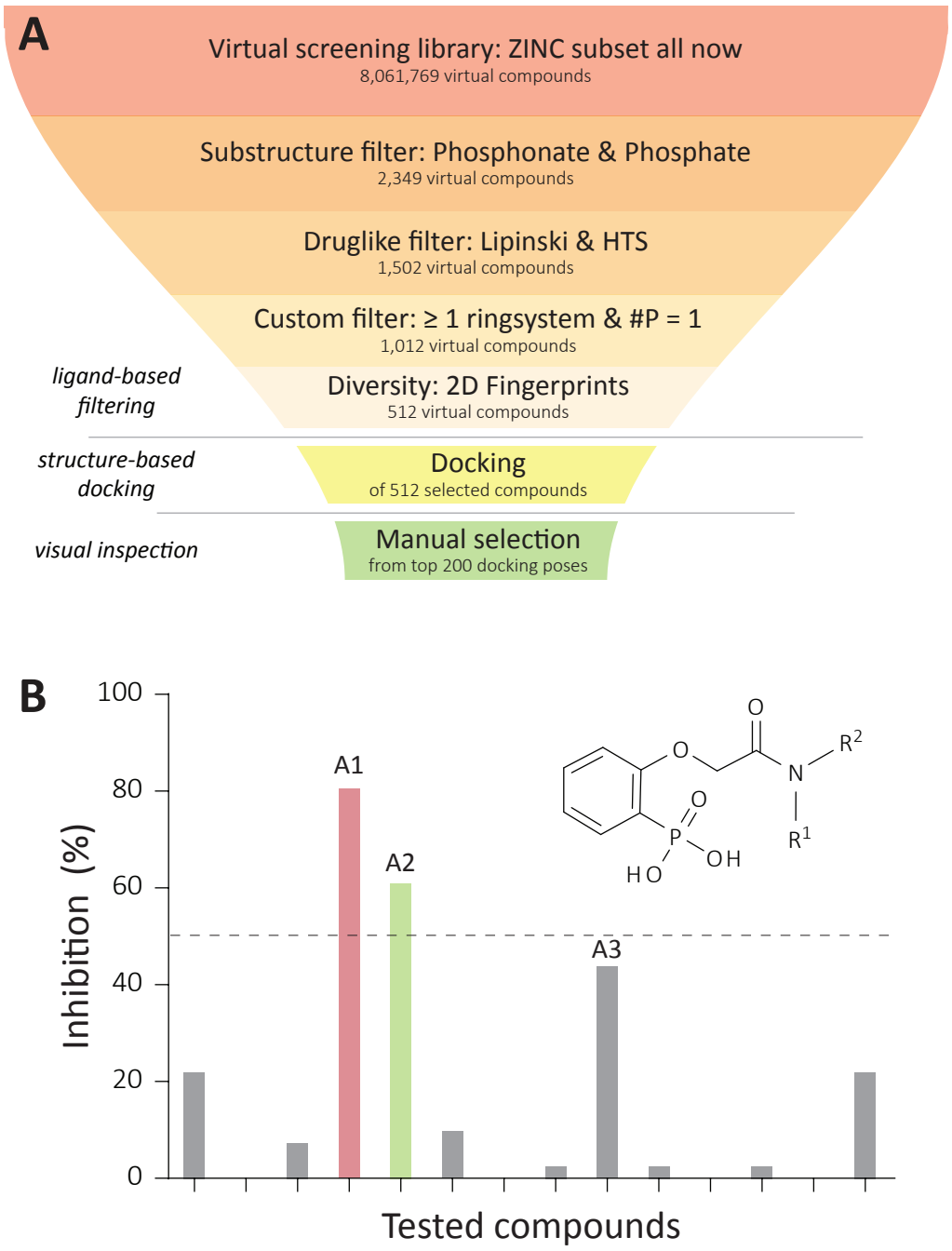


Figure 6.3: Virtual screening and experimental validation. **(A)** Results of the ligand-based filtering workflow, docking, and manual selection. **(B)** Experimental testing of 14 selected compounds from VS. The bar chart shows normalized percent-inhibition for the compounds' potency to disrupt the binding of a phosphopeptide from C-Raf to 14-3-3ζ at a concentration of 250 μM. Our cutoff for hit selection was a 50 % inhibition (dashed grey line). We identified A1 and A2 as active inhibitors, whose common scaffold is also shown.

and is thus well suited to validate the results of FP measurements. This homogeneous time resolved fluorescence (HTRF) assay was also performed by Dr. Lars Röglin, CGC Dortmund.

Table 6.4: Active inhibitors from VS validated by *in vitro* screening.

Compound	R^1	R^2	IC_{50} (μM)
A1	H	2-(trifluoromethyl)phenyl	30
A2	H	2,5-dimethoxyphenyl	116

6.3.3 Crystallography and Structure-Activity Relationships

To validate our VS hypothesis and to gain deeper insight into the compounds' mode of action we tried to obtain structures of compounds A1 and A2 in complex with 14-3-3σ. Using the procedure described in Section 6.2.3, we successfully grew crystals which diffracted up to a resolution of 1.6 Å. The initial models resulting from the single MR solutions already showed extensive residual density in the $F_o - F_c$ electron density maps for both ligands. Iterative cycles of model building and refinement allowed unambiguous interpretation of compound A1. Compound A2 has an overall occupancy of only 75 % and reveals two distinct conformations for its dimethoxybenzene moiety. Both conformations have a an occupancy of 37 % and are thus equally distributed.

The crystal structures of A1 and A2 are shown in Fig. 6.4. As expected, the phosphonate groups are coordinated via five hydrogen bonds to Arg56, Arg129, and Tyr130. Interestingly, the compound orientations are perpendicular to the peptide backbones and the R^2 substituents bind to a region that is not occupied by any currently known peptide. Furthermore, the compounds form an intramolecular and charge-assisted hydrogen bond between a phosphonate oxygen and the anilide nitrogen. The importance of this interaction is underlined by the observation that all tertiary amides are inactive (A4–A9). Additionally, *para*-substitutions seem detrimental for 14-3-3 binding due to spatial limitations (A10–A12). In our structures, the 14-3-3 protein undergoes no major conformational changes. However, an important observation is the elongated side-chain conformation of Arg60. Depending on the binding partner of 14-3-3, this residue also occurs in a bent conformation where the guanidinium group occupies the pocket where the inhibitors' variable R^2 moieties are located. Since we used a 14-3-3 structure with bent Arg60 for docking, the predicted docking poses only match the phenylphosphonate moiety when compared to the observed binding geometry. Based on these preliminary SARs, we evaluated 31 additional compounds containing the validated scaffold featuring a secondary amide for internal hydrogen bonding (B1–B31). Using the FP assay, we found nine additional compounds with IC_{50} values below 200 μM. The most potent compounds

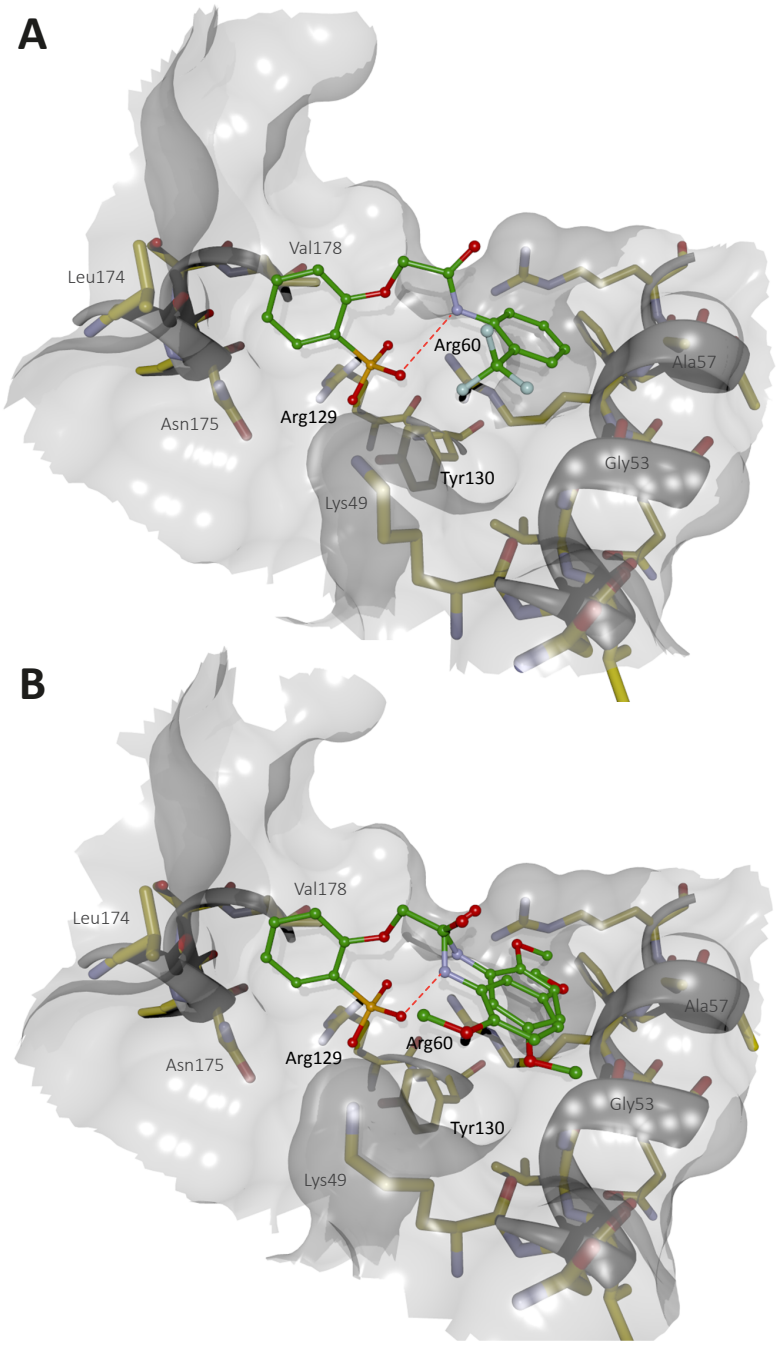


Figure 6.4: Crystal structures of identified inhibitors bound to 14-3-3σ (grey semi-transparent SES and cartoon, residues with yellow carbons). **(A)** Compound A1 (green carbons). **(B)** Compound A2 (green carbons). Both compounds display a similar binding mode for their phenylphosphonic moieties. The main interacting residues of 14-3-3 are labelled in boldface. The 2,5-dimethoxybenzoic moiety of compound A2 binds in two different conformations. The intramolecular hydrogen bond is highlighted as dashed orange line.

B1 and B2 show IC_{50} values of 5 μM and 15 μM, respectively These inhibitory concentrations are close to those of known phosphopeptide ligands.

We also tried to structurally characterize compounds B1–B9 and successfully solved their complex structures with resolutions below 1.8 Å. All nine inhibitors share the conserved orientation of the phenylphosphonic moiety, which is stabilized by its intramolecular hydrogen bond and the conserved hydrogen bonds of phosphonate oxygens to side-chains Arg56, Arg129, and Tyr130. Additionally, the phosphonic phenylring makes hydrophobic contacts to Val178. The inhibitors' variable R^2 moieties fill a shallow subpocket enclosed by Arg56, Arg60, Lys49, and the protein backbone. Due to the variety of chemotypes binding into this pocket, a mainly hydrophobic contribution can be suggested. This is exemplified by the 14-3-3 complexes of the two most active compounds B1 and B2, which are shown in Fig. 6.5A and 6.5B. As shown in Fig. 6.5B, similar van der Waals volumes (V_{vdw}) are occupied by the aromatic R^2 moiety (2,3-dichlorophenyl, $V_{vdw} = 111$ Å^3) as well as by the aliphatic R^2 moiety (cyclohexyl, $V_{vdw} = 102$ Å^3).

Notably, the data collection statistics of some data sets presented in this section suggest, that the resolution cutoffs used for trimming have been chosen too conservative. This is especially true for the data sets collected in-house. However, the cutoffs of these data sets reflect the instrumentation limits, which in these cases led to a preset limit for the minimum outer resolution.

Structure-Activity Relationships

In summary, we analyzed 43 compounds which share the validated inhibitor scaffold shown in Fig. 6.3B. These compounds already include a wide range of substitutions and provide deeper insight into SARs. We were able to group these compounds into 11 different structure-activity classes, which are annotated in the appendix (Table E.5):

Group 1 comprises all derivatives with tertiary amides in the linker, that is $R^2 = H$. The secondary amide B2 (15 μM) and its methylated derivative A9 (inactive) demonstrate that tertiary amides are not tolerated.

Group 2 comprises all aromatic substitutes with a substitution in *para*-position, which is not tolerated as shown by B4 (27 μM) and its *para*-substituted derivative B30 (inactive). Even the smallest possible *para*-fluorophenyl leads to an inactive compound (B11). The bicyclic compounds A14 and B28 fail for the same reason.

Group 3 derivatives are inactive due to an extended linker. Even the elongation by a single methylene group is deleterious as demonstrated by B4 (27 μM) and B10 (inactive).

Group 4 comprises the acyclic aliphatic compounds, which are all inactive.

Group 5 contains the active alicycles B2 and B8. Here, the larger cyclohexane of B2 is about eightfold more active than the cyclopentane of B8.

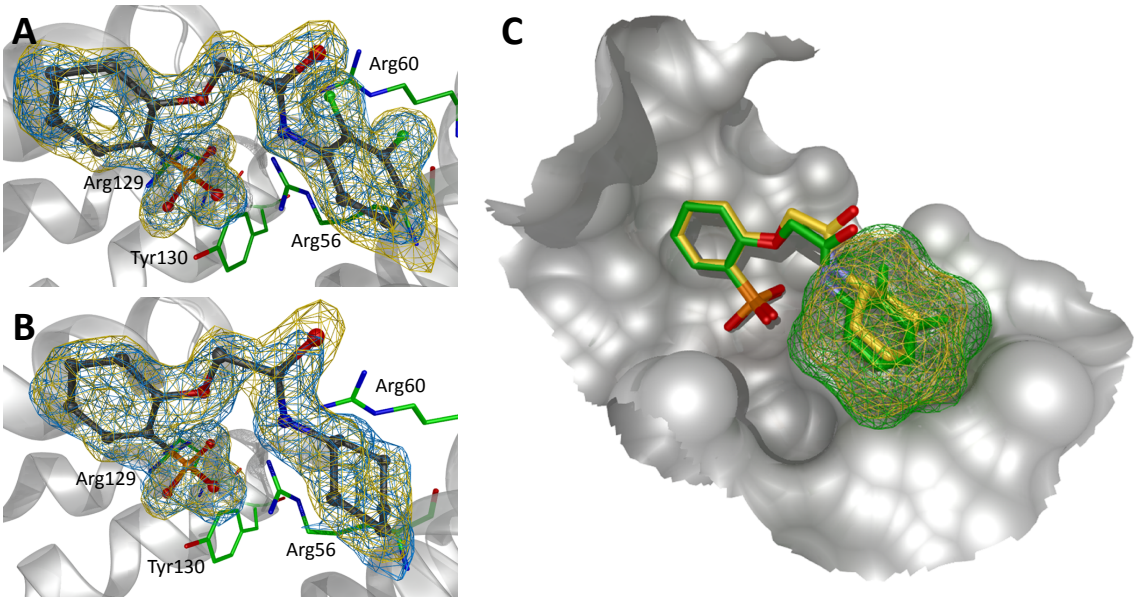


Figure 6.5: Superimposed crystal structures of B1 and B2 bound to 14-3-3σ. **(A)** Crystal structure of B1 bound to 14-3-3σ (grey cartoon; gold mesh: $mF_o - DF_c$ density map at 3.0σ contour level after protein refinement; blue mesh: $2mF_o - DF_c$ simulated annealing composite-omit map at 1.2σ contour level of the final model) **(B)** Crystal structure of B2 bound to 14-3-3σ (same coloring and map contour levels as for B1) **(C)** Superimposed B1 (green carbons) and B2 (yellow carbons). Meshes represent the SES of the compounds' R^2 moieties including hydrogen atoms, which occupy a comparable volume (coloring according to ligand carbons).

Group 6 compounds feature small, electron-withdrawing substituents in *ortho*-position (B4: 27 μM), which show a subtle advantage over the *meta*-position of group 7 compounds (B5: 32 μM). A combination of both shows an additive effect (group 8, B1: 5 μM). However, the larger and electron-donating B26 is inactive (group 9).

Finally, group 10 contains 2,5-substituted compounds, which are less active than a 2,4-substitution present in compound B3 (group 11).

6.3.4 Inhibition of the 14-3-3σAminopeptidase N Interaction

Evaluation of the inhibitors' potency to suppress 14-3-3 triggered upregulation of MMP-1 mRNA levels was performed in a cell-based assay using human IMR90 lung fibroblasts (Nicole Meissner and Svenja Schäfers, CGC Dortmund). In brief, the influence of extracellular 14-3-3 on MMP-1 mRNA transcription in human lung fibroblasts was analyzed by real-time quantitative PCR.[231] As shown in Fig. 6.6A, treatment with 14-3-3σ for 24 h led to a three-fold increase of MMP-1 levels compared to untreated cells. When the cells were treated with 14-3-3σ in combination with 200 μM of compound B1 or B2, the 14-3-3-induced transcrip-

tional increase was suppressed. A subsequent titration experiment of the more potent inhibitor B1 is shown in Fig. 6.6B and yielded an IC_{50} value of 81 ± 15 µM.

