

A consistent cheminformatics framework for automated virtual screening

Cumulative Dissertation

to receive the degree

Dr. rer. nat.

at the Faculty of Mathematics, Informatics and Natural Sciences
University of Hamburg

submitted to the
Department of Informatics of
the University of Hamburg

Sascha Urbaczek

born in Mannheim

Hamburg, August 2014

1. Reviewer: Prof. Dr. Matthias Rarey
2. Reviewer: JProf. Dr. Tobias Schwabe
3. Reviewer: Prof. Dr. Andreas Hildebrandt

Date of thesis defense: 23.01.2015

Für meine Lena

Acknowledgements

At this point, I would like to acknowledge the people who supported and encouraged me during my doctoral thesis. First and foremost, I extend my sincere gratitude to my supervisor Prof. Matthias Rarey for entrusting me with this interesting research project and providing guidance and support every step of the way. I also thank the Beiersdorf AG for funding the project and Dr. Inken Groth and Prof. Stefan Heuser in particular for their constant support as well as their great interest in my work. Furthermore, I thank the BiosolveIT GmbH for the productive working relationship during and especially after my time at the Center for Bioinformatics.

As for the current and former colleagues from the Center of Bioinformatics (which are far too many to list here); I am grateful to have had the opportunity to work in such a remarkably pleasant and productive environment. I really appreciate everyone's hard work and dedication during these (very successful) last years. In particular I owe a great deal of gratitude to Dr. Tobias Lippert from whom I learned a lot and who has become a very dear friend to me. Last but not least I also thank JProf. Tobias Schwabe and to Prof. Alexander Hildebrandt for reviewing my thesis.

Kurzfassung

Virtuelles Screening hat sich zu einem integralen Bestandteil der industriellen und akademischen Arzneimittelforschung entwickelt. Es wird eingesetzt, um sehr große Substanzdatenbanken mit der Hilfe von computerbasierten Methoden auf eine überschaubare Zahl vielversprechender Kandidaten zu reduzieren. Um eine möglichst hohe Vorhersagegenauigkeit zu erreichen, wird hierbei versucht, so viele Informationen wie möglich über das Zielprotein und seine bekannten Bindungspartner in die Berechnungen einfließen zu lassen. Die resultierenden Arbeitsprozesse umfassen deshalb häufig mehrere Schritte, in denen die verschiedenen Informationen berücksichtigt werden. Aufgrund der Vielzahl verfügbarer Methodiken und des starken Einflusses jedes einzelnen Schrittes auf das Ergebnis, ist virtuelles Screening eine Aufgabe von enormer Komplexität und mit vielen Fallstricken, die für gewöhnlich nur von Spezialisten durchgeführt werden kann.

Das Ziel der vorliegenden Arbeit war der Aufbau einer verlässlichen Basis für die Entwicklung von Software, die eine Integration von Medizinalchemikern in die computergestützte Wirkstoffsuche fördert. Das Ergebnis dieser Bemühungen ist ein neues chemieinformatisches Softwareframework (NAOMI) für virtuelles Screening, welches speziell auf die damit verbundenen Anforderungen angepasst wurde. Erstens erlaubt NAOMI Medizinalchemikern, basierend auf innovativen und intuitiv verständlichen Konzepten, ihre Erfahrung und ihr Spezialwissen an den Stellen der Berechnungen einzubringen, die für den Erfolg ihrer Projekte maßgeblich sind. Zweitens umfasst NAOMI zahlreiche neue chemieinformatische Methoden, die auf einem konsistenten internen chemischen Modell aufbauen und eine effiziente Ausführung der einzelnen Schritte des virtuellen Screenings erlauben. Die so erreichte Effizienz ist jedoch nicht nur eine notwendige Voraussetzung für Hochdurchsatz-Screening, sondern spielt auch eine zentrale Rolle in interaktiven Anwendungen, die in Kombination mit einer intuitiven Benutzerschnittstelle eine Schlüsselrolle in der Integration von Medizinalchemikern in die computergestützte Wirkstoffsuche spielen. Drittens wurde viel Wert darauf gelegt sicherzustellen, dass die Ergebnisse der verschiedenen Rechenschritte chemisch sinnvoll sind und dass ein hoher Grad an Konsistenz zwischen den verschiedenen Komponenten des Prozesses sichergestellt ist. Dies ist von entscheidender Bedeutung, da viele Schritte automatisiert werden müssen, um zu erreichen, dass Medizinalchemiker

sich auf die für sie relevanten Teilaspekte konzentrieren können. Wie in den Publikationen dieser kumulativen Dissertation gezeigt wird, sind zahlreiche Methoden in NAOMI selbst relevante Beiträge in ihren jeweiligen Anwendungsgebieten und ihre Kombination erlaubt den Aufbau effizienter, komplett automatisierter und hoch-adaptiver Screening-Prozesse.

Abstract

Virtual screening has long since become an integral part of the drug discovery process in both industry and academia. Its main purpose is to reduce huge compound databases to a manageable number of promising drug candidates with the help of computational methods. In order to provide reliable predictions, screening campaigns always aim to incorporate as much knowledge as possible about the target protein and its known binding partners. For that reason, the resulting workflows generally comprise multiple consecutive stages in which the different types of information are processed. The multitude of available methodologies and the strong dependency of the results on each individual step make virtual screening a complex task with numerous pitfalls which is usually only performed by specialized computational chemists.

The aim of the presented work was to provide a reliable basis for the development of software applications supporting the inclusion of medicinal chemists in computer-aided drug design activities. The result of this effort is a new cheminformatics framework (NAOMI) for virtual screening which has been specifically designed to meet the demanding requirements imposed by this scenario. First, based on both innovative and intuitively understandable concepts, NAOMI enables medicinal chemists to contribute their expert knowledge and experience at those points of the calculations that are crucial for the success of their current projects. Second, building on the consistent internal chemical model NAOMI comprises numerous novel cheminformatics methods enabling an efficient execution of each individual step of the virtual screening workflow. Apart from being a crucial factor in high-throughput calculations, computational speed is also of the essence in interactive applications which, in combination with intuitive user interfaces, are a key factor in involving medicinal chemists in computational endeavors. Third, great care was taken to ensure that the results of each stage are chemically reasonable and that a high degree of consistency is maintained between the different components of the pipeline. This is important as many steps of the process have to be automated in order to allow medicinal chemists to focus on those aspects relevant for their projects. As is shown in the publications presented in this work, many individual methods included in the NAOMI framework are in themselves relevant contributions to their specific field of application and their combination results in efficient, completely automated, and highly adaptable screening workflows.

Contents

1	Introduction	1
1.1	Protein-Ligand Interactions	3
1.2	Rational Drug Design	5
1.3	Computer-Aided Drug Design	7
1.4	Virtual Screening	9
1.5	Cheminformatics	11
1.6	Motivation	12
2	Screening Library	15
2.1	Representation of Molecules	16
2.2	Representation and Perception of Rings	20
2.3	Interpretation of Molecules from Chemical File Formats	25
2.4	Storage of Molecules in Databases	29
2.5	Selecting Sets of Molecules	33
2.6	Generation and Selection of Protomers	37
3	Protein Structure	43
3.1	Representation of Protein Structures	44
3.2	Structural Data of Protein-Ligand Complexes	46
3.3	Protomers in Protein-Ligand Complexes	50
4	Virtual Screening	55
4.1	Representation of Molecular Interactions	56
4.2	Molecular Docking	60
4.3	Structure-Based Pharmacophores	64
4.4	Protomers in Docking	68
4.5	Virtual High-Throughput Screening	70
4.6	Inverse Virtual Screening	72

CONTENTS

4.7 Comparison of Binding Pockets	73
5 Summary and Outlook	75
Bibliography	79
Bibliography of this Dissertation's Publications	95
Appendices	97
A Publications and conference contributions	97
A.1 Publications in scientific journals	97
A.2 Conferences	102
B Additional Data	103
B.1 Valence States	103
B.2 Corresponding Pairs of Valence States	104
B.3 Substructure Patterns	106
C Software Architecture	111
C.1 Software Libraries	111
C.2 Application Software	117
C.3 Software Testing	117
D Software Tools	121
E Journal articles	123

1

Introduction

Over the course of the last few decades, computers have gained more and more importance throughout all fields of chemistry. They are the only viable means to store and process the huge amount of experimental data produced by chemical research and to perform demanding theoretical calculations with a sufficient level of accuracy. The ever-growing need for more advanced computer-based approaches and improved application programs eventually lead to the establishment of new disciplines, which exclusively focused on the use and development of computational methods to solve chemical problems. The software produced in these fields was initially designed to be used by trained specialists in a supporting capacity. This has, however, considerably changed over the last ten to fifteen years. Nowadays, there is a computer on the desk of nearly every chemist and scientific software is routinely used to help with the day-to-day work. The typical user has evolved from a skilled computer professional with intricate knowledge of the underlying algorithmic and conceptual details to a computationally literate researcher without specific education in computer science. This change in clientele had an enormous impact on the development and design of scientific software, as aspects such as graphical user interface, usability and interactivity shifted more and more into focus. The disciplines of medicinal and pharmaceutical chemistry are no exceptions from this general trend. Researchers in these fields routinely use computers to guide them in decisions regarding the most promising strategies for the development of new and improved drugs. As the associated scientific problems are far too complex to be solved by any algorithm in an automated fashion, computers are mainly used to organize, analyze, and visualize the available experimental data. Using different types of heuristics, computer programs additionally can generate recommendations for the best course of action on the basis of the available data, e.g., which molecules are promising candidates for experimental testing. These application scenarios pose certain requirements

1. INTRODUCTION

on scientific software developed for computer-aided drug design. First, the underlying algorithms must be efficient enough to make interactive workflows possible. Second, the recommendations made by the programs must be both sensible and reliable. And third, the application programs must allow medicinal chemists to contribute their expert knowledge and experience at those points of the calculations that are crucial for the success of their current projects.

In the present thesis, a new cheminformatics framework for structure-based virtual screening (NAOMI) will be introduced. The main goal was to establish a robust and intuitive workflow which could serve as a basis for the development of scientific software that both enables and motivates medicinal chemists to participate in computer-aided drug design activities. For that purpose, the individual components of the pipeline have been developed under careful consideration of the three central requirements mentioned above, namely efficiency, reliability, and interactivity. At the heart of NAOMI is a robust and consistent description of chemical compounds which was developed in the course of this work. NAOMI comprises various novel methods ranging from standard cheminformatics operations such as the conversion of chemical file formats to advanced applications such as the prediction of molecular interactions in protein-ligand complexes. These methods are the building blocks of the above-mentioned screening pipeline. The project was a cooperation between the Center for Bioinformatics Hamburg and Beiersdorf¹. The prime objective was the development and application of virtual screening methodologies for the discovery of bioactive natural products which play an important role as active ingredients in the cosmetic industry. The presented work benefited largely from the close cooperation with an industrial partner as the developed methods and computer programs were directly applied by their intended users. The resulting feedback has positively influenced both the design and the functionality of the software.

In the following, the basic principles and main applications of computer-aided drug design are presented. The first two sections introduce the molecular basis of the complex interactions between biological targets and drugs and explain how this knowledge can be used for the rational design of drugs. The next sections emphasize the role computers currently play in the drug development process with the focus being on virtual screening methodologies. Afterwards, a cursory introduction to the fairly broad field of cheminformatics is presented. Finally, a motivation for the project based on the current problems and challenges in drug design is provided and the overall structure of the thesis is outlined.

¹Beiersdorf AG, Research Active Ingredients, Hamburg

1.1 Protein-Ligand Interactions

The specific interactions between proteins and small molecules are the molecular basis for the therapeutic effect of most common drugs. Proteins are biological macromolecules which play an essential role in virtually all cellular processes in living organisms, e.g., metabolism, signal transduction, and energy transfer. They are composed of one or multiple chains of amino acids, their molecular building blocks, whose respective linear sequence is determined by the genetic code. Of the more than 500 naturally occurring amino acids currently known [1], only 22 are proteinogenic, i.e., protein building. These comprise an identical backbone part, the α amino acid substructure, and a side-chain which is unique for each type and determines its characteristic physicochemical properties. Specific interactions between amino acids in a particular sequence determine the overall three-dimensional structure of the protein which in turn is responsible for its biological function. The latter is directly connected to one of the most characteristic features of proteins, the ability to bind other molecules at a specific location called binding site. The shape and the physicochemical properties of this region, determined by the side chains of the surrounding amino acids, regulate the specificity and tightness with which the ligands are bound. The interaction with other molecules gives rise to the vast array of functions proteins can serve in organisms including cell signaling (scaffold proteins), forming connective tissue (structural proteins), and contracting muscle fibers (motor proteins). The following discussion will be restricted to the binding of small molecules which is of high importance in the context of drug design. A comprehensive overview of the structure and function of proteins can be found in standard biochemistry textbooks [2, 3].

One of the most important roles of proteins in cells is that of a catalyst for chemical reactions. Such specialized proteins are called enzymes and the respective reaction partners substrates. Many of the chemical transformations mediated by enzymes are essential steps in metabolism and are thus vital for living organisms. The actual reaction takes place in the active site, usually a small cavity in the protein, and involves only a small number of the protein's amino acids directly. Their three-dimensional arrangement forms the molecular basis for the enzyme's biological function by determining the shape and the physicochemical properties of its reactive center. The amino acids are arranged in such a way that the transition state, the configuration with the highest potential energy on the reaction coordinate, is stabilized. This generally results in a considerably higher reaction rate compared to the uncatalyzed process under the same conditions. Furthermore, the three-dimensional structure also accounts for the

1. INTRODUCTION

enzyme-substrate specificity as only molecules with complementary shape and physicochemical properties can interact favorably with the atoms in the active site. This concept is known as the “lock-and-key” principle, which was first introduced by Emil Fischer [4]. It considers both protein and molecule as rigid entities which must fit together as lock and key. In view of the fact that proteins are flexible structures whose conformations often change when interacting with other molecules, many experimental observations are not in accordance with this model. In 1958, Daniel Koshland [5] presented the induced-fit theory, according to which the initial binding of the substrate induces conformational changes in the enzyme which are essential for its catalytic activity. Even today, the driving forces for the recognition of small molecules by proteins are not fully understood and it is still not possible to predict the function of a protein from its structure alone.

Nonetheless, many effects influencing the binding of small molecules to proteins have been identified [6–8] and can be used for the rational design of drugs. Major contributing factors are hydrogen bonding, ionic interactions, metal-coordination, van-der-Waals forces, and the hydrophobic effect. With exception of the latter, these contributions are subsumed under the term electrostatic interactions and are frequently the predominant reason for the specificity of molecular recognition [9]. A hydrogen bond can be interpreted as a dipole-dipole attraction between an electronegative atom and a hydrogen atom attached to a second electronegative atom [10]. Nitrogen, oxygen and fluorine are, due to their high electronegativities, the most common elements to be involved in this type of chemical bonding. As the strength of hydrogen bonds strongly depends on the relative arrangement of the involved atoms, they are often characterized as directed interactions. All standard amino acids can partake in hydrogen bonding owing to their α amino acid substructure. Additionally, there are functional groups in particular side chains which also have this ability, e.g. the hydroxy group of serine. Attractive ionic interactions, sometimes called “salt bridges”, are a result of coulomb forces between atoms or groups of atoms with opposite charges. They are quite typical for amino acids with side-chain functional groups which are ionized under physiological conditions, e.g., arginine and aspartic acid. Hydrogen bonds and ionic interactions often occur at the same time, as positively charged groups usually also have hydrogen atoms attached to them. If a protein has bound metal ions in its active site, the coordination by ligand atoms can also have an important stabilizing effect. Since metal atoms have a preference for specific coordination geometries, their interactions with ligand atoms can be categorized as directed. Van-der-Waals forces is a collective term for various undirected attractive and repulsive dipole-dipole interactions which are usually weaker than the aforementioned contributions. The interactions between nonpolar portions

of a molecule and nonpolar amino acid side chains are subsumed under the term hydrophobic interactions. These are generally undirected as their strength only depends on the distance between the respective substructures. Their contribution to stability is mainly determined by the change in nonpolar surface of both protein and ligand before and after binding. This has been summarized as the hydrophobic effect [11, 12], which describes the tendency of nonpolar compounds to aggregate in aqueous solution in order to avoid unfavorable interactions with water molecules.

1.2 Rational Drug Design

Historically, pharmaceutical research has mostly been governed by empirical observations, accidental discoveries, and trial and error methods [13]. With the advent of molecular and structural biology and the associated increase in understanding of cellular processes and pathways as well as the simultaneous development and establishment of advanced experimental methods, more systematic and targeted approaches to the problem became available. These are often subsumed under the term rational drug design which describes the process of inventing new medicines on the basis of the available knowledge of a particular biological target, i.e. the protein targeted by the drug. This target is in some way specific to a disease condition, e.g. an enzyme involved in vital cell processes of microbial pathogens, and its manipulation therefore results in a therapeutic benefit. In most cases, the drug is a small organic molecule which inhibits the target protein from fulfilling its biological function. There are, however, also numerous examples for conditions in which the activity of the target protein actually needs to be enhanced. The principles applied to the design of new drugs are based on the knowledge about protein-ligand interactions described in the previous section. If the specific way a molecule interacts with a particular protein is known, then it is possible to devise other compounds which behave in a similar fashion. According to the “lock-and-key” principle such molecules must have a comparable shape and compatible physicochemical properties. A typical way to design an enzyme inhibitor, for instance, is to find a molecule that binds tightly to the protein but does not undergo the catalyzed chemical reaction. In this way the protein is no longer available for the transformation of its actual biological substrates.

Depending on the amount of available experimental data, there are two complementary strategies for the rational design of drugs. In case of structure-based methods, the three-dimensional structure of the target protein is used. Information of that kind is typically obtained through X-ray crystallography [14] or NMR spectroscopy which yield

1. INTRODUCTION

three-dimensional positions of protein atoms as well as those of potentially bound ligands and water molecules. Both methods provide immediate insight into the structure of the protein's binding pocket with respect to shape and potential interactions sites. In the best case, the resolved structure also contains a bound ligand so that the respective binding mode can be directly inspected and interpreted. In case no experimental protein structure data is available, ligand-based techniques are applied. These evaluate the properties of molecules which are already known to interact with the target. The basic idea of this indirect approach is to develop an idea of the potential binding mode in the respective protein by analysis of the similarities between these molecules.

Rational design techniques are typically applied during the drug discovery stage [15], the first step of the drug development process. The aim of drug discovery is to identify biologically active small molecules with suitable characteristics for the approval as drugs by the appropriate agencies. The high requirements of the latter constitute a considerable difficulty for the task and can only be met with enormous research efforts involving a large variety of experimental procedures. As rational drug design is focused on the inhibition or activation of a specific biological entity, the identification of a suitable drug target is naturally the first step of the process. This requires a thorough understanding of the disease mechanisms and the roles particular proteins play in them. In order to be worthy of consideration, a potential drug target must meet a number of different criteria. First, it must be druggable, i.e. it must be possible to manipulate the targets' biological activity by drug molecules. Second, engaging the target does result in a statistically relevant therapeutic benefit. And third, the pharmacological modulation of the target does not compromise the safety of the patient even in long-term clinical usage. After a valid target has been established, experimental methods for the measurement of the target's biological activity, so called assays, need to be developed. These are essential for assessing the inhibiting or activating effect molecules have on the target protein. Assays form the basis for the subsequent hit discovery phase in which molecules with the desired activity, called hits, are identified by using one or multiple compound screening techniques. These processes are often completely automated, especially in pharmaceutical companies, and can be used to test up to millions of compounds. From the often large number of initial screening hits, promising candidates need to be selected as starting points for the following hit-to-lead phase. This selection can be based on results of additional experimental procedures, e.g., concerning pharmaco-kinetic properties, or considerations of aspects of chemical synthesis, such as ease of preparation. The selected hit molecules generally do not fulfill all the necessary requirement imposed on drugs and need to be further optimized with respect to multiple parameters. At this stage structure-based and ligand-based methods can

play an important role as they help in developing structure-activity relationships. If the structural basis for the interaction with the target protein is known and understood, compounds can be systematically modified in order to optimize both their potency and selectivity. Apart from the ability to interact with a target protein, there are additional pharmaco-kinetic properties which are essential requirements for potential drug candidates. The most relevant aspects are absorption, distribution, metabolism, and excretion (ADME). When administered to a patient, the active molecule must find its way to the intended target protein in the human body to take effect. This means it must be absorbed into the bloodstream and transported to the respective effector site without being inactivated by metabolic processes. Afterwards the drug must be completely removed from the body to avoid an accumulation which in turn could result in adverse effects on normal metabolism. All of the above-mentioned parameters must be optimized simultaneously in order to transform a hit molecule into an acceptable lead structure. In the final phase, lead optimization, the lead from the previous step is further characterized and improved, typically using more advanced experimental methods, until it is finally ready to be declared as preclinical candidate. In an industrial setup, the initial screening typically starts out with several hundred thousand compounds which are reduced to a few hundred during the hit-to-lead phase. From these only one or two compounds are eventually submitted to the clinical phases[15].

1.3 Computer-Aided Drug Design

The ultimate goal of pharmaceutical research is the development of new medicines for the treatment of diseases. The process of discovering novel drugs and converting them into products ready for the market is extremely complex and involves a wide variety of experimental procedures as was discussed in the previous section. This makes drug development a generally time-consuming and, above all, cost-intensive endeavor. The approximate time frame for the completion of a drug development cycle is 13 years and the associated costs are estimated at 1.8 billion U.S.\$ [16].

Computer-aided drug design (CADD) can help to considerably reduce the high costs of the drug development process by replacing expensive experimental procedures with cost-effective computations. This applies in particular to the early stages of drug discovery, where computational methods can make valuable contributions to each of the individual steps introduced in the previous section [17]. Bioinformatics approaches can help in identifying and selecting potential disease targets as they provide the means for the analysis and utilization of the vast amounts of heterogeneous data and information gathered from diverse experiments, patents and literature sources [18]. Moreover,

1. INTRODUCTION

computational methods are an important tool to assess the druggability of potential target proteins [19]. This prediction can either be based solely on the protein's amino acid sequence [20] or additionally incorporate information about its three-dimensional structure [21]. In the context of the latter, the automated identification of protein binding sites [22] often also plays an important role. Virtual screening methodologies have long since become an integral part of the hit discovery phase as they help to reduce the number of compounds which need to be tested in expensive experimental screening campaigns [23]. Over the time, numerous efficient algorithms and methods have been developed reflecting the multitude of different scenarios medicinal chemists are confronted with. If the three-dimensional structure of the target protein is available, molecular docking [24] can be applied to find molecules with geometrical and chemical properties that are complementary to those of the protein's binding pocket. Otherwise, ligand-based approaches, such as pharmacophore-based screening [25] and 3D-QSAR [26], are available which only rely on the properties of already known active molecules. If the experimental screening setup is based on the combinatorial synthesis of compounds, computational methods provide the means to select and combine only those reagents that will result in libraries with a desired physicochemical profile [27]. Such multiobjective optimization tasks are but one example of the many different problems which cannot be efficiently solved without the use of computers. Apart from these typical screening applications, computational methods can also be used to optimize and focus compound libraries in a more general sense, e.g using parameters such as diversity [28] and drug-likeness [29]. In this way, compounds which are either too similar or have unfavorable physicochemical properties can be excluded even before the experimental screening phase. Many of the methods mentioned above are, however, not only useful when it comes to the design and optimization of screening libraries. Docking, for instance, can also be used in the context of lead optimization in order to verify if particular structural modifications in the molecule will result in a valid binding mode. Moreover, fragment-based approaches, such as fragment growing and scaffold hopping [30], provide a systematic way to identify sensible chemical transformations under consideration of additional constraints imposed by the protein's binding pocket. Since eliminating unsuitable candidates as early as possible in the discovery pipeline is one of the mayor concerns of medicinal chemists, the *in silico* prediction of drug metabolism and toxicity [31, 32] is an important element of CADD, especially considering that these properties are experimentally assessed at late stages of the process. In this context, aspects such as polypharmacology, adverse effects and drug promiscuity are of great importance and have also been addressed by computational methods [33–35].

As the previous introduction shows, the number and kinds of applications of CADD is diverse and includes a wide variety of methods from different fields, including cheminformatics, bioinformatics, and computational chemistry. The following sections will only focus on virtual screening methodologies as only those are relevant in the context of the presented thesis. A more comprehensive overview of the whole field of CADD can be found in standard text books [17, 36].