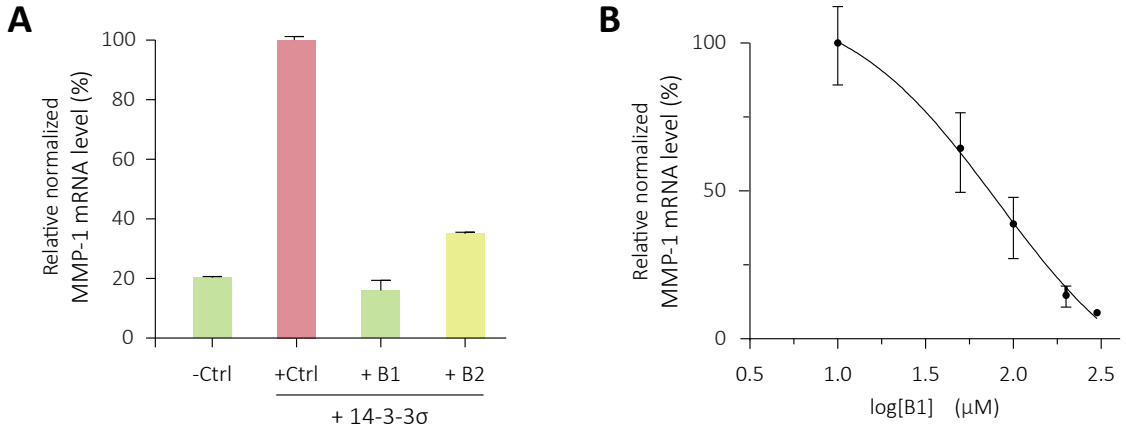


Figure 6.6: Cell-based assay on human lung fibroblasts for MMP-1 mRNA expression. **(A)** Treatment of human lung fibroblasts with 14-3-3σ alone as a positive control (+Ctrl) and co-treated with 200 µM of compounds B1 and B2. As a negative control (-Ctrl) cells were only treated with DMSO. **(B)** Relative abundance of 14-3-3σ-stimulated MMP-1 levels were down-regulated upon treatment with compound B1 in a concentration-dependent manner and yields an IC_{50} of 81 ± 15 µM.

Membrane Permeability

To exclude intracellular effects of compounds B1 and B2, which possibly could interfere with 14-3-3's extracellular application, the compounds ability to permeate cell membranes using the parallel artificial membrane permeation assay (PAMPA) was analyzed by the LDC Dortmund. The measured flux values for compounds B1 and B2 were -10.6 % and -8.7 %, respectively. Flux values below 5 % classify compounds as having a low membrane permeability, which is the case for both compounds.

6.3.5 Covalent Inhibition of 14-3-3 PPIs

In the initial validation screen shown in Fig. 6.3B, a third compound (A3) also inhibited the monitored PPI with lower potency. As this compound does not share the previously described phenylphosphonate scaffold, we were also interested in obtaining a structure of 14-3-3 in complex with A3. However, multiple crystallization trials as described in Section 6.2.2 were not successful and no crystal growth was observed. We speculated that the low binding affinity of A3 and 14-3-3 – as indicated by its lower potency – is too weak to yield a homogeneous complex solution, which enables crystal formation.

We thus tried an alternative approach by incubating existing crystals consisting of complexes between 14-3-3 and phosphopeptides from p53 with A3 in order to displace bound

peptide by compound. Crystals of 14-3-3σp53 were obtained as described previously.[123] We directly added A3 from a 20 mM DMSO stock solution to hanging-drops containing grown crystals and incubated them for three days. After this incubation period a crystal was directly flash-frozen in liquid nitrogen and used for a diffraction experiment.

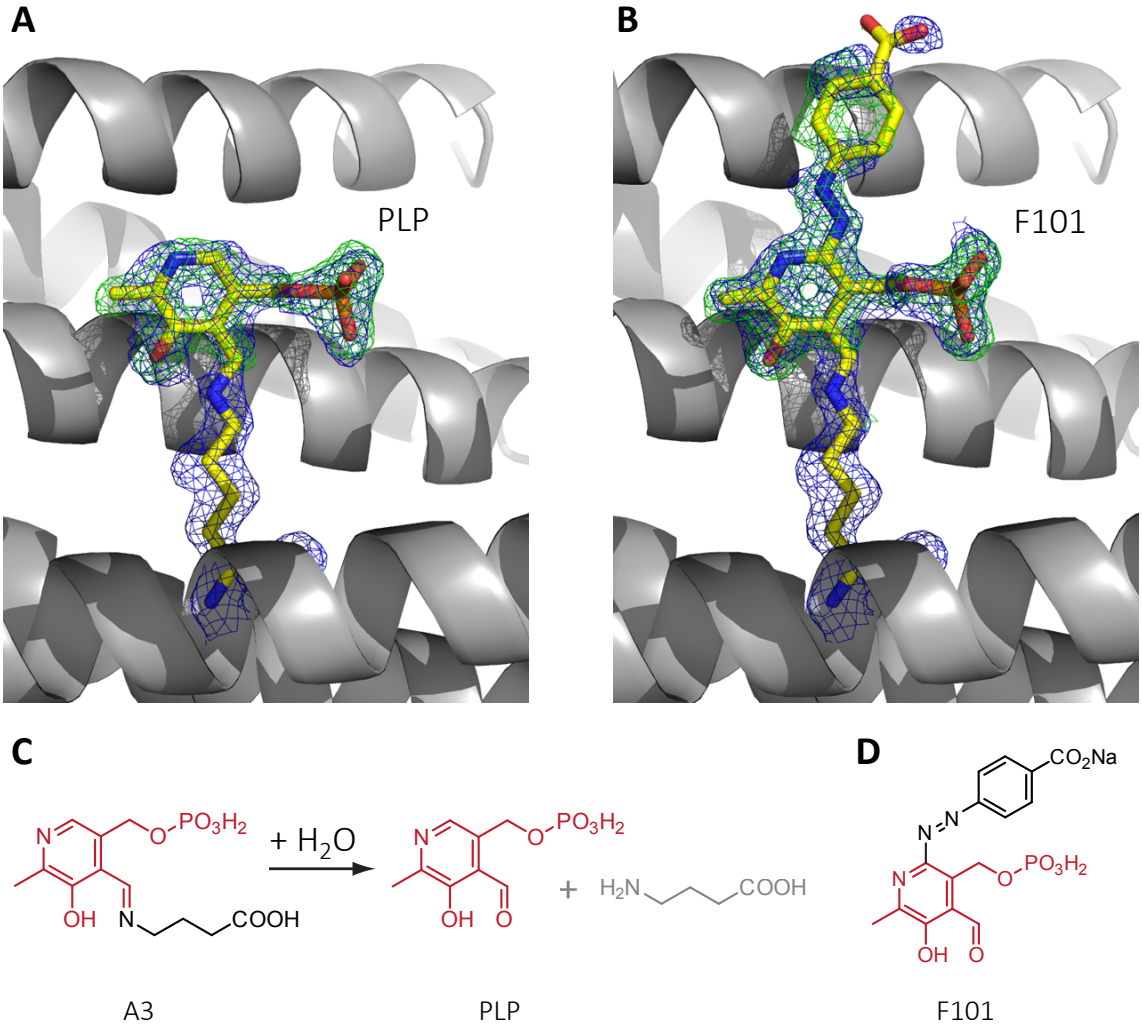


Figure 6.7: Crystal structures of covalent PLP binders to 14-3-3. **(A)** A3 covalently bound to Lys122 of 14-3-3σ (PDB ID: 3u9x, 14-3-3: grey cartoon, Lys122 with bound compound: stick-model colored by element, blue mesh: simulated annealing composite-omit map calculated with PHENIX at 1.3σ contour level, green mesh: $mF_o - DF_c$ density map at 3.0σ contour level). The $mF_o - DF_c$ density map has been calculated after protein refinement without ligand bias. **(B)** F101 covalently bound to Lys122 of 14-3-3σ (same coloring and contour levels as in A). **(C)** A3 featuring an imine substructure and its hydrolysis in aqueous solution to PLP and a primary amine. The PLP substructure is highlighted in red. **(D)** F101 that already possesses the aldehyde necessary for reacting with Lys122. The PLP substructure is also highlighted in red. (A) and (B) are taken from Röglin *et al.* (*Proc. Natl. Acad. Sci. USA* (2012), **109**, E1051-3).[208]

Processing of the collected data set yielded a resolution of 1.8 Å. The initial electron density after MR indicated that the phosphopeptide was displaced by compound A3 for which unmodelled $F_o - F_c$ electron density appeared. To our surprise, the compound occupied another volume in the amphipathic groove and the phosphate moiety was not coordinated as observed before. In fact, the compound covalently binds to Lys122 of 14-3-3σ by forming a *Schiff base* as shown in Fig. 6.7A. The most likely mechanism for this observation is a reversible imine formation between the lysines primary amine and an aldehyde. As shown in Fig. 6.7C, compound A3 does not possess an aldehyde group but a secondary imine substructure. This imine can be hydrolyzed in a weak acidic environment resulting in pyridoxal-phosphate (PLP), which is one of the most important physiological co-factors.[232] PLP can reversibly bind to lysines and form the modified residue N6-PLP-L-lysine (PSI-MOD 128), which is currently present in 229 PDB structures (accessed 26/6/2013). The importance of this mechanism is underlined by the observation, that more than 95 % of PLP in plasma is reversibly linked to a lysine residue in human serum albumin, which serves as a reservoir and transport system.[233]

Interestingly, the afore-described 14-3-3 inhibitor F101 is also a PLP derivative, but the published crystal structure shown in Fig. 6.1 and the proposed reaction mechanism differ markedly from our observations. Notably, this derivative already features the aldehyde necessary for covalent linking to Lys122 of 14-3-3, as can be seen in Fig. 6.1D. As the high quality of our X-ray data hardly supports any other covalent binding hypothesis we decided to obtain a sample of F101 and tried to solve a complex structure with 14-3-3σ.[208] Based on our experiences with A3 we also performed soaking experiments. As the compound is highly soluble in aqueous solution, we dissolved it in crystallization buffer. From this solution we directly added to a hanging-drop containing crystals of 14-3-3σ and p53 phosphopeptide complexes to reach 5 mM compound concentration. After an incubation period of eight days, crystals were frozen in liquid nitrogen and used for a diffraction experiment. The resulting data set is of comparable quality as for the A3 complex and was refined up to 1.65 Å resolution. The final model is shown in Fig. 6.1B and F101 shows the same covalent attachment to Lys122 as A3, which supports our observations. We also observed interpretable electron density for the *para*-amino-benzoate moiety despite its high flexibility, which argues against radiation induced cleavage of F101. Additionally, further unmodelled difference electron density nearby Lys49 indicated the presence of a second covalently bound F101 molecule. To confirm these observations, we performed an electrospray ionization mass spectrometry (ESI-MS) experiment (Dr. Lars Röglin, CGC Dortmund) and verified the imine formation between 14-3-3ζ and F101 in solution (Fig. D.5). Overnight incubation followed by dilution revealed two, three, and four molecules F101 attached to one monomer of 14-3-3.

6.4 Discussion

We demonstrated a successful *in silico* screening approach for the identification of 14-3-3 PPI inhibitors. We were able to identify the first non-covalent and non-peptidic inhibitors of this important protein family, whose mode of action is completely characterized by X-ray crystallography and supported by initial SAR studies. As an interesting *in vivo* assay, we showed the ability of the identified inhibitors to suppress the upregulation of MMP-1 mRNA levels, which is normally induced by addition of extracellular 14-3-3σ.

Analysis of nine 14-3-3 complex structures extracted from the PDB yielded insights into the binding mode of different phosphopeptides to 14-3-3. These X-ray structures form an ensemble of naturally occurring protein conformations for 14-3-3 and especially for its bound peptides. For 14-3-3, which we used as the receptor for structure-based docking, the analysis of its conformations showed only marginal flexibility independent of the bound phosphopeptide. Thus, using one of the compared structures as a docking template formed a useful and valid hypothesis. Additionally, the analysis of the bound phosphopeptides revealed only a relatively small sequence stretch with low flexibility, which we could use to derive a pharmacophore hypothesis. As expected, this spatially most conserved moiety is the phosphorylated residue and most prominently its phosphate group.

Interestingly, in the pharmacophore model used for virtual ligand screening created by Corradi *et al.*, the phosphate was not represented as a group of three spatially related hydrogen bond acceptors, but as a spherical negative charge.[220] To the best of our knowledge, neither *in vitro* assay data nor a structure of their inhibitor – BV02 – in complex with 14-3-3 were published yet. Thus, it is possible that BV02 does not directly interact with 14-3-3 but that it hits an off-target, which is responsible for the observed cellular effects. This would underline the importance of the spatial arrangement of the phosphate group and suggests the same importance of the phosphate's geometry as its negative charge for binding to 14-3-3.

Using a generalized phosphonate in our ligand-based VS led to a tremendous reduction of the input library comprising more than eight million compounds. Subsequent filtering steps and diverse selection yielded a manageable number of compounds for protein-ligand docking. The final step in our virtual screening approach was docking into a high-resolution crystal structure of 14-3-3σ. We employed docking to calculate sterically and physicochemically meaningful poses. We chose the size of the receptor bounding box larger than the default values to encompass the majority of 14-3-3 's amphipathic groove, centered at the phospho-coordination site. This enabled the docking algorithm to sample compound orientations in both directions of the amphipathic groove. Despite of these settings, only 70 different compounds out of 512 passed the Glide docking funnel. We used the 200 top-ranked docking poses for visual inspection of the placement of the phosphonate moiety, for which various positions were observed. Rejecting compounds that were not able to establish the conserved hydrogen bonds

to Arg56, Arg129, Tyr130 via their phosphonate oxygens was a further reasonable restriction, especially because the overall quite high docking scores gave no further clue for selection. Two different *in vitro* assays to monitor the interaction of 14-3-3 and a phosphorylated peptide from C-Raf yielded two compounds with inhibitory activity above 50 %.

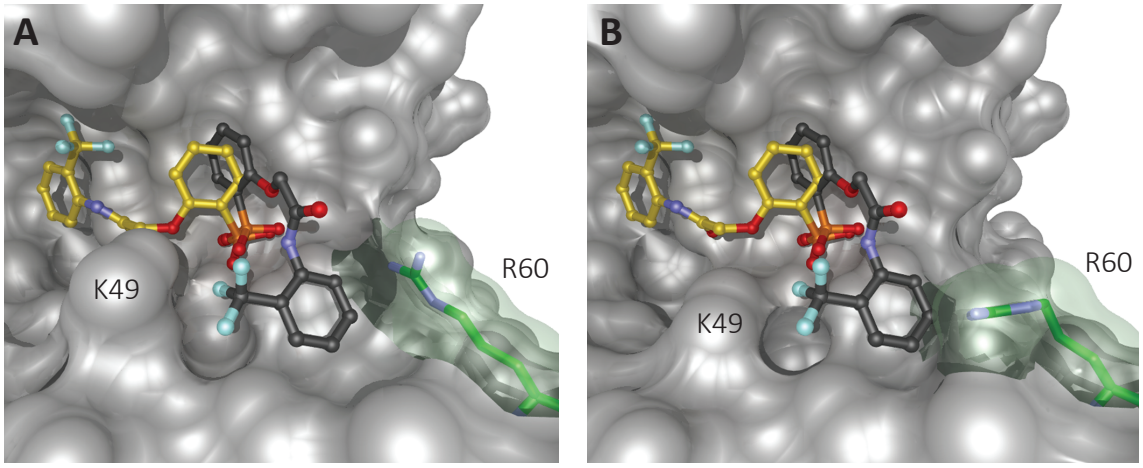


Figure 6.8: Comparison of docking pose (ball-and-stick model, yellow carbons) and X-ray conformation (ball-and-stick model, grey carbons) of compound A1. **(A)** 14-3-3σ (grey SES) from complex structure with A1 (PDB ID: 3t0l). Residue Arg60 is highlighted (green, semi-transparent SES and stick model). **(B)** 14-3-3σ structure (grey SES, PDB ID: 3p1n) used as receptor for docking. Residue Arg60 is highlighted accordingly.

Successful crystallization of these 14-3-3 inhibitor complexes revealed binding modes and enabled us to validate our virtual screening approach. The most important observation is the placement of the compounds' phosphonate groups, which matches our hypothesis. Also important is the observation that the inhibitors do not bind in an elongated conformation within the amphipathic groove but cross it. Thereby, both ligands occupy a quite shallow subpocket at the upper rim of the groove. These binding modes markedly differ from the selected docking poses. Superposition of the docked pose of A1 and its crystallized geometry is shown in Fig. 6.8. Subfigure A shows the superimposed inhibitor conformations in the 14-3-3 structure, which has been crystallized with A1. Most notable is the conformation of Arg60, which is elongated and forms one boundary of the shallow subpocket. In contrast, Fig. 6.8B shows the compound conformations in the 14-3-3 structure, which was used as the docking receptor (PDB ID: 3p1n). Here, Arg56 adopts a bent conformation where its guanidinium group points into the shallow subpocket, thereby reducing the volume of the latter. Interestingly, in only one out of the 14 available and analyzed 14-3-3∘phosphopeptide structures Arg60 was in a bent conformation. Since we accessed these structures in 2010, the number of deposited 14-3-3 entries has nearly doubled and comprises three additional complexes with bent Arg60. Furthermore, the highly flexible Lys49 is slightly displaced in the crystallized 14-3-3 conformation, which also enlarges the subpocket. These two subtile structural changes were responsible for incorrect placement

of ligands by docking, which underlines the importance of our decision to solely focus on the correct placement of the phosphonate. It remains speculation, whether the phosphonate of the identified inhibitors alone is responsible for binding to 14-3-3 and their effect. However, the inactivity of various tested compounds which should be flexible enough to adopt a suitable position in the amphipathic groove, contradict this hypothesis.

Crystal structures and *in vitro* data enabled us to speculate about SAR and led to the identification of further active inhibitors sharing the conserved scaffold. These additional structures extended the SAR but the explored chemical space is limited yet and further optimizations might lead to inhibitors with submicromolar activity. First, chemical modifications should focus on the phenylphosphonate ring because surrounding 14-3-3 offers unexplored interaction possibilities. For example, polar interactions to Asn175 and Asn226 or extended hydrophobic contacts to Leu174, Val178 and to the protein backbone are possible. Second, introduction of a ring system in the linker region to decrease conformational flexibility might be entropically beneficial. Here, the high quality 14-3-3$\sigma\circ$inhibitor structures we presented in this work form an excellent resource to accompany both optimization strategies with *in silico* methods. For example, fragment linking tools could suggest alternative chemical linkers between the phenylphosphonate ring and the R^2 moiety.

We further presented an interesting application where we demonstrated the ability of the compounds to inhibit the overexpression of MMP-1 in human fibroblasts via extracellular 14-3-3 stimulation. Knowledge of the extracellular tasks of 14-3-3 proteins is continuously growing and this research area can greatly benefit from our results. As chemical biology tools the most potent inhibitors are already useful to analyze the binding events between 14-3-3 and its surface receptor APN. The mode of action of our inhibitors suggests a phosphorylation-dependent binding of APN and possibly facilitates identification of the interacting sequence.

A surprising discovery was PLP derivative A3, which binds covalently to a conserved lysine in the amphipathic groove. The compound's ability to attach to the primary amine of lysines is not directly obvious from its structure because preceding hydrolysis is necessary to form an aldehyde as the corresponding reaction partner. Structure determination of the covalent complex revealed its true mode of action. Additionally, we were able to elucidate the true binding mode of another PLP derivative, which has recently been published.[107,208] Finally, our results on binding of PLP derivatives to 14-3-3 allows us to speculate if this mechanism is an as yet unknown storage or transport function of 14-3-3 proteins. Binding to an entirely conserved lysine and the high abundance of 14-3-3 proteins in cells argues for this speculation. Furthermore, it is already known that PLP is stored and transported in the blood via this mechanism.[233]

Chapter 7

Conclusion and Outlook

The presented thesis covers a broad range of methods in computer-assisted as well as experimental drug discovery. The continuously growing demands by steadily increasing amounts of structural data in this discipline and the necessity to discover novel ways in drug discovery were the incitement of this work. The first two chapters are focused on studying theoretical problems from chemoinformatics. Here, we introduced novel methods that enable meaningful and efficient classification of large chemical spaces. The obtained results were integrated in the second part of this work, which focuses on applying computational methods on emerging targets and its associated problems in modern drug discovery. Accompanied by continuous *in vitro* validation and structure elucidation, this work is strongly interdisciplinary.

In Chapter 3, we presented a novel chemoinformatics clustering method for efficient and fast processing of large chemical spaces. The core of our method is formed by an extremely fast and parallel algorithm to calculate similarities between 2D binary fingerprints. It was one of our goals that the entire method is architecture-independent and that it runs on standard hardware without using specialized instruction sets. Our benchmarking results demonstrate, that our implementation is as fast as state-of-the-art hardware-dependent methods and can calculate up to 365 million similarities per second on a desktop computer. This performance enabled clustering of the available chemical space comprising 17 million compounds in less than three days. The method is a useful alternative to existing and frequently used tools, especially because it is a deterministic approach taking all pairwise similarities into account.

The performance of our method creates further interesting perspectives. A possible future development could be an incremental clustering algorithm. An obvious use case for such a variant would be the regular update of an already clustered compound database. New compounds are continuously inserted into the database, which makes an update of the clustering necessary. For this task, the presented algorithm is perfectly suited because inserted compounds can be merged into a new inverted index block. The latter can easily be processed against the already

existing library blocks. As our algorithm reaches already 75 % of its peak performance at a block size of 100, a weekly or even daily update would be feasible.

Additionally, our method can possibly be transferred to other problem domains in bioinformatics dealing with large amounts of data that can be represented as binary feature vectors. A possible application could be the clustering of sparsely encoded mass spectra in computational proteomics. As the implementation is independent of the chemoinformatics layer it could be adapted with only little efforts.

In Chapter 4, we introduced *scaffold families* as a concept to generalize the static scaffold representation of small molecules. Here, our goal was to merge different scaffolds that are related in a medicinal chemistry way into a single *scaffold family*. For selected examples, we showed that connected component decomposition of binary fingerprint-based similarity networks yields meaningful *scaffold families*. The method can be used to pre-cluster large chemical spaces or to estimate the medicinal chemistry diversity in compound data sets.

In the second part of this thesis, computational tools were developed and utilized to study the modulation of PPIs. In Chapter 5, we studied the small molecule-induced stabilization of PPIs. Examples from nature and the recently published proof of concept for rational *in vitro* design of a PPI stabilizer inspired us to analyze their structural details. Our goal was to gain knowledge about this fascinating mode of action and to develop novel approaches for PPI stabilizer-focused VS. We used the developed tools to screen the PDB and indeed found unrecognized PPI stabilizers. With that knowledge, we performed a stabilizer VS using the 14-3-3∘Task3 complex as our model system. Using our tools, we finally selected 89 candidates from VS for *in vitro* testing. Different biochemical assays to test stabilizing activity and to exclude artifacts finally left one hit compound. However, we were not able to solve a crystal structure of the ternary complex with 14-3-3 and Task3 until now and further experiments will be necessary. Nevertheless, these results are encouraging. The developed tools and procedures will be of great benefit for the structure-based drug design community. Applying them on other PPIs is highly desirable. Further theoretical work and research should focus on strategies to rescore predicted PPI-ligand complexes. Our current structure-based filter does not evaluate binding-free energy changes of the target PPI upon ligand binding. Incorporating such energetic contributions, however, could serve as a final rescoring step and their calculation is possibly feasible by means of sophisticated molecular dynamics simulations. Studies performing such simulations have not been published yet and thus, our work serves as an excellent starting point for such research projects.

In Chapter 6 we examined the inhibition of 14-3-3 PPIs by small molecules. The PDB provides a large number of crystal structures of 14-3-3 in complex with different protein binding partners. We used this source of structural information to analyze the binding modes and to infer a minimal pharmacophore. Based on the latter we set up a VS and indeed identified

small molecule inhibitor candidates that were validated *in vitro*. Successful crystallization experiments enabled us to elucidate complex structures and to perform a SAR studies, which led to the identification of improved inhibitors. Furthermore, we were able to demonstrate the activity of our inhibitors in cell-based experiments. The identified compounds are the first biochemically and structurally characterized reversible small molecule 14-3-3 inhibitors. As an important signaling hub, 14-3-3 is an attractive target for drug design research. Thus, our results will be of great interest for researchers working with 14-3-3.

In summary, the results presented in this thesis will be of interest for theoretical and experimental scientists in chemoinformatics, structural bioinformatics, structural biology, and drug design communities. Interesting questions and starting points for further research – experimental as well as theoretical – arise from this work and will contribute to the integration of PPIs into the druggable genome.

Bibliography

[1] Moll A., Hildebrandt A., Lenhof H.-P., and Kohlbacher O. (2005). BALLView: an object-oriented molecular visualization and modeling framework. *J. Comput.-Aided Mol. Des.*, 19(11):791–800. xi

[2] Hildebrandt A., et al. (2010). BALL–biochemical algorithms library 1.3. *BMC Bioinformatics*, 11(1):531. xi, 37, 67, 141

[3] Westendorf W. *Handbuch der altägyptischen Medizin; 1*. Brill (1999). 1

[4] Lander E. S., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921. 1

[5] Venter J. C., et al. (2001). The sequence of the human genome. *Science*, 291(5507):1304–51. 1

[6] Hopkins A. L. and Groom C. R. (2002). The druggable genome. *Nat. Rev. Drug Discovery*, 1(9):727–30. 1

[7] Imming P., Sinning C., and Meyer A. (2006). Drugs, their targets and the nature and number of drug targets. *Nat. Rev. Drug Discov.*, 5(10):821–34. 1

[8] Overington J. P., Al-Lazikani B., and Hopkins A. L. (2006). How many drug targets are there? *Nat. Rev. Drug Discovery*, 5(12):993–6. 1

[9] Goode D. R., Totten R. K., Heeres J. T., and Hergenrother P. J. (2008). Identification of promiscuous small molecule activators in high-throughput enzyme activation screens. *J. Med. Chem.*, 51(8):2346–9. 2

[10] Matschinsky F. M. (2009). Assessing the potential of glucokinase activators in diabetes therapy. *Nat. Rev. Drug Discov.*, 8(5):399–416. 2

[11] Lewis Phillips G. D., et al. (2008). Targeting HER2-positive breast cancer with trastuzumab-DM1, an antibody-cytotoxic drug conjugate. *Cancer Res.*, 68(22):9280–90. 2

[12] Whitty M. G., Adrian, Alley S. C., Okeley N. M., and Senter P. D. (2010). Antibody-drug conjugates: targeted drug delivery for cancer. *Curr. Opin. Chem. Biol.*, 14(4):529–537. 2

[13] Rao D. D., Vorhies J. S., Senzer N., and Nemunaitis J. (2009). siRNA vs. shRNA: similarities and differences. *Adv. Drug Delivery. Rev.*, 61(9):746–59. 2

[14] Garzon R., Marcucci G., and Croce C. M. (2010). Targeting microRNAs in cancer: rationale, strategies and challenges. *Nat. Rev. Drug Discov.*, 9(10):775–89. 2

[15] Thomas J. R. and Hergenrother P. J. (2008). Targeting RNA with small molecules. *Chem. Rev.*, 108(4):1171–224. 2

[16] Zhang Q. C., et al. (2012). Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature*, 490(7421):556–60. 3

[17] Berg T. (2003). Modulation of protein-protein interactions with small organic molecules. *Angew. Chem. Int. Ed. Engl.*, 42(22):2462–81. 3, 61

[18] Yin H. and Hamilton A. D. (2005). Strategies for targeting protein-protein interactions with synthetic agents. *Angew. Chem. Int. Ed. Engl.*, 44(27):4130–63. 61

[19] Jubb H., Higueruelo A. P., Winter A., and Blundell T. L. (2012). Structural biology and drug discovery for protein-protein interactions. *Trends Pharmacol. Sci.*, 33(5):241–8. 3

[20] Wells J. A. and McClendon C. L. (2007). Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature*, 450(7172):1001–9. 3, 24, 61, 91

[21] Mullard A. (2012). Protein-protein interaction inhibitors get into the groove. *Nat. Rev. Drug Discovery*, 11(3):173–175. 3, 24

[22] Thiel P., Kaiser M., and Ottmann C. (2012). Small-molecule stabilization of protein-protein interactions: an underestimated concept in drug discovery? *Angew. Chem. Int. Ed. Engl.*, 51(9):2012–8. 3, 7, 22, 23, 26, 27, 61, 62, 65, 90

[23] Block P., Weskamp N., Wolf A., and Klebe G. (2007). Strategies to search and design stabilizers of protein-protein interactions: a feasibility study. *Proteins*, 68(1):170–86. 3, 62, 63, 93

[24] Gao M. and Skolnick J. (2012). The distribution of ligand-binding pockets around protein-protein interfaces suggests a general mechanism for pocket formation. *Proc. Natl. Acad. Sci. U. S. A.*, 109(10):3784–9. 3, 63

[25] Rose R., et al. (2010). Identification and structure of small-molecule stabilizers of 14-3-3 protein-protein interactions. *Angew. Chem. Int. Ed. Engl.*, 49(24):4129–32. 3, 26, 61, 64, 90, 93

[26] Directory of computer-aided Drug Design tools - Swiss Institute of Bioinformatics (SIB). `http://www.click2drug.org`. Accessed: January 2014. 3

[27] Cormen T. H., Stein C., Rivest R. L., and Leiserson C. E. (2001). Introduction to Algorithms. 7, 8

[28] Sibeyn J. F. External Connected Components. In *Algorithm Theory - SWAT 2004*, pages 468–479 (2004). 8

[29] Engel T. (2006). Basic overview of chemoinformatics. *J. Chem. Inf. Model.*, 46(6):2267–77. 9

[30] Bajorath J. and Vogt M. (2012). Chemoinformatics: A view of the field and current trends in method development. *Bioorg. Med. Chem.*, 20(18):5317–5323. 9, 36

[31] Bayer HealthCare. Small and large molecules: drugs on a chemical and biological basis. (http://www.bayerpharma.com/en/research-and-development/technologies/small-and-large-molecules/index.php). 9

[32] Lipinski C. A., Lombardo F., Dominy B. W., and Feeney P. J. (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery. Rev.*, 46(1-3):3–26. 9

[33] Weininger D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.*, 28(1):31–36. 10

[34] Gasteiger J., editor. *Handbook of Chemoinformatics*. Wiley-VCH Verlag GmbH, Weinheim, Germany (2003). 10

[35] Daylight Theory Manual, Daylight Chemical Information Systems, Inc. `http://www.daylight.com/dayhtml/doc/theory/`. Version: 4.9. 10, 11

[36] Bender A., Mussa H. Y., Glen R. C., and Reiling S. (2004). Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. *J. Chem. Inf. Comput. Sci.*, 44(5):1708–18. 11

[37] Rogers D. and Hahn M. (2010). Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, 50(5):742–54. 11, 42, 99

[38] Jaccard P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547 – 579. 11

[39] Tanimoto T. T. IBM Internal Report. Technical report (1957). 11

[40] Leach A. R. and Gillet V. J. *An Introduction to Chemoinformatics*. Springer, 1st edition (2007). 12, 13

[41] Forgy E. W. (1965). Cluster analysis of multivariate data: efficiency vs interpretability of classifications. *Biometrics*, 21:768–769. 12

[42] MacQueen J. B. Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Symp. on Math. Statist. and Prob.*, pages 281–297 (1967). 12

[43] Hodes L. (1989). Clustering a large number of compounds. 1. Establishing the method on an initial sample. *J. Chem. Inf. Comput. Sci.*, 29(2):66–71. 13

[44] Jarvis R. and Patrick E. (1973). Clustering Using a Similarity Measure Based on Shared Near Neighbors. *IEEE Trans. Comput.*, C-22(11):1025–1034. 13

[45] Böhm H.-J. and Schneider G., editors. *Virtual Screening for Bioactive Molecules - Wiley Online Library*. WILEY-VCH Verlag GmbH (2000). 13

[46] Ward J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.*, 58(301):236–244. 13, 73

[47] Downs G. M., Willett P., and Fisanick W. (1994). Similarity Searching and Clustering of Chemical-Structure Databases Using Molecular Property Data. *J. Chem. Inf. Model.*, 34(5):1094–1102. 13, 35

[48] Brown R. and Martin Y. (1996). Use of Structure-Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Model.*, 36(3):572–584. 13

[49] Bemis G. W. and Murcko M. A. (1996). The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.*, 39(15):2887–93. 14, 51

[50] Bemis G. W. and Murcko M. A. (1999). Properties of known drugs. 2. Side chains. *J. Med. Chem.*, 42(25):5095–9.