1.4 Virtual Screening

Virtual screening, as the name already suggests, can be considered as the *in silico* analogue of experimental screening. Its main purpose is to analyze large compound databases with the help of computational methods in order to identify possible new drug candidates [37]. This is achieved by either prioritizing molecules on the basis of a calculated score value, which reflects, at least to a certain extent, the associated probability of activity against a particular target, or by eliminating unsuitable candidates using various types of filter criteria. Depending on the kind and amount of structural and bioactivity data included in the prediction, virtual screening methodologies can be classified into different categories [38]. As was already mentioned above, there is a fundamental distinction between structure-based and ligand-based methodologies. Structure-based approaches explicitly take the three-dimensional structure of the target protein into account and try to identify promising candidates on the basis of their geometric, and in many cases also physicochemical, complementarity to the respective binding pocket. Molecular docking [39] is the most prominent example for the structure-based approach. It is the computationally most demanding screening technique and has high requirements with respect to the quality of the provided structural data. Ligand-based methods [40], on the other hand, make use of information derived from one or multiple known bioactive molecules and are thus based on similarity rather than complementarity. The underlying algorithmic strategies can range from simple similarity searches which are applied in case of a minimal information basis, e.g., a small number of known actives, to sophisticated machine learning techniques. The results are generally coarser than those of structure-based calculations but in contrast to those ligand-based methods are also applicable if only little is known about the target, e.g., at the early stages of projects when no protein structure is available and only a few active molecules have been identified. Pharmacophore-based screenings [41] can be counted among either of the two approaches as the underlying, often three-dimensional, pharmacophore description can be generated from the protein's binding pocket as well

1. INTRODUCTION

as the superposition of active ligands. With respect to ligand-based methods, pharmacophores are often applied when multiple active ligands are known so that their common features can help to find similar molecules. Naturally, it is also possible to combine all types of approaches according to the requirements of the respective drug design problem.

Virtual screening has long since become an established technology for the discovery of new lead structures and is widely applied in both academical and pharmaceutical research [42, 43]. It is a viable alternative to cost-intensive and time-consuming experimental techniques and allows to perform large-scale drug discovery campaigns without the need to maintain an expensive specialized screening laboratory. This fact can be of particular value to smaller pharmaceutical companies and academic institutions which often do not have access to automated high-throughput screening facilities in order to identify new hit and lead structures. While the accuracy of current screening tools is certainly by far not high enough to replace experimental research [44]—and maybe never will be considering the complexity of the involved problems—there are numerous examples for the successful combination of both approaches [23, 45]. In most cases, virtual screening is used as a complement to experimental testing and serves as an efficient prefilter with the aim to generate more focused and target-specific compound libraries by explicitly taking the structure and physicochemical properties of the target protein into account. Moreover, it can serve as an idea generator for the identification of novel biologically active molecular scaffolds which, in the context of combinatorial chemistry, may even include compounds that have not yet been synthesized.

The virtual screening process can be roughly divided into four individual steps [46], all of which have a significant impact on the quality of the results. These are the compilation of a small-molecule screening library, the selection and preparation of an appropriate reference system, e.g., the target protein structure or the pharmacophore model, the actual screening calculation, and the postprocessing of the obtained results. Since the structure-based screening pipeline and its different stages will be presented in detail in the following chapters of this thesis, the discussion at this point will focus on more general aspects. Due to the fact that virtual screening has a strong tendency to produce false-positive results [47], the postprocessing step is of great importance. The aim is to minimize the number of false-positives and at the same time propagate the true hits to the top of the result lists. This is usually hard to realize when relying only on a single approach. Therefore, the combination of different screening approaches which are applied sequentially according to their respective level of complexity [48] has become a very popular strategy. In this way all available information about both target and active ligands can be used to identify true hits and lead structures. However, even

today there is still no general process for the application of virtual screening to specific drug design problems. Determining which type of method or which combination is best suited requires a fundamental understanding of the respective system and a careful analysis of all available information.

Virtual screening methods are designed to process large databases containing up to millions of compounds. In order to provide the necessary throughput for such a demanding task, screening methods, in addition to the use of highly efficient algorithms, generally need to rely on considerable simplifications concerning the physicochemical description of molecules, proteins, and their respective interactions. For that reason, the technology underlying the different approaches for virtual screening are largely based on methods from the field of cheminformatics which will be introduced in the next section.

1.5 Cheminformatics

When the term “cheminformatics” (synonymously used with chemoinformatics) was first introduced in the literature fifteen years ago [49], computational methods had already been used for decades [50, 51] to solve problems in different areas of chemistry. The multitude of sometimes even unrelated fields of application lead to a lot of controversy [51] about the actual scope of this young discipline and its clear distinction from related fields. Even today, this controversy is not yet completely resolved and the discussion about its distinctive underlying models and concepts is still ongoing [52]. The very broad definition of cheminformatics given by one of its pioneers, Johann Gasteiger [50], is a testimony to its expansive scope:

Chemoinformatics is the use of informatics methods to solve chemical problems.

In many alternative definitions of cheminformatics given over the years [49, 53, 54], the term “chemical information” plays a preponderant role. Cheminformatics methods are used to store, organize, visualize, and analyze chemical information. Data associated with particular chemical compounds or chemical reactions, typically experimental results, is used to expand our chemical knowledge and to improve our understanding of the relationship between chemical structure and properties. A comprehensive overview about all applications of cheminformatics methods is well beyond the scope of this work and the following discussion will focus on those aspects relevant for the methods developed in the course of this thesis. A detailed introduction spanning the whole field can be found in *The Handbook of Chemoinformatics* [50].

1. INTRODUCTION

As the initial definition by Brown [49] suggests, cheminformatics is closely connected to drug design and many of its methods have been specifically designed to assist computational medicinal chemists in their day-to-day work. The chemical problems encountered in this field are mostly related to organic molecular chemistry and more specifically to the relationship between molecular structure and its associated properties. Typical tasks involve the design of molecules with predetermined properties or the identification of molecules with similar properties. The enormous complexity of the structure-property relationship usually makes it impossible to solve most common medicinal chemical problems from first principles, so that cheminformatics has to make use of different molecular models than physics-based disciplines such as quantum chemistry [52]. The main goal is not the a priori calculation of molecular properties to chemical accuracy, but the manipulation and analysis of large sets of molecules [55]. Since its inherent models are not based on current physical theory, the appropriate representation of molecular structures in itself is one of the most basic aspects of cheminformatics. Its conclusions are usually derived from the statistical analysis of large amounts of data rather than rigorous physical concepts, so that the underlying inference mechanism can be best described as inductive [52]. Apart from the rapidity of such calculations, which make the processing of large molecule sets possible in the first place, cheminformatics is generally able to predict any type of molecular property assuming the existence of sufficient experimental data.

As a discipline dealing with large numbers of molecules, the concept of chemical space [56], which comprises all possible small organic molecules, is central to cheminformatics. Representing and, more importantly, navigating this literally infinite space is of crucial importance for the identification of new bioactive scaffolds and compound classes. For that purpose, molecules are generally represented by either chemical graphs [57] or descriptor vectors [58]. Both types of descriptions provide the basis for the efficient handling of typical cheminformatics tasks, including similarity searching, scaffold classification, structure clustering, or building QSAR models. Many theoretical and practical aspects concerning the representation of molecules in the context of virtual screening will be discussed in chapter 2. A more detailed and comprehensive discussion of the application of cheminformatics methods in drug discovery can be found in [59].

1.6 Motivation

Computational methods have become an integral part of the drug discovery pipeline in industry and academia and often play an important role in the decision making process concerning the generation and selection of new lead structures [60]. The literature is

full of accounts of successful applications of CADD and virtual screening [61, 62] for the design of new lead compounds and drug molecules. Furthermore, the existence of five scientific journals dealing exclusively with CADD is a testimony to the general relevance of cheminformatics methods [63]. Nevertheless, in recent years there has also been a lot of criticism concerning the quality, reliability, and general applicability of *in silico* approaches, especially with respect to virtual screening methodologies [44, 64, 65]. Their predictive power is a long way from being sufficient enough to provide accurate and reliable results that can be used without the careful analysis and interpretation by experts. One of the main challenges in the future of CADD is the fundamental improvement of the current screening approaches, including, above all, the scientific foundation of the underlying scoring functions [66]. However, considering the complexity of common drug design problems, the notion of a fully automated and universally reliable virtual screening platform seems more or less illusory.

For that reason, it has been proposed that instead of focusing solely on the development and application of new concepts and methods, “future success depends on the proper integration of new promising technologies with the experience and strategies of classical medicinal chemistry” [67]. An important step in this direction could be the motivation of medicinal chemists to undertake CADD activities for themselves [68]. In this way, experimental scientists could benefit from a better understanding of the three-dimensional aspects of protein-ligand interactions including, for instance, the implications of the conformational degrees of freedom of their synthesized molecules. Moreover, computational methods could provide them with the means to efficiently formulate and comprehensibly validate specific hypotheses based on their individual experiences. The inclusion of medicinal chemists in CADD activities will, however, have considerable influence on the way cheminformatics software has to be designed. What is needed are application programs which are “well-thought-out, suitable for their needs, [and] able to generate useful, timely and valid results” [68]. Such requirements, particularly the design of well-thought-out interfaces and the maintenance of the corresponding software, are extremely time-consuming and generally also not in the focus of academic research groups. Furthermore, complex software projects are virtually always carried out by highly interactive teams rather than isolated scientists which is rather contradictory to the current structure in academia. For that reason the main development in this field is done by professional software vendors [68, 69].

The goal of the presented work was the development of a new cheminformatics framework for virtual screening which explicitly enables medicinal chemists to partake in this particular stage of the computer-aided drug design process. This was done under consideration of the requirements for scientific software stated in [68]. The challenges

1. INTRODUCTION

of such an endeavor are manifold and involve problems from many different areas. In order to be “suitable for their needs” CADD tools need to both reflect the problems and tasks medicinal chemists are typically confronted with and allow them to bring their experience and knowledge to bear in the respective calculations. This usually requires that the models and concepts underlying the respective methods are sufficiently intuitive to be applied in the general context of medicinal chemistry. On the other hand, drug design software should provide reliable automated procedures for all those processes and operations which are not in the focus of the current projects but nevertheless have effects on the quality of the obtained results. Additionally, the tools must be “well-thought-out”, meaning that they should be equipped with a well designed and intuitive user interface which allows medicinal chemists to propagate their expert knowledge in an appropriate way. Since “timely” results are a prerequisite for any type of interactive application the underlying software must make use of efficient algorithms and sophisticated concepts. Consistent chemical models, on the other hand, ensure that the results are both scientifically “valid” and “useful”. In order to realize the above-mentioned requirements, the NAOMI framework was implemented with the focus on creating a reliable basis for the implementation of state-of-the-art virtual screening platforms and other typical cheminformatics applications in both an academic and a professional setup. The underlying software library should not only provide the means for the development of innovative new methodologies but also the design of software tools that are suitable for the professional use in the field of CADD.

In the following the conceptual and algorithmic contributions to the field of CADD included in the NAOMI framework are presented which constitute the components of the virtual screening pipeline and essentially correspond to the publications comprising this cumulative dissertation. These include the interpretation of molecular structures from chemical file formats, the processing, storage, and querying of large compound collections, the prediction and evaluation of intermolecular interactions in the context of protein-ligand complexes, and the identification of promising chemical structures by an index-based docking approach. The text is organized into three chapters corresponding to particular stages of the structure-based screening process as described in section 1.4. Each chapter begins with a short introduction to the general goals, requirements and preconditions of the respective phase and is followed by multiple sections reflecting problems and challenges typically encountered at this particular point of the process. These sections comprise both a discussion of the respective topic in the general context of drug design and cheminformatics as well as the presentation of the solution developed in the NAOMI framework. Finally, the thesis finishes with a summary of the work and an outlook to possible expansions.

2

Screening Library

The compilation of a small-molecule screening library is the initial step of the virtual screening pipeline. The starting point is usually a large compound collection, e.g., a vendor catalog or an in-house database, from which a subset of molecules with properties suitable for the respective drug design problem is selected. Many different aspects have to be considered in order to assess the suitability of a particular compound, including its physicochemical properties and structural features, its commercial availability or synthetic accessibility and the current patent situation. In an optimal case, the resulting library does only contain those compounds which are considered as realistic candidates for the subsequent drug discovery pipeline as this both reduces the runtime of the screening calculation and avoids any unnecessary effort during the potentially time-consuming analysis of the obtained results. Apart from these more general considerations, which, in principle, equally apply to the selection of compounds for experimental testing, there are numerous technical issues which are specific to the virtual screening approach. On the one hand, there are various aspects concerning data maintenance. The validity of the chemical data provided by the primary sources has to be ensured, as invalid data will inevitably lead to scientifically invalid and thus useless results. Additionally, the chemical structure data must be normalized and organized, for instance by registration into an existing compound management system, in order to eliminate duplicates, avoid inconsistencies, and enable different types of filter and search capabilities. Furthermore, the molecules in the dataset usually have to be subjected to specific ligand preparation routines ensuring that all the information required by the subsequent docking routines, e.g., three-dimensional atomic coordinates or explicit positions of hydrogen atoms, is available.

The following subsections will highlight numerous aspects which play an important role during the processing of the chemical data which typically stands at the beginning

2. SCREENING LIBRARY

of the virtual screening pipeline. A more comprehensive discussion and additional literature references can be found in [70].

2.1 Representation of Molecules

The explicit representation of atoms and their respective connectivity is the foundation of a large number of common cheminformatics methods, e.g., substructure searching, 2D structure depiction, or comparing molecules for identity. For this purpose, cheminformatics employs the structural formula as a mathematical model of chemical structure. This conception of molecules dates back to Gilbert N. Lewis [71] and is still today the predominant notion among experimental chemists [72]. The structural formula represents molecules as an undirected graph, in which atoms are vertices labeled by their elements and bonds are edges labeled by their bond order. The connectivity of such chemical graphs is restricted by valence rules which state the number and the types of bonds a particular chemical element can form considering its valence electrons. The description of molecules as topological entities rather than geometrical ones has several advantages with respect to typical cheminformatics problems, the most important ones probably being the well-defined concept of molecular identity and a tangible notion of structural similarity. Moreover, it allows to formulate, using both the terminology and methods of chemical graph theory [57], many common chemical concepts, e.g., isomerism, tautomerism, and substructures, with mathematical stringency. An excellent discussion of the topological description of molecules and its inherent implications for chemistry, including the comparison to the physics-based concepts of quantum chemistry, can be found in [72]. A general introduction to the structure of molecules and their associated representation is included in Linus Pauling’s standard work, the “Nature of the chemical bond” [10]. According to Pauling, the term valence bond structure will be used as a synonym for lewis structure, structural formula, and kekule structure in the following.

The representation of molecules by a single valence bond structure is a common practice throughout chemistry. Be it as a structural diagram in a research paper or as a single entry in a chemical database, this description offers a simple and comprehensible way to communicate to other chemists which molecule is actually meant. However, in the context of cheminformatics, where the associated chemical graphs are used as a mathematical description of molecules, this approximation is in many cases no longer sufficient. The simple fact that the same molecule can be represented by different valence bond structures can easily lead to inconsistencies and needs to be addressed

explicitly by cheminformatics software systems. Two prototypical cases are shown in Figure 2.1.

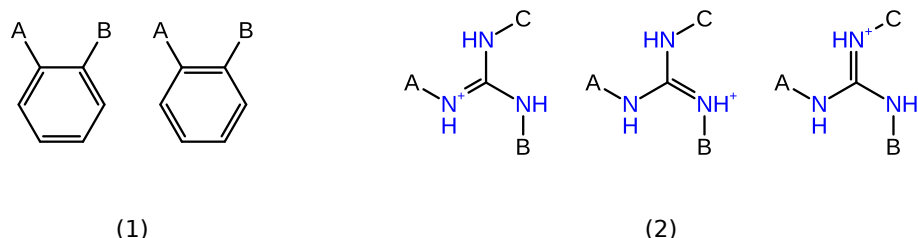


Figure 2.1: Two examples for the ambiguities resulting from the description of molecules by valence bond structures. In case of asymmetric substitution at the aromatic ring, the two valence bond structures of example (1) are not identical due to the different locations of the double bonds. The same applies for the three structures in (2) in which both the double bond and the positive formal charge change positions.

Example (1) comprises two representations of a six-membered aromatic ring in which the double bonds are occupying different positions. These are usually referred to as Kekulé structures. Example (2) shows three representations of a guanidinium group in which both the positive formal charge and the double bond have changed their locations. Both cases actually correspond to the exact same chemical entities and the existence of multiple distinct valence bond structures is an artifact of valence theory. Although these limitations clearly show that the valence bond description does not accurately model all aspects of chemical bonding, cases such as the ones discussed above can, with the help of a few additional concepts, still be represented with sufficient accuracy for the vast majority of cheminformatics applications. This is due to the fact that such molecules, even though the description is ambiguous, can be described in terms of valence theory at all. In contrast to that, there are different types of chemical species for which this does not hold true. Typical examples are electron-deficient compounds in which the chemical bond comprises more centers than electrons, e.g. boranes, and organometallic compounds with haptic bonds such as ferrocene. The bonding situation of such structures can only be accurately described using molecule orbital theory. Although there are a few approaches trying to extend the usual valence bond model of cheminformatics software systems, e.g. RAMSES [73], such compounds are in most cases simply ignored. This is, apart from the extreme difficulties of modeling such compounds without recourse to quantum-theoretical methods, probably also due to their rather low significance in the field of medicinal chemistry and drug design.

In order to make use of the valence bond model and its numerous advantages in cheminformatics software applications, molecules have to be translated into a computer-

2. SCREENING LIBRARY

readable representation. The most common choice is the connection table [74] which basically corresponds to a topological graph description of the molecule. Although virtually all chemical software systems are based on a valence bond representation, their underlying chemical models often vary with respect to particular aspects [75]. First, there is the internal valence model which defines the allowed valences of elements and thus determines which types of molecules can be handled. It plays an important role during chemical validity checks and is also needed for the calculation of implicit hydrogen atoms, e.g., when reading molecules from files as will be explained further down. Second, there is the internal atom type model which extends the valence bond description in order to provide a more precise chemical description based on the atom's environment. Atom types can be considered as an additional annotation of chemical information which often play a crucial role in the prediction of physicochemical properties and are an essential part of force field calculations [76]. Third, there is the internal aromaticity perception which is often needed to compensate for the inherent shortcomings and ambiguities of the valence bond model on a structural level. The consistent handling of the alternating bonds from Example (1) in Figure 2.1 is a typical example. Additionally, aromaticity is an important physicochemical property since aromatic rings are an ubiquitous feature of drug molecules. Unfortunately, there is, to the best of the author's knowledge, no review, publication, or other document discussing the respective chemical models of commonly used cheminformatics toolkits in detail. The only definitive sources of information remain the documentations of the respective software libraries or their source code in case of open source projects (a list containing a selection of popular cheminformatics software systems with corresponding references can be found in Table 2.1). Although the study performed by Sayle [75] can only be considered as a first step in the systematic investigation of the chemical models of cheminformatics toolkits, the provided benchmark calculations give at least a general insight into the existing inconsistencies and problems.

The chemical model used for the representation of molecules in the NAOMI framework [D1] is essentially based on a graph description. It comprises three distinct layers of chemical information which have been designed to fulfill the different requirements imposed on cheminformatics systems by typical application scenarios in the context of CADD (see Figure 2.2). The element layer (A) offers the most basic level of description and essentially reflects the underlying graph structure of the molecule. It comprises the element identities of all atoms in combination with their respective connectivity. The valence state layer (B) extends the graph properties by providing valence states for atoms and bond orders for bonds thus corresponding directly to a valence bond description of the molecule. Valence states represent valid bond order distributions for

2.1 Representation of Molecules

Table 2.1: Compilation of popular cheminformatics software libraries and frameworks including online resources and literature references where available.

Toolkit	Open source	Publication	Homepage
CACTVS	yes	[77]	[78]
CDK	yes	[79, 80]	[81]
MOE	no	-	[82]
OpenBabel	yes	[83]	[84]
OEChem	no	-	[85]
PerlMol	yes	-	[86]
Pipeline Pilot	no	-	[87]
RDKit	yes	-	[88]

atoms in valence bond structures and can be considered as their atomic building blocks. As will be described in the following sections, they are the basis of the valence model of the NAOMI software system and play a central role in many algorithms and methods. The main purpose of the atom type layer (C) is to circumvent the limitations of valence theory with regard to conjugated systems. On the one hand, this means providing a united representation for cases in which multiple equivalent valence bond structures could be formulated. On the other hand, effects such as aromaticity and planarity of particular atoms due to hybridization effects need to be reflected. This is exemplified in Figure 2.2 for two prototypical cases. First, there is a six-membered aromatic ring which is drawn in circle notation indicating both the aromaticity of the system and the ambiguous location of the double bonds. This is handled in the NAOMI model by assigning an aromatic atom type (C-Aro) to the atoms of the ring and by marking the ring itself as having alternating bonds. Second, there is an amidinium group with a delocalized positive charge. In this case both nitrogen atoms, despite their differing valence states, have an identical atom type (N-Deloc(+)) which labels them as conjugated, and thus sp² hybridized, with a positive partial charge. Additionally, all bonds of the group are marked as delocalized.

The hierarchical chemical model is the heart of the NAOMI framework and has been specifically designed for the accurate and efficient handling of molecules relevant in the context of drug design, i.e., organic compounds. One of its most important purposes is to provide all the necessary structural information and chemical descriptors needed for the development and implementation of diverse cheminformatics methods and algorithms. In this respect, the separation of the valence bond description from the handling of hybridization effects is an important concept underlying the NAOMI

2. SCREENING LIBRARY

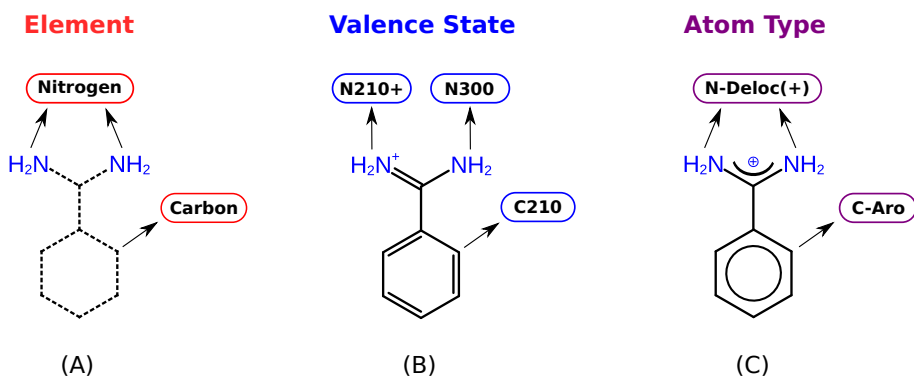


Figure 2.2: The three layers comprising the chemical model of the NAOMI framework. The Element layer (A) represents the graph structure of the molecule, the valence state layer (B) corresponds to a valence bond description, and the atom type layer (C) describes the effects resulting from the delocalization of electrons. The atomic descriptors associated with each layer are shown for three atoms.

model. Both types of descriptions can be useful in different contexts and thus need to be available when working with molecules. The valence state layer is often needed for the modification of molecules or when their validity has to be ensured whereas the atom type layer plays an important role during the calculation of physicochemical properties. The independent handling of the perception of aromaticity based on Hueckel’s rule and the identification of rings with alternating double bonds is another example for the application of this concept. Both properties are relevant in different contexts and must be available independently when needed in order to provide a more adaptable description of molecules. For the reasons explained above, there is no feasible way to directly compare the chemical models of different cheminformatics software systems with respect to quality and generality. However, the investigation based on the conversion of file formats presented in [D1] can be considered at least as an indication that the internal chemical model of the NAOMI framework is indeed more robust than those underlying other commonly used cheminformatics tools. Additionally, many successful applications to problems from both cheminformatics and CADD, which will be presented in the following, demonstrate its suitability in the context of drug design.