[51] Schuffenhauer A., et al. (2006). The scaffold tree–visualization of the scaffold universe by hierarchical scaffold classification. *J. Chem. Inf. Model.*, 47(1):47–58. 14, 51

[52] Kitchen D. B., Decornez H., Furr J. R., and Bajorath J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.*, 3(11):935–49. 15

[53] Böhm H. J. (1994). The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comput.-Aided Mol. Des.*, 8(3):243–56. 15

[54] Friesner R. A., et al. (2004). Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.*, 47(7):1739–49. 15, 74, 100

[55] Muegge I. and Martin Y. C. (1999). A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J. Med. Chem.*, 42(5):791–804. 15

[56] Gohlke H., Hendlich M., and Klebe G. (2000). Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.*, 295(2):337–56. 15

[57] Rupp B. *Biomolecular Crystallography*. Garland Science, 1st edition (2010). 16

[58] Drenth J. *Principles of Protein X-ray Crystallography*. Springer-Verlag New York, Inc., 1st edition (1994). 16

[59] McGuffee S. R. and Elcock A. H. (2010). Diffusion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm. *PLoS Comput. Biol.*, 6(3):e1000694. 22

[60] Clamp M., et al. (2007). Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl. Acad. Sci. U. S. A.*, 104(49):19428–33. 22

[61] Pontén F., et al. (2009). A global view of protein expression in human cells, tissues, and organs. *Mol. Syst. Biol.*, 5:337. 22

[62] Zimmerman S. B. and Trach S. O. (1991). Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of Escherichia coli. *J. Mol. Biol.*, 222(3):599–620. 22

[63] Arkin M. R. and Wells J. A. (2004). Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. *Nat. Rev. Drug Discovery*, 3(4):301–17. 22, 24

[64] Nooren I. M. A. and Thornton J. M. (2003). Diversity of protein-protein interactions. *EMBO J.*, 22(14):3486–92. 23

[65] Bogan A. A. and Thorn K. S. (1998). Anatomy of hot spots in protein interfaces. *J. Mol. Biol.*, 280(1):1–9. 24

[66] Conte L. L., Chothia C., Janin J., and Lo Conte L. (1999). The atomic structure of protein-protein recognition sites. *J. Mol. Biol.*, 285(5):2177–2198. 24

[67] Janin J., Bahadur R. P., and Chakrabarti P. (2008). Protein-protein interaction and quaternary structure. *Q. Rev. Biophys.*, 41(2):133–80. 24

[68] Jones S. and Thornton J. M. (1996). Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. U. S. A.*, 93(1):13–20. 24

[69] Sperandio O., Reynès C. H., Camproux A.-C., and Villoutreix B. O. (2010). Rationalizing the chemical space of protein-protein interaction inhibitors. *Drug Discov. Today*, 15(5-6):220–9. 24

[70] Villoutreix B. O., et al. (2008). In silico-in vitro screening of protein-protein interactions: towards the next generation of therapeutics. *Curr. Pharm. Biotechnol.*, 9(2):103–22. 24, 61

[71] Reynès C., et al. (2010). Designing focused chemical libraries enriched in protein-protein interaction inhibitors using machine-learning methods. *PLoS Comput. Biol.*, 6(3):e1000695. 24

[72] Czarna A., et al. (2010). Robust generation of lead compounds for protein-protein interactions by computational and MCR chemistry: p53/Hdm2 antagonists. *Angew. Chem. Int. Ed. Engl.*, 49(31):5352–6. 24

[73] Koch O., et al. (2013). Identification of M. tuberculosis Thioredoxin Reductase Inhibitors Based on High-Throughput Docking Using Constraints. *J. Med. Chem.*, 56(12):4849–59. 25

[74] Oltersdorf T., et al. (2005). An inhibitor of Bcl-2 family proteins induces regression of solid tumours. *Nature*, 435(7042):677–81. 25

[75] Eyrisch S. and Helms V. (2007). Transient pockets on protein surfaces involved in protein-protein interaction. *J. Med. Chem.*, 50(15):3457–64. 25

[76] Inoué S. and Sato H. (1967). Cell motility by labile association of molecules. The nature of mitotic spindle fibers and their role in chromosome movement. *J. Gen. Physiol.*, 50(6):Suppl:259–92. 25

[77] Desai A. and Mitchison T. J. (1997). Microtubule polymerization dynamics. *Annu. Rev. Cell Dev. Biol.*, 13:83–117. 25

[78] Jordan M. A. and Wilson L. (2004). Microtubules as a target for anticancer drugs. *Nat. Rev. Cancer*, 4(4):253–65. 26

[79] Wani M. C., Taylor H. L., Wall M. E., Coggon P., and McPhail A. T. (1971). Plant antitumor agents. VI. The isolation and structure of taxol, a novel antileukemic and antitumor agent from Taxus brevifolia. *J. Am. Chem. Soc.*, 93(9):2325–7. 26

[80] Nogales E., Wolf S. G., Khan I. A., Ludueña R. F., and Downing K. H. (1995). Structure of tubulin at 6.5 A and location of the taxol-binding site. *Nature*, 375(6530):424–7. 26

[81] Thomson A. W., Turnquist H. R., and Raimondi G. (2009). Immunoregulatory functions of mTOR inhibition. *Nat. Rev. Immunol.*, 9(5):324–37. 26

[82] Oecking C., Eckerskorn C., and Weiler E. W. (1994). The fusicoccin receptor of plants is a member of the 14-3-3 superfamily of eukaryotic regulatory proteins. *FEBS Lett.*, 352(2):163–6. 26

[83] Würtele M., Jelich-Ottmann C., Wittinghofer A., and Oecking C. (2003). Structural view of a fungal toxin acting on a 14-3-3 regulatory complex. *EMBO J.*, 22(5):987–94. 26

[84] Klausner R. D., Donaldson J. G., and Lippincott-Schwartz J. (1992). Brefeldin A: insights into the control of membrane traffic and organelle structure. *J. Cell Biol.*, 116(5):1071–80. 28

[85] Peyroche A., et al. (1999). Brefeldin A acts to stabilize an abortive ARF-GDP-Sec7 domain protein complex: involvement of specific residues of the Sec7 domain. *Mol. Cell*, 3(3):275–85. 28

[86] Renault L., Guibert B., and Cherfils J. (2003). Structural snapshots of the mechanism and inhibition of a guanine nucleotide exchange factor. *Nature*, 426(6966):525–30. 28

[87] Mossessova E., Corpina R. A., and Goldberg J. (2003). Crystal structure of ARF1*Sec7 complexed with Brefeldin A and its implications for the guanine nucleotide exchange mechanism. *Mol. Cell*, 12(6):1403–11. 28

[88] Viaud J., et al. (2007). Structure-based discovery of an inhibitor of Arf activation by Sec7 domains through targeting of protein-protein complexes. *Proc. Natl. Acad. Sci. U. S. A.*, 104(25):10370–5. 28

[89] Seamon K. B., Padgett W., and Daly J. W. (1981). Forskolin: unique diterpene activator of adenylate cyclase in membranes and in intact cells. *Proc. Natl. Acad. Sci. U. S. A.*, 78(6):3363–7. 28

[90] Hurley J. H. (1999). Structure, Mechanism, and Regulation of Mammalian Adenylyl Cyclase. *J. Biol. Chem.*, 274(12):7599–7602. 28

[91] Sunahara R. K., Dessauer C. W., Whisnant R. E., Kleuss C., and Gilman A. G. (1997). Interaction of Gsalpha with the cytosolic domains of mammalian adenylyl cyclase. *J. Biol. Chem.*, 272(35):22265–71. 28

[92] Zhang G., Liu Y., Ruoho A. E., and Hurley J. H. (1997). Structure of the adenylyl cyclase catalytic core. *Nature*, 386(6622):247–53. 28

[93] Tesmer J. J. (1997). Crystal Structure of the Catalytic Domains of Adenylyl Cyclase in a Complex with GsaGTPgS. *Science*, 278(5345):1907–1916. 28

[94] Moore B. W. and Perez V. J. Specific acidic proteins of the nervous system. In Carlson F. D., editor, *Physiological and Biochemical Aspects of Nervous Intergration*, pages 343–359. Prentice-Hall, Woods Hole (1967). 28

[95] Aitken A. (2006). 14-3-3 proteins: a historic overview. *Semin. Cancer Biol.*, 16(3):162–72. 29

[96] Johnson C., et al. (2010). Bioinformatic and experimental survey of 14-3-3-binding sites. *Biochem. J.*, 427(1):69–78. 29, 31

[97] Fu H., Subramanian R. R., and Masters S. C. (2000). 14-3-3 proteins: structure, function, and regulation. *Annu. Rev. Pharmacol. Toxicol.*, 40:617–47. 29

[98] Fu H., Coburn J., and Collier R. J. (1993). The eukaryotic host factor that activates exoenzyme S of Pseudomonas aeruginosa is a member of the 14-3-3 protein family. *Proc. Natl. Acad. Sci. U. S. A.*, 90(6):2320–4. 29

[99] Seimiya H., et al. (2000). Involvement of 14-3-3 proteins in nuclear localization of telomerase. *EMBO J.*, 19(11):2652–61.

[100] Patel A., et al. (2006). Host protein interactions with enteropathogenic Escherichia coli (EPEC): 14-3-3tau binds Tir and has a role in EPEC-induced actin polymerization. *Cell. Microbiol.*, 8(1):55–71. 29

[101] Bridges D. and Moorhead G. B. G. (2004). 14-3-3 proteins: a number of functions for a numbered protein. *Sci. STKE*, 2004(242):re10. 29

[102] Notredame C., Higgins D. G., and Heringa J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, 302(1):205–17. 29

[103] Biegert A., Mayer C., Remmert M., Söding J., and Lupas A. N. (2006). The MPI Bioinformatics Toolkit for protein sequence analysis. *Nucleic Acids Res.*, 34(Web Server issue):W335–9. 29

[104] The UniProt Consortium (2012). Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, 40(Database issue):D71–5. 29

[105] Dougherty M. K. and Morrison D. K. (2004). Unlocking the code of 14-3-3. *J. Cell Sci.*, 117(Pt 10):1875–84. 29

[106] Hermeking H. and Benzinger A. (2006). 14-3-3 proteins in cell cycle regulation. *Semin. Cancer Biol.*, 16(3):183–92. 29

[107] Zhao J., et al. (2011). Discovery and structural characterization of a small molecule 14-3-3 protein-protein interaction inhibitor. *Proc. Natl. Acad. Sci. U. S. A.*, 108(39):16212–16216. 30, 95, 97, 116

[108] Fantl W. J., et al. (1994). Activation of Raf-1 by 14-3-3 proteins. *Nature*, 371(6498):612–4. 30

[109] Dumaz N. and Marais R. (2003). Protein kinase A blocks Raf-1 activity by stimulating 14-3-3 binding and blocking Raf-1 interaction with Ras. *J. Biol. Chem.*, 278(32):29819–23. 30

[110] Molzan M., et al. (2010). Impaired binding of 14-3-3 to C-RAF in Noonan syndrome suggests new approaches in diseases with increased Ras signaling. *Mol. Cell. Biol.*, 30(19):4698–711. 30, 72, 92

[111] Lee V. M., Goedert M., and Trojanowski J. Q. (2001). Neurodegenerative tauopathies. *Annu. Rev. Neurosci.*, 24:1121–59. 30

[112] Yuan Z., Agarwal-Mawal A., and Paudel H. K. (2004). 14-3-3 binds to and mediates phosphorylation of microtubule-associated tau protein by Ser9-phosphorylated glycogen synthase kinase 3beta in the brain. *J. Biol. Chem.*, 279(25):26105–14. 30

[113] Hashiguchi M., Sobue K., and Paudel H. K. (2000). 14-3-3zeta is an effector of tau protein phosphorylation. *J. Biol. Chem.*, 275(33):25247–54. 30

[114] Klein D. C., et al. (2002). 14-3-3 Proteins and photoneuroendocrine transduction: role in controlling the daily rhythm in melatonin. *Biochem. Soc. Trans.*, 30(4):365–73. 30

[115] Obsil T., Ghirlando R., Klein D. C., Ganguly S., and Dyda F. (2001). Crystal Structure of the 14-3-3ζ:Serotonin N-Acetyltransferase Complex. *Cell*, 105(2):257–267. 30

[116] Zheng W. and Cole P. A. (2002). Serotonin N-acetyltransferase: mechanism and inhibition. *Curr. Med. Chem.*, 9(12):1187–99. 30

[117] Peschke E., et al. (2011). The insulin-melatonin antagonism: studies in the LEW.1AR1-iddm rat (an animal model of human type 1 diabetes mellitus). *Diabetologia*, 54(7):1831–40. 30

[118] Kostelecky B., Saurin A. T., Purkiss A., Parker P. J., and McDonald N. Q. (2009). Recognition of an intra-chain tandem 14-3-3 binding site within PKCepsilon. *EMBO Rep.*, 10(9):983–9. 30

[119] Molzan M. and Ottmann C. (2012). Synergistic binding of the phosphorylated S233- and S259-binding sites of C-RAF to one 14-3-3ζ dimer. *J. Mol. Biol.*, 423(4):486–95. 30

[120] Muslin A. J., Tanner J. W., Allen P. M., and Shaw A. S. (1996). Interaction of 14-3-3 with signaling proteins is mediated by the recognition of phosphoserine. *Cell*, 84(6):889–97. 31

[121] Yaffe M. B., et al. (1997). The structural basis for 14-3-3:phosphopeptide binding specificity. *Cell*, 91(7):961–71. 31

[122] Ganguly S., et al. (2005). Melatonin synthesis: 14-3-3-dependent activation and inhibition of arylalkylamine N-acetyltransferase mediated by phosphoserine-205. *Proc. Natl. Acad. Sci. U. S. A.*, 102(4):1222–7. 31

[123] Schumacher B., Mondry J., Thiel P., Weyand M., and Ottmann C. (2010). Structure of the p53 C-terminus bound to 14-3-3: implications for stabilization of the p53 tetramer. *FEBS Lett.*, 584(8):1443–8. 32, 112

[124] Fujita N., Sato S., Katayama K., and Tsuruo T. (2002). Akt-dependent phosphorylation of p27Kip1 promotes binding to 14-3-3 and cytoplasmic localization. *J. Biol. Chem.*, 277(32):28706–13. 32

[125] Thiel P., Sach-Peltason L., Ottmann C., and Kohlbacher O. (2014). Blocked Inverted Indices for Exact Clustering of Large Chemical Spaces. *J. Chem. Inf. Model.*, 54:2395–2401. 33, 141

[126] ChemAxon User's Guide 6.1.3. 33

[127] Irwin J. J., Sterling T., Mysinger M. M., Bolstad E. S., and Coleman R. G. (2012). ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.*, 52(7):1757–1768. 34, 42, 99

[128] Blum L. C. and Reymond J.-L. (2009). 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.*, 131(25):8732–3. 34

[129] Zahoránszky L. A., et al. (2009). Breaking the hierarchy–a new cluster selection mechanism for hierarchical clustering methods. *Algorithms Mol. Biol.*, 4(1):12. 35

[130] Haque I. S., Pande V. S., and Walters W. P. (2011). Anatomy of High-Performance 2D Similarity Calculations. *J. Chem. Inf. Model.*, 51(9):2345–2351. 36, 43

[131] Zobel J. and Moffat A. (2006). Inverted files for text search engines. *ACM Comput. Surv.*, 38(2):6–es. 36

[132] Nasr R. J., Vernica R., Li C., and Baldi P. (2012). Speeding Up Chemical Searches Using the Inverted Index: the Convergence of Chemoinformatics and Text Search Methods. *J. Chem. Inf. Model.*, 52(4):891–900. 36

[133] Kristensen T. G., Nielsen J., and Pedersen C. N. S. (2011). Using inverted indices for accelerating LINGO calculations. *J. Chem. Inf. Model.*, 51(3):597–600. 36

[134] Peltason L. *Clustering auf großen Strukturdatenbanken*. Studienarbeit, Universtät Tübingen (2004). 36, 141

[135] Siek J. G., Lee L.-Q., and Lumsdaine A. *The Boost Graph Library*. Addison-Wesley Professional, 1 edition (2001). 40

[136] Murtagh F. (1983). A Survey of Recent Advances in Hierarchical Clustering Algorithms. *Comput. J.*, 26(4):354–359. 41

[137] Murtagh F. *Multidimensional Clustering Algorithms*. Physica-Verlag Würzburg-Wien, compstat l edition (1984). 41

[138] Kelley L. A., Gardner S. P., and Sutcliffe M. J. (1996). An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies. *Protein Eng.*, 9(11):1063–5. 41

[139] Pipeline Pilot, Accelrys Software Inc. `http://accelrys.com/products/pipeline-pilot/`. Version: 6.1.5.0. 42, 99

[140] OEChem, OpenEye Scientific Software Inc. `http://www.eyesopen.com`. Version: 1.7.4. 42

[141] JChem, ChemAxon. `http://www.chemaxon.com`. Version: 5.8.0. 42, 46

[142] Lepp Z., Huang C., and Okada T. (2009). Finding key members in compound libraries by analyzing networks of molecules assembled by structural similarity. *J. Chem. Inf. Model.*, 49(11):2429–43. 48

[143] Williams A. and Tkachenko V. (2014). The Royal Society of Chemistry and the delivery of chemistry data repositories for the community. *J. Comput. Aided Mol. Des.* 49

[144] Singh N., et al. (2009). Chemoinformatic analysis of combinatorial libraries, drugs, natural products, and molecular libraries small molecule repository. *J. Chem. Inf. Model.*, 49(4):1010–24. 51, 53

[145] Li R., Stumpfe D., Vogt M., Geppert H., and Bajorath J. (2011). Development of a method to consistently quantify the structural distance between scaffolds and to assess scaffold hopping potential. *J. Chem. Inf. Model.*, 51(10):2507–14. 51

[146] Möcklinghoff S., et al. (2011). Design and evaluation of fragment-like estrogen receptor tetrahydroisoquinoline ligands from a scaffold-detection approach. *J. Med. Chem.*, 54(7):2005–11. 51

[147] Langdon S. R., Ertl P., and Brown N. (2010). Bioisosteric Replacement and Scaffold Hopping in Lead Generation and Optimization. *Mol. Inform.*, 29(5):366–385. 51

[148] Krier M., Bret G., and Rognan D. (2006). Assessing the scaffold diversity of screening libraries. *J. Chem. Inf. Model.*, 46(2):512–24. 51

[149] Oefner C., et al. (2009). Increased hydrophobic interactions of iclaprim with Staphylococcus aureus dihydrofolate reductase are responsible for the increase in affinity and antibacterial activity. *J. Antimicrob. Chemother.*, 63(4):687–98. 51, 52

[150] Langdon S. R., Brown N., and Blagg J. (2011). Scaffold diversity of exemplified medicinal chemistry space. *J. Chem. Inf. Model.*, 51(9):2174–85. 51, 53

[151] Pipeline Pilot, Accelrys Software Inc. `http://accelrys.com/products/pipeline-pilot/`. Version: 8.5.0.200. 53, 67, 69

[152] Gaulton A., et al. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, 40(Database issue):D1100–7. 53

[153] Oates P. J. (2008). Aldose reductase, still a compelling target for diabetic neuropathy. *Curr. Drug Targets*, 9(1):14–36. 58

[154] Mylari B. L., et al. (1992). Potent, orally active aldose reductase inhibitors related to zopolrestat: surrogates for benzothiazole side chain. *J. Med. Chem.*, 35(3):457–65. 58

[155] Borensztajn K. and Spek C. A. (2011). Blood coagulation factor Xa as an emerging drug target. *Expert Opin. Ther. Targets*, 15(3):341–9. 59

[156] Guilford W. J., et al. (1999). Synthesis, Characterization, and Structure-Activity Relationships of Amidine-Substituted (Bis)benzylidene-Cycloketone Olefin Isomers as Potent and Selective Factor Xa Inhibitors 1,2. *J. Med. Chem.*, 42(26):5415–5425. 59

[157] Grossmann K. (2010). Auxin herbicides: current status of mechanism and mode of action. *Pest Manag. Sci.*, 66(2):113–20. 61

[158] Ray S. S., Nowak R. J., Brown R. H., and Lansbury P. T. (2005). Small-molecule-mediated stabilization of familial amyotrophic lateral sclerosis-linked superoxide dismutase mutants against unfolding and aggregation. *Proc. Natl. Acad. Sci. U. S. A.*, 102(10):3639–44. 62

[159] Kiernan M. C., et al. (2011). Amyotrophic lateral sclerosis. *Lancet*, 377(9769):942–955. 62

[160] Wheeler R. A., et al. (2008). PubChem: Integrated Platform of Small Molecules and Biological Activities. *Annual Reports in Computational Chemistry*, 4:217–241. 63

[161] Jiang H., et al. (2009). Stabilizers of the Max homodimer identified in virtual ligand screening inhibit Myc function. *Mol. Pharmacol.*, 76(3):491–502. 63

[162] Berg T. (2011). Small-molecule modulators of c-Myc/Max and Max/Max interactions. *Curr. Top. Microbiol. Immunol.*, 348:139–49. 63

[163] Anders C., et al. (2013). A Semisynthetic Fusicoccane Stabilizes a Protein-Protein Interaction and Enhances the Expression of K+ Channels at the Cell Surface. *Chem. Biol.*, 20(4):583–593. 64, 99

[164] Richter A., Rose R., Hedberg C., Waldmann H., and Ottmann C. (2012). An optimised small-molecule stabiliser of the 14-3-3-PMA2 protein-protein interaction. *Chem.–Eur. J.*, 18(21):6520–7. 64

[165] Tan X., et al. (2007). Mechanism of auxin perception by the TIR1 ubiquitin ligase. *Nature*, 446(7136):640–5. 64

[166] Sheard L. B., et al. (2010). Jasmonate perception by inositol-phosphate-potentiated COI1-JAZ co-receptor. *Nature*, 468(7322):400–5. 66

[167] Ghose A. K. and Crippen G. M. (1986). Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships I. Partition Coefficients as a Measure of Hydrophobicity. *J. Comput. Chem.*, 7(4):565–577. 67

[168] Ghose A. K. and Crippen G. M. (1987). Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions. *J. Chem. Inf. Model.*, 27(1):21–35.

[169] Ghose A. K., Pritchett A., and Crippen G. M. (1988). Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships III: Modeling hydrophobic interactions. *J. Comput. Chem.*, 9(1):80–90. 67

[170] Leach A. R. *Molecular Modelling: Principles and Applications*. Prentice Hall, 2nd edition (2001). 67

[171] Glide: A complete solution for ligand-receptor docking. `http://www.schrodinger.com/productpage/14/5/`. Version: 5.8.518. 70, 71, 74

[172] Maestro: A powerful, all-purpose molecular modeling environment. `http://www.schrodinger.com/productpage/14/12/`. Version: 9.3.518. 70, 73

[173] LigPrep: Versatile generation of accurate 3D molecular models. `http://www.schrodinger.com/productpage/14/10/`. Version: 4.0.518. 71, 73

[174] Mu D., et al. (2003). Genomic amplification and oncogenic properties of the KCNK9 potassium channel gene. *Cancer Cell*, 3(3):297–302. 72

[175] Bittner S., Budde T., Wiendl H., and Meuth S. G. (2010). From the background to the spotlight: TASK channels in pathological conditions. *Brain Pathol.*, 20(6):999–1009. 72

[176] Nayal M. and Honig B. (2006). On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins*, 63(4):892–906. 72

[177] Coot: Tool for macromolecular model building. `http://www2.mrc-lmb.cam.ac.uk/personal/pemsley/coot/`. Version: 0.7, Revision: 4459. 72

[178] Emsley P., Lohkamp B., Scott W. G., and Cowtan K. (2010). Features and development of Coot. *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 66(Pt 4):486–501. 72, 76, 103

[179] Krissinel E. and Henrick K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 60(12):2256–68. 72

[180] Glide: A complete solution for ligand-receptor docking. `http://www.schrodinger.com/productpage/14/5/`. Version: 5.6.107. 73, 100

[181] Halgren T. A., et al. (2004). Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.*, 47(7):1750–9. 74, 100

[182] Shape-it, Silicos-it. `http://www.silicos-it.com/software/software.html`. Version: 1.0.1. 74

[183] Grant J. A., Gallardo M. A., and Pickup B. T. (1996). A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *J. Comput. Chem.*, 17(14):1653–1666. 74

[184] Du Y., Masters S. C., Khuri F. R., and Fu H. (2006). Monitoring 14-3-3 protein interactions with a homogeneous fluorescence polarization assay. *J. Biomol. Screen.*, 11(3):269–76. 74, 97, 101

[185] Schumacher B., Skwarczynska M., Rose R., and Ottmann C. (2010). Structure of a 14-3-3σ-YAP phosphopeptide complex at 1.15 A resolution. *Acta Crystallogr., Sect. F: Struct. Biol. Cryst. Commun.*, 66(Pt 9):978–84. 75, 101, 160

[186] Kabsch W. (2010). XDS. *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 66(Pt 2):125–32. 76, 102

[187] Winn M. D., et al. (2011). Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.*, 67(Pt 4):235–42. 76, 102

[188] McCoy A. J., et al. (2007). Phaser crystallographic software. *J. Appl. Crystallogr.*, 40(Pt 4):658–674. 76, 102

[189] Murshudov G. N., Vagin A. A., and Dodson E. J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 53(Pt 3):240–55. 76, 103

[190] Watson P. J., Fairall L., Santos G. M., and Schwabe J. W. R. (2012). Structure of HDAC3 bound to co-repressor and inositol tetraphosphate. *Nature*, 481(7381):335–40. 79

[191] Paoletti P., Bellone C., and Zhou Q. (2013). NMDA receptor subunit diversity: impact on receptor properties, synaptic plasticity and disease. *Nat. Rev. Neurosci.*, 14(6):383–400. 80

[192] Gotti B., et al. (1988). Ifenprodil and SL 82.0715 as cerebral anti-ischemic agents. I. Evidence for efficacy in models of focal cerebral ischemia. *J. Pharmacol. Exp. Ther.*, 247(3):1211–21. 80

[193] Karakas E., Simorowski N., and Furukawa H. (2011). Subunit arrangement and phenylethanolamine binding in GluN1/GluN2B NMDA receptors. *Nature*, 475(7355):249–53. 80

[194] Goley E. D. and Welch M. D. (2006). The ARP2/3 complex: an actin nucleator comes of age. *Nat. Rev. Mol. Cell Biol.*, 7(10):713–26. 80

[195] Nolen B. J., et al. (2009). Characterization of two classes of small molecule inhibitors of Arp2/3 complex. *Nature*, 460(7258):1031–4. 80

[196] le Maire A., et al. (2010). A unique secondary-structure switch controls constitutive gene repression by retinoic acid receptor. *Nat. Struct. Mol. Biol.*, 17(7):801–7. 80

[197] Ranaivoson F. M., Gigant B., Berritt S., Joullié M., and Knossow M. (2012). Structural plasticity of tubulin assembly probed by vinca-domain ligands. *Acta Crystallogr. D Biol. Crystallogr.*, 68(Pt 8):927–34. 80

[198] Zhang Y., et al. (2004). Structural and virological studies of the stages of virus replication that are affected by antirhinovirus compounds. *J. Virol.*, 78(20):11061–9. 80

[199] Wang R., Lu Y., and Wang S. (2003). Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.*, 46(12):2287–303. 83

[200] Perola E., Walters W. P., and Charifson P. S. (2004). A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins*, 56(2):235–49.

[201] Repasky M. P., et al. (2012). Docking performance of the glide program as evaluated on the Astex and DUD datasets: a complete set of glide SP results and selected results for a new scoring function integrating WaterMap and glide. *J. Comput.-Aided Mol. Des.*, 26(6):787–99. 83

[202] Pan Y., Huang N., Cho S., and MacKerell A. D. (2002). Consideration of molecular weight during compound selection in virtual target-based database screening. *J. Chem. Inf. Comput. Sci.*, 43(1):267–72. 87

[203] Carta G., Knox A. J. S., and Lloyd D. G. (2007). Unbiasing scoring functions: a new normalization and rescoring strategy. *J. Chem. Inf. Model.*, 47(4):1564–71. 87

[204] Sackett D. L. and Sept D. (2009). Protein-protein interactions: making drug design second nature. *Nat. Chem.*, 1(8):596–7. 91

[205] Grüneberg S., Wendt B., and Klebe G. (2001). Subnanomolar Inhibitors from Computer Screening: A Model Study Using Human Carbonic Anhydrase II. *Angew. Chem., Int. Ed.*, 40(2):389–393. 92

[206] Doman T. N., et al. (2002). Molecular Docking and High-Throughput Screening for Novel Inhibitors of Protein Tyrosine Phosphatase-1B. *J. Med. Chem.*, 45(11):2213–2221.

[207] Varady J., et al. (2003). Molecular modeling of the three-dimensional structure of dopamine 3 (D3) subtype receptor: discovery of novel and potent D3 ligands through a hybrid pharmacophore- and structure-based database searching approach. *J. Med. Chem.*, 46(21):4377–92. 92

[208] Röglin L., Thiel P., Kohlbacher O., and Ottmann C. (2012). Covalent attachment of pyridoxal-phosphate derivatives to 14-3-3 proteins. *Proc. Natl. Acad. Sci. U. S. A.*, 109(18):E1051–1053. 95, 112, 113, 116, 142, 148

[209] Thiel P., et al. (2013). Virtual screening and experimental validation reveal novel small-molecule inhibitors of 14-3-3 protein-protein interactions. *Chem. Commun.*, 49(76):8468–70. 95, 142

[210] Steinacker P., Aitken A., and Otto M. (2011). 14-3-3 proteins in neurodegeneration. *Semin. Cell Dev. Biol.*, 22(7):696–704. 95

[211] Sluchanko N. N. and Gusev N. B. (2011). Probable participation of 14-3-3 in tau protein oligomerization and aggregation. *J. Alzheimers Dis.*, 27(3):467–76. 95

[212] Ghahary A., et al. (2005). Differentiated keratinocyte-releasable stratifin (14-3-3 sigma) stimulates MMP-1 expression in dermal fibroblasts. *J. Invest. Dermatol.*, 124(1):170–7. 95

[213] Medina A., Ghaffari A., Kilani R. T., and Ghahary A. (2007). The role of stratifin in fibroblast-keratinocyte interaction. *Mol. Cell. Biochem.*, 305(1-2):255–64. 96

[214] Visse R. and Nagase H. (2003). Matrix metalloproteinases and tissue inhibitors of metalloproteinases: structure, function, and biochemistry. *Circ. Res.*, 92(8):827–39. 96

[215] Ghaffari A., Li Y., Kilani R. T., and Ghahary A. (2010). 14-3-3 sigma associates with cell surface aminopeptidase N in the regulation of matrix metalloproteinase-1. *J. Cell Sci.*, 123(Pt 17):2996–3005. 96

[216] Mina-Osorio P. (2008). The moonlighting enzyme CD13: old and new functions to target. *Trends Mol. Med.*, 14(8):361–71. 96

[217] Wang B., et al. (1999). Isolation of high-affinity peptide antagonists of 14-3-3 proteins by phage display. *Biochemistry (Mosc.)*, 38(38):12499–504. 96

[218] Masters S. C., et al. (2002). Survival-promoting functions of 14-3-3 proteins. *Biochem. Soc. Trans.*, 30(4):360–5. 96

[219] Wu H., Ge J., and Yao S. Q. (2010). Microarray-assisted high-throughput identification of a cell-permeable small-molecule binder of 14-3-3 proteins. *Angew. Chem. Int. Ed. Engl.*, 49(37):6528–32. 96

[220] Corradi V., et al. (2010). Identification of the first non-peptidic small molecule inhibitor of the c-Abl/14-3-3 protein-protein interactions able to drive sensitive and Imatinib-resistant leukemia cells to apoptosis. *Bioorg. Med. Chem. Lett.*, 20(20):6133–7. 97, 114

[221] Corradi V., et al. (2011). Computational techniques are valuable tools for the discovery of protein-protein interaction inhibitors: the 14-3-3σ case. *Bioorg. Med. Chem. Lett.*, 21(22):6867–71. 97

[222] Altschul S. F., Gish W., Miller W., Myers E. W., and Lipman D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–10. 98

[223] Molecular Operating Environment (MOE), Chemical Computing Group Inc. `http://www.chemcomp.com/`. Version: 2010.10. 98

[224] Maestro: A powerful, all-purpose molecular modeling environment. `http://www.schrodinger.com/productpage/14/12/`. Version: 9.1.107. 100

[225] LigPrep: Versatile generation of accurate 3D molecular models. `http://www.schrodinger.com/productpage/14/10/`. Version: 2.4.107. 100

[226] Friesner R. A., et al. (2006). Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem.*, 49(21):6177–96. 100

[227] Kansy M., Senner F., and Gubernator K. (1998). Physicochemical high throughput screening: parallel artificial membrane permeation assay in the description of passive absorption processes. *J. Med. Chem.*, 41(7):1007–10. 101

[228] Lovell S. C., et al. (2003). Structure validation by Calpha geometry: phi,psi and Cbeta deviation. *Proteins*, 50(3):437–50. 103, 147, 148

[229] Adams P. D., et al. (2010). PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 66(Pt 2):213–21. 103

[230] Crooks G. E., Hon G., Chandonia J.-M., and Brenner S. E. (2004). WebLogo: a sequence logo generator. *Genome Res.*, 14(6):1188–90. 104

[231] Asdaghi N., et al. (2012). Extracellular 14-3-3 from human lung epithelial cells enhances MMP-1 expression. *Mol. Cell. Biochem.*, 360(1-2):261–70. 110

[232] Eliot A. C. and Kirsch J. F. (2004). Pyridoxal phosphate enzymes: mechanistic, structural, and evolutionary considerations. *Annu. Rev. Biochem.*, 73:383–415. 113

[233] Bohney J. P., Fonda M. L., and Feldhoff R. C. (1992). Identification of Lys190 as the primary binding site for pyridoxal 5'-phosphate in human serum albumin. *FEBS Lett.*, 298(2-3):266–8. 113, 116

[234] Diederichs K. and Karplus P. A. (1997). Improved R-factors for diffraction data analysis in macromolecular crystallography. *Nat. Struct. Biol.*, 4(4):269–275. 161, 162, 163, 164, 165

Appendix A

Abbreviations

Amino acids in a general context are not abbreviated. If we refer to a particular amino acid within a peptide or protein, standard abbreviations in 3-letter code are used with the residue position as suffix. Thus, an arbitrary serine at position 256 is abbreviated by Ser256. Within amino acid sequences, standard abbreviations in 1-letter code are used.