2.2 Representation and Perception of Rings

Rings are an ubiquitous structural motif in synthetic and natural compounds and generally have large influence on their respective physicochemical properties. Ring strain often enhances the reactivity of molecules and its relief can be one of the driving forces

2.2 Representation and Perception of Rings

of chemical reactions. Cyclic delocalization gives rise to the important effect of aromatic stabilization which profoundly changes the chemical behavior of the respective compounds compared to similar non-aromatic systems. The inherent rigidity of cyclic structures results in a limited conformational flexibility which is often exploited by medicinal chemists to establish and lock a specific spatial arrangement of atoms or functional groups in a molecule. Furthermore, the replacement of acyclic parts in lead compounds with ring structures is a common strategy in drug design to gain an entropy-driven increase in receptor-ligand binding energy. Rings also have a strong impact on a compound's synthetic accessibility [89] and constitute a very important factor in the assessment of molecular complexity. Aromatic and heteroaromatic rings, for instance, are a common feature of drug molecules partly because there is a large number of established methodologies for their synthesis and modification [90]. In the context of cheminformatics, knowledge about the number of rings contained in a molecule or the number of rings an atom is part of is crucial for a wide range of applications. Many physicochemical properties and chemical features such as aromaticity can simply not be determined without this type of information. Ring counts, e.g., the total number of rings or the number of rings having a specific size, and ring sizes, the smallest or largest ring for instance, are efficient prefilters for the preparation of screening libraries and can be very useful for the elimination of unwanted molecules. Ring membership is also a commonly used property for atoms and bonds in substructure queries as it allows to specify their respective chemical environment more precisely. Furthermore, the subdivision of molecules into cyclic and acyclic parts is a common heuristic strategy in cheminformatics algorithms and workflows, including the generation of structure diagrams [91] and three-dimensional atomic coordinates [92].

The perception and classification of ring systems and their individual rings is thus a crucial aspect of the description of molecules and plays a major role in many cheminformatics applications. It is not surprising, that a large number of different concepts for their representation and algorithms for their perception have been developed over the years. A thorough and comprehensive review of these approaches can be found in [93]. In general, ring systems are represented by a set of individual rings resulting from the detection of cycles within their underlying graph structure. However, the process of identifying cycles in a graph is ambiguous so that both the size and the members of the resulting set are determined by the respective perception strategy. The criteria which are applied to decide whether a particular set is chemically useful are, as often in cheminformatics, a compromise between chemical intuition and technical requirements and also strongly depend on the respective context of application. The following discussion will be restricted to three instructive examples of ring descriptors, the set of all

2. SCREENING LIBRARY

rings (Ω), the smallest set of smallest rings (SSSR), and the relevant cycles (RC). These will be used to explain the relevance of those aspects that played the most important role during the development of the ring perception method used in the NAOMI system. A detailed and comprehensive discussion of possible criteria for the determination of the chemical relevance of ring sets and a thorough investigation of their satisfaction by previously described methods can be found in [94].

The set of all rings (Ω) [95] is based on an exhaustive ring perception meaning that all cycles contained in the ring system are used as a representation of the same. This concept fulfills one important criterion for chemically useful ring sets, namely uniqueness. The resulting descriptor does neither depend on the concrete algorithm used for its generation nor on the ordering of the vertices and edges of the graph it has been derived from. On the other hand, the often huge number of rings that need to be calculated and stored in order to obtain this descriptor is one of the method's major disadvantages. This property is usually referred to as exponential size, meaning that for particular types of graphs the number of members in the ring set grows exponentially with the number of vertices. Apart from problems with runtime and memory during the actual calculation, the consideration of all rings can also be problematic for other cheminformatics applications which depend on information about rings for their internal heuristics. This is especially true if these heuristics have been developed on the basis of different ring perception concepts, e.g., a decomposition into a set of smaller cycles, as explained further down. Additionally, The set of all rings (Ω) is not necessarily consistent with chemical intuition concerning the number of rings an atom is part of as is shown for a simple example in Figure 2.3.

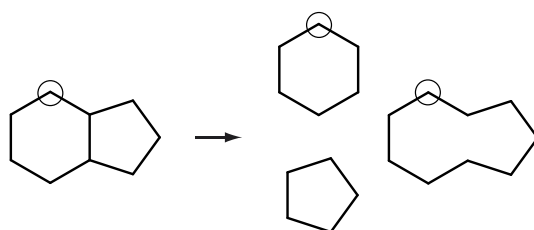


Figure 2.3: The set of all rings (Ω) for a bicyclic ring system. According to this perception method, the circled atom is part of both a six-membered and a nine-membered ring.

The SSSR is without doubt the most common ring perception method in cheminformatics, presumably because it can both be efficiently computed [96], i.e. in polynomial time, and is easy to implement. The basic idea is to describe the ring system in terms of a relatively small set of rings, a minimum cycle basis, from which all other cycles can be constructed by defined mathematical operations. This has the advantage, that the

2.2 Representation and Perception of Rings

descriptor is always polynomial in size and thus does not pose any problems concerning both calculation and storage. Additionally, the subdivision of the ring system into its smallest rings generally matches chemical intuition and has been the basis of most other cheminformatics heuristics. Despite these significant advantages, there is a serious flaw in the SSSR concept. Minimum cycle bases are generally ambiguous, i.e., there might be multiple equivalent SSSRs for a particular ring system. Since the descriptor is not unique, it is not independent from both implementational details and input data which can lead to artificial results in cheminformatics calculations [97]. Additionally, this ambiguity, in some cases, does also lead to incompatibility with chemical intuition as shown for cubane in Figure 2.4. Only five of the six sides of the cube can be part of an SSSR, meaning that atoms from the side which is not included are members of two rings whereas all other atoms are part of three. Which of the six faces is not included is, however, arbitrary.

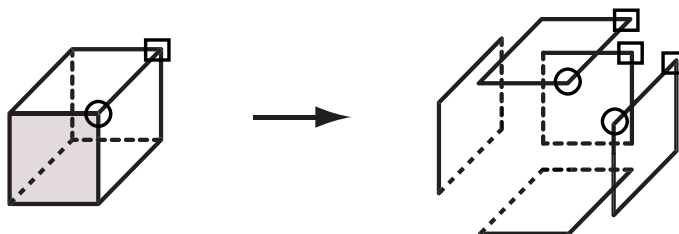


Figure 2.4: SSSR of the cubane molecule. As only five sides of the cube can be part of a SSSR, the descriptor is not unique in this case. If, for instance, the front side (grey) is excluded, as shown on the right hand side, the circled atom is a member of two rings whereas the squared atom is part of three.

The RC [98] are conceptionally similar to the SSSR in that ring systems are described by a set of smallest cycles. But instead of arbitrarily selecting one of the multiple minimum cycle bases the RC are defined as their union. In that way the problem of ambiguity can be circumvented altogether. In case of the cubane molecule mentioned above, for instance, all six sides of the cube will be part of the set. Using the union instead of the smallest set of rings, however, brings back the problem of exponential size for particular types of ring systems. The most relevant example for chemistry are presumably cyclophane-type molecules as the one shown in Figure 2.5. Additionally, the algorithm needed to calculate the RC [98] is rather complex and considerably more difficult to implement than the ones described for the calculation of the SSSR.

The ring perception method in the NAOMI framework is based on the concept of Unique Ring Families (URF) [D2]. The URF are a further development of the RC with a particular focus on avoiding the exponential size for all types of ring systems.

2. SCREENING LIBRARY

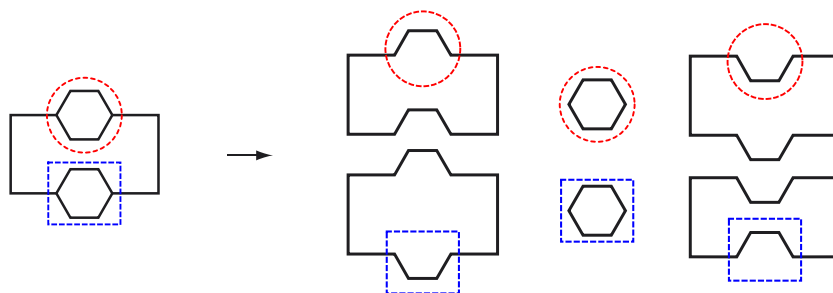


Figure 2.5: RC of [2.2]paracyclophane including two six-membered and four twelve-membered rings. The number of relevant cycles for paracyclophanes grows exponentially with an increasing number of additional para-linked six-membered ring inserted into the chain. As indicated by the circles and squares, each additional six-membered ring introduces two distinct paths in the larger cyclic structure thus increasing the number of large rings by a factor of two.

This is achieved by merging the unintuitively large number of relevant cycles associated with particular types of ring systems into a unified ring description as indicated for two simple examples in Figure 2.6. A comprehensive introduction into the terminology, a stringent derivation of the respective mathematical theorems, and a thorough description of the algorithmic details can be found in the original publication [D2].

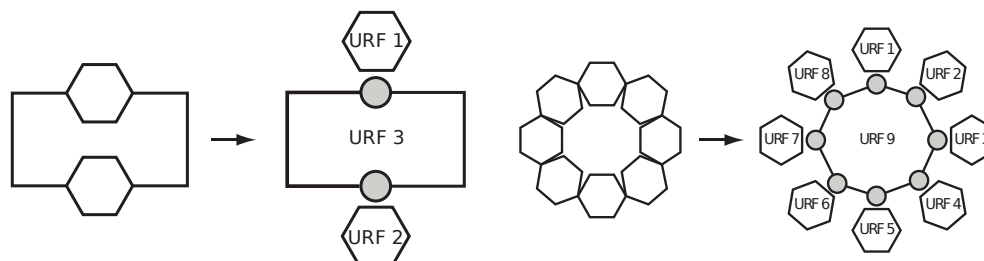


Figure 2.6: URF for two prototypical types of ring systems. The multitude of rings resulting from alternating paths through the different six-membered rings are merged into a single ring family thus providing a unified description. Figure adapted from [D2].

The concept of URFs does, however, not only solve the problem of the exponential size, but also provides a more intuitive description of the ring membership of particular atoms compared to other approaches (see Figure 2.7 for examples). By being unique, polynomial in size, and intuitive the URF is the first published cheminformatics ring descriptor fulfilling all three criteria at the same time. Additionally, it can be calculated in polynomial time and its generation does not pose problems with respect to computing times even in complicated cases [D2].

2.3 Interpretation of Molecules from Chemical File Formats

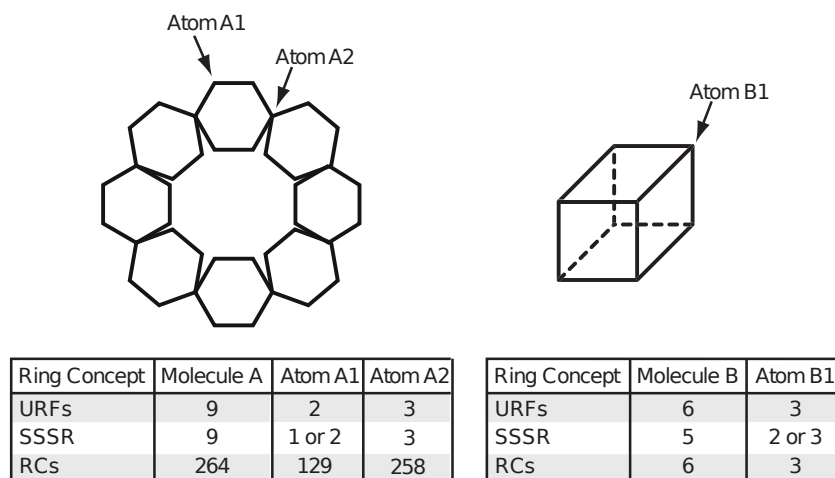


Figure 2.7: Comparison of different ring concepts with respect to ring membership of particular atoms. Figure reprinted from [D2].

2.3 Interpretation of Molecules from Chemical File Formats

Cheminformatics methods are generally designed to process large numbers of molecules. These are, in the vast majority of cases, not generated by the methods themselves, but are compiled from external sources such as vendor catalogs and chemical databases. Due to this inherent dependency on external data, the exchange of chemical information, and of molecules in particular, plays a fundamental role in cheminformatics. Over the years many specialized chemical file formats, the open babel project [83] currently supports more than 100 different types, have been developed for that particular purpose. Three of those are especially widely used, namely Tripos MOL2 [99], Symyx SDF [100, 101], and SMILES [102], with the latter two being the de facto standards. The following discussion is restricted to these three cases which are instructive to the problems that arise during the interpretation of molecules from chemical file formats. A more comprehensive discussion of file formats can be found in [74, 103]. Most cheminformatics systems, if not all, are based on a description of molecules by valence bond structures. This is naturally reflected in the file formats used to transfer data between these systems. As a consequence, each format includes a certain way to specify element identities, formal charges, and bond types. The graph structure is stored using either a connection table (SDF, MOL2) or a specialized character string (SMILES) [74]. In addition to these rather obvious technical differences, there are also subtle dissimilarities with respect to the underlying chemical models. Although essentially being based on

2. SCREENING LIBRARY

a valence bond description, each format implements a different strategy to circumvent the inherent shortcomings associated with this representation. The following discussion will focus on those aspects that can easily lead to ambiguity and whose resolution requires a robust and consistent internal chemical model. A more detailed description of the concrete format-specific models and conventions is provided in [D1].

The reason for most problems encountered during the interpretation of molecules from file formats is the omission of redundant chemical information. For instance, it is very common to exclude hydrogen atoms in order to save disc space (SDF, MOL2) or to obtain a more compact string representation (SMILES). This alone does not pose a problem since the number and connectivity of missing hydrogen atoms can be unambiguously derived from the bond orders and formal charges of the remaining non-hydrogen atoms. Another example is the introduction of an additional bond type, called aromatic bond, as it is done in both SMILES and MOL2. This extension is intended to resolve the problem of arbitrary single and double bond positions in aromatic rings (see Example (1) in Figure 2.1) such as benzene. In this case bond orders are omitted and must be derived from the remaining data. The last example mentioned at this point is the implicit handling of formal charges by specialized atom types such as they are used in the MOL2 format. This does provide a unified description of delocalized charges (see Example (2) in Figure 2.1), which otherwise could only be represented by multiple mesomeric valence bond structures. Although each of these strategies is not problematic on its own, ambiguities can arise when several are used at the same time. This will be illustrated with the help of the three examples shown in Figure 2.8. Example (1) shows a five-membered carbon ring in which all bonds are annotated with an aromatic type and which is stripped of all hydrogen atoms. In contrast to its six-membered counterpart, there is no way to formulate a neutral valence bond structure containing only sp^2 hybridized carbon atoms for this particular case. The structure could be interpreted as cyclopentadiene by assigning an sp^3 hybridization to a particular atom or as a cyclopentadienyl-anion by addition of a negative charge. This means, however, that the molecule resulting from this input is ill-defined and its final structure depends on the correction mechanisms of the respective cheminformatics software system. Example (2) shows an unsymmetrical imidazole moiety which could be perceived as either one of its two distinct tautomeric forms since the position of the hydrogen atom bound to one of the nitrogen atom is not explicitly specified. Again, the result will ultimately depend on the underlying algorithms of the respective software system. The last example (3) shows a six-membered ring containing a nitrogen and an oxygen atom. In this case the structure could correspond to two different oxidation

2.3 Interpretation of Molecules from Chemical File Formats

forms, either the neutral 1,4-oxazine with an additional hydrogen at the nitrogen atom or the pyrylium ion with a positively charged oxygen atom.

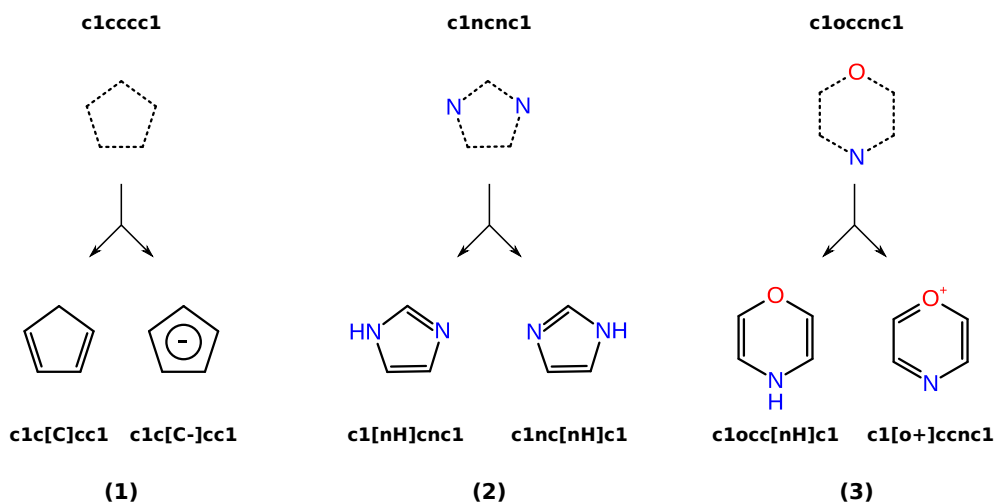


Figure 2.8: Three examples for the ambiguities resulting from the omission of chemical information in chemical file formats. The ambiguous molecules are shown in the form of a SMILES string (at the top) and a structural diagram with dashed lines indicating aromatic bonds. The SMILES strings and valence bond structures below represent possible interpretations of the respective input.

Despite the fact that all cheminformatics tools must provide functionality for the interpretation of molecules from different file formats and offer ways for their respective conversion, either explicitly or implicitly, this topic has not received much attention in the scientific literature. This is probably due to the fact that these procedures are generally considered as either trivial or too technical. Even comprehensive textbooks restrict themselves to the mere enumeration of file format converters, with OpenBabel being the most commonly cited tool [83]. Both the comparison of different commonly used tools performed during the course of this thesis [D1] and the study by Sayle [75] indicate that there is indeed a large potential for errors. Considering the fact that such errors and inconsistencies almost certainly will have detrimental effects on the results of downstream algorithms and calculations, it is the author's belief that not enough attention is paid to this basic first step in cheminformatics. Although the problem has been recognized as a pitfall of virtual screening workflows [65, 104], it has, to the best of the author's knowledge, neither been systematically investigated nor conceptually addressed elsewhere.

The NAOMI framework enforces a very strict adherence to valence rules on the basis of the valence state model presented in the previous sections. Molecules are considered

2. SCREENING LIBRARY

as invalid and rejected if the valence state layer could not be consistently constructed. The criterion for validity is that a valence state could be successfully assigned to each atom and that a distribution of bond orders could be found which is in accordance with this assignment. In the absence of aromatic bonds this process is straightforward. The number of multiple bonds in combination with a formal charge are sufficient to unambiguously identify valence states. Obscured bond types can make the assignment more complicated since multiple valence states may be compatible with the atom's local topology. In this case, bond localization routines must be used after the initial assignment in order to ensure the molecule's validity. The complete workflow for the interpretation of molecules from different file formats is shown in Figure 2.9. A more detailed discussion of the individual steps can be found in the original publication [D1].

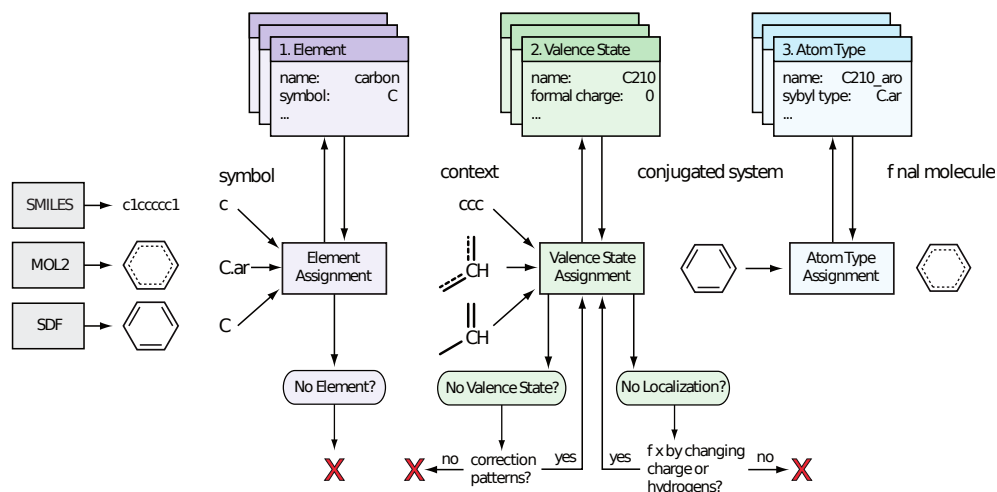


Figure 2.9: Schematic workflow for the interpretation of molecules from different file formats in the NAOMI framework. Figure reprinted from [D1].

As shown by different evaluation procedures, the NAOMI framework [D1] is highly consistent with respect to the interpretation and conversion of chemical file formats both internally and in comparisons to other commonly used cheminformatics software tools. By relying on a robust and consistent chemical model, the valence state layer in particular, NAOMI is able to completely maintain the integrity of the data provided by file formats with different underlying representations even after multiple conversion runs. Furthermore, the presented workflow is highly efficient and shows better performance than comparable methods. The validation procedures described in [D1] played an important part in achieving this level of consistency and are now a permanent part of the internal test suite (see Appendix C). With respect to virtual screening applications, the processing of input data from chemical file formats is a mandatory first step

and typically needs to be performed for a large number of molecules. This means that apart from being efficient, the process must be completely automated and unsupervised which in turn requires that data from chemical file formats is accurately interpreted and reliably propagated. On the one hand, the NAOMI framework has a strict rejection policy with respect to chemical data which is either considered as invalid or cannot be handled, so that a reasonable description of molecules can be guaranteed at all times. On the other hand, it employs a number of correction mechanisms which help working with input data containing inconsistencies, e.g., by not conforming to format specifications. Both strategies are perfectly suited for rather inexperienced users which are not working with manually curated datasets.

2.4 Storage of Molecules in Databases

As already mentioned in Chapter 1, data management is one of the most central applications of cheminformatics methods and naturally also plays a crucial role in medicinal chemistry and drug discovery. In order to make decisions at different stages of the drug development process, medicinal chemists need easy access to the vast amounts of available information on the compounds in question, e.g., physicochemical properties, literature references, and experimental results. The ability to handle and retrieve this data in an efficient and comprehensive manner can only be realized with the help of chemical information systems. These comprise a database backend which provides the possibility to assign different types of data from various sources to the same chemical entities and a set of application programs for the formulation of search queries. In the context of virtual screening, the storage of molecules in specialized database systems also offers many technical advantages over the direct use of chemical data files. First, the required disc space can be considerably reduced due to the omission of redundant information. For instance, when working with multiple conformations of the same molecules, which is common when using particular screening methods, the topological information needs to be stored only once. The standard chemical file formats, on the other hand, are not designed for this scenario so that each conformation corresponds to an individual molecule entry. Second, the setup times are usually considerably lower due to the possible reuse of data in different projects, meaning that compounds need to be registered only once and become part of a global compound collection including all annotated data. When working with chemical data files all entries in the file need to be processed again for each application. Third, the fast and flexible data retrieval capability associated with modern database systems allows to perform even complex

2. SCREENING LIBRARY

queries with high efficiency. As will be discussed in the next section, the latter is of central importance for the compilation of screening libraries.

In the context of chemical databases, the well-defined concept of molecular identity, as one of the most striking features of the representation of molecules as topological entities [72], is of vital importance. The underlying chemical graphs provide both a mathematically sound and computationally tractable way to determine whether two molecules are identical independent from their rather fuzzy and ambiguous geometrical properties. This important feature of valence bond structures makes it possible to both formulate and perform queries based on molecular structure in a straightforward and comprehensible way. It is thus not surprising that the valence bond description of molecules is the foundation of virtually all chemical information systems. From a technical point of view, any cheminformatics software system must provide some way to compare molecules for identity in order to avoid duplicates when registering compounds into a database. This task is, however, not trivial, considering that the description of molecules by graphs or connection tables is in itself neither unambiguous nor unique. In general, there are multiple ways to order its vertices and edges resulting in a multitude of distinct but equally acceptable graph representations. For that reason, the atoms and bonds of the molecule must be canonically ordered, or numbered, in advance to make a reliable comparison possible. Although multiple approaches have been developed to derive a canonical ordering over the years, the Morgan algorithm [105] and its variations, e.g. the CANON algorithm [106] for the generation of unique SMILES, are still the most common choices. For the storage of molecules in chemical databases, the canonical molecule graphs are normally converted into unique string representations which can be both efficiently compared and indexed by computers. The most widely used examples are the above mentioned unique SMILES string [106] and more recently the InChI identifier [107]. However, canonicalization does not only play an important role in the context of data management, but every cheminformatics algorithm can benefit from the enhanced degree of consistency when working with canonical structures, e.g. if the result of the algorithm depends on graph traversal.