2D	*Two-dimensional*
3D	*Three-dimensional*
A	
AANAT	*Arylalkylamine N-acetyltransferase*
AC	*Adenylyl cyclase*
APN	*Aminopeptidase N*
ARF1	*ADP ribosylation factor 1*
ARP	*Actin-related protein*
ASA	*Accesible surface area*
B	
BALL	*Biochemical algorithms library*
bII	*Block inverted index*
BSA	*Buried surface area*
C	
CADD	*Computer-aided drug design*
CC	*Connected component*
D	
DAPK2	*Death-associated protein kinase 2*
DHFR	*Dihydrofolate reductase*

DI	*Druggability index*
DMSO	*Dimethyl sulfoxide*

E

EC_{50}	*Effective concentration 50 %*
ECFP	*Extended connectivity fingerprints*
ESI-MS	*Electrospray ionization mass spectrometry*
EM	*Extracellular matrix*

F

F101	*Fobisin101*
FP	*Fluorescence polarization*

H

HDAC3	*Histone deacetylase 3*
HTS	*High-throughput screening*

I

IC_{50}	*Inhibitory concentration 50 %*
iiDS	*Inverted index data structure*
IP_4	*Inositol tetraphosphate*

M

Max	*Myc-associated factor X*
MD	*Molecular dynamics*
MOE	*Molecular Operating Environment*
MMP	*Matrix metalloproteinase*
MR	*Molecular replacement*
MS	*Mass spectrometry*
MSA	*Multiple sequence alignment*
MT	*Microtubules*
MW	*Molecular weight*

N

NMDA	*N-methyl-D-aspartate*
NMR	*Nuclear magnetic resonance*
NRCoRep1	*Nuclear receptor co-repressor 1*

P

PAMPA	*Parallel artificial membrane permeation assay*
PBFP	*Path-based fingerprints*

PDB	*Protein Data Bank*
PEG	*Polyethylene glycol*
PLP	*Pyridoxal phosphate*
PMA2	*Plasma membrane H^+-ATPase 2*
Population count	*Popcount*
PPI	*Protein-protein interaction*

R

RARα	*Retinoic acid receptor α*
RMSD	*Root-mean-square deviation*
RNN	*Reciprocal nearest neighbor*
Ro5	*Lipinski's Rule-of-five*

S

SAR	*Structure-activity relationship*
SASA	*Solvent-accessible surface area*
SES	*Solvent-excluded surface*
SMILES	*Simplified Molecular Input Line Entry System*
SOD1	*Superoxide dismutase 1*

T

Task3	*TWIK-related acid-sensitive potassium (K^+) channel 3*
Tps	*Tanimoto calculations per second*

V

VS	*Virtual Screening*

Y

YAP	*Yes-associated protein*

Appendix B

Contributions

All ideas, approaches and results presented in this work were developed and discussed with my supervisors Prof. Dr. Oliver Kohlbacher (OK) and Prof. Dr. Christian Ottmann (CO). The following co-workers also contributed to the different projects:

- Maria Bartel (MB)
- Dr. Sven Hennig (SH)
- Nicole Meißner (NM)
- Dr. Manuela Molzan (MM)
- Dr. Lars Röglin (LR)
- Dr. Lisa Sach-Peltason (LSP)
- Svenja Schäfers (SS)
- Dr. Benjamin Schumacher (BS)
- Dr. Malgorzata Skwarczynska (MS)

Chapter 3: Deterministic Clustering of Large Chemical Spaces

The project was designed by myself, OK, and CO. Additionally, LSP contributed to this project in the context of her *Studienarbeit*.[134] The inverted index method was reimplemented and parallelized by myself. All other C++ code was newly developed by myself and embedded into the software framework BALL.[2] Computational experiments were performed by myself. Data was analyzed and interpreted by myself and OK. The manuscript of the published article that arose from this work was written by myself.[125]

Chapter 4: Scaffold Families

The project was designed by myself, OK, and CO. Computational experiments were performed by myself. Data was analyzed and interpreted by myself and OK.

Chapter 5: In Silico Analysis of Protein-Protein Interaction Stabilization

The project was designed by myself, OK, and CO. Computational experiments and protein crystallography were performed by myself. Protein for *in vitro* experiments was kindly provided by MM, BS, and MS. The *in vitro* assays were performed by MB. Results and data were analyzed and interpreted by myself, OK, and CO.

Chapter 6: Virtual Screening for 14-3-3 Protein-Protein Interaction Inhibitors

The project was designed by myself, LR, SH, OK, and CO. Computational experiments and protein crystallography were performed by myself. Protein for *in vitro* experiments was kindly provided by MM, BS, and MS. The *in vitro* assays were performed by LR. He also contributed the ESI-MS experiments. Cellular assays were performed by NM, SS, and SH. PAMPA measurements were provided by the Lead Discovery Center Dortmund. Results and data were analyzed and interpreted by myself, LR, SH, OK, and CO. Published manuscripts were written by myself and LR.[208,209]

Appendix C

Publications

Accepted manuscripts

Hildebrandt A.K., Stöckel D., Fischer N.M., de la Garza Trevino L., Krüger J., Nickels S., Röttig M., Schärfe C., Schumann M., **Thiel P.**, Lenhof H.-P., Kohlbacher O., Hildebrandt A. "ballaxy: web services for structural bioinformatics." *Bioinformatics* DOI: 10.1093/bioinformatics/btu574

2014

Thiel P., Sach-Peltason L., Ottmann C., and Kohlbacher O. "Blocked Inverted Indices for Exact Clustering of Large Chemical Spaces." *J. Chem. Inf. Model.* **54**, 2395-401.

2013

Thiel P., Röglin L., Meissner N., Hennig S., Kohlbacher O. and Ottmann C. (2013) "Virtual Screening and Experimental Validation Reveal Novel Small-Molecule Inhibitors of 14-3-3 protein-protein interactions." *Chem. Commun.* **49**, 8468-70.

Anders C., Higuchi Y., Koschinsky K., Bartel M., Schumacher B., **Thiel P.**, Nitta H., Preisig-Müller R., Schlichthörl G., Renigunta V., Ohkanda J., Daut J., Kato N. and Ottmann C. (2013) "A Semisynthetic Fusicoccane Stabilizes a Protein-Protein Interaction and Enhances the Expression of K^+ Channels at the Cell Surface." *Chem. Biol.* **109**, 583-593.

Thiel P., Peltason L., Ottmann C., and Kohlbacher O. (2013) "Deterministic clustering of the available chemical space." *J. Cheminf.* **5** (Suppl. 1), 53.

2012

Röglin L., **Thiel P.**, Kohlbacher O. and Ottmann C. (2012) "Covalent attachment of pyridoxal-phosphate derivatives to 14-3-3 proteins." *Proc. Natl. Acad. Sci. USA* **109**, E1051-3.

Thiel P., Kaiser M. and Ottmann C. (2012) "Small-Molecule Stabilization of Protein-Protein Interactions: An Underestimated Concept in Drug Discovery?" *Angew. Chem. Int. Ed.* **51**, 2012-8.

2010

Molzan M., Schumacher B., Ottmann C., Baljuls A., Polzien L., Weyand M., **Thiel P.**, Rose R., Rose M., Kuhenne P., Kaiser M., Rapp U. R., Kuhlmann J. and Ottmann C. (2010) "Impaired binding of 14-3-3 to C-RAF in Noonan syndrome suggests new approaches in diseases with increased Ras signaling." *Mol. Cell. Biol.* **30**, 4698-711.

Schumacher B., Mondry J., **Thiel P.**, Weyand M. and Ottmann C. (2010) "Structure of the p53 C-terminus bound to 14-3-3: implications for stabilization of the p53 tetramer." *FEBS Lett.* **584**, 1443-8.

2009

Feldhahn M., Dönnes P., **Thiel P.** and Kohlbacher O. (2009) "FRED — a framework for T-cell epitope detection." *Bioinformatics* **25**, 2758-9.

2008

Feldhahn M.*, **Thiel P.***, Schuler M. M., Hillen N., Stevanovic S., Rammensee H. G. and Kohlbacher O. (2008) "EpiToolKit — a web server for computational immunomics." *Nucleic Acids Res.* **36**, W519-22. (* contributed equally)

Appendix D

Supporting Figures

	β	ε	η	γ	σ	τ	ζ	PDB ID
14-3-3 β	0.00	0.79	0.57	0.55	0.88	0.73	0.57	2bq0
14-3-3 ε	0.79	0.00	0.89	0.92	1.13	0.72	0.73	2br9
14-3-3 η	0.57	0.89	0.00	0.39	0.70	0.72	0.52	2c63
14-3-3 γ	0.55	0.92	0.39	0.00	0.75	0.79	0.58	3uzd
14-3-3 σ	0.88	1.13	0.70	0.75	0.00	0.78	0.77	3lw1
14-3-3 τ	0.73	0.72	0.72	0.79	0.78	0.00	0.65	2btp
14-3-3 ζ	0.57	0.73	0.52	0.58	0.77	0.65	0.00	4fj3

Figure D.1: Pairwise RMSD matrix of all human 14-3-3 homologs. Structure and sequence alignment was performed using Protein Superpose from MOE (version 2012.10). The RMSD values were calculated from superposed C_α backbones of 14-3-3 monomers. N-terminal tag residues and C-terminal overhangs were trimmed after sequence alignment. The following optional settings were enabled: Optimize Gap penalties for Superposition, Accent Secondary Structure Matches. The averaged RMSD value is 0.74 Å.

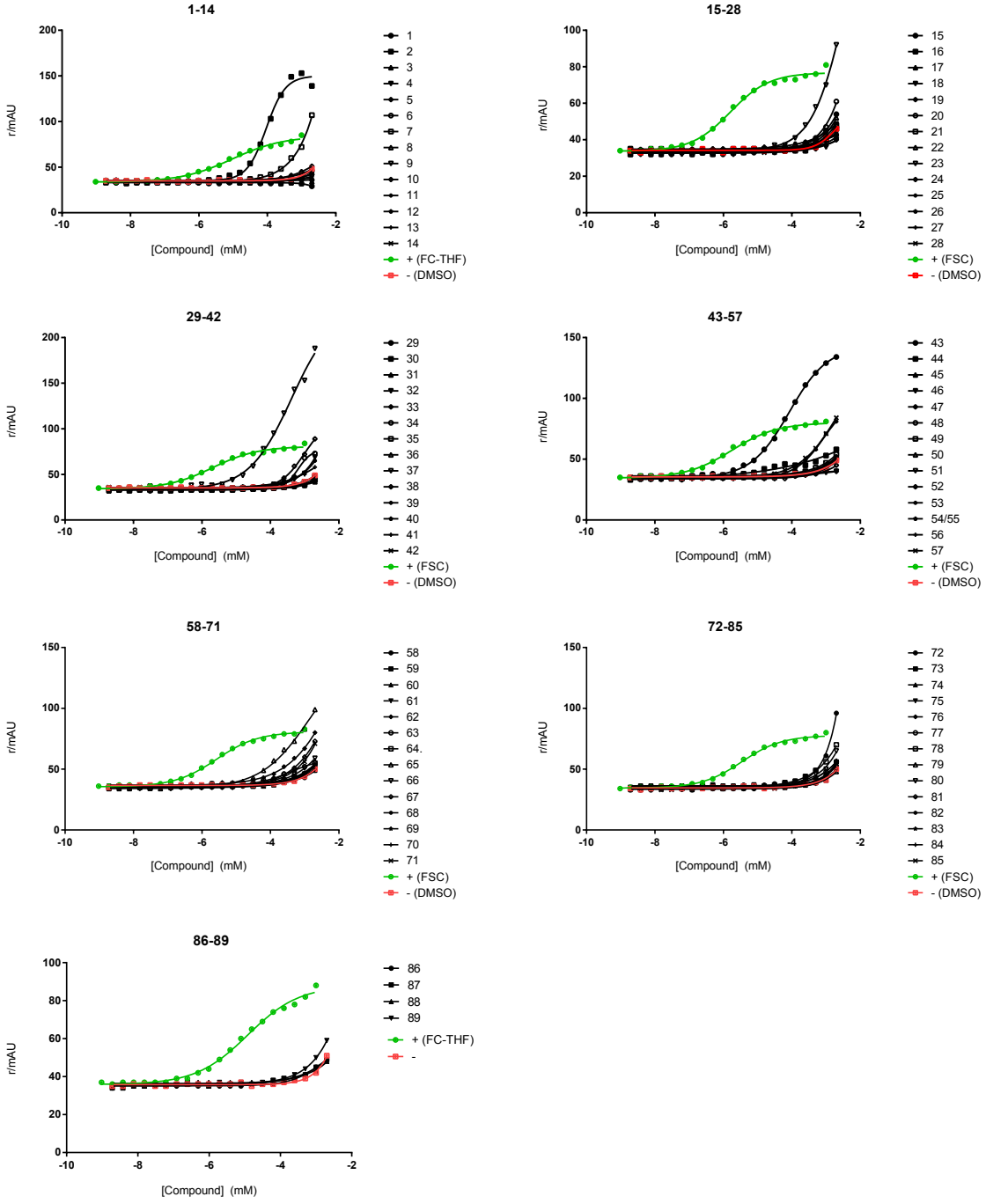


Figure D.2: FP-based assay for *in vitro* validation of selected stabilizer candidates for the 14-3-3σ∘Task3 interaction. The experiments were performed by Maria Bartel from TU Eindhoven.

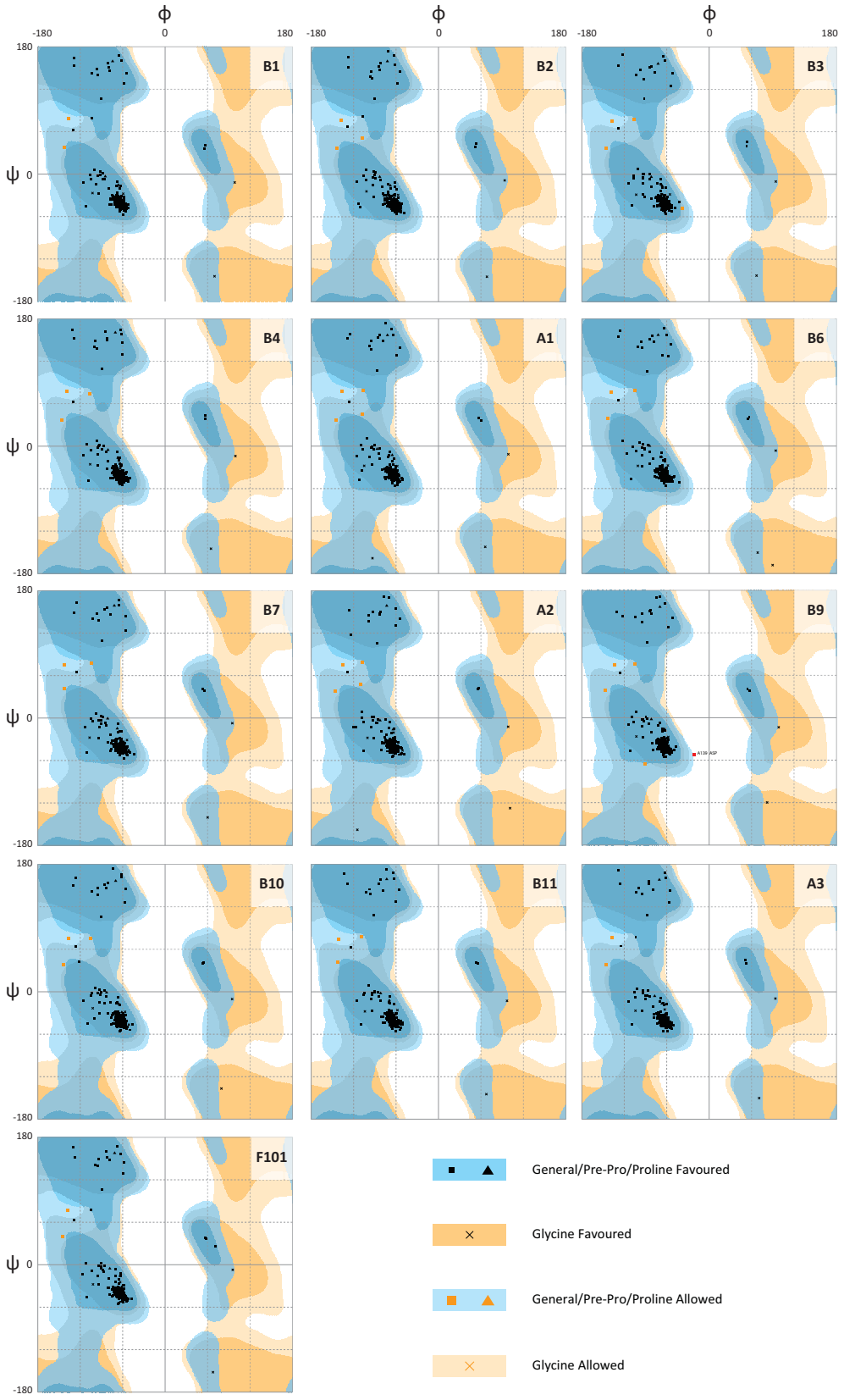


Figure D.3: Ramachandran plots of 14-3-3∘inhibitor complexes generated using RAMPAGE.[228]

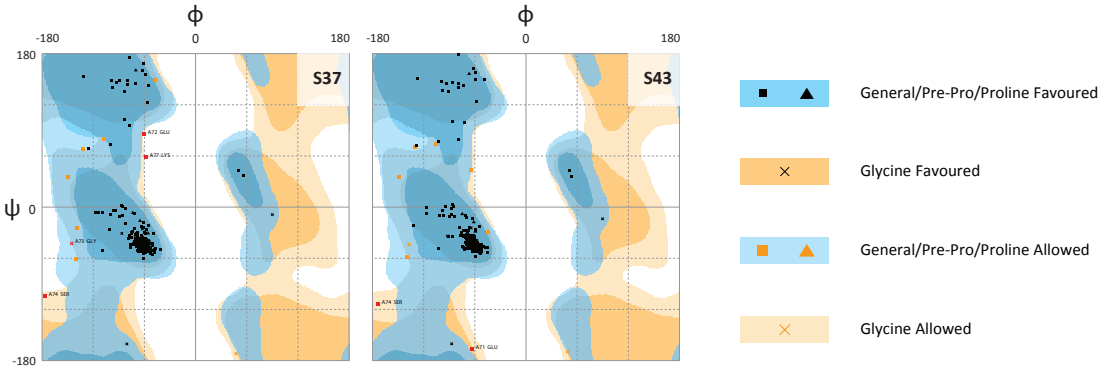


Figure D.4: Ramachandran plots of 14-3-3∘Task3 complexes soaked with stabilizer candidates S37 and S43 from VS. Plots were generated using RAMPAGE.[228]