Considering the inherent ambiguities of the valence bond description discussed in the previous sections, it becomes apparent that the canonicalization of the chemical graph is not necessarily sufficient to provide a unique representation for a particular compound. In case multiple valence bond structures, and thus different underlying chemical graphs, can be formulated for a molecule, the result of the comparison will ultimately depend on the ones that have been provided for the compounds in the respective context. This problem is not only restricted to artificial cases such as the two examples shown in Figure 2.1. Despite their different physicochemical properties,

tautomers and protonation states are generally also considered to be the same chemical compound. This interpretation gives rise to a number of additional problems with respect to the storage, searching and retrieval of molecules. First, it complicates both the concept and the determination of equality. Second, it eventually makes it necessary to choose a suitable representation for a particular molecule, e.g. during visualization or export to chemical file formats. In order to deal with these problems a general methodology for the canonical generation and selection of valence bond structures for molecules is needed. Such methods are a fundamental requirement for the appropriate description of molecules and ignoring them will almost certainly lead to inconsistent and thus unreliable results. The ramifications of this fact have been extensively reviewed recently [108].

The functionality for the storage of molecules in the NAOMI framework [D3] is based on a relational SQL database referred to as MolDB in the following. As most other chemical database systems the MolDB establishes the identity of molecules on the basis of their respective topology using an internal canonical string identifier (MolString). One of the most important concepts of the MolDB is the distinction between molecules and instances. The former term describes the actual compound, represented by its topology, whereas the latter refers to its occurrence in a data set. Depending on the context the interpretation of instances can vary. On the one hand, a particular compound can be present in two distinct chemical files so that one of the instances can be considered as a duplicate of the other. On the other hand, instances can correspond to different conformations (see Figure 2.10). Upon registration into the database each entry of a file is assigned both a MoleculeKey based on its topology and an unique InstanceKey. The former plays an important role during the management of screening sets as described in the next section, whereas the latter is mostly needed to manage either data from different input files or docking poses depending on the context.

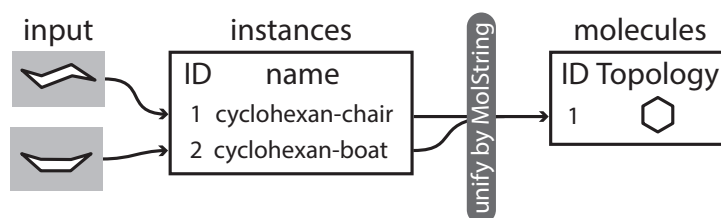


Figure 2.10: Two conformations, boat and chair, of the six-membered ring on the left correspond to different instances of the cyclohexane molecule. Both conformations receive an unique InstanceKey and an identical MoleculeKey. Figure reprinted from [D3].

2. SCREENING LIBRARY

MolStrings are used for the assignment of instances to their corresponding molecules and basically correspond to the valence state layer introduced in Section 2.1. Thus they are closely related to the internal description of molecules in the NAOMI system. This allows to rapidly convert them back to the internal molecule representation which is needed for numerous applications of the MolDB. MolStrings are generated following the typical workflow for unique molecule descriptors mentioned above. First, the molecule graph is canonicalized and then the string identifier is generated on the basis of the resulting canonical structure. The procedure used for the canonicalization on the basis of the current valence bond structure corresponds with a few minor modifications to the CANON algorithm [106]. The methods for the generation of canonical protomers will be discussed in Section 2.6.

The requirements on chemical databases with respect to consistency are quite similar to the ones discussed for the interpretation of molecules from different file formats. It has to be ensured that compounds stored in the database can be retrieved without modification or loss of information. This has been validated for the MolDB on the basis of different large public datasets [D3] and the associated procedures have all become part of the automated test framework. The consistency and robustness of the canonicalization methods, in particular with respect to different valence bond structures, have also been thoroughly investigated [D4]. In spite of the fact that different topological and physicochemical properties are calculated (see next section), molecules are subjected to canonicalization procedures, string identifiers are generated, and a check for duplicates is performed, the registration of compounds still remains very efficient [D3]. However, since the database can be saved and reloaded almost instantly, the runtime are not of central importance in this particular case.

The general technical advantages of chemical database systems have already been discussed above. There are, however, many additional benefits which can be gained from integrating the MolDB system directly into a screening pipeline. First, it provides a clearer and more intuitive structure for the management of compounds and the compilation of screening libraries than chemical file formats. The automated detection of duplicates and an intuitive storage of multiple ligand conformations are but two relevant features in this respect. Additionally, the visualization of molecules using two-dimensional structure diagrams is an important way to inspect compound collections [D3]. Second, the binding poses and scores resulting from screening runs can be handled in an efficient and consistent manner by storing them as instances. Since the database can be saved and reloaded, the results of time-consuming docking calculations can thus be easily stored without the use of chemical file formats as intermediates. Moreover, instances can be recreated from the MolDB without much effort, so that the

associated three-dimensional coordinates can be easily accessed, for instance for the three-dimensional visualization of binding poses. Third, results of screening calculations are directly integrated into the compound management system thus making it possible to perform typical database operations such as selecting, sorting, and filtering under consideration of the respective scores. The MolDB thus provides an intuitive and comprehensible way to manage both screening libraries and results of virtual screening applications at the same time.

2.5 Selecting Sets of Molecules

As the processing of large numbers of molecules is one of the main characteristics of cheminformatics, it is not surprising that its methods are generally used when it comes to the selection or prioritization of compounds from large collections. The task of reducing large sets of molecules to a manageable number of promising candidates is quite common in drug design. Often, the number of available compounds exceeds the capacities of the screening facilities so that many decisions have to be made prior to the actual measurement and thus without the aid of experimental data. The careful compilation of a screening library, for instance, is an important task for medicinal chemists as it has substantial influence on the success of the associated screening campaign. The main goal is to assemble a collection which is both enriched with molecules having favorable property profiles and at the same time depleted of all undesirable compounds. This increases the chances of finding viable lead structures with a high potential for becoming actual drugs. The criteria which can be applied for that purpose are quite diverse and depend on the respective application context. Physicochemical properties are generally a very important aspect as they have a strong influence on the bioavailability of compounds [109, 110] which plays an important role in the later steps of the drug development pipeline. Particular structural features, i.e. functional groups or substructures, are associated with toxicological effects [111] and are thus often excluded from the drug development process in advance. Furthermore, there are compounds which are known to interfere with experimental screening methods and thus tend to produce false positive results for reasons other than specific activity towards the target protein [112], so called screening artifacts. These also need to be excluded from screening sets as well as pharmacologically promiscuous compounds which are prone to interact with different types of unrelated targets.

Although the associated costs are substantially lower than those of their experimental counterparts, the well-considered selection of molecules for screening libraries is nevertheless also advisable for virtual screening applications. On the one hand, if

2. SCREENING LIBRARY

Table 2.2: Filter types commonly used during the preparation of compound libraries for virtual screening calculations.

Filter Type	Example Application
Element	Halogenated compounds
Topology	Compounds with complex ring systems
Property	Unpolar compounds (logP)
Substructure	'Pan Assay Interference Compounds' (PAINS) [118]
Molecule	Compounds protected by patents

particular compounds cannot be tested in an experimental setup or are not viable as a drug for any reason, the results of the calculations will ultimately have no practical value. On the other hand, considering their role as prefilters for experimental screenings, a lack of diversity [113] can easily lead to the omission of relevant classes of compounds in subsequent experiments. Different preparation schemes for virtual screening data sets have been described in the literature [114, 115]. An overview of the respective publications and a broad introduction to the topic can be found in [70]. The filters routinely employed to select wanted and eliminate unwanted compounds can vary widely and serve different purposes (see Table 2.2). These basic types are often combined into more complex filters reflecting the desired physicochemical profile of the remaining compounds. Lipinski's 'Rule of five' [116] and the criteria proposed by Oprea [117] are prominent examples for this approach.

In order to perform the essential operations for the preparation of screening data sets cheminformatics software offers two different approaches. On the one hand, there are so-called workflow tools [119] which allow to create and maintain complex processing protocols on the basis of isolated and configurable building blocks. These are called "components" or "nodes" and directly correspond to typical cheminformatics tasks such as filtering or substructure searches. Pipeline Pilot [87] and KNIME [120] are the most commonly used tools for that purpose. One important advantage of this approach is the complete automation of the process once the pipeline has been established. This is especially useful if the same operations are performed repeatedly with varying input data. On the other hand, there are chemical database systems [121] which are also often used to store and manage compound collections as discussed in the previous section. In many cases a predefined set of properties, both topological and physicochemical, are calculated during the registration of molecules and stored in the database. This data can be rapidly accessed and thus provides the basis for the implementation of efficient filter mechanisms. One major disadvantage of chemical databases is the laborious

2.5 Selecting Sets of Molecules

installation and initialization of the associated server-based system. A third and quite different option is to go without the help of specialized software tools by directly using data sets which were specifically prepared for virtual screening applications such as ZINC [122].

The functionality for the management of screening libraries in the NAOMI framework is based on the MolDB which was introduced in the previous section. One of its central concepts is the molecule set which is defined as a collection of pairwise different molecules (not instances) determined on the basis of their topology. Molecule sets usually represent subsets of the complete compound database resulting from the application of specific filter criteria. The MolDB offers different operations for the generation and modification of molecule sets which are useful for the compilation of screening libraries (see Figure 2.11).

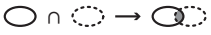
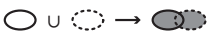
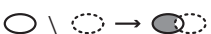
operations	signature	description	
	$\text{intersect}(S_1, S_2, \dots, S_n) \longrightarrow S_r$	set S_r contains molecules that are in all original sets	
	$\text{union}(S_1, S_2, \dots, S_n) \longrightarrow S_r$	set with all molecules contained in any of the original sets	
	$\text{difference}(S_1, S_2) \longrightarrow S_r$	contains molecules from S_1 that are not in S_2	
filter set	$\text{filter}(S) \longrightarrow S_r$	available filters	example
		physico chemical property	molecular weight > 200
		chemical element	contains oxygen, no nitrogen
		functional group	contains Pyridol
		smarts	contains c1ccccc1
visual select subset	$\text{select}(S) \longrightarrow S_r$	select subset of S by picking structure diagrams	
split subset	$\text{split}(S, n) \longrightarrow (S_1, S_2, \dots, S_n)$	split set S into n equally sized parts	

Figure 2.11: Overview of the operation on molecule sets supported by the MolDB. Figure reprinted from [D3].

First, there are the mathematical set operations, namely union, intersection, and difference, which are relevant when working with two distinct molecule sets. The difference operation, for instance, can be used to eliminate predefined collections, e.g., lists of compounds protected by patents, by subtracting them from the potential screening set. Second, the MolDB supports various kinds of filter operations routinely used in the field of CADD in order to eliminate compounds with unsuitable properties. The necessary data for the respective queries is calculated during the registration of compounds into the database [D3]. The only exception is substructure searching based on SMARTS as the respective strings are defined externally and thus cannot be generated in advance. Furthermore, the MolDB supports the concept of filter chains, meaning that different elementary filters can be logically combined which allows to implement complex criteria

2. SCREENING LIBRARY

such as the 'Rule of five'. Finally, the splitting of molecule sets into smaller units is also supported and can be helpful, for instance, when preparing chemical files for test cases or external tools.

Molecule sets are stored as simple lists of MoleculeKeys in the database. Since all the mathematical set operations are solely based on molecular identities, i.e., the comparison of MoleculeKeys, they can be realized directly by the database using SQL statements and are thus very efficient. The same is essentially true for filter operations with the exception of SMARTS queries. The precalculated values are stored in the database and can be accessed using its built-in functionality. In case of SMARTS, molecules have to be recreated from the MolString representation and internally evaluated against the substructure pattern. This makes them less efficient than other type of queries. Filter chains are in most cases as efficient as the contained elementary filter operations. Only if tolerances are specified, meaning that only a part of the specified filters need to be passed, this is not the case since multiple SQL statements are necessary. The runtime associated with different kinds of operations has been thoroughly investigated (see Figure 2.12). As internal consistency is an important feature of the NAOMI framework the different operations provided by the MolDB have also been extensively tested in various validation procedures [D3].

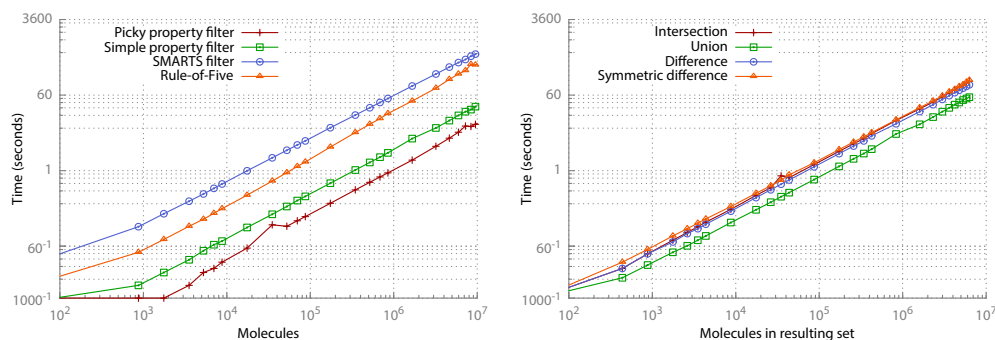


Figure 2.12: Runtimes for different operations supported by the MolDB. The diagram on the left shows a comparison of different types of filters, the diagram on the right for set operations. Figure reprinted from [D3].

The presented approach to the manipulation and compilation of screening libraries based on molecule sets is very intuitive and thus perfectly suited for the use by medicinal chemists. Furthermore, the operations of the MolDB are efficient enough to ensure the interactivity of workflows with up to one million compounds. This is an important feature, since in many contexts the best suited combination of filters for the current drug design problem are not known in advance. Often, there are limitations with

respect to the number of compounds that can be experimentally tested, so that the dataset must be reduced using more and stricter filters until a specific size is reached. In this scenario, the processing of molecule datasets becomes an interactive process in which filters are iteratively applied and adapted according to the results of the previous steps. MONA [D3] is an application offering this kind of functionality. It is an example of how the NAOMI framework can be used to implement efficient and interactive software with intuitive and sophisticated user interfaces for medicinal chemists. The included set operations and filter options cover a wide variety of typical tools needed for the identification of suitable candidates. Furthermore, the visualization of molecule sets using structure diagrams provides an easy way to inspect the obtained results. Moreover, the precalculated molecular properties are well known to chemists. The generation of SMARTS expressions is facilitated by the inclusion of the SMARTSeditor [123], a graphical approach to pattern design developed in the same research group.

2.6 Generation and Selection of Protomers

The representation of molecules by a single valence bond structure, as already discussed with respect to consistency for the registration of compounds into databases, can pose significant problems for typical cheminformatics applications. This is especially true when valence bond structures are not only used for the identification of compounds but as a mathematical model of molecular structure. Numerous methods in the field of cheminformatics are based on the general idea of calculating molecular properties as a sum of contributions from atoms or larger structural units [76]. The individual increments associated with these fragments are usually derived from the analysis of experimentally measured values of series of compounds using multilinear regression. In solution, which is the typical environment for these measurements, molecules can, however, undergo rapid transformations such as acid-base reactions or tautomeric rearrangements. These lead to a set of new chemical species which also contribute to the macroscopic properties of the system. This fact gives rise to a number of difficulties concerning both the consistency of the respective approaches and the accuracy of their predictions. Since multiple species can contribute to a macroscopic property, the result of the calculation should not depend on the provided representation of the respective compound, i.e. as input for the method. Readily interconvertible tautomeric forms of a molecule, for instance, should not lead to different predictions when it comes to physicochemical properties such as logP. This problem of alternative molecule forms, however, is not only encountered at the actual prediction but already during the parametrization of the increments. Although the issues with consistency could be circumvented

2. SCREENING LIBRARY

by transforming the input structure into a canonical form, the mere canonicalization is nevertheless not a sufficient strategy in this scenario. Cheminformatics methods are designed to predict properties of unknown compounds on the basis of their topological features. Consequently, if the molecule is not well described by the provided representation, for instance because the chosen form does not correspond to the most energetically stable or otherwise prevalent one, the estimation could be inaccurate [124]. For that reason, a procedure called normalization, which not only produces a unique but also a preferential representation, is usually applied as a preprocessing step. Alternatively, the procedure could intrinsically work with an ensemble of reasonable structures. In both cases a scoring scheme is needed which is able to reliably identify the prevalent forms.

Virtual screening techniques such as molecular docking are another type of application in which the restriction to a single valence bond structure can be problematic. Docking methodologies usually rely on the explicit evaluation of intermolecular interactions such as hydrogen bonds and salt bridges between the binding pocket of a protein and its bound ligand. Each tautomeric form and protonation state of a molecule corresponds to a different topological and spatial distribution of hydrogen atoms and thus can interact differently with the amino acids of the protein. The limitation to a single tautomer or protonation state, even a normalized one, can therefore easily lead to false negative predictions. On the other hand, working with a large ensemble of alternative tautomers and protonation states including unrealistic species can cause false positive results [125, 126]. Virtual screening is thus another application in which a reliable scoring scheme is required.

The necessity to consider tautomers and protonation states, which will be subsumed under the term protomers from this point on, has been ignored in the field of computer-aided drug design for a long time [127]. It is only in the last ten years that the subject of reliable protomer prediction has been starting to move to the forefront of cheminformatics concerns and that efforts have been undertaken investigating its influence on the quality of the results of various virtual screening applications [125, 126, 128]. One major obstacle for the development of such methods was and still is the lack of knowledge on how to predict relative stabilities of protomers in solution even when using high level *ab initio* calculations [129]. This is aggravated by the fact that cheminformatics methods are subject to restrictions concerning their runtime and generally cannot rely on time intensive quantum chemical calculations. Despite these considerable difficulties, multiple protomer generation methods have been developed and published in recent years [124, 128, 130–134]. Two aspects play a major role for the handling of protomers in a cheminformatics context. On the one hand, there is the technical

2.6 Generation and Selection of Protomers

or algorithmic problem of efficiently and consistently enumerating a set of protomeric forms on the basis of the valence bond description. On the other hand, there is the need for a scoring scheme which is able to separate energetically inaccessible species from realistic ones. A discussion of both aspects and the associated problems can be found in [135]. According to Sayle [135], the approaches for the enumeration of protomers can be roughly divided into two categories, local approaches [124, 128, 130, 133] and global approaches [131, 132, 134]. The former rely on predefined transformation patterns which are successively applied to a molecule in order to generate new structures, whereas global approaches systematically enumerate alternatives in previously determined substructures. Both strategies are, in principle, suitable to enumerate sets of protomers for virtual screening approach.

In his publication, Sayle [135] additionally defined five specific tasks associated with the handling of protomers in the context of cheminformatics, namely comparison (#1), canonicalization (#2), enumeration (#3), selection (#4), and prediction (#5). The first two tasks are usually closely related as the generation of a unique representation can be considered as a preprocessing step for the determination of molecular identity. Enumeration corresponds to the mere generation of alternative valence bond structures, whereas selection and prediction typically involve additional scoring procedures. The aim of selection is to generate a set of reasonable but otherwise unordered protomers whereas prediction additionally involves the determination of their relative ratios.

The methods for the handling of protomers employed in the NAOMI framework are based on the valence state combination (VSC) model [D4] developed in the course of this thesis. It provides the basis for the realization of the specific tasks defined by Sayle [135]. The normalization and canonicalization routines of the NAOMI framework reflect the first two tasks. The main difference between the two is that in case of canonicalization, the structure is arbitrary and does not necessarily correspond to an energetically favorable or otherwise preferential structure. Normalization, on the other hand, is used when a preferential representation is needed and thus involves scoring. The other three tasks are part of the protomer generation routines whose aim is to generate a set of reasonable protomers for typical cheminformatics applications. The general principle of the VSC approach is outlined in Figure 2.13, a more detailed description of the concepts, algorithms, and assumptions can be found in the original publication [D4]. As a global approach, the method starts with a subdivision of the molecule into non-overlapping substructures which are then treated independently. In each of those partitions atoms which can change their respective valence states are identified and these additional states are used to enumerate valid valence bond

2. SCREENING LIBRARY

structures. In the final step, a fragment-based scoring scheme is used to identify the best solutions for the respective substructures.

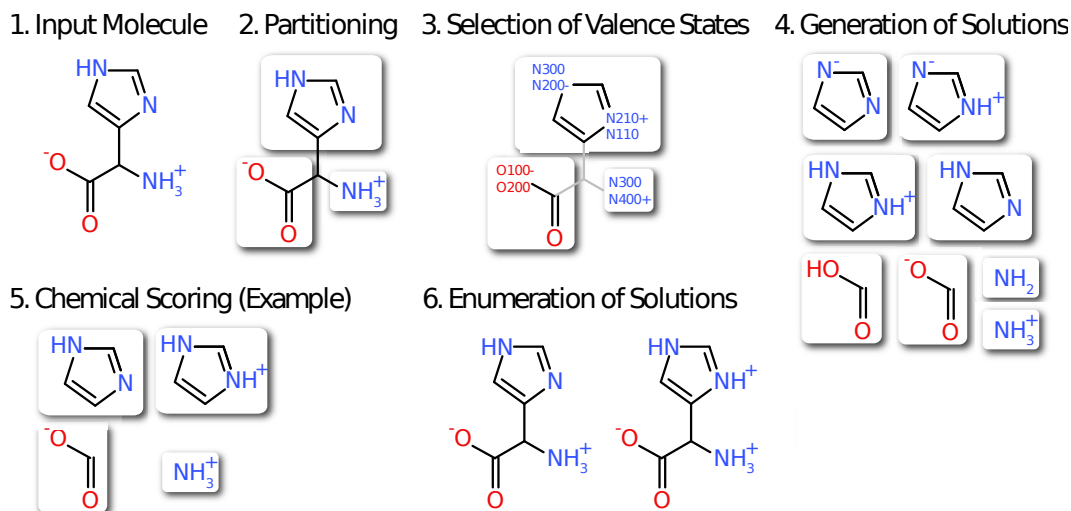


Figure 2.13: General workflow of the VSC approach. Figure adapted from [D4].

In contrast to many other approaches, the VSC model allows to consider all relevant aspects of the problem, namely mesomerism, tautomerism, and ionization in a consistent and comprehensive manner. By offering the means to systematically categorize the different types of transformations, the valence state description provides the conceptual framework for this task. Additionally, it also plays an important part in the generation step where the chemical validity of different valence bond structures needs to be tested. The four steps of the procedure are completely decoupled and can be individually customized in order to solve different tasks. For that reason, the VSC model is not only used for the handling of protomers but also plays an important role in many other applications of the NAOMI framework. It is used during the construction of the atom type layer when equivalent resonance forms (see Figure 2.2) need to be identified. Additionally, it is part of the correction mechanisms of the file formats in case of aromatic bonds. It also plays a pivotal role for the interpretation of molecules from three dimensional coordinates presented in the next chapter.

Several validation procedures show that the general concept is internally consistent and, more importantly, that methods developed on this basis are independent of the valence bond structure used as initial input [D4]. Furthermore, the comparison of the generated results with valence bond representations found in curated datasets shows the excellent performance of the associated scoring scheme [D4]. In combination with

2.6 Generation and Selection of Protomers

the reasonable number of additional protomers and the methods high efficiency it is perfectly suited for application in the context of cheminformatics [D4].

3

Protein Structure

The three-dimensional structure of the target protein is, besides the screening library, the second essential prerequisite for the structure-based virtual screening approach. The prediction of realistic binding poses by any docking engine is contingent upon an accurate and detailed insight into the structure of the protein's binding site on an atomic level. Although there are multiple ways to obtain the necessary structural data for that purpose, e.g., NMR or homology modeling, X-ray crystallography [14] is still the standard method of choice in the context of CADD. The Protein Data Bank (PDB) [136], with currently more than 90000 entries, is by far the most important source of experimentally determined structures of proteins and protein-ligand complexes in the public domain. Additionally, many academic institutions and pharmaceutical companies maintain their own structural biology departments which can provide crystal structures of specific targets that are needed in drug design projects.