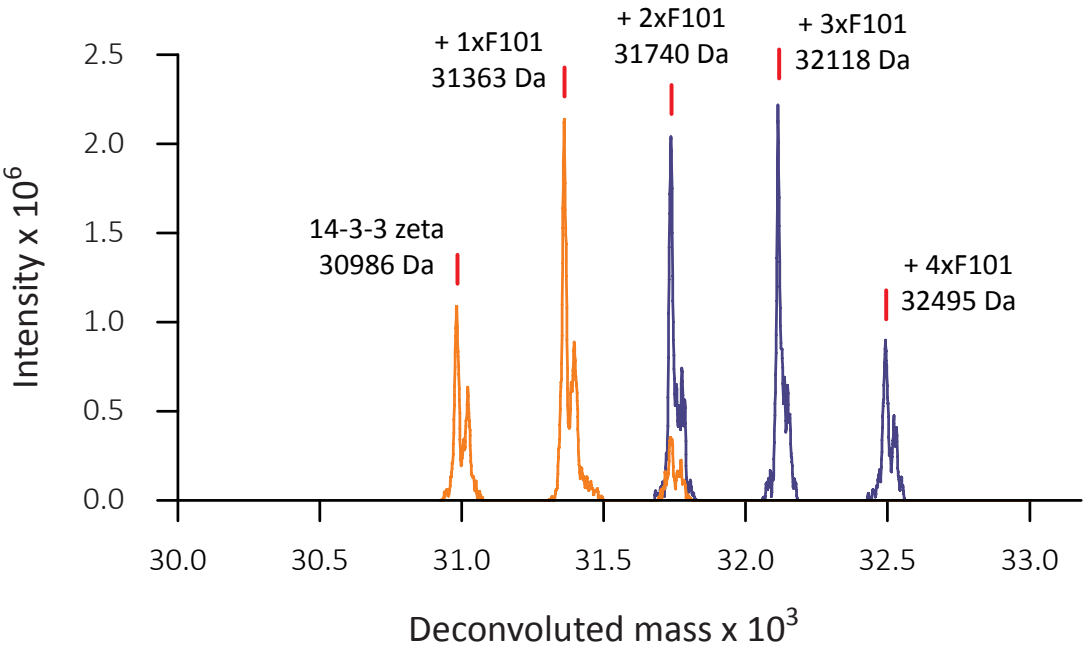


Figure D.5: Deconvoluted mass spectrum of 14-3-3ζ after treatment with compound F101 (Dr. Lars Röglin, CGC Dortmund). A solution of 1 mM 14-3-3ζ was mixed with F101 (final concentration 5 mM), incubated overnight in the presence of 25 mM tris(2-carboxyethyl)phosphine (TCEP) and analyzed using ESI-MS after 1:50 dilution with water (blue peaks). One aliquot was diluted before overnight incubation and ESI-MS analysis (orange peaks). The resulting mass spectra were deconvoluted using Mag-Tran. Red marks show the calculated masses for 14-3-3ζ with 0-4 additional F101 moieties covalently linked to 14-3-3 via imine formation. Figure is adapted from Röglin *et al.* (*Proc. Natl. Acad. Sci. USA* (2012), **109**, E1051-3).[208]

Appendix E

Supporting Tables

Table E.1: Composition of immediately available compound library sorted by compounds per supplier.

Supplier Name	*Compounds*
Enamine	1,622,458
Uorsy	1,295,301
ChemDiv	1,208,146
Vitas-M Laboratory	846,810
ChemBridge	568,409
Asinex	307,164
TimTec	213,395
Princeton BioMolecular Research	198,537
Life Chemicals	126,358
InterBioScreen	108,216
Total	6,494,794

Table E.2: Complex structures of 14-3-3σ and mode III phosphopeptides, which were analyzed in course of the stabilizer VS in Section 5.2.5. Structures obtained from the PDB were accessed 6/3/2013. In-house structures which have not yet been deposited (n.d.) with the PDB were also analyzed. The ligand IDs are the identifiers from the Chemical Components Dictionary of the PDB.

PDB ID	*Resolution* (Å)	*Ligand ID*	*Construct*	*Peptide*	C-terminus
3p1s	1.65	FSC	C138V, N166H	Task3	Val (+1)
3smm	2.00	FJA	C138V, N166H	Task3	Val (+1)
3smo	1.80	FJA	C138V, N166H	Task3	Val (+1)
3spr	1.99	FC7	C138V, N166H	Task3	Val (+1)
3ux0	1.75	0DV	wild-type	Task3	Val (+1)
n.d.	1.65	FSC	wild-type	Task3	Val (+1)
n.d.	1.60	FSC	wild-type	HAP1a	Ile (+1)
n.d.	1.48	FSC	wild-type	DAPK2	Ser (+1)
3iqv	1.20	FSC	wild-type	C-Raf 6-mer	Thr (+1)

Table E.3: Non-mammalian complex structures of 14-3-3 and mode III phosphopeptides, which were analyzed in course of the stabilizer VS in Section 5.2.5. Structures obtained from the PDB were accessed 6/3/2013. The ligand IDs are the identifiers from the Chemical Components Dictionary of the PDB.

PDB ID	*Resolution*	*Ligand ID*	*Construct*	*Peptide*	C-terminus
1o9f	2.70 Å	FSC	*N. tabacum* 14-3-3C	PMA2 5-mer	Val (+1)
2o98	2.70 Å	FSC	*N. tabacum* 14-3-3C	PMA2 35-mer	Ile (+1)

Table E.4: Selected compounds from stabilizer VS described in Chapter 5.

ID	*MolPort ID*	*Supplier*	*Catalog No.*
S1	MolPort-002-545-968	Vitas-M Laboratory	STK700033
S2	MolPort-002-736-096	Vitas-M Laboratory	STK686363
S3	MolPort-009-115-861	ENAMINE Ltd.	Z968784634
S4	MolPort-000-035-955	Asinex	ASN 04060288
S5	MolPort-000-757-021	Vitas-M Laboratory	STK662812
S6	MolPort-000-852-152	Vitas-M Laboratory	STK622053
S7	MolPort-000-119-910	Asinex	ASN 11172668
S8	MolPort-003-006-350	Alinda Chemical	IBS-0010522
S9	MolPort-002-578-667	InterBioScreen Ltd.	STOCK3S-13326
S10	MolPort-020-083-188	ENAMINE Ltd.	Z1136425863
S11	MolPort-008-346-818	InterBioScreen Ltd.	STOCK1N-23077
S12	MolPort-002-580-849	InterBioScreen Ltd.	STOCK3S-23906
S13	MolPort-001-537-445	Vitas-M Laboratory	STK129741
S14	MolPort-002-593-689	InterBioScreen Ltd.	STOCK3S-87770
S15	MolPort-009-363-011	ENAMINE Ltd.	Z88619050
S16	MolPort-000-758-979	Vitas-M Laboratory	STL055571
S17	MolPort-000-006-616	Specs	AG-690/10758045
S18	MolPort-000-068-565	Asinex	ASN 05338572
S19	MolPort-000-068-574	Asinex	ASN 05338581
S20	MolPort-000-728-502	Vitas-M Laboratory	STL147606
S21	MolPort-019-745-957	Specs	AG-690/34036035
S22	MolPort-006-395-331	Vitas-M Laboratory	STL146393
S23	MolPort-009-758-814	InterBioScreen Ltd.	STOCK1N-76189
S24	MolPort-000-484-898	ChemDiv, Inc.	D074-0284
S25	MolPort-000-748-046	Vitas-M Laboratory	STK069487
S26	MolPort-002-322-601	Vitas-M Laboratory	STK370217
S27	MolPort-000-409-821	Vitas-M Laboratory	STK789593
S28	MolPort-001-975-340	Specs	AO-082/13829007
S29	MolPort-002-697-608	Vitas-M Laboratory	STK672779
S30	MolPort-002-722-969	Vitas-M Laboratory	STK773272

Table E.4: Continued.

ID	*MolPort ID*	*Supplier*	*Catalog No.*
S31	MolPort-006-811-597	Vitas-M Laboratory	STK648895
S32	MolPort-009-421-945	ENAMINE Ltd.	Z196494254
S33	MolPort-002-972-593	Vitas-M Laboratory	STK199399
S34	MolPort-004-195-644	ENAMINE Ltd.	Z31373283
S35	MolPort-005-658-213	ENAMINE Ltd.	Z168814872
S36	MolPort-009-357-856	ENAMINE Ltd.	Z992226530
S37	MolPort-007-567-493	ChemDiv, Inc.	5275-0079
S38	MolPort-005-911-044	InterBioScreen Ltd.	STOCK1N-72125
S39	MolPort-007-640-377	ChemDiv, Inc.	C530-0369
S40	MolPort-003-116-786	Life Chemicals Inc.	F2325-0245
S41	MolPort-005-650-834	ENAMINE Ltd.	Z46493886
S42	MolPort-023-141-229	ENAMINE Ltd.	Z1082917684
S43	MolPort-001-953-171	Vitas-M Laboratory	STK885541
S44	MolPort-009-294-226	ENAMINE Ltd.	Z368533022
S45	MolPort-009-284-060	ENAMINE Ltd.	Z356761160
S46	MolPort-009-266-368	ENAMINE Ltd.	Z324816790
S47	MolPort-000-051-348	Asinex	ASN 04642642
S48	MolPort-003-881-792	Vitas-M Laboratory	STK367994
S49	MolPort-008-276-745	Asinex	SYN 15028367
S50	MolPort-016-642-207	Asinex	ADD 13552951
S51	MolPort-016-642-832	Asinex	ADD 14244518
S52	MolPort-016-676-781	Asinex	AOP 22040411
S53	MolPort-016-680-954	Asinex	ART 22837893
S54	MolPort-016-712-663	Asinex	LEG 16295076
S55	MolPort-016-717-984	Asinex	SYN 22855182
S56	MolPort-016-720-760	Asinex	SYN 22962998
S57	MolPort-016-721-543	Asinex	SYN 22982454
S58	n.a.	ENAMINE Ltd.	Z26388880
S59	n.a.	ENAMINE Ltd.	Z1247351969
S60	n.a.	ENAMINE Ltd.	Z1191626709

Table E.4: Continued.

ID	*MolPort ID*	*Supplier*	*Catalog No.*
S61	n.a.	ENAMINE Ltd.	Z195915686
S62	n.a.	ENAMINE Ltd.	Z32662649
S63	n.a.	ENAMINE Ltd.	Z56909743
S64	n.a.	ENAMINE Ltd.	Z1222029545
S65	MolPort-004-985-406	ChemBridge Corp.	43480567
S66	MolPort-016-587-281	ChemBridge Corp.	7776956
S67	MolPort-016-628-317	ChemBridge Corp.	94488956
S68	MolPort-005-078-830	ChemBridge Corp.	43408944
S69	MolPort-005-031-693	ChemBridge Corp.	21944477
S70	MolPort-020-225-776	ChemBridge Corp.	84937785
S71	MolPort-021-767-739	ChemBridge Corp.	73024304
S72	MolPort-005-089-370	ChemBridge Corp.	49023446
S73	MolPort-016-618-320	ChemBridge Corp.	71629761
S74	MolPort-019-801-389	ChemBridge Corp.	18869608
S75	MolPort-008-364-330	ChemBridge Corp.	28313054
S76	MolPort-020-215-374	ChemBridge Corp.	65587844
S77	MolPort-019-822-802	ChemBridge Corp.	90809011
S78	MolPort-020-227-977	ChemBridge Corp.	89013059
S79	MolPort-002-152-386	ChemBridge Corp.	5476872
S80	MolPort-019-901-874	ChemBridge Corp.	95035132
S81	MolPort-005-020-915	ChemBridge Corp.	17996934
S82	MolPort-019-892-017	ChemBridge Corp.	56037516
S83	MolPort-005-021-540	ChemBridge Corp.	18217624
S84	MolPort-020-204-818	ChemBridge Corp.	46200622
S85	MolPort-005-132-482	ChemBridge Corp.	75500017
S86	MolPort-021-751-748	ChemBridge Corp.	26210901
S87	MolPort-020-214-644	ChemBridge Corp.	64233937
S88	MolPort-016-618-878	ChemBridge Corp.	72905872
S89	MolPort-005-049-413	ChemBridge Corp.	29279862

Table E.5: Compounds from 14-3-3 inhibitor development described in Chapter 6 with supplier information (IBS: InterBioScreen, S.-Aldr.: Sigma-Aldrich). In addition to the compounds' supplier IDs an internal identifier is provided (ID), which is used in this thesis. The IC_{50} values were determined using an FP-based competition assay by Dr. Lars Röglin ('n.a.': > 500 µM, '-': not measured). The SAR column indicates the membership of the compound to one out of 11 different SAR groups. The SAR groups are discussed in Section 6.3.3. The table is divided in SAR groups. The first block contains all active inhibitors sorted by increasing IC_{50} values.

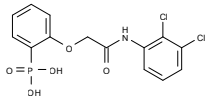
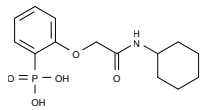
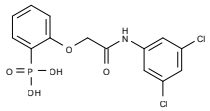
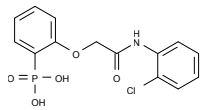
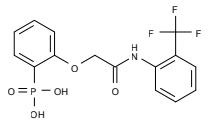
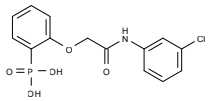
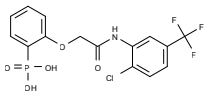
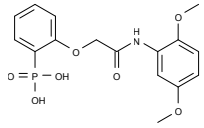
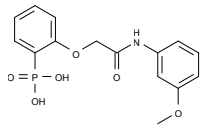
Molecule	*ID*	*Rank*	*XP Score*	*Supplier (ID)*	*IC_{50}* (µM)	SAR
	B1	n.a.	n.a.	IBS (6S-30541)	5	8
	B2	n.a.	n.a.	IBS (6S-21429)	15	5
	B3	n.a.	n.a.	IBS (6S-31619)	16	11
	B4	n.a.	n.a.	IBS (6S-31676)	27	6
	A1	44	-8.64	IBS (6S-23126)	30	6
	B5	n.a.	n.a.	IBS (6S-30099)	32	7
	B6	n.a.	n.a.	IBS (6S-26034)	36	10
	A2	85	-8.29	IBS (6S-23618)	116	10
	B7	n.a.	n.a.	IBS (6S-35412)	118	7

Table E.5: Continued.

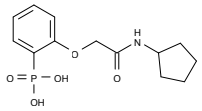
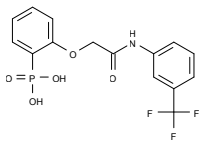
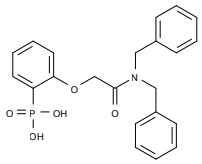
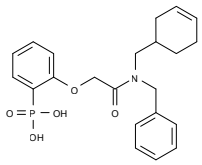
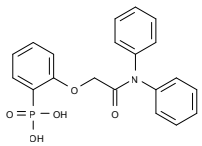
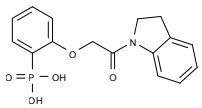
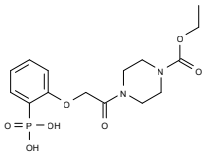
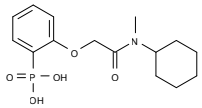
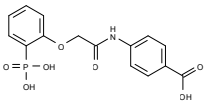
Molecule	*ID*	*Rank*	*XP Score*	*Supplier* (*ID*)	IC_{50} (µM)	SAR
	B8	n.a.	n.a.	IBS (6S-26169)	128	5
	B9	n.a.	n.a.	IBS (6S-25642)	165	7
	A4	83	-8.31	IBS (6S-19901)	n.a.	1
	A5	123	-8.05	IBS (6S-22115)	n.a.	1
	A6	64	-8.43	IBS (6S-26036)	n.a.	1
	A7	69	-8.41	IBS (6S-39852)	n.a.	1
	A8	165	-7.82	IBS (6S-39246)	n.a.	1
	A9	57	-8.52	IBS (6S-30315)	n.a.	1
	A10	27	-8.83	IBS (6S-27243)	n.a.	2

Table E.5: Continued.