The model of the protein used as a basis for virtual screening applications is generally treated as a direct three-dimensional image of its structure by cheminformatics software. Therefore, it has a strong influence on the quality of the results of the calculations and multiple aspects have to be considered in order to ensure an optimal performance. First, there is the quality of the respective crystal structure data. Being an experimental method, X-ray crystallography is subjected to measurement errors and uncertainties which will inevitably be reflected in the generated model. Additionally, the process of resolving the structure of a protein needs a lot of manual intervention and computer-assisted refinement steps which both can easily lead to inaccuracies due to misinterpretations of electron densities or insufficient parameterization of chemical models [137]. Second, there are the conditions under which the structure was determined. Even if the provided model is reasonable error free, crystal structures can

3. PROTEIN STRUCTURE

nevertheless be unsuitable for specific screening contexts. The conditions during the actual measurement could, for instance, be extremely different from the ones encountered *in vivo* so that the protein will most likely behave differently in its natural environment. The same also applies for proteins with and without bound ligands. Upon ligand binding proteins often change their conformations which essentially means the unbound form does not reflect the binding situation that needs to be modeled by the software. Another problem are packing effects which are a result of the experimental method and do not reflect the behavior of the protein *in vivo*. All in all, the evaluation and selection of a suitable structural model for proteins is a complicated process with many pitfalls and generally requires a lot of experience.

Before the actual docking calculation, protein structures generally need to be subjected to preprocessing routines by the docking software in order to generate the necessary data for the prediction of the binding poses. The first step is the definition of the binding site, which can be done either manually or automatically using specialized algorithms [138]. Afterwards, the properties of the binding site have to be calculated including, for instance, a representation of its shape, potential interaction centers, and hydrophobic regions. The specifics of this step strongly depend on the respective docking strategy and the underlying software library. Two aspects usually play an important role during that process. First, there is the assignment of protomeric states for the side chains of the protein's residues, which will be discussed in more detail further down, and the treatment of solvent molecules, first and foremost water. A very common approach with respect to the latter is to simply let the user decide which solvent molecules to consider in the actual docking calculation. The automated classification of water molecules in the binding site as being displaceable or conserved can, however, be helpful in this respect [139].

The following subsections will highlight some of the aspects which play an important role during the processing of crystal structure data for the use as models of the binding site in docking calculations. A more thorough discussion of the topics including additional references to relevant literature can be found in [140].

3.1 Representation of Protein Structures

As was already mentioned in Chapter 1 proteins are biological macromolecules consisting of chains of amino acids and both their description and classification play a central role in the molecular life sciences. Depending on the application context there are multiple aspects which are relevant for the characterization of protein structures,

3.1 Representation of Protein Structures

including the linear sequence of the amino acids (primary structure), the general three-dimensional form of local segments (secondary structure) the overall geometric shape (tertiary structure), or the arrangement of multiple protein chains (quartary structure). In the context of virtual screening, where proteins are in most cases reduced to their binding site, the most common approach is to rely directly on the three-dimensional coordinates of the respective atoms and to generate various algorithm-specific representations starting from this description, e.g., grid representations.

The representation of proteins in the NAOMI software system is based on the same principles that are used for the description of small molecules presented in the previous section. All three layers of the chemical model are constructed for the atoms of the protein chains and can be used in downstream methods and algorithms. The only difference is an additional layer containing information about the individual residues of the protein, e.g. their respective type, a decomposition into side chain and backbone atoms, and the bonds connecting them to other residues. This subdivision is quite common when working with proteins and provides the basis for a number of algorithmic strategies taking advantage of the constant recurrence of the same standard amino acids. Another important difference to small molecules is the fact that, due to insufficient resolution of crystallographic measurements, it is not uncommon for certain portions, e.g., groups of atoms or complete residues, to be missing from the respective structural data. Since proteins are, however, considerably larger than their bound ligands this does not mean that the respective structure is not suitable for cheminformatics applications. In order to ensure a consistent construction of the layers of the chemical model, missing atoms of known residues are topologically added to the protein structure. Such atoms are, however, marked as artificial and can be easily excluded from subsequent calculations if necessary. As binding pockets are in the main focus of virtual screening applications, their representation naturally also plays an important role in the NAOMI software system. Internally, they are composed of a set of residues and a set of molecules which are not covalently bound to the protein chain. Prosthetic groups and covalently bound ligands are treated as residues of unknown type.

Additionally, the NAOMI software system provides a database scheme for the efficient storage of both proteins and binding sites definitions which will be referred to as ProteinDB in the following. The ProteinDB [D9] is based on the same technology as the MolDB introduced in the previous section but has been extended in order to cope with the polymeric nature of proteins. The molecules of the MolDB correspond, in the context of the ProteinDB, to so-called residue templates which represent the topologies of the different types of residues. Since proteins consist of only a small number of different amino acids, this has the advantage that the storage of redundant information

3. PROTEIN STRUCTURE

about the chemical composition of repeating subunits can be avoided. One important difference to the MolDB is, however, that residue templates are only fragments rather than complete molecules so that the recreation of protein structures from the stored data is less straightforward. For that purpose, two additional description layers are needed, one of which corresponds conceptionally to the instances of the MolDB. The occurrences of particular residues in the protein, so-called residue instances, are stored in a separate table (for details see [D9]) and include information about their corresponding residue template, chain, sequence index, and three-dimensional coordinates. As instances in the MolDB, residue instances have different meanings depending on the context. They can either correspond to residues of the same type at different positions in the protein chain or to different conformations of the same residue. Both residue templates and residue instances receive unique keys upon registration into the database which can be used for the retrieval of the associated data. The third description layer, the residue connection, has no counterpart in the MolDB. It reflects the covalent bonds between different residue instances comprising the chain of the protein. As was discussed above, binding pockets in the NAOMI system are represented as lists of residues and molecules. This is directly reflected in the ProteinDB by using lists of residue instances and molecule instances for the storage in the database.

By being based on the same hierarchical chemical model as small molecules, the associated concepts, algorithms, and methods presented in the previous section can be directly transferred to protein structures. According to the concept of the separation of chemical information, the description of substructures as residues is merely an additional layer of the model. On the one hand, this means that the same type of chemical descriptions, e.g., valence states and atom types, are available when working with atom of proteins or residues in the NAOMI framework. Therefore, in many cases it is not even relevant if the respective atoms are part of proteins or small molecules. The handling of protomers in binding pockets presented further down is an example for that. On the other hand, additional information about residues is also available and can be accessed when needed.

3.2 Structural Data of Protein-Ligand Complexes

Crystal structures of protein-ligand complexes play an important role in the drug development process. They provide valuable insights into how and where small molecules interact with the active site of a protein and thus often serve as a starting point for both the development of new or the optimization of already known active molecules. Furthermore, they are an important resource for the statistical analysis of geometrical

3.2 Structural Data of Protein-Ligand Complexes

data, e.g. favorable torsion angles or optimal interaction geometries, which are essential for the parametrization of various common cheminformatics methods, e.g., the generation of small-molecule conformations or molecular docking. They also provide the structural basis for a large number of different computational tasks including the identification of putative binding sites, the prediction of protein function, and the generation of pharmacophores. In the context of structure-based drug design, the interface between protein and ligands, i.e., the binding pocket, is of particular importance. Unfortunately, for a long time the main efforts of crystallographers were directed towards the generating reasonable protein conformations whereas bound molecules did not receive the same attention [141]. As a result, ligand structures in crystallographic data are often poorly modeled and can even contain incorrect geometries [141, 142]. In order to compensate for these shortcomings, robust cheminformatics methods are needed which can help with the interpretation of questionable molecular models or even correct them if necessary. With respect to certain questions, the analysis small-molecule crystal structures, for which the Cambridge Structural Database (CSD) [143] is by far the most relevant repository, can also be a viable alternative.

As was already mentioned above, the PDB is the most important source for crystallographic data of protein-ligand complexes. The structural data in the PDB can be accessed via specialized file formats [144] containing only element identities and three-dimensional atomic coordinates as reliable chemical information. This clearly distinguishes them from chemical file formats described in previous sections, which were based on the representation of molecules by valence bond structures. Consequently, fundamentally different procedures are necessary for the interpretation of the chemical information provided by these formats. The main task is the translation of the three-dimensional data into a valid valence bond structure, which involves, simply put, the identification of covalent bonds between atoms and the subsequent assignment of bond types. Although this might seem straightforward at first glance, there are three aspects that considerably complicate this task. First, there are the general shortcomings of valence bond structures (for examples see Figure 2.1) which prevent a simple mapping of bond lengths to bond orders. Benzene with its alternating bonds of equal length is a prototypical example for that. Second, there is the requirement of chemical validity which essentially means that certain bond order distributions are not possible for existing molecules. This must be considered in order to avoid the generation of completely artificial valence bond structures. Third, the atomic coordinates provided by the files are a result of experimental procedures and thus can contain a certain degree of uncertainty. Therefore, deviations from ideal geometries are to be expected and must be tolerated to a certain extent. In case of biopolymers such as proteins and nucleic

3. PROTEIN STRUCTURE

acid the interpretation of data from PDB files can be considerably facilitated by relying on predefined structural templates for the relatively small number of standard residues. The necessary information for the unambiguous assignment of individual atoms to particular residues is also provided in the coordinate section of the file format. After the construction of the isolated monomers, these need to be connected in order to obtain the complete polymer chain. Since the type of functional group connecting residues is also known in advance, e.g., an amide group in case of proteins, this can also be realized with recourse to predefined patterns. Since this procedure is, however, obviously not feasible in the absence of templates for a particular type of residue, which is often the case for non-standard residues, cofactors and bound ligands, the above mentioned coordinate-based routines are essential.

Since the interpretation of small molecules from three-dimensional coordinates is a necessary task when working with protein crystal structures from the PDB, multiple methods have been developed for that purpose [145–150]. These can be classified into two separate categories depending on how the assignment of hybridizations and bond orders are handled. In approaches from the first class, both assignments are handled separately which usually requires an additional step in which potential inconsistencies are resolved. This can, for instance, be realized with the help of substructure matching and predefined patterns. In contrast to that, approaches from the second class derive bond orders from a previous assignment of atomic hybridization using different optimization strategies and bond localization routines. Although the published methods already perform well on many different classes of molecules found in PDB entries, one important aspect is generally neglected, namely the possibility to represent molecules by different valence bond structures. As was explained above, the problem is complicated by the fact that this usually involves the adherence to certain conventions for the representation of molecules as well as the consideration of the inherent stabilities of particular protomers. In combination with the above mentioned problems with respect to the quality of ligands structures in PDB data, this essentially means that the best suited valence bond structure often cannot be reliably determined on the basis of three-dimensional coordinates alone, which is, however, the working assumption of most of the published methods.

The individual steps of the procedure for the perception of molecules from three-dimensional atomic coordinates included in the NAOMI framework are shown in Figure 3.1. A more detailed description of the method can be found in the original publication [D5]. In the first step, covalent bonds are identified on the basis of interatomic distances. Based on the number of these bonds and the geometric arrangement of the connected atoms potential valence states are identified for each atom. For each of these

3.2 Structural Data of Protein-Ligand Complexes

assignments a confidence score is calculated reflecting the agreement to the atom's local environment including both geometric and chemical criteria. Based on these alternative valence states multiple possible valence bond structures are calculated and scored using a similar procedure as the one described for the generation of protomers in [D4].

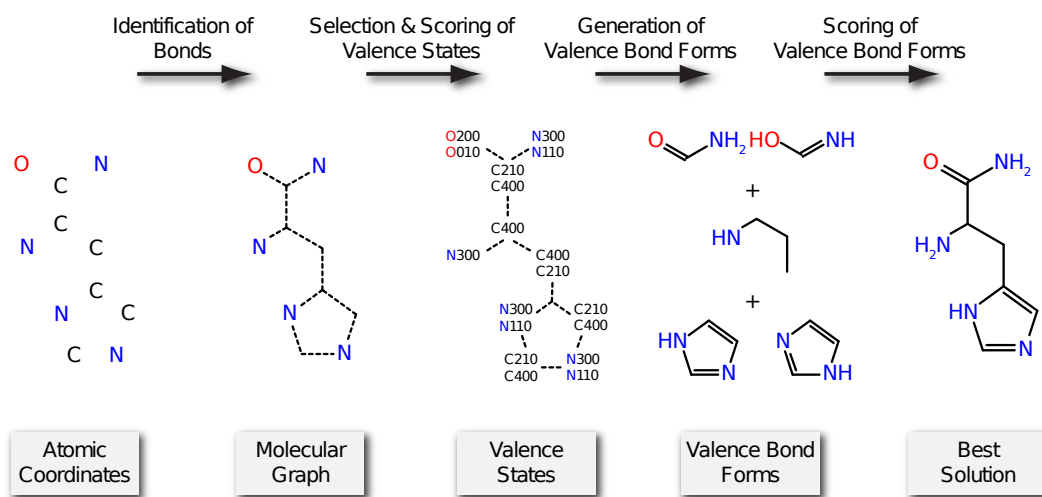


Figure 3.1: General workflow of the perception of molecules from three-dimensional coordinates. The figure was taken from [D5].

As with many other methods in the NAOMI framework, the valence state concept plays a central role for the interpretation of molecules from three-dimensional coordinates. During the perception of covalent bonds it helps to ensure the chemical validity of the resulting molecular scaffolds. If no valence state could be found which is in agreement with the number of partners bound to a particular atom, the molecule is discarded. In this way errors resulting from severe geometric distortions, which often result in the formation of highly cyclic structures, can be avoided in most cases. Additionally, the enumeration of valence bond structures based on the valence state combination methodology guarantees that only valid molecules will be generated by the procedure. While the assignment of confidence values reflecting the compatibility with the atom's local environment enforces a good agreement with the provided three-dimensional coordinates, the subsequent scoring procedure makes sure that the resulting structures are chemically reasonable at the same time. In contrast to previously published approaches the possible existence of alternative representations is explicitly considered and handled on the basis of the same methods which were developed for the selection of reasonable protomers.

The perception workflow of the NAOMI framework has been evaluated by comparison of the resulting molecules to reference structures in two different contexts [D5]. On

3. PROTEIN STRUCTURE

the one hand, an extensive analysis involving all small molecules deposited in the PDB shows that the presented method is able to reproduce the expected results in the vast majority of the cases even when distorted geometries are encountered. The reason for deviations have been carefully investigated and could in most cases be attributed to inconsistencies in the provided coordinates which could not be unambiguously resolved. On the other hand, the comparison to the results from other previously published methods shows that many of their errors can be avoided by the NAOMI approach. In addition to the better performance in respect to quality, the method is also very efficient and allows the processing of all small molecules contained in the PDB in mere minutes.

3.3 Protomers in Protein-Ligand Complexes

The formation of hydrogen bonds plays a crucial role in stabilizing the three-dimensional structure of biological macromolecules and is also an important factor governing the interactions of proteins with their bound ligands. Knowledge about their frequency, strength, and internal geometry, especially with respect to the involved atoms and functional groups, is therefore of vital importance for the understanding of molecular recognition as well as for the rational design of new drugs. Unfortunately, as light atoms generally display only weak contributions to diffraction, the usual resolution of X-ray protein crystallography is not sufficient to reliably determine the positions of hydrogen atoms. This information, however, is necessary for the systematic investigation of hydrogen bonds and their geometric properties in different proteins and their respective complexes. Additionally, the presence of hydrogen atoms allows to determine the protonation states and tautomeric forms of both the residues of the binding site and the bound ligand which cannot be reliably derived from the coordinates of heavy atoms. Although the PDB does contain a small number of protein structures from high-resolution measurements which include resolved hydrogen coordinates, the vast majority of the deposited structures does only provide information about the positions of heavy atoms. In order to be able to work with this kind of data, the automated addition of hydrogen atoms has become a very common step in crystallographic structure refinement.

Apart from the statistical analysis of protein-ligand complexes, the explicit representation of hydrogen atoms and their positions is also a necessary prerequisite for all cheminformatics application dealing with hydrogen bonding, e.g., docking and other virtual screening approaches. Due to the large number of degrees of freedom resulting from different types of functional groups, this task is anything but trivial. Diverse

3.3 Protomers in Protein-Ligand Complexes

aspects such as freely rotatable terminal groups, tautomers and protonation states, alternative water orientations and terminal side chain flips have to be considered in order to obtain a realistic prediction (see Figure 3.2 for examples). In particular, the inclusion of ionization and tautomerism of ligands, due to their larger structural variety compared to the small number of standard residues, makes the problem complicated from a chemical point of view. Furthermore, there are many situations in which the orientation or even presence of hydrogen atoms can only be deduced by consideration of the surrounding chemical moieties. This in turn can lead to a high degree of mutual dependency which additionally adds to the complexity of the problem. Typically, algorithms start with assigning initial hydrogen positions on the basis of idealized geometries and then try to optimize their orientation, for instance by maximizing the number of hydrogen bonds and by reducing internal clashes on the basis of either energy-based functions or heuristic strategies [151].

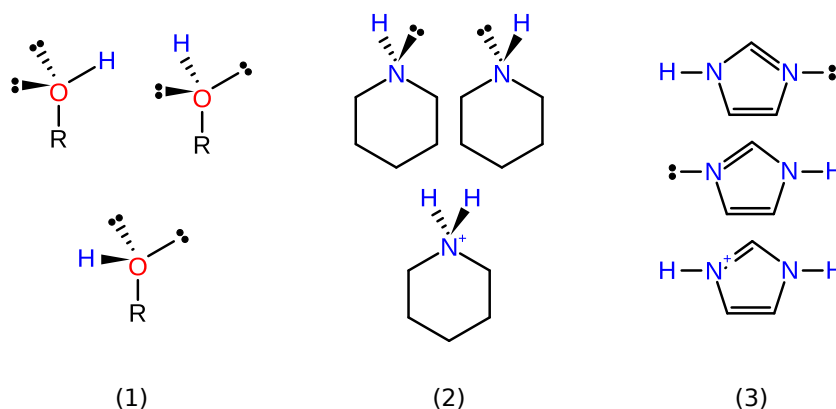


Figure 3.2: Functional groups with variable hydrogen positions. Alcohols (1) are an example for freely rotatable groups. The cyclic secondary amine in (2) can either be protonated or neutral. In case of the latter the attached hydrogen atom can occupy two distinct positions. The imidazole ring (3) has two tautomeric form in its neutral form but can also be protonated under physiological conditions. The figure was taken from [D6].

Due to the importance of the automated processing of protein structures in CADD, it is not surprising that a large number of different methods for the prediction of hydrogen positions in protein complexes have been developed. A thorough review of those can be found in [151]. Despite considerable differences in their subjective functions and optimization algorithms, the degrees of freedom covered by these approaches are rather similar. Typically, the treatment of amino acids in proteins is quite comprehensive, whereas variations in the protonation states and tautomeric forms of the ligands are often ignored. Only a few more recent methods consider this aspect to some extent

3. PROTEIN STRUCTURE

[152, 153]. Considering the fact that different protomers can interact differently with the residues of the protein, ignoring them, or, more precisely, choosing the wrong one, can have detrimental effects on the results of all subsequent calculations. The prediction of suboptimal hydrogen bonding networks or the introduction of hydrogen clashes are but a few possible consequences (see Figure 3.3 for examples).

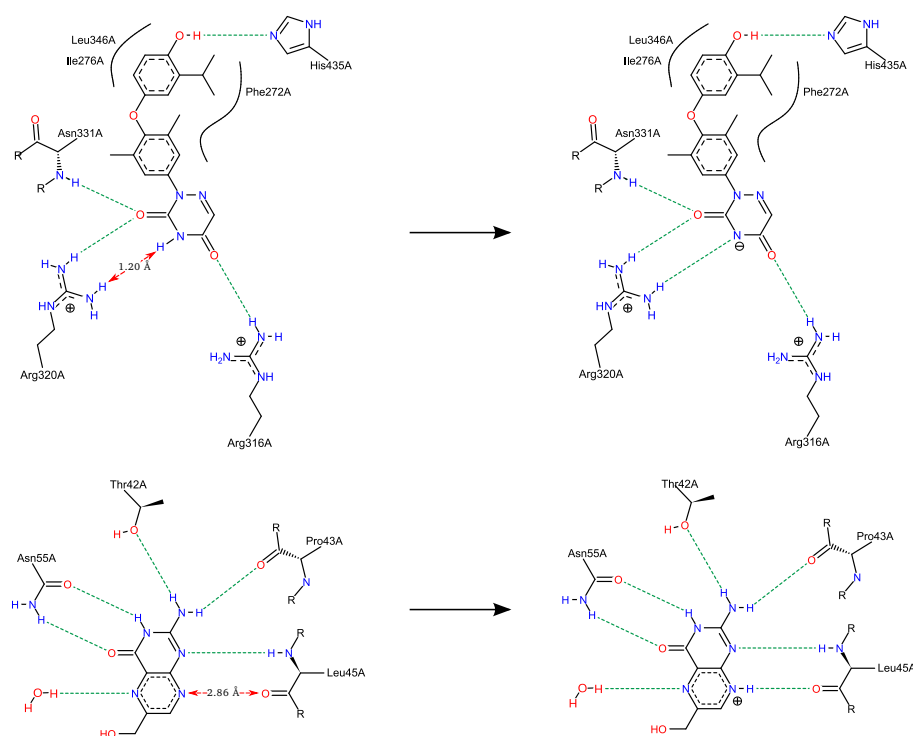


Figure 3.3: Two examples for protein-ligand-complexes in which the omission of relevant ligand protomers leads to hydrogen clashes or suboptimal hydrogen bonding-networks. The red arrows on the left side indicate unfavorable interactions resulting from either hydrogen clashes (top) or acceptor contacts (bottom). The optimal hydrogen network is shown on the right side. Both examples were taken from the evaluation study presented in [D6].

The functionality for the prediction of hydrogen positions in protein-ligand complexes in the NAOMI framework [D6] builds on the Protoss methodology previously developed in the same research group [154]. Apart from being completely based on the chemical and interaction model of the NAOMI framework, which the initial version of Protoss was not, the most important extension is the comprehensive treatment of alternative protomeric forms on the ligand side. The aim of the extended version is thus to predict an optimal hydrogen bonding network under consideration of the relative stabilities of the involved chemical moieties. An overview of the associated workflow

3.3 Protomers in Protein-Ligand Complexes

is shown in Figure 3.4, a more detailed description can be found in the original publication [D6]. The first step is almost identical to the procedures for the generation of protomers described in the previous section. The residues of the protein and the ligand molecule are partitioned into independent substructures for which both alternative hydrogen positions (see Figure 3.2) and reasonable protomeric states are generated. For each state in each substructure potential hydrogen bonding patterns are identified on the basis of interaction surfaces which will be discussed in more detail in the next chapter. An optimization procedure based on a dynamic programming approach [154] is then used to find the best distribution of hydrogen atoms under consideration of both the quality of the formed hydrogen bonds and the stability of the respective protomers.

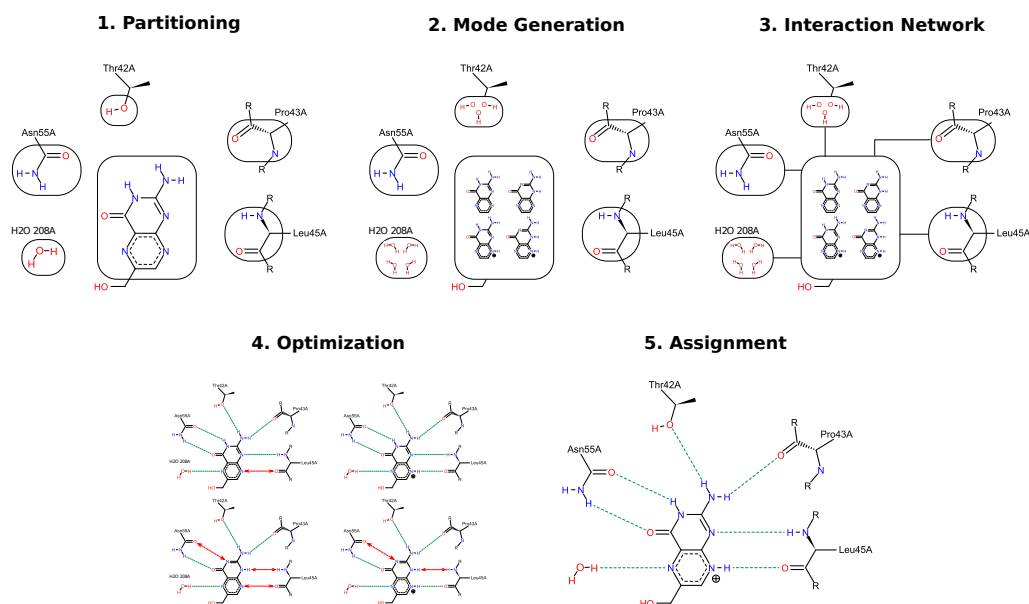


Figure 3.4: General workflow of the identification of the optimal hydrogen bonding network in protein-ligand complexes.

Protoss can be considered as a typical cheminformatics application in which the enumeration of a reasonable set of protomers without an exact prediction of the associated ratios in solution is needed. The actual states of both ligands and residues strongly depend on the respective local environments and are predicted from a combination of both hydrogen bond strengths and rather general stability considerations. By relying on the VSC model [D4], Protoss is the only existing approach which is able to handle the wide variety of chemical moieties contained in ligands in a generic manner. The gen-

3. PROTEIN STRUCTURE

eral necessity of such a treatment was demonstrated by a thorough analysis of different tautomerizable and ionizable substructures contained in small molecules deposited in the PDB [D6]. The performance of Protoss has been extensively investigated and compared to previously published approaches on the basis of different validation procedures [D6]. The most important criteria for the assessment of the quality of hydrogen prediction tools are the frequency of undesirable interactions, e.g., hydrogen-hydrogen clashes and acceptor-acceptor contacts, and the deviation from the expected protomeric forms. In both respects, Protoss outperforms existing approaches which is mostly due to the fact that it is the only approach which is able to handle protomers comprehensively. Additionally, it is highly efficient and thus allows interactive workflows.

4

Virtual Screening

After the compound library and the binding pocket have been carefully chosen and prepared, the actual screening calculation can be performed. Molecular docking forms the methodical core of the structure-based virtual screening approach. Its aim is to predict the way in which molecules bind to the target protein, the so-called binding mode, in combination with a score value reflecting the corresponding binding affinity. As the predictive power of different docking methods strongly depends on the target protein, which has been shown in various comparative studies [155–157], an appropriate strategy must be carefully chosen in advance. This can be done, for instance, by evaluating case studies found in the literature or by performing simple validation tests such as redocking experiments. If the method in question is not able to reproduce the binding mode of the molecule observed in the crystal structure of the protein-ligand complex, the chances of finding reliable new leads are rather low. Depending on the target, aspects such as the treatment of protein flexibility and solvation effects by the respective method can also play an important role in this decision.

The number of molecules that can be processed by virtual screening approaches usually exceeds by far the available capacities for experimental testing so that the results often need to be further prioritized in order to select the most promising candidates. Unfortunately, the scores provided by the internal scoring functions of docking methods are in general not sufficiently accurate to permit decisions of that kind. It has been shown in different studies [158, 159] that the virtual screening efficacy, i.e., the ability to discriminate true binders from inactive molecules, is far from ideal and the intrinsic propensity for the generation of false-positive results is relatively high [65]. In order to cope with these limitations, the hit lists from docking are usually subjected to additional postprocessing steps. Typical strategies are the rescoring on the basis of more advanced scoring schemes [160], the application of consensus scoring [161], the use of

4. VIRTUAL SCREENING

additional filter criteria [162], and the optimization of the binding poses using specialized force-fields [163]. Alternatively, information about known binders can be used to reduce the number of hits by filtering binding poses on the basis of pharmacophoric constraints [164]. However, despite all the progress that has been made with regard to the development of automated procedures, one of the most important postprocessing steps is still the visual inspection of the results by an experienced computational medicinal chemist.

In general, structure-based screening is an extremely complex endeavor involving a large number of different algorithms and concepts. The following subsections will only highlight a small fraction of the problems associated with the prediction of accurate binding poses in protein-ligand complexes. A more thorough introduction including discussions of topics not covered here, e.g., protein flexibility, scoring, and fragment-based approaches, can be found in [165].

4.1 Representation of Molecular Interactions

Medicinal chemists seek to design biologically active molecules by optimizing their potential interactions with the binding pocket of target proteins on the basis of its three-dimensional structure. In order to do so, extensive knowledge about the geometries and the individual affinity contributions of particular types of interactions is of the highest relevance. The major contributing factors for the binding of molecules to proteins have already been introduced and discussed in Chapter 1. Considering the multitude of fundamentally different effects governing the formation of protein-ligand complexes, it becomes obvious that the accurate prediction of the respective association energy is an extremely complex problem. In order to provide a useful estimation for applications in the context of structure-based drug design the introduction of specific approximations cannot be avoided. The most important concepts are the reduction of the considered effects to a well-defined set of dominant contributions by particular interaction types and the assumption of their respective additivity. The quality of the interactions in a protein-ligand complex is usually assessed using a scoring function which combines the various contributions in combination with their weights into a functional form. Although not all aspects of protein-ligand binding, e.g., desolvation or weak interactions, are modeled by such functions, they provide a reasonable starting point for the identification of potentially bioactive molecules.

There are three different approaches to modeling the interactions of proteins and ligands in cheminformatics applications which have been extensively reviewed in the literature [6, 166, 167]. Force-field based methods rely on classical mechanics and try

4.1 Representation of Molecular Interactions

to describe the various aspects of ligand binding using specific individual contributions including bond-stretching, angle-bending, torsional strain, electrostatic, and van-der-Waals terms. The necessary parameters for the respective terms are usually derived from physical measurements or *ab initio* calculations. Knowledge-based approaches use information about frequently observed pairs of atoms to assess the quality of interactions in protein-ligand complexes. The underlying potential, the so-called potential of mean force, is derived from the frequency distribution of particular pairs found in protein crystal structures using Boltzmann statistics. Empirical scoring methodologies are based on the evaluation of localized and chemically intuitive interaction types with a particular focus on the evaluation of their structural and geometric properties. Since only the latter approach plays a role in the NAOMI framework, the following discussion will be restricted to the empirical description of interactions.

Due to their fundamental role in biological recognition processes, the modeling of hydrogen bonds is a central problem in docking and screening applications. As is known from the analysis of crystal structure data, they usually adhere to strict rules with respect to their geometric properties including both distances and angular distributions. For that reason, hydrogen bonds are often scored on the basis of individual donor-acceptor pairs which fall into a given distance and angle range favorable for hydrogen bonding. The score for such an individual interaction is additionally scaled by a function that accounts for deviations from idealized standard values. Some scoring schemes also distinguish between hydrogen bonds involving different types of atoms or functional groups. The same general principles can also be applied to the evaluation of ionic and metal interactions. The former are often handled as charge-assisted hydrogen bonds which in some cases are associated with higher weight in order to reflect the additional electrostatic attraction. With respect to the latter some scoring functions also consider the specific coordination geometries associated with particular metal ions. Hydrophobic contributions to affinity are usually estimated on the basis of the proximity of specific types of atoms, often called hydrophobic contacts, in the protein-ligand complex. The foundation of this procedure is the classification of atoms by their hydrophilic or hydrophobic character and the subsequent identification of matching pairs in close vicinity of each other. In some cases the unfavorable contributions of mismatched contacts are also considered. In order to enable a more specific evaluation, the scoring scheme can be based on surfaces rather than individual atoms. In this way the contact area buried upon complex formation can be estimated, thus providing a more accurate measure. Empirical scoring functions can additionally include terms such as lipophilic and aromatic contributions, loss of ligand flexibility, and in some cases also desolvation effects. The individual terms of the scoring function are scaled to explain

4. VIRTUAL SCREENING

experimentally determined dG values. A more thorough overview of different scoring concepts is provided in [167], a list of the most important scoring functions with the corresponding original publications can be found in [168].

The representation and evaluation of hydrogen bonds in the NAOMI framework is based on an empirical description of molecular interactions. The underlying concept is derived from the interaction model of the FlexScore scoring function [169] and has been further adapted to suit the needs of the different screening applications presented in this thesis. Hydrogen bonding is described in terms of interaction surfaces which are assigned to hydrogen bond donors and acceptors and whose mutual orientation in space provides the basis for the assessment of the respective quality. Each interaction surface corresponds to either one specific hydrogen atom or free electron pair thus reflecting the ability of atoms to partake in multiple hydrogen bonds simultaneously (see Figure 4.1). While the number of donor interactions is obviously determined by the number of bound hydrogens, the number of acceptor surfaces needs to be derived from the hybridization of the respective atom. Such information is provided by the atom types of the chemical model.

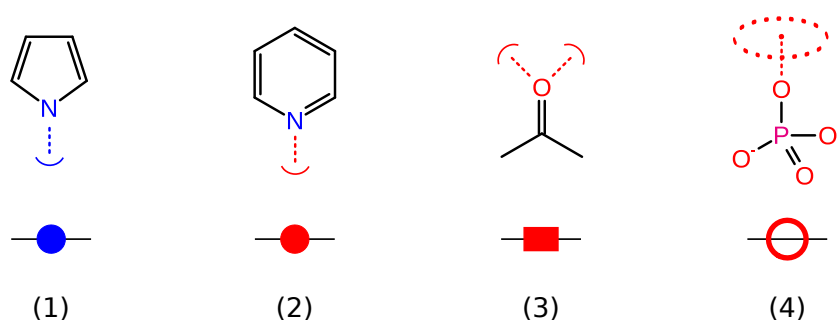


Figure 4.1: Interaction surfaces for different types of atoms. Donor interactions are depicted in blue, acceptor interactions in red. The geometrical shape of the respective surfaces in reference to the plane of the drawing is indicated below the structural diagrams.

The general scheme for the evaluation of the quality of hydrogen bonds between donors and acceptors is shown in Figure 4.2. The main direction, which corresponds to the orientation of either a hydrogen atom (donor) or a free electron pair (acceptor), determines the location of the respective binding partner in an ideal hydrogen bond. Deviations from this optimal arrangement are penalized using a function including terms for both distance and directions. During the calculation of interaction scores, three chemical types of hydrogen bond acceptors are differentiated. First, there are cases which are essentially treated as hydrogen bond donors. A typical example is the aromatic nitrogen in pyridines (2). Second, there are cases in which an additional

4.1 Representation of Molecular Interactions

reference direction is needed in order to reflect additional constraints imposed by the the ideal interaction geometries. For instance, the electron pairs of the oxygen atom in carbonyl groups (3) lie, due to its sp^2 hybridization, in the same plane as the atoms of the functional group. In this case the scoring function needs to contain an additional term penalizing deviations from the plane. Third, there are acceptors with an sp^3 hybridization in which the orientation of electron pairs is arbitrary (4). These can be considered as 'rotatable', meaning that any orientation on the orbit of the respective electron pairs is acceptable. The chemical type of an interaction is also annotated in the interaction surface and is considered during the scoring procedure. Interactions of metal atoms are, in principle, handled in the same way as hydrogen bond donors. The associated directions are derived from the coordination geometry determined by the analysis of the surrounding heavy atoms.

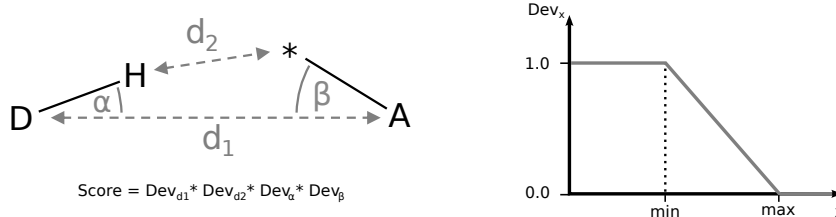


Figure 4.2: Scheme for the evaluation of interactions. The final score is calculated as the product of deviation factors derived from four different geometric parameters. Each deviation factor can have a value between zero and one which is determined using the step function depicted on the right side.

Not every atom in a molecule or protein has a propensity to form stable hydrogen bonds, which naturally must be reflected during the assignment of interaction surfaces. In most applications it is sufficient to restrict the evaluation to the strong hydrogen bonds involving nitrogen and oxygen atoms as it is done, for instance, in Protoss. Although the NAOMI interaction model is conceptionally able to handle additional types of hydrogen bonds, there is currently no case in which this is actually needed.

The presented interaction model is the foundation of applications in which directed polar interactions need to be evaluated and thus plays a key role in virtually all screening methodologies of the NAOMI framework. As was the chemical model with respect to molecules and proteins, the interaction model has been designed to provide relevant structural information and chemical descriptions needed for the development and implementation of different methods and concepts involving polar interactions. The clear separation between the description of individual interaction partners and the evaluation of the quality of hydrogen bonds is an important decision in this respect. Although

4. VIRTUAL SCREENING

the model constitutes a particular concept for the description of hydrogen bonding, the data provided is still flexible enough to be used in different contexts, e.g., the evaluation of hydrogen bonding networks or the generation of interactions triangles.

4.2 Molecular Docking

As has already been mentioned above, molecular docking is at the center of the structure-based screening approach. The underlying algorithmic problem is to find a spatial orientation of the molecule which geometrically fits into the cavity of the binding site and which at the same time corresponds to a potential bioactive conformation, i.e. the actual conformation of the ligand in its bound state. This involves the simultaneous consideration of the inherent degrees of freedom resulting from the translation and rotation of molecules within the binding pocket, their inherent conformational flexibility and the effects governing favorable interactions between proteins and their bound ligands. In order to fulfill these requirements, docking engines generally comprise two components, a search strategy and a scoring function, which, depending on the underlying algorithmic concepts, may be strongly intertwined. Specialized search strategies are needed in order to sample the search space with optimal efficiency, as an exhaustive exploration, even if the flexibility of the protein is completely ignored, is generally not possible with current computing resources. The approaches used to cope with the inherent flexibility of small molecules in the context of molecular docking are quite diverse and will be introduced further down. The flexibility of proteins, although it plays an important part during ligand binding, especially when considering the induced-fit mechanism (see Chapter 1), is in many cases ignored due to the enormous increase in algorithmic and conceptual complexity this additional degree of freedom entails. The number of docking tools explicitly dealing with this difficult problem has, however, been growing over the last years [170]. The purpose of the scoring function is to predict whether the current orientation of the molecule does in fact correspond to a realistic binding mode considering potential interactions with the target protein. Scoring is a central aspect of any docking approach since the geometric fit of the compound into the protein's binding pocket alone is generally not sufficient to constitute a bioactive conformation. Assessing the stability of protein-ligand complexes, as was explained in Chapter 1, is extremely complex since a wide variety of fundamentally different effects needs to be accurately modeled in order to produce reliable results. In the context of screening applications this situation is additionally complicated by the fact that large numbers of molecules need to be processed which considerably reduces the acceptable time frame of such computations. Due to the fundamental importance of scoring functions in the

docking process, it is not surprising that a large number of different approaches have been developed over the years, an overview of which can be found in [167, 171, 172]. The trade-off between accuracy and efficiency, which is necessary to efficiently calculate scores for hundreds of thousands of compounds in a reasonable time, often only permits the use of rather basic scoring functions during the actual docking calculations. As a consequence, the obtained results need to be carefully analyzed and possibly reworked in a postprocessing step. Since the scoring of protein-ligand complexes is not in the scope of the present thesis, the following discussion will be focused on strategies for the placement of ligands. A thorough discussion of scoring functions and additional literature references can be found in [167].

In case of fragment-based docking methods, molecules are partitioned into rigid fragments which are then reconstructed inside the binding pocket of the protein. During the addition of each new fragment, the conformational space of the joined components is explored under explicit consideration of the surrounding residues of the protein. In this way, the generation of a bioactive conformation can be guided directly by the geometric and physicochemical properties of the binding pocket. The most common algorithmic strategies applied for this approach are incremental construction [169, 173] and place-and-join [174, 175]. Stochastic methods start with the initial placement of an arbitrary ligand conformation in the binding site. This provides a starting point for a series of random translations, rotations, and variations of torsion angles with the aim of finding a bioactive conformation of the molecule. These random changes are generated using different strategies, the most important among them being the Monte Carlo method [176] and genetic algorithms [177, 178]. Multiconformer methods enumerate a set of conformers prior to the actual docking calculation, thus completely separating the handling of flexibility from the placement of the ligand in the binding pocket. The latter is realized in a second step using rigid docking, e.g., on the basis of shape complementarity [179]. Due to the strong dependency on the quality of the precalculated conformations, post-optimization routines are a common means to further improve the initial results [180].

The docking engine based on the NAOMI framework presented in the following sections is a further development of the TriX approach [181, 182] previously developed in the same research group. As the underlying screening procedure relies on a descriptor-based bitmap search, the technology is referred to as rapid index-based screening engine (RAISE). Its most prominent feature is the interaction triangle descriptor [D7], called RAISE descriptor in the following, which plays a central role in various screening applications. RAISE descriptors are derived from interactions surfaces of polar atoms (see 4.1) and undirected hydrophobic interaction sites by forming triangles on the basis of

4. VIRTUAL SCREENING

different combinations of triplets. Each corner of the descriptor triangle corresponds to one particular interaction site and encodes information about its type (donor, acceptor, hydrophobic) and its directions in case of polar interactions. As metals generally interact with acceptor type atoms, they are internally handled as hydrogen bond donors. Each triangle additionally includes a description of its local geometric shape based on eighty rays originating from its center (see Figures 4.3 and 4.4).

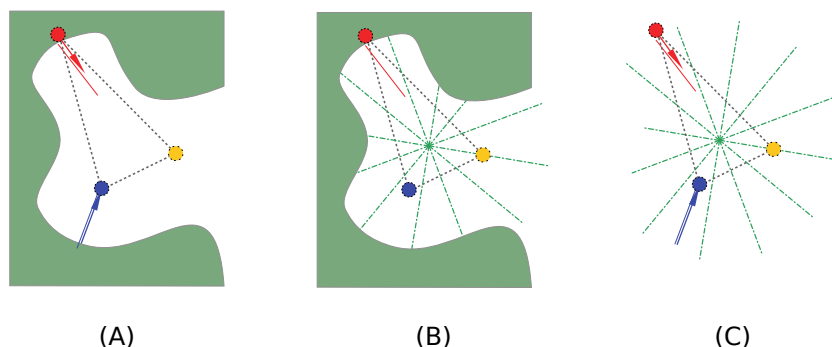


Figure 4.3: RAISE descriptors for binding sites. Each corner of the triangle corresponds to one particular interaction type and has an associated direction in case of polar interactions (A). The rays of the shape descriptor originate from the geometric center of the triangle and represent the shape of the cavity (B). The final descriptor (C) is an abstract representation of interaction patterns and does not encode information about the structure it was derived from.

The positions of the triangle corners depend on the respective interaction types and the chemical objects they are generated from (see Figures 4.3 and 4.4). In case of hydrogen bond donors they are placed at a distance of an idealized hydrogen bond (2.8 Å) away from the heavy atom in the direction of the corresponding hydrogen atom. For hydrogen bond acceptors they reside on the corresponding heavy atom. The directions associated with the triangle corners of polar interactions correspond to the main directions of interaction surfaces (see 4.1). While hydrogen bond donor corners still represent one particular hydrogen atom and thus one particular direction, all acceptors interaction surfaces of the respective heavy atom are contracted into a single triangle corner with multiple directions. The positions of hydrophobic interaction sites are determined differently for molecules and binding pockets [D7]. In case of the former they are placed on aromatic rings, aliphatic carbons, and halogen atoms and are thus situated inside the respective molecule. For the latter they correspond to points inside the cavity which are mainly surrounded by hydrophobic atoms and are identified using a grid-based procedure. The different location of interaction triangles is directly reflected in the rays of the associated shape descriptors. For molecules they correspond

to the van-der-Waals volume, while they represent the interior volume of the cavity for binding sites.

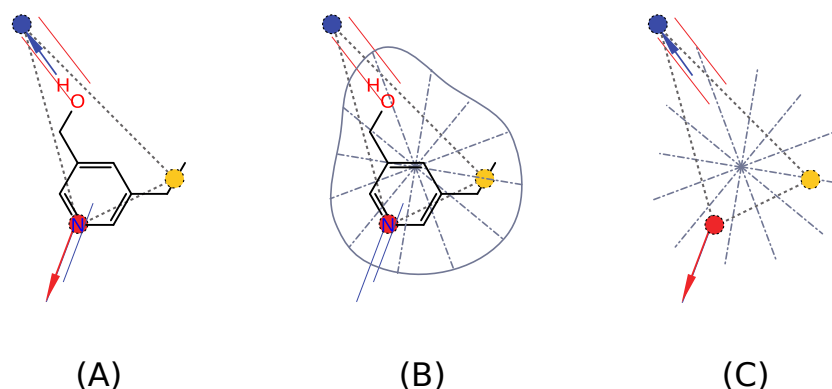


Figure 4.4: RAISE descriptors for molecules. Each corner of the triangle corresponds to one particular interaction type and has an associated direction in case of polar interactions (A). The rays of the shape descriptor originate from the geometric center of the triangle and represent the van-der-Waals volume of the compound (B). The final descriptor (C) is an abstract representation of interaction patterns and does not encode information about the structure it was derived from.

The decision whether a molecule fits into the binding pocket of a protein is based on a comparison between the respective RAISE descriptors. Since hydrogen bonds can only be formed between donors and acceptors, triangle corners representing polar interactions must have both complementary types and opposite directions. Additionally, the side lengths of the triangles and their associated shape descriptors must be compatible. With respect to the latter this means that all rays from the molecule descriptor must be shorter than the corresponding ones from the site descriptor in order to avoid steric overlap with the atoms of the protein. The individual steps are summarized in Figure 4.5, a more detailed description of the matching process can be found in [182]. Ligand poses are generated by superposing matching triangles and applying the resulting affine transformation to the respective coordinates of the molecule. As the latter is rigidly placed into the binding pocket of the protein and the shape descriptor only captures local features of both interaction partners a number of additional procedures are necessary to ensure a reasonable binding mode. For this purpose, a hierarchical pose filtering and scoring scheme is applied which, on the one hand, efficiently eliminates poses with sparse contacts or clashes and, on the other, rapidly assesses the quality of the fit between ligand and receptor. A detailed description of the scoring procedure can be found in [D7].

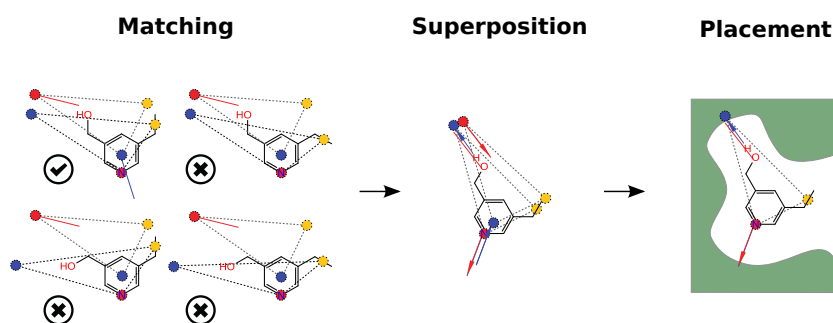


Figure 4.5: General workflow for the generation of docking poses.

The RAISE technology and the associated docking method (cRAISE) are, in contrast to the older TrixX engine, completely based on the NAOMI framework and can thus benefit from its consistent description of both molecules and proteins. Although RAISE has adopted the general principles of the old approach, there are considerable differences with respect to the underlying interaction model. The most prominent one is the integration of the inherent flexibility of rotatable terminal groups into the triangle descriptors for both binding pockets and molecules. In case of donors, i.e., rotatable hydrogens, multiple triangle corners are generated by the discretization of the orbit resulting from the rotation of the hydrogen atom around the main axis of the associated hetero atom. The associated procedure is quite similar to the generation of modes in the context of Protoss. Acceptors, on the other hand, are contracted into a single corner with multiple directions. Furthermore, the strategy for the generation of hydrophobic corners in the binding pocket of proteins has been completely exchanged. Although the influence of each individual modification has not been investigated, the poses generated using the RAISE [D7] technology are generally of higher quality than those of the TrixX approach [182]. As redocking experiments and enrichment studies show, the docking performance of RAISE is comparable to those of common docking tools [D7]. One has to keep in mind, however, that the intended purpose of the RAISE technology is not high-precision docking but large-scale virtual screening which will be presented in the next sections.