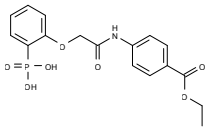
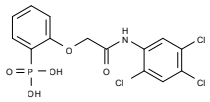
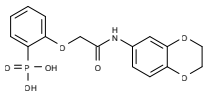
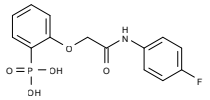
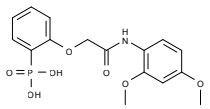
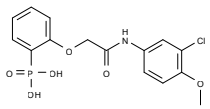
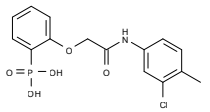
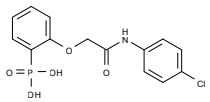
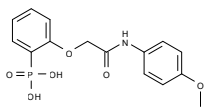
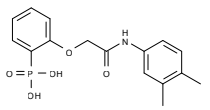
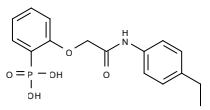
Molecule	*ID*	*Rank*	*XP Score*	*Supplier* (*ID*)	IC_{50} (μM)	SAR
	A11	39	-8.67	IBS (6S-20951)	n.a.	2
	A12	30	-8.81	IBS (6S-30980)	n.a.	2
	A14	38	-8.68	IBS (6S-35347)	n.a.	2
	B11	n.a.	n.a.	IBS (6S-23301)	n.a.	2
	B12	n.a.	n.a.	IBS (6S-25681)	n.a.	2
	B14	n.a.	n.a.	IBS (6S-33871)	n.a.	2
	B15	n.a.	n.a.	IBS (6S-39499)	n.a.	2
	B17	n.a.	n.a.	IBS (6S-23881)	n.a.	2
	B19	n.a.	n.a.	IBS (6S-26928)	n.a.	2
	B21	n.a.	n.a.	IBS (6S-29765)	n.a.	2
	B22	n.a.	n.a.	IBS (6S-29992)	n.a.	2

Table E.5: Continued.

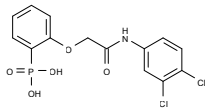
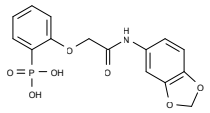
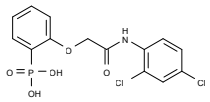
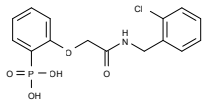
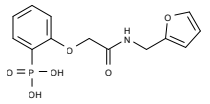
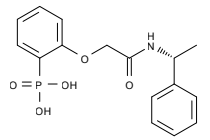
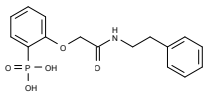
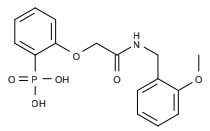
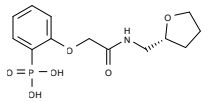
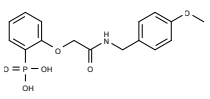
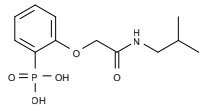
Molecule	*ID*	*Rank*	*XP Score*	*Supplier* (*ID*)	IC_{50} (µM)	SAR
	B23	n.a.	n.a.	IBS (6S-30569)	n.a.	2
	B28	n.a.	n.a.	IBS (6S-39700)	n.a.	2
	B30	n.a.	n.a.	IBS (6S-31316)	n.a.	2
	B10	n.a.	n.a.	IBS (6S-30286)	n.a.	3
	B16	n.a.	n.a.	IBS (6S-23822)	n.a.	3
	B18	n.a.	n.a.	IBS (6S-24610)	n.a.	3
	B20	n.a.	n.a.	IBS (6S-28426)	n.a.	3
	B24	n.a.	n.a.	IBS (6S-32535)	n.a.	3
	B25	n.a.	n.a.	IBS (6S-34461)	n.a.	3
	B31	n.a.	n.a.	IBS (6S-37412)	n.a.	3
	B13	n.a.	n.a.	IBS (6S-37402)	n.a.	4

Table E.5: Continued.

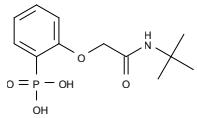
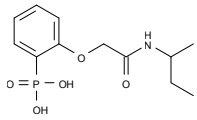
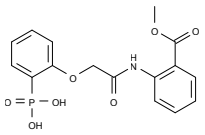
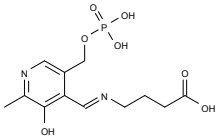
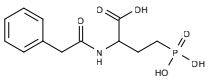
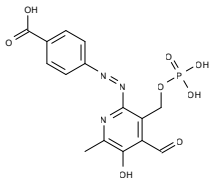
Molecule	*ID*	*Rank*	*XP Score*	*Supplier* (*ID*)	IC_{50} (μM)	SAR
	B27	n.a.	n.a.	IBS (6S-37663)	n.a.	4
	B29	n.a.	n.a.	IBS (6S-43021)	n.a.	4
	B26	n.a.	n.a.	IBS (6S-36239)	n.a.	9
	A3	25	-8.85	IBS (1N-29704)	n.a.	12
	A13	142	-7.92	IBS (1N-27073)	n.a.	12
	F101	n.a.	n.a.	Sigma-Aldrich (MRS 2159)	-	-

Table E.6: Materials used for wet lab experiments.

	Product	*Supplier*
24-well plates	Linbro® Tissue Culture Plates	ICN Biomedicals
96-well plates	Greiner 609171, CrystalQuick™	Jena Bioscience
Pipettes	Eppendorf Research®	Eppendorf, Hamburg
Pipetting system	Liquidator96©	Steinbrenner Laborsysteme
Adhesive foil	HDClear™ Packaging Tape	ShurTech Brands
Imaging System	Rock Imager 1000	Formulatrix, Waltham

Table E.7: Complexation buffer to dilute 14-3-3σΔC○inhibitor stock solutions for crystallization.

	Unit	*Concentration*
HEPES	mM	20.0
$MgCl_2$	mM	2.0
2-Mercaptoethanol	mM	2.0
	pH	7.5

Table E.8: Crystallization condition for 14-3-3σ complexes with short phosphopeptides reported by Schumacher *et al.*.[185] 2D grid variation of this condition used for crystallization experiments of 14-3-3∘inhibitor complexes. The original condition corresponds to Qiagen® initial screening buffer no. 16 from the JCSG Core I Suite.

	95 mM Na-HEPES (pH)	*PEG 400* (% v/v)	*Glycerol* (% v/v)	$CaCl_2$ (mM)
Original	7.5	26.6	5.0	190
C1	7.1	23.0	5.0	190
C2	7.3	24.0	5.0	190
C3	7.5	25.0	5.0	190
C4	7.7	26.0	5.0	190
C5	7.1	27.0	5.0	190
C6	7.3	28.0	5.0	190
C7	7.5	23.0	5.0	190
C8	7.7	24.0	5.0	190
C9	7.1	25.0	5.0	190
C10	7.3	26.0	5.0	190
C11	7.5	27.0	5.0	190
C12	7.7	28.0	5.0	190
C13	7.1	23.0	5.0	190
C14	7.3	24.0	5.0	190
C15	7.5	25.0	5.0	190
C16	7.7	26.0	5.0	190
C17	7.1	27.0	5.0	190
C18	7.3	28.0	5.0	190
C19	7.5	23.0	5.0	190
C20	7.7	24.0	5.0	190
C21	7.1	25.0	5.0	190
C22	7.3	26.0	5.0	190
C23	7.5	27.0	5.0	190
C24	7.7	28.0	5.0	190

Table E.9: Data processing and refinement statistics. Values for outermost resolution shells in parentheses.

		Compound Information	
ID		S43	F101
PDB ID		not deposited	not deposited
		Crystal Parameters	
Dimensions	a,b,c (Å)	81.7, 111.6, 62.4	82.3, 112.3, 62.8
Angles	α,β,γ (°)	90.0, 90.0, 90.0	90.0, 90.0, 90.0
Space group		$C222_1$	$C222_1$
		Data Collection Statistics	
Beamline		Rigaku	Bruker
Wavelength (Å)		1.5418	1.5418
Resolution (Å)		29.75-1.85 (2.00-1.85)	19.60-1.65 (1.75-1.65)
Measured reflections		100449 (20245)	191999 (21100)
Unique reflections		24713 (5079)	34969 (5475)
Completeness		99.9 (100.0)	99.1 (97.6)
Redundancy		4.1 (4.0)	9.1 (6.4)
I/σ(I)		14.6 (4.1)	20.1 (4.6)
R_{meas} **(%)** [a]		7.5 (39.6)	6.3 (38.6)
		Refinement Statistics	
Resolution (Å)		29.76-1.85 (1.89-1.85)	19.60-1.65 (1.69-1.65)
Number of atoms		2171	2511
R_{work} **(%)**		21.7 (26.0)	14.4 (18.4)
R_{free} **(%)**		25.8 (37.6)	19.1 (25.4)
RMS bond lengths (Å) [b]		0.020	0.021
RMS bond angles (°) [b]		1.929	1.972
Averaged B-factors (Å^2)			
Total		21.2	19.6
Compound		n.a.	32.2
Ramachandran plot residues			
Favoured regions (%)		95.7	98.2
Allowed regions (%)		3.4	1.8
Disallowed regions (%)		0.9	0.0

[a] Redundancy independent R-factor (intensities).[234]
[b] RMSD from ideal geometry values.

Table E.9: Continued.

Compound Information				
ID		A1	A2	B1
PDB ID		3t0l	3t0m	4dhu
Crystal Parameters				
Dimensions	a,b,c (Å)	82.4, 112.8, 62.7	82.1, 112.3, 62.4	82.3, 112.5, 62,5
Angles	α,β,γ (°)	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0
Space group		$C222_1$	$C222_1$	$C222_1$
Data Collection Statistics				
Beamline		SLS	SLS, Rigaku	Rigaku
Wavelength (Å)		0.9778	0.9778, 1.5418	1.5418
Resolution (Å)		45.62-1.6 (1.70-1.60)	45.42-1.62 (1.70-1.62)	19.54-1.67 (1.75-1.67)
Measured reflections		338818 (54028)	245634 (21990)	262477 (21104)
Unique reflections		38584 (6363)	36467 (4759)	33635 (4129)
Completeness		99.2 (99.4)	98.6 (96.7)	98.8 (94.1)
Redundancy		8.8 (8.5)	6.7 (4.6)	7,8 (5.1)
I/σ(I)		23.9 (5.9)	13.9 (4.0)	45.0 (13.2)
R_{meas} **(%)** [a]		5.4 (43.0)	9.7 (43.7)	3.4 (13.3)
Refinement Statistics				
Resolution (Å)		45.62-1.60 (1.64-1.60)	45.42-1.62 (1.66-1.62)	19.55-1.67 (1.71-1.67)
Number of atoms		2460	2467	2525
R_{work} **(%)**		15.7 (19.2)	15.5 (19.5)	13.3 (16.8)
R_{free} **(%)**		18.9 (22.7)	19.8 (24.8)	17.6 (24.6)
RMS bond lengths (Å) [b]		0.021	0.026	0.019
RMS bond angles (°) [b]		1.997	2.223	1.718
Averaged B-factors (Å^2)				
Total		22.3	20.8	16.9
Compound		25.8	30.6	17.8
Ramachandran plot residues				
Favoured regions (%)		96.2	95.2	97.1
Add. allowed regions (%)		3.8	4.8	2.9
Gen. allowed regions (%)		0.0	0.0	0.0
Disallowed regions (%)		0.0	0.0	0.0

[a] Redundancy independent R-factor (intensities).[234]
[b] RMSD from ideal geometry values.

Table E.9: Continued.

Compound Information				
ID		B2	B3	B4
PDB ID		4dht	4dhs	4dhr
Crystal Parameters				
Dimensions	a,b,c (Å)	82.3, 112.3, 62.5	81.2, 112.0, 62.3	82.2, 112.3, 62.5
Angles	α,β,γ (°)	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0
Space group		$C222_1$	$C222_1$	$C222_1$
Data Collection Statistics				
Beamline		SLS	Rigaku	SLS
Wavelength (Å)		0.9778	1.5418	0.9778
Resolution (Å)		45.52-1.80 (1.95-1.80)	19.47-1.74 (1.80-1.74)	45.49-1.40 (1.50-1.40)
Measured reflections		103583 (20590)	257320 (15990)	275724 (39272)
Unique reflections		26906 (5697)	28634 (2647)	56179 (10309)
Completeness		98.7 (99.1)	96.1 (93.1)	98.2 (97.7)
Redundancy		3.8 (3.6)	8.9 (6.0)	6.9 (5.4)
I/σ(I)		13.4 (4.1)	30.9 (6.8)	23.0 (4.1)
R_{meas} **(%)** [a]		8.4 (36.0)	6.7 (31.9)	3.9 (34.8)
Refinement Statistics				
Resolution (Å)		45.62-1.80 (1.85-1.80)	19.47-1.74 (1.78-1.74)	45.49-1.40 (1.44-1.40)
Number of atoms		2390	2417	2421
R_{work} **(%)**		16.4 (22.4)	15.3 (21.7)	12.7 (18.1)
R_{free} **(%)**		20.6 (29.8)	19.3 (26.8)	15.8 (24.0)
RMS bond lengths (Å) [b]		0.024	0.018	0.026
RMS bond angles (°) [b]		1.924	1.646	2.301
Averaged B-factors (Å^2)				
Total		20.2	16.4	19.5
Compound		19.8	20.2	21.3
Ramachandran plot residues				
Favoured regions (%)		96.2	97.1	97.6
Add. allowed regions (%)		3.8	2.9	2.4
Gen. allowed regions (%)		0.0	0.0	0.0
Disallowed regions (%)		0.0	0.0	0.0

[a] Redundancy independent R-factor (intensities).[234]
[b] RMSD from ideal geometry values.

Table E.9: Continued.

Compound Information				
ID		B5	B6	B7
PDB ID		4dhq	4dhp	4dho
Crystal Parameters				
Dimensions	a,b,c (Å)	82.3, 112.4, 62.6	82.3, 112.2, 62.4	81.9, 111.7, 62.3
Angles	α,β,γ (°)	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0
Space group		$C222_1$	$C222_1$	$C222_1$
Data Collection Statistics				
Beamline		Rigaku	Rigaku	Rigaku
Wavelength (Å)		1.5418	1.5418	1.5418
Resolution (Å)		19.56-1.75 (1.85-1.75)	19.53-1.75 (1.85-1.75)	19.46-1.70 (1.85-1.70)
Measured reflections		206270 (20537)	234914 (22063)	166297 (23894)
Unique reflections		29498 (4413)	28885 (4268)	31209 (6750)
Completeness		99.3 (98.4)	97.9 (95.9)	98.0 (96.2)
Redundancy		10.0 (4.7)	10.6 (6.8)	5.3 (3.5)
I/σ(I)		39.1 (13.7)	28.3 (7.3)	29.0 (8.0)
R_{meas} **(%)** [a]		3.8 (11.5)	7.0 (27.7)	4.2 (17.0)
Refinement Statistics				
Resolution (Å)		19.56-1.75 (1.80-1.75)	19.53-1.75 (1.80-1.75)	19.46-1.70 (1.74-1.70)
Number of atoms		2543	2475	2403
R_{work} **(%)**		15.1 (19.3)	15.6 (20.2)	16.2 (19.5)
R_{free} **(%)**		18.9 (23.7)	19.6 (25.2)	20.2 (23.1)
RMS bond lengths (Å) [b]		0.017	0.015	0.016
RMS bond angles (°) [b]		1.571	1.488	1.571
Averaged B-factors (Å^2)				
Total		16.3	15.7	16.4
Compound		15.5	21.2	25.6
Ramachandran plot residues				
Favoured regions (%)		96.1	95.6	96.6
Add. allowed regions (%)		3.9	4.4	3.4
Gen. allowed regions (%)		0.0	0.0	0.0
Disallowed regions (%)		0.0	0.0	0.0

[a] Redundancy independent R-factor (intensities).[234]
[b] RMSD from ideal geometry values.

Table E.9: Continued.

Compound Information				
ID		B8	B9	A3
PDB ID		4dhn	4dhm	3u9x
Crystal Parameters				
Dimensions	a,b,c (Å)	82.2, 111.8, 62.4	82.2, 112.3, 62.4	82.2, 112.1, 62.7
Angles	α,β,γ (°)	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0
Space group		$C222_1$	$C222_1$	$C222_1$
Data Collection Statistics				
Beamline		Rigaku	Bruker	Rigaku
Wavelength (Å)		1.5418	1.5418	1.5418
Resolution (Å)		45.41-1.80 (1.80-1.90)	19.52-1.70 (1.85-1.70)	45.54-1.80 (2.0-1.8)
Measured reflections		34486 (14026)	176142 (23320)	157300 (26966)
Unique reflections		26775 (3757)	31580 (6598)	26966 (7058)
Completeness		99.0 (94.4)	98.2 (93.1)	99.0 (97.0)
Redundancy		5.0 (3.7)	5.6 (3.5)	5.8 (3.8)
I/σ(I)		32.4 (12.0)	30.7 (9.7)	35.8 (16.0)
R_{meas} **(%)** [a]		3.8 (11.7)	3.9 (15.1)	4.9 (9.0)
Refinement Statistics				
Resolution (Å)		45.41-1.80 (1.85-1.80)	19.52-1.70 (1.74-1.70)	45.54-1.80 (1.85-1.80)
Number of atoms		2412	2474	2548
R_{work} **(%)**		15.0 (18.4)	15.4 (21.8)	13.6 (17.3)
R_{free} **(%)**		18.5 (24.0)	18.7 (23.1)	18.8 (22.6)
RMS bond lengths (Å) [b]		0.024	0.022	0.017
RMS bond angles (°) [b]		1.904	1.882	1.584
Averaged B-factors (Å^2)				
Total		18.1	17.0	17.6
Compound		20.7	22.5	21.4
Ramachandran plot residues				
Favoured regions (%)		97.1	96.2	96.0
Add. allowed regions (%)		2.9	3.8	4.0
Gen. allowed regions (%)		0.0	0.0	0.0
Disallowed regions (%)		0.0	0.0	0.0

[a] Redundancy independent R-factor (intensities).[234]
[b] RMSD from ideal geometry values.