4.3 Structure-Based Pharmacophores

The concept of the pharmacophore plays an important role in rational drug design and provides the basis for a number of established virtual screening techniques. The currently accepted definition was given by Wermuth et al. [183]: “A pharmacophore is

4.3 Structure-Based Pharmacophores

the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger or block its biological response”. Pharmacophore models are thus a compact representation of the chemical features which are thought to be relevant for the interactions of proteins with their respective ligands. In general, they comprise elements such as hydrophobic regions, aromatic rings, hydrogen bond donors and acceptors, and positively and negatively charged groups. Pharmacophores are often used to find compounds which can form similar interactions with a particular target without restricting them to specific molecular structures (scaffold-hopping). As has been mentioned in Chapter 1 pharmacophores can be derived from both superpositions of multiple active ligands (ligand-based) and interaction patterns observed in protein-ligand complexes (structure-based). A general introduction to pharmacophore methods is beyond the scope of this thesis and the following discussion will be restricted to their applications in structure-based virtual screening. A comprehensive overview of different methodologies and algorithms can be found in [41, 184].

In the context of structure-based virtual screening, pharmacophore models can be useful in many different ways. Since they essentially incorporate the same information as molecular docking, e.g., hydrogen bonds and hydrophobic interactions, pharmacophores provide the basis for specialized screening methodologies of their own [185]. These are, due to the reduced description of the relevant properties in the binding pockets, in most cases considerably more efficient than workflows involving docking calculations. Therefore, they can serve as an efficient prefilter to reduce the size of the screening library prior to the application of computationally more demanding techniques. Pharmacophoric constraints can, however, also be used as a postfilter after docking calculations with the aim of prioritizing poses with particular chemical features [164, 186]. This integration of both approaches usually results in better binding mode predictions and improved enrichment of active molecules. Instead of performing both calculations separately there are also a few approaches [187] which incorporate pharmacophoric constraints into the docking calculation. In this way the search space can be effectively reduced resulting in much more efficient calculations as unsuitable compounds are removed directly during the generation of poses.

The RAISE engine supports the specification structure-based pharmacophoric queries on the basis of inclusion and exclusion features whose evaluation is directly incorporated into the generation of binding poses. The building blocks for the definition of pharmacophore hypotheses essentially correspond to the interaction types modeled by the triangle corners (see section 4.2) and are represented as spheres with a tolerance radius. This includes hydrogen bond donors, hydrogen bond acceptors, hydrophobic

4. VIRTUAL SCREENING

regions and, additionally, the presence of any atom. In case of polar interactions, the locations of ligand atoms can be further constrained by specifying the respective directionality of either the hydrogen atom or the free electron pair. A compilation of all supported types can be found in [D7]. Inclusion features require the presence of ligand atoms at specific regions in the binding site of the protein, whereas exclusion features explicitly prevent it. The final pharmacophore hypothesis comprises an arbitrary combination of both feature types in combination with the number of inclusion features (N_e) that need to be fulfilled simultaneously (see Figure 4.6).

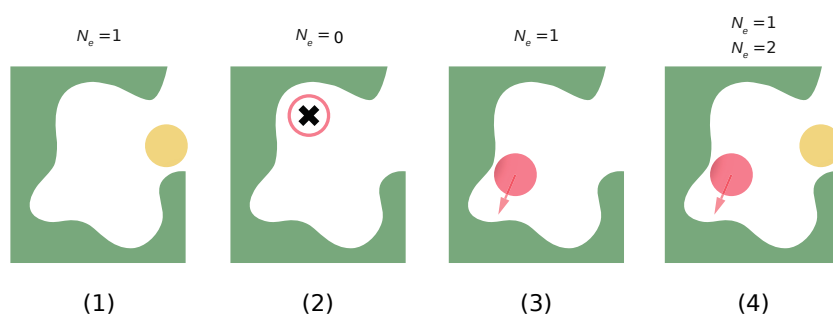


Figure 4.6: Definition of pharmacophores in the context of cRAISE. Example (1) contains one hydrophobic inclusion feature meaning that at least one hydrophobic atom of the ligand has to be located inside the area marked by the circle. In this case N_e has to be one in order for the pharmacophore to have any effect. Example (2) contains an exclusion feature of the acceptor type meaning that no acceptor atom of the ligand must be placed in the area marked by the circle. In this case N_e is not relevant as no inclusion features are defined. Example (3) contains an inclusion feature of acceptor type including the specification of its directionality. Example (4) contains two inclusion features which depending on N_e need to be fulfilled simultaneously.

Pharmacophoric constraints are evaluated at two different stages of the docking procedure [D7]. Inclusion features can be already taken into consideration during the generation of RAISE descriptors. If an interaction triangle does not contain at least one corner which fulfills one of the essential features defined in the pharmacophore (see (B) in Figure 4.7), it can be omitted. In this way the number of query triangles can be reduced which leads to an considerable decrease in runtime. As RAISE descriptors only capture a local environment of the binding pocket enforcing more than one inclusion feature could easily lead to false negative predictions and is therefore not supported. Since the remaining constraints cannot be evaluated without recourse to the concrete orientation of the molecule inside the binding pocket they have to be processed during the generation of ligand poses. Exclusion features can be considered as inaccessible regions and are thus incorporated directly into the clash testing routines of the pose

filtering step. By handling them as clashes, violating poses can be eliminated very early in the process. The fulfillment of all essential features is checked after the placement of the molecule in the binding pocket (see (C) in Figure 4.7).

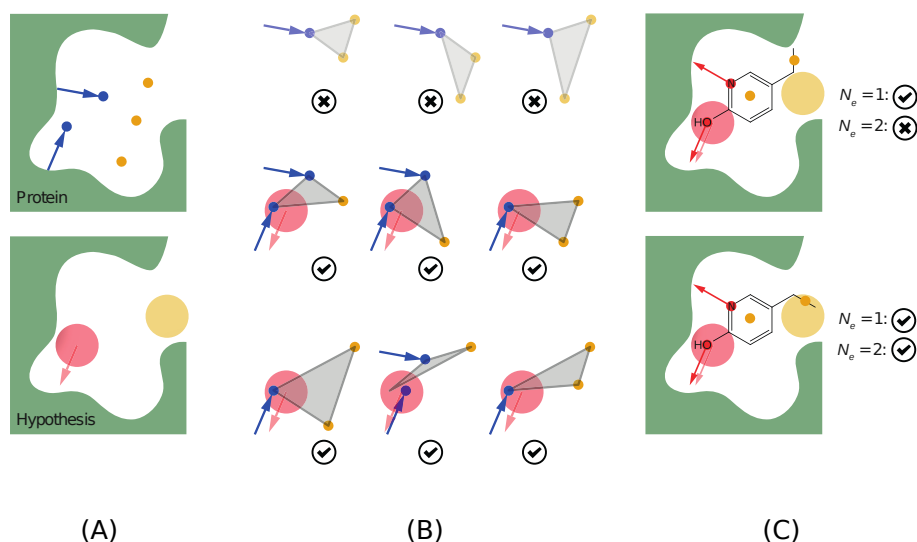


Figure 4.7: Evaluation of Pharmacophores in RAISE. The definition of a pharmacophore hypothesis for a particular binding pocket is shown in (A). During the generation of RAISE descriptors the pharmacophore can be used to exclude particular triangles (B). After the placement of the ligand, the remaining constraints are evaluated (C).

The chemical features contained in the presented pharmacophore concept are well known to medicinal chemists so that the resulting models are intuitively understandable to them. The reduction of the complex interactions between ligands and proteins to a set of clearly defined building blocks facilitates their access to the otherwise complex screening methodologies. Even without intricate knowledge of the underlying technology it is possible to formulate and evaluate hypotheses about potential binding modes and important contributions to affinity. Additionally, pharmacophores provide the means to directly influence the screening process with external knowledge and thus pave the way for the integration of medicinal chemists in the screening process. As both redocking experiments and enrichment studies show [D7], the docking and screening performance of cRAISE can be considerably improved by the inclusion of pharmacophore constraints. In contrast to other approaches, this can also result in a considerable decrease in runtime which essentially means the method gets faster and more accurate at the same time.

4.4 Protomers in Docking

Structure-based virtual screening is an application in which the explicit handling of protomers is generally required (see Section 2.6). As the presence and the positions of hydrogen atoms change, so do the potential hydrogen bond networks which in turn can result in alternative binding modes. If this fact is neglected during the screening calculation the risk of false negative results increases. One possible way to avoid this is to transform every chemical moiety into a representation which corresponds to the form assumed to be the most stable under the respective conditions. As has been shown by different studies, this procedure is generally preferable to a more or less exhaustive enumeration of protomers as this can easily lead to a high rate of false positive predictions [125, 126, 128]. This results from the fact that current scoring functions are not equipped to reliably differentiate reasonable from unreasonable protomers. There are, however, a number of functional groups and ringsystems with protomers of similar stability for which this selection is rather arbitrary [188]. Typical examples are aromatic heterocycles such as pyrazole and imidazole. In these cases the consideration of multiple protomers is necessary in order to maximize the chances of finding the best possible binding mode.

Although the same types of moieties, namely imidazole in case of histidine, can be found in the side chains of amino acids, the consideration of different protomers in virtual screening is generally restricted to the ligand side. This is probably due to the fact that protomers of small molecules can be enumerated and sequentially processed without the need to modify the underlying docking procedure. If their respective number is additionally restricted to a reasonable size, the associated increase in runtime is usually acceptable. The same strategy, however, is not applicable with respect to proteins or binding pockets. In case of enumeration multiple docking runs would have to be performed for the complete library which in turn would result in an considerable increase of runtime. Considering the fact that binding pockets can include a large number of residues, the number of relevant protomeric states can easily become very large. Thus, the best course of action is to incorporate the handling of protomers into the docking procedure.

The RAISE engine is able to implicitly handle multiple protomeric forms of both ligands and residues using the interaction triangle description introduced in Section 4.2. This is realized by generating triangle corners of both hydrogen bond donor and acceptor type for atoms changing their state in different protomers. The latter are generated using the methods for the generation of protomers based on the VSC model (see section

2.6). A general scheme for the procedure is shown in Figure 4.8, a more detailed description can be found in [D8]. First, normalization is applied to the functional groups and rings in order to transform them into their most stable form as was suggested by the previous studies [125, 126, 128]. Afterwards, reasonable protomers are generated and used to identify atoms which change their role from donor to acceptor or vice versa. During the generation of interaction triangles these atoms are considered as both donor and acceptor and the respective triangles are created for both cases. The same procedure is applied to the atoms of the binding site. Both the triangle matching and the subsequent generation of binding poses are identical to the procedure explained in Section 4.2. However, after the placement of the ligand the optimal hydrogen bonding network needs to be generated using the Protoss approach (see section 3.3).

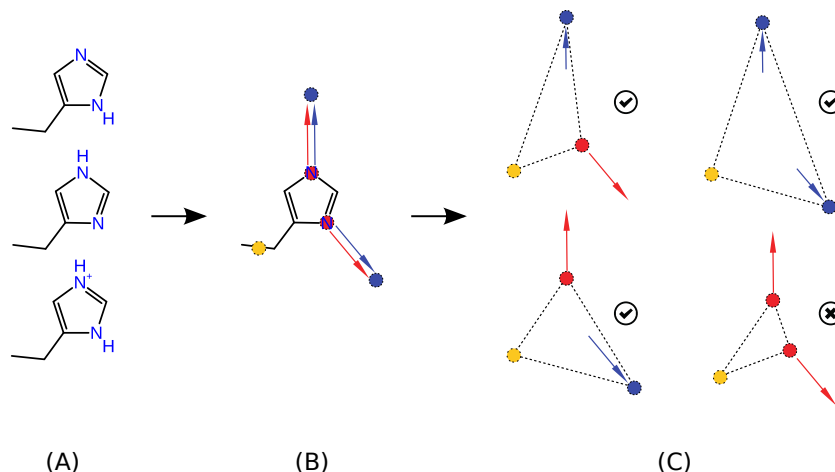


Figure 4.8: Handling of protomers in the context of cRAISE. The relevant protomers of the imidazole group (A) are transformed into a unified representation including multiple interaction surfaces (B). These are converted into individual RAISE descriptors under consideration of the possible combinations (C).

By relying on both the VSC model [D4], Protoss [D6] and the efficient RAISE technology the presented method for the handling of protomers in docking calculations has many advantages compared to other approaches. For the screening library, the drawbacks associated with an explicit enumeration of states can be entirely avoided. In this case conformations would need to be calculated for each individual protomer and consequently RAISE descriptors for each conformation. This in turn would lead to a large number of redundant triangles considering the large conformational overlap between the resulting structures. By generating triangles using the procedure introduced above their resulting number can be considerably reduced. This means that the

4. VIRTUAL SCREENING

consideration of additional protomers do not lead to an linear increase of runtime as it does with other approaches. The treatment of protomers on the protein side is unique in its way and cannot be realized by other current approaches. As the comparison to the calculation without protomers shows, the respective treatment does not result in a large increase in runtime [D8].

4.5 Virtual High-Throughput Screening

Structure-based virtual screening is a complex and computationally intensive procedure. Predicting the binding modes of millions of compounds generally requires considerable computing capacities and is also often coupled with a high storage complexity. Even today, large-scale screenings can be prohibitively time-consuming without access to an advanced computer infrastructure. Only improvements in hardware performance and massive parallelization have made it possible to keep track with the ever growing runtime and storage requirements of large structure-based screening projects. Additionally, constant algorithmic and methodical optimizations further increased the efficiency of many contemporary docking tools. In some cases, modes of different complexity are available in order to find the right balance between runtime and accuracy for screening experiments of different scales. The Glide [180] software, for instance, offers an high-throughput screening mode using a less advanced scoring function for the sake of runtime optimization. Despite all these efforts and strategies, there are still many scenarios extending the boundaries of the possible. Extensive ensemble docking, i.e., modeling the flexibility of proteins by an iterative screening against multiple protein conformations, is but one typical example.

An alternative way to enhance the efficiency of virtual screening approaches is to abandon the sequential screening paradigm as has been realized in the TrixX approach [181, 182]. Instead of iteratively docking each potential ligand into the binding site, TrixX relies on a combination of pharmacophoric constraints, docking, and bitmap indexing techniques in order to overcome the linear dependency on the size of the screening library. This concept forms the basis of the RAISE strategy implemented using the NAOMI framework. At its core is the ability to precalculate and store RAISE descriptors using a compressed bitmap index and to realize their comparison using the native and thus efficient database functionality. The main advantage of this approach is the separation of the screening procedure into two independent phases, the preprocessing of the screening library and the generation of binding poses. The former has only to be performed once and the resulting index can be used any number of times.

4.5 Virtual High-Throughput Screening

The individual steps of the preprocessing procedure are shown in Figure 4.9. First, the entire screening library is registered into the MolDB (see Section 2.4), in order to ensure both a consistent and efficient storage of the respective molecules. Afterwards, conformations are generated using CONFGEN [189], a tool for the enumeration of bioactive conformations based on the NAOMI framework, and stored in the MolDB as instances. Then, RAISE descriptors are calculated and registered to the descriptor index for each of these conformations. A detailed description of the indexing techniques as well as the associated query procedures can be found in [182]. Since the preprocessing is performed without a specific target, the respective molecules usually constitute the complete compound database.

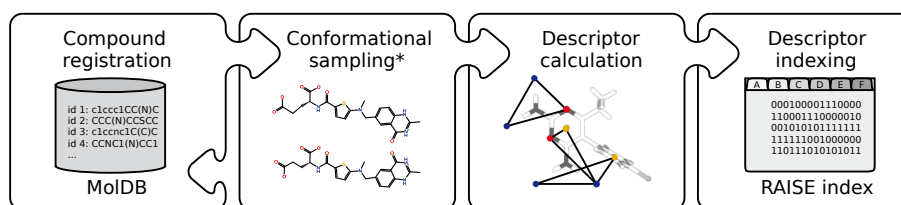


Figure 4.9: Workflow for the preprocessing of screening libraries.

The screening procedure (see Figure 4.10) starts with the generation of interaction triangles for the binding pocket, which in turn provide the basis for the formulation of queries to the bitmap index. If pharmacophore constraints have been specified these are used to reduce the number of query triangles as described in Section 4.3. Since the static database generated in the preprocessing step has not been optimized for a specific purpose, the MolDB can be used to tailor the screening library to the optimal target profile using the filter mechanisms presented in Section 2.5. The subsequent triangle matching is only performed for molecules in the active screening set. In this context the MoleculeKeys assigned by the MolDB play a central role. These are also stored in the bitmap index and are the connection between the two databases. In case of matches the respective conformations are fetched from the MolDB using the respective InstanceKeys and placed into the binding site as described in 4.2.

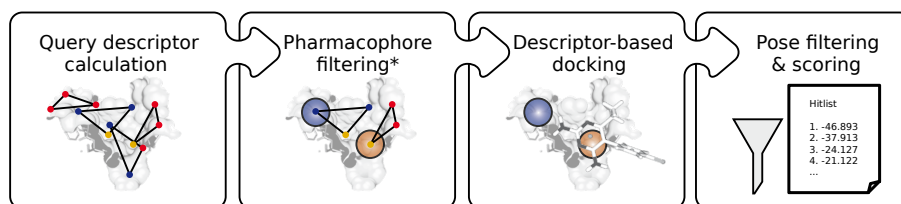


Figure 4.10: Workflow for virtual screening.

4. VIRTUAL SCREENING

The presented structure-based pipeline essentially combines all methods and concepts developed in this thesis. The result is a highly efficient screening workflow which is, in principle, completely automated, but allows interventions at different relevant stages. Both the compilation of the screening library and the formulation of pharmacophore hypotheses are intuitive procedures which can be easily performed by medicinal chemists.

Additional Applications

The models, concepts, and methods presented in the previous sections are not only relevant in the context of structure-based virtual screening but can also be successfully applied to other common tasks from the field of CADD. These are inverse virtual screening and the comparison of binding sites of different proteins. Since the contributions of the author to the development and the subsequent evaluation of the corresponding methodologies are not as central as in the previously described cases, only a cursory overview of the underlying problems and a short description of the solution will be presented. The focus will be on those parts the author was mainly concerned with. Additional details can be found in the respective publications [D9,D 10].

4.6 Inverse Virtual Screening

Binding selectivity is one of the most important aspects in the discovery and development of drug molecules [190]. A lot of effort is usually put into the design and optimization of molecules which bind with the highest affinity to their intended target proteins. In fact, most of the methods presented in this thesis up to this point have been developed to that purpose. However, the problem of avoiding or restricting the potential interactions with other components of the biological system, e.g., other enzymes, is at least as important. Unexpected interactions can easily result in adverse side effects which are one of the reasons for the high attrition rates of potential candidates in late stages of the drug development process. On the other hand, polypharmacology can also have positive aspects, e.g., drug repurposing [191] or multi-targeted drugs [192]. Considering the multitude of different chemical species which could serve as potential interaction partners, gaining the complete target profile of potential drug candidates is significantly more complex than optimizing their interactions with a single target. A description of existing methods and a more detailed introduction to the topic including literature references can be found in the original publication [D9].

Although serving a completely different purpose, the fundamental challenges of inverse screening approaches are very similar in many respects to those of conventional structure-based screening approaches. The underlying problem is still to predict the binding pose of a ligand in the binding pocket of a protein. For that reason, the inverse screening pipeline of the NAOMI software system is based on the same concepts as the previously presented methods. Matching interaction triangles are identified and subsequently used to generate the respective ligand poses in the binding pocket in the same way as described above. The major difference is that these binding poses are not generated for multiple ligands in the same binding pocket, but for a single ligand in binding sites of different proteins. For that reason the pipeline presented in the previous section has to be inverted. This means that the RAISE descriptors of different binding pockets are registered into the bitmap index and those of the ligand are used for the generation of the respective queries. In this context the MolDB from the conventional pipeline is replaced by the ProteinDB introduced in Section 3.1. The scoring function is also modified as its purpose is now to rank different poses in different binding pockets instead of ranking poses in a single binding pocket. A detailed description of the associated procedures and a discussion of the results can be found in [D9].

4.7 Comparison of Binding Pockets

The annotation of protein function is an important task in many different fields including biology, biotechnology and pharmaceutical research. Although structural genomics projects have provided access to large amounts of protein sequence and structure data, current experimental methods are not able to keep up with the sheer quantity of uncharacterized cases. For that reason, computational approaches are of high practical relevance in this context. The function of unknown structures is generally inferred from either sequence or structural similarity to already annotated proteins. This methodology is, however, not only restricted to whole proteins but can also be applied to the comparison of different binding sites which in turn can help to gain insight into aspects such as substrate specificity or potential mutation sites for enzyme optimization. The number of computational approaches for the comparison of binding sites is quite large and an overview can be found in the respective original publication [D10].

The concepts, methods and components underlying the structure-based virtual screening pipeline presented in the previous section are the foundation for the efficient comparison of protein binding sites in the NAOMI software system. In this case also the comparison of interaction triangles forms the core of the approach. In contrast to the prediction of binding poses compatible descriptors need to represent similar

4. VIRTUAL SCREENING

rather than complementary features. For that reason, the matching procedure needs to be slightly modified with respect to the evaluation of both the types of triangle corners and the associated interaction directions. Furthermore, as the shape descriptors of the triangles in this context are not intended to avoid steric overlap but rather to assess structural similarity the matching procedure has to be adapted accordingly. This is realized by allowing partial bulk matches. If compatible triangles have been identified, the respective binding pockets are aligned in the same way as described for molecules. An affine transformation is determined by superposing the respective triangles which is then applied to the coordinates of the atoms of the binding site. The superimposed binding pockets are then scored based on this overlap. The complete matching procedure is described in more detail in the original publication [10]. In analogy to the structure-based screening pipeline the comparison of binding pockets is divided into two separate steps. First, descriptors for a collection of binding pockets are calculated and stored in a triangle index which can be reused indefinitely. Then the actual screening calculation can be performed. A detailed description of the associated procedures and a discussion of the results can be found in [D10].

5

Summary and Outlook

In the presented work a consistent framework for automated virtual high-throughput screening (NAOMI) has been introduced. Based on a robust chemical description, NAOMI incorporates numerous innovative concepts and methods which together form the foundation of the RAISE screening pipeline. Each of its individual components has been designed in order to provide an intuitive and comprehensible way to perform screening calculations and was subsequently investigated with respect to its suitability for completely automated workflows. It could be shown in multiple evaluation studies that the underlying models and algorithms are both efficient and reliable thus allowing a balanced combination of automatic and interactive steps. This in turn is an important prerequisite for the implementation of highly adaptive screening workflows for different contexts of applications. The handling of molecules and proteins is based on a consistent and robust chemical model which was shown to produce reliable results in many different cheminformatics contexts. The newly developed methods represent significant improvements over existing approaches in their respective fields. This includes the conversion of file formats, the interpretation of molecules from three-dimensional coordinates, the canonicalization of molecules, and the generation of protomers. Additionally, many of the presented concepts can be considered as important contributions to the accurate solution of cheminformatics problems, e.g., the representation of rings by URF and the handling of protomers based on the VSC Model. The databases for both molecules (MolDB) and proteins (ProteinDB) allow an efficient storage and processing of the respective structures and provide the basis for the interactive functionality for the compilation of screening libraries. The consistent handling of ionization and tautomerism allows the comprehensive treatment of protomers in the context of protein-ligand complexes thus making Protoss and the inclusion of protomers in

5. SUMMARY AND OUTLOOK

structure-based virtual screening unique solutions in their respective fields of application. The presented RAISE technology is the foundation for a very efficient screening engine which also supports the formulation of intuitive pharmacophoric constraints. The additional RAISE applications, inverse screening and comparison of binding pockets, show the generic nature of the approach and the high potential for the solution of different cheminformatics problems.

Although the presented pipeline marks an important milestone in the development of an interactive virtual screening approach, there is still room for improvements. On the one hand, the performance of the docking method with respect to the reliable ranking of binding poses could be enhanced by using a subsequent optimization routine. As has already been demonstrated in combination with the docking tool FlexX, a procedure based on the HYDE scoring function, which was also developed in this group, would be perfectly suitable for that purpose [193]. The integration of both approaches has already been realized on the software level with the help of the author. However, the respective results need to be carefully evaluated and it is to be expected that modifications on both sides will be necessary in order to achieve an optimal performance. Considering the high efficiency of the RAISE approach and, in particular, the ability to handle cases of high computational and storage complexity, the treatment of protein flexibility, which is still one of the unsolved problems of virtual screening, could also be considered as a prospective application. As the NAOMI framework already contains the necessary functionality for the generation of side chain conformations, the step towards putting it to use in a flexible docking and screening scenarios seems promising. Additionally, the RAISE technology is not necessarily restricted to structure-based approaches and could also serve as a basis for the development of a ligand-based screening methodology. This in turn could build on the strategies developed for the comparison of binding pockets. The internal screening pipeline would largely profit from ligand-based components.

As the NAOMI framework provides a solid basis for the development of methods and algorithms involving small molecules, proteins, or protein-ligand complexes, a large variety of future applications are quite conceivable. The robust chemical model could be used to analyze crystal structure data of protein-ligand complexes in order to identify potential inconsistencies. New concepts for the visualization of molecule sets based on structural similarity or common molecular scaffolds would further increase the value of the MONA software with respect to the analysis of compound databases. The generic interaction model could provide the basis for the systematic investigation and classification of interaction patterns in protein-ligand complexes. Another part of the NAOMI framework which has not been discussed in this thesis are fragment spaces.

In this context both the consistent chemical model and the extremely efficient procedures developed for the handling of molecules can be of great value. It is the author's believe that, particularly in combination with the MolDB, there are a lot of promising applications in this direction.

In addition to the establishment of a fully automated screening workflow, one of the main goals during the development of NAOMI was creating the necessary conditions for the inclusion of medicinal chemists in CADD. By providing reliable and efficient methods for that purpose, NAOMI is a robust foundation for the implementation of intuitive and interactive software tools. MONA, as a solution for the preparation of screening libraries, can be considered as a first step in this direction. The SeeSAR software developed by the BiosolveIT is another example for an application program based on the NAOMI framework. Both examples show the inherent potential of the NAOMI framework as the basis for the development of sophisticated drug-design software in both an academic and an industrial setup.

Bibliography

- [1] I. Wagner and H. Musso. New naturally occurring amino acids. *Angew. Chem. Int. Ed.*, 22 (11):816–828, 1983.
- [2] D. Voet and J.G. Voet. *Biochemistry*. John Wiley & Sons, 2002.
- [3] J. M. Berg, J. L. Tymoczko, and L. Stryer. *Biochemistry*. WH Freeman and Company, New York, 2002.
- [4] E. Fischer. *Berichte der Deutschen chemischen Gesellschaft zu Berlin*, 27:2985–2993, 1894.
- [5] D. E. Koshland. Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci. USA*, 44(2):98–104, 1958.
- [6] H. Gohlke and G. Klebe. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew. Chem. Int. Ed.*, 41 (15):2644–2676, 2002.
- [7] C. Bissantz, B. Kuhn, and M. Stahl. A medicinal chemist’s guide to molecular interactions. *J. Med. Chem.*, 53(14):5061–5084, 2010.
- [8] H. Gohlke. *Protein-ligand interactions: Methods and principles in medicinal chemistry*, volume 53. Wiley-VCH, 2012.
- [9] A. Fersht. *Enzyme Structure and Mechanism*. Freeman, New York, 1985.
- [10] L.C. Pauling. *The Nature of the Chemical Bond and the Structure of Molecules and Crystals: An Introduction to Modern Structural Chemistry*. Cornell Univ. Press, Ithaca N.Y., 1960.
- [11] N. T. Southall, K. A. Dill, and A. D. J. Haymet. A view of the hydrophobic effect. *J. Phys. Chem. B*, 106(3):521–533, 2002.
- [12] D. Chandler. Interfaces and the driving force of hydrophobic assembly. *Nature*, 437(7059): 640–647, 2005.
- [13] G. Klebe. Drug research: Yesterday, today, and tomorrow. In *Drug Design - Methodology, Concepts, and Mode-of-Action*, pages 3–22. Springer, 2013.

BIBLIOGRAPHY

- [14] A.M. Davis, S.J. Teague, and G.J. Kleywegt. Application and limitations of x-ray crystallographic data in structure- based ligand and drug design. *Angew. Chem. Int. Ed.*, 42(24): 2718–2736, 2003.
- [15] J.P. Hughes, S. Rees, S.B. Kalindjian, and K.L. Philpott. Principles of early drug discovery. *Br. J. Pharmacol.*, 162(6):1239–1249, 2011.
- [16] S.M. Paul, D.S. Mytelka, C.T. Dunwiddie, Charles C. Persinger, B. H Munos, S. R Lindborg, and A.L. Schacht. How to improve r&d productivity: The pharmaceutical industry’s grand challenge. *Nat. Rev. Drug. Discov.*, 9:203–214, 2010.
- [17] D.C. Young. *Computational Drug Design: A Guide for Computational and Medicinal Chemists*. John Wiley & Sons, 2009.
- [18] Y. Yang, S.J. Adelstein, and A.I. Kassis. Target discovery from data mining approaches. *Drug. Discov. Today*, 14(3-4):147–154, 2009.
- [19] U. Egner and R.C. Hillig. A structural biology view of target drugability. *Expert Opin. Drug Discov.*, 3(4):391–401, 2008.
- [20] L.Y. Han, C.J. Zheng, B. Xie, J. Jia, X.H. Ma, F. Zhu, H.H. Lin, X. Chen, and Y.Z. Chen. Support vector machines approach for predicting druggable proteins: Recent progress in its exploration and investigation of its usefulness. *Drug. Discov. Today*, 12(7-8):304–313, 2007.
- [21] E.B. Fauman, B.K. Rai, and E.S. Huang. Structure-based druggability assessment – identifying suitable targets for small molecule therapeutics. *Curr. Opin. Chem. Biol.*, 15(4): 463–468, 2011.
- [22] Laurie A.T. and Jackson R.M. Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screening. *Curr Protein Pept Sci.*, 7(5): 395–406, 2006.
- [23] C. McInnes. Virtual screening strategies in drug discovery. *Curr. Opin. Chem. Biol.*, 11 (5):494–502, 2007.
- [24] D.B. Kitchen, H. Decornez, J.R. Furr, and J. Bajorath. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nat. Rev. Drug. Discov.*, 3:935–949, 2004.
- [25] Q. Gao, L. Yang, and Y. Zhu. Pharmacophore based drug design approach as a practical process in drug discovery. *Curr. Comput. Aided. Drug. Des.*, 6(1):37–49, 2010.
- [26] J. Verma, V.M. Khedkar, and E.C. Coutinho. 3d-qsar in drug design – a review. *Curr. Top. Med. Chem.*, 10(1):95–115, 2010.
- [27] G. Schneider. Trends in virtual combinatorial library design. *Curr. Med. Chem.*, 9(23): 2095–2101, 2002.

- [28] P.M. Petrone, A.M. Wassermann, E. Lounkine, P. Kutchukian, B. Simms, J. Jenkins, P. Selzer, and M. Glick. Biodiversity of small molecules – a new perspective in screening set selection. *Drug Discov. Today*, 18(13-14):674–680, 2013.
- [29] W.P. Walters and M.A. Murcko. Prediction of 'drug-likeness'. *Adv. Drug Deliv. Rev.*, 54(3):255–271, 2002.
- [30] K. Loving, I. Alberts, and W. Sherman. Computational approaches for fragment-based and de novo design. *Curr. Top. Med. Chem.*, 10(1):14–32, 2010.
- [31] D.R Hawkins. Comprehensive expert systems to predict drug metabolism. In John B Taylor and David J Trigg, editors, *Comprehensive Medicinal Chemistry {II}*, pages 795–807. Elsevier, Oxford, 2007.
- [32] A. Roncaglioni, A.A. Toropov, A.P. Toropova, and E. Benfenati. In silico methods to predict drug toxicity. *Curr Opin Pharmacol.*, 13(5):802–806, 2013.
- [33] L. Xie, J. Wang, and P.E. Bourne. In silico elucidation of the molecular mechanism defining the adverse effect of selective estrogen receptor modulators. *PLoS Comput. Biol.*, 3(11):e217, 2007.
- [34] S.L. Kinnings, N. Liu, N. Buchmeier, P.J. Tonge, L. Xie, and P.E. Bourne. Drug discovery using chemical systems biology: Repositioning the safe medicine Comtan to treat multi-drug and extensively drug resistant tuberculosis. *PLoS Comput. Biol.*, 5(7):e1000423, 2009.
- [35] L. Xie and P.E. Bourne. Structure-based systems biology for analyzing off-target binding. *Curr. Opin. Struct. Biol.*, 21(2):189–199, 2011.
- [36] K.M. Merz Jr., D. Ringe, and C.H. Reynolds. *Drug Design - Structure- and Ligand-Based Approaches*. Cambridge University Press, 2010.
- [37] W.P. Walters, M.T. Stahl, and M.A. Murcko. Virtual screening - an overview. *Drug Discov. Today*, 3(4):160–178, 1998.
- [38] D. Wilton and P. Willett. Comparison of ranking methods for virtual screening in lead-discovery programs. *J. Chem. Inf. Model.*, 43(2):469–474, 2003.
- [39] D. Rognan. Docking methods for virtual screening: Principles and recent advances. In C. Sotriffer, editor, *Virtual Screening - Principles, Challenges, and Practical Guidelines*, volume 48 of *Methods and Principles in Medicinal Chemistry*, pages 153–176. Wiley-VCH, Weinheim, 2011.
- [40] H. Koeppen, J. Kriegl, U. Lessel, C. S. Tautermann, and B. Wellenzohn. Ligand-based virtual screening. In C. Sotriffer, editor, *Virtual Screening - Principles, Challenges, and Practical Guidelines*, volume 48 of *Methods and Principles in Medicinal Chemistry*, pages 61–86. Wiley-VCH, Weinheim, 2011.

BIBLIOGRAPHY

- [41] P. Markt, D. Schuster, and T. Langer. Pharmacophore models for virtual screening. In C. Sotriffer, editor, *Virtual Screening - Principles, Challenges, and Practical Guidelines*, volume 48 of *Methods and Principles in Medicinal Chemistry*, pages 115–152. Wiley-VCH, Weinheim, 2011.
- [42] B.K. Shoichet. Virtual screening of chemical libraries. *Nature*, 432(7019):862–865, 2004.
- [43] G. Klebe. Virtual ligand screening: Strategies, perspectives and limitations. *Drug Discov. Today*, 11(13-14):580–594, 2006.
- [44] G. Schneider. Virtual screening: An endless staircase? *Nat. Rev. Drug Discov.*, 9(4):273–276, 2010.
- [45] A.S. Reddy, S. P. Pati, P. P. Kumar, H. N. Pradeep, and G. N. Sastry. Virtual screening in drug discovery – a computational perspective. *Curr. Protein Pept. Sci.*, 8(4):329–351, 2007.
- [46] P.D. Lyne. Structure-based virtual screening: An overview. *Drug Discov. Today*, 7(20):1047–1055, 2002.
- [47] M. Kontoyianni, G. S. Sokol, and L. M. McClellan. Evaluation of library ranking efficacy in virtual screening. *J. Comput. Chem.*, 26(1):11–22, 2005.
- [48] K.H. Bleicher, H.-J. Böhm, K. Müller, and A.I. Alanine. Hit and lead generation: Beyond high-throughput screening. *Nat. Rev. Drug Discov.*, 2(5):369–378, 2003.
- [49] F.K. Brown. Chemoinformatics: What is it and how does it impact drug discovery. *Annu. Rep. Med. Chem.*, 33:375–384, 1998.
- [50] J. Gasteiger and T. Engel, editors. *Chemoinformatics: A Textbook*. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, 2003.
- [51] T. Engel. Basic overview of chemoinformatics. *J. Chem. Inf. Model.*, 46(6):2267–2277, 2006.
- [52] A. Varnek and I. I. Baskin. Chemoinformatics as a theoretical chemistry discipline. *Mol. Inf.*, 30:20–32, 2011.
- [53] D.K. Agrafiotis, D. Bandyopadhyay, J. K. Wegner, and H. van Vlijmen. Recent advances in chemoinformatics. *J. Chem. Inf. Model.*, 47(4):1279–1293, 2007.
- [54] J.L. Faulon and A. Bender. *Handbook of Chemoinformatics Algorithms*. Chapman and Hall CRC,, 2010.
- [55] A.R. Leach and V.J. Gillet. *An Introduction To Chemoinformatics*. Springer Netherlands, 2007.

- [56] C.M. Dobson. Chemical space and biology. *Nature*, 432(7019):824–828, 2004.
- [57] D. Bonchev and D.H. Rouvray. *Chemical Graph Theory: Introduction and Fundamentals*. Gordon and Breach, New York, 1991.
- [58] R. Todeschini and C. Viviana. *Molecular Descriptors for Chemoinformatics*, volume 41 of *Methods and Principles in Medicinal Chemistry*. Wiley-VCH, Weinheim, 2000.
- [59] J. Bajorath, editor. *Chemoinformatics for Drug Discovery*. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, 2014.
- [60] V. Schnecke and J. Boström. Computational chemistry-driven decision making in lead generation. *Nat. Rev. Drug Discov.*, 11(1-2):43–50, 2006.
- [61] Talele T.T., S.A. Khedkar, and A.C. Rigby. Successful applications of computer aided drug discovery: Moving drugs from concept to the clinic. *Curr. Top. Med. Chem.*, 10(1): 127–141, 2010.
- [62] H. Matter and C. Sotriffer. Applications and success stories in virtual screening. In C. Sotriffer, editor, *Virtual Screening - Principles, Challenges, and Practical Guidelines*, volume 48 of *Methods and Principles in Medicinal Chemistry*, pages 319–358. Wiley-VCH, Weinheim, 2011.
- [63] S.W. Muchmore, J.J. Edmunds, K.D. Stewart, and P.J. Hajduk. Cheminformatic tools for medicinal chemists. *J. Med. Chem.*, 53(13):4830–4841, 2010.
- [64] P. Ripphausen, B. Nisius, L. Peltason, and J. Bajorath. Quo vadis, virtual screening? a comprehensive survey of prospective applications. *J. Med. Chem.*, 53(24):8461–8467, 2010.
- [65] T. Scior, A. Bender, G. Tresadern, J.L. Medina-Franco, K. Martinez-Mayorga, T. Langer, and D. K. Cuanalo-Contreras, K. and Agrafiotis. Recognizing pitfalls in virtual screening: A critical review. *J. Chem. Inf. Model.*, 52(4):867–881, 2012.
- [66] Heikamp K. and Bajorath J. The future of virtual compound screening. *Chem. Biol. Drug Des.*, 81(1):33–40, 2013.
- [67] H. Kubinyi. Drug research: Myths, hype and reality. *Nat. Rev. Drug Discov.*, 2(8): 665–668, 2003.
- [68] T.J. Ritchie and I.M. McLay. Should medicinal chemists do molecular modelling? *Drug Discov. Today*, 17(11-12):534–537, 2012.
- [69] C. Liao, M. Sitzmann, A. Pugliese, and M.C. Nicklaus. Software and resources for computational medicinal chemistry. *Future Med. Chem.*, 3(8):1057–1085, 2011.

BIBLIOGRAPHY

- [70] M.D. Cummings, E. Arnoult, C. Buyck, G. Tresadern, A.M. Vos, and J.K. Wegner. Preparing and filtering compound databases for virtual and experimental screening. In C. Sotriffer, editor, *Virtual Screening - Principles, Challenges, and Practical Guidelines*, volume 48 of *Methods and Principles in Medicinal Chemistry*, pages 35–54. Wiley-VCH, Weinheim, 2011.
- [71] G.N. Lewis. The atom and the molecule. *J. Am. Chem. Soc.*, 38(4):762–785, 1916.
- [72] D.H. Rouvray. A rationale for the topological approach to chemistry. *J. Mol. Struct. - THEOCHEM*, 336(2-3):101–114, 1995.
- [73] S. Bauerschmidt and J. Gasteiger. Overcoming the limitations of a connection table description: A universal representation of chemical species. *J. Chem. Inf. Comput. Sci.*, 37(4):705–714, 1997.
- [74] W. A. Warr. Representation of chemical structures. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 1(4):557–579, 2011.
- [75] R. Sayle. Cheminformatics toolkits: A personal perspective. http://rdkit.org/UGM/2012/Sayle_RDKitPerspective.pdf. (accessed February 25, 2014).
- [76] J. Gasteiger. Calculation of physical and chemical data. In J. Gasteiger and T. Engel, editors, *Chemoinformatics: A Textbook*, pages 319–337. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, 2003.
- [77] W.D. Ihlenfeldt, Y. Takahashi, H. Abe, and S. Sasaki. Computation and management of chemical properties in cactvs: An extensible networked approach toward modularity and compatibility. *J. Chem. Inf. Comput. Sci.*, 34(1):109–116, 1994.
- [78] Cactvs. <http://www2.ccc.uni-erlangen.de/software/cactvs/>. (accessed February 25, 2014).
- [79] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen. The chemistry development kit (cdk): An open-source java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.*, 43(2):493–500, 2003.
- [80] C. Steinbeck, C. Hoppe, S. Kuhn, M. Floris, R. Guha, and E. L. Willighagen. Recent developments of the chemistry development kit (cdk) - an open-source java library for chemo- and bioinformatics. *Curr. Pharm. Des.*, 12(17):2111–2120, 2006.
- [81] The chemistry development kit (cdk). <http://cdk.sourceforge.net/>. (accessed February 25, 2014).
- [82] Molecular operating environment (moe). http://www.chemcomp.com/MOE-Molecular_Operating_Environment.htm. (accessed February 25, 2014).
- [83] N. O’Boyle, M. Banck, C. James, C. Morley, T. Vandermeersch, and G. Hutchison. Open babel: An open chemical toolbox. *J. Cheminform.*, 3(1):33, 2011.

- [84] Open babel: The open source chemistry toolbox. <http://www.openbabel.org>. (accessed February 25, 2014).
- [85] Oechem tk - programming library for chemistry and cheminformatics. <http://www.eyesopen.com/oechem-tk>. (accessed February 25, 2014).
- [86] Perlmol - perl modules for molecular chemistry. <http://www.perlmol.org/>. (accessed February 25, 2014).
- [87] Pipeline pilot. <http://accelrys.com/products/pipeline-pilot/>. (accessed February 25, 2014).
- [88] Rdkit: Open-source cheminformatics. <http://www.rdkit.org>. (accessed February 25, 2014).
- [89] P. Ertl and A. Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.*, 1(1):8, 2009.
- [90] T.R. Ritchie and S.J.F. Macdonald. The impact of aromatic ring count on compound developability - are too many aromatic rings a liability in drug design? *Drug Discov. Today*, 14(21-22):1011–1020, 2009.
- [91] A. M. Clark, P. Labute, and M. Santavy. 2D structure depiction. *J. Chem. Inf. Model.*, 46(3):1107–1123, 2006.
- [92] J. Sadowski and J. Gasteiger. From atoms and bonds to three-dimensional atomic coordinates: Automatic model builders. *Chem. Rev.*, 93(7):2567–2581, 1993.
- [93] G. M. Downs, V. J. Gillet, J. D. Holliday, and M. F. Lynch. Review of ring perception algorithms for chemical graphs. *J. Chem. Inf. Comput. Sci.*, 29(3):172–187, 1989.
- [94] F. Berger, C. Flamm, P. M. Gleiss, J. Leydold, and P. F. Stadler. Counterexamples in chemical ring perception. *J. Chem. Inf. Comput. Sci.*, 44(2):323–331, 2004.
- [95] T. Hanser, P. Jauffret, and G. Kaufmann. A new algorithm for exhaustive ring perception in a molecular graph. *J. Chem. Inf. Comput. Sci.*, 36(6):1146–1152, 1996.
- [96] J. Figueras. Ring perception using breadth-first search. *J. Chem. Inf. Comput. Sci.*, 36(5):986–991, 1996.
- [97] G. Carta, V. Onnis, A.J.S. Knox, D. Fayne, and D.G. Lloyd. Permuting input for more effective sampling of 3d conformer space. *J. Comput.-Aided Mol. Des.*, 20(3):179–190, 2006.
- [98] P. Vismara. Union of all the minimum cycle bases of a graph. *Electron. J. Comb.*, 4:1–15, 1997.

BIBLIOGRAPHY

- [99] Tripos mol2 file format, . <http://www.tripos.com/data/support/mol2.pdf>. (accessed February 25, 2014).
- [100] Ctf file formats, . <http://download.accelrys.com/freeware/ctfile-formats/ctfile-formats.zip>. (accessed February 25, 2014).
- [101] A. Dalby, J.G. Nourse, W.D. Hounshell, A.K.I. Gushurst, D.L. Grier, B.A. Leland, and J. Laufer. Description of several chemical structure file formats used by computer programs developed at molecular design limited. *J. Chem. Inf. Comput. Sci.*, 32(3):244–255, 1992.
- [102] D. Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28(1):31–36, 1988.
- [103] T. Engel. Representation of chemical compounds. In J. Gasteiger and T. Engel, editors, *Chemoinformatics: A Textbook*, pages 40–52. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, 2003.
- [104] J. Kirchmair, P. Markt, S. Distinto, G. Wolber, and T. Langer. Evaluation of the performance of 3d virtual screening protocols: Rmsd comparisons, enrichment assessments, and decoy selection - what can we learn from earlier mistakes? *J. Comput.-Aided Mol. Des.*, 22(3-4):213–228, 2008.
- [105] H. L. Morgan. The generation of a unique machine description for chemical structures - a technique developed at chemical abstracts service. *J. Chem. Doc.*, 5:107–113, 1965.
- [106] D. Weinigner, A. Weininger, and Weininger J. L. Smiles. 2. algorithm for generation of unique smiles notation. *J. Chem. Inf. Comput. Sci.*, 29(2):97–101, 1989.
- [107] S. Heller, A. McNaught, S. Stein, D. Tchekhovskoi, and I. Pletnev. Inchi - the worldwide chemical structure identifier standard. *J. Cheminform.*, 5(1):7, 2013.
- [108] W. A. Warr. Tautomerism in chemical information management systems. *J. Comput.-Aided Mol. Des.*, 24(6-7):497–520, 2010.
- [109] O. H. Chan and B. H. Stewart. Physicochemical and drug-delivery considerations for oral drug bioavailability. *Drug. Discov. Today*, 1(11):461–473, 1996.
- [110] D. F. Veber, S. R. Johnson, H.-Y. Cheng, B. R. Smith, K. W. Ward, and K. D. Kopple. Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.*, 45(12):2615–2623, 2002.
- [111] G.F. Smith. Designing drugs to avoid toxicity. *Prog. Med. Chem.*, 50:1–47, 2011.
- [112] J. Hüser, E. Lohrmann, B. Kalthof, N. Burkhardt, U. Brüggemeier, and M. Bechem. *High-Throughput Screening in Drug Discovery*, volume 35. Wiley-VCH, Weinheim, 2006.

- [113] Y.C. Martin. Diverse viewpoints on computational aspects of molecular diversity. *J. Comb. Chem.*, 3(3):231–250, 2001.
- [114] C.G. Bologa, M.M. Olah, and T.I. Oprea. Chemical database preparation for compound acquisition or virtual screening. In Richard S. Larson, editor, *Bioinformatics and Drug Discovery*, volume 316 of *Methods in Molecular Biology*, pages 375–388. Humana Press Inc., New York, 2006.
- [115] M.D. Cummings, A.C. Gibbs, and R.L. DesJarlais. Processing of small molecule databases for automated docking. *Med. Chem.*, 3(1):107–113, 2007.
- [116] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug. Deliv. Rev.*, 46(1-3):3–26, 2001.
- [117] T.I. Oprea, A.M. Davis, S.J. Teague, and P.D. Leeson. Is there a difference between leads and drugs? a historical perspective. *J. Chem. Inf. Comput. Sci.*, 41(5):1308–1315, 2001.
- [118] J.B. Baell and G.A. Holloway. New substructure filters for removal of pan assay interference compounds (pains) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.*, 53(7):2719–2740, 2010.
- [119] W. A. Warr. Scientific workflow systems: Pipeline pilot and knime. *J. Comput.-Aided Mol. Des.*, 26(7):801–804, 2011.
- [120] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel, and B. Wiswedel. Knime - the konstanz information miner: Version 2.0 and beyond. *SIGKDD Explor. Newsl.*, 11(1):26–31, 2009.
- [121] E. Martin, A. Monge, J.-A. Duret, F. Gualandi, M. Peitsch, and P. Pospisil. Building an r&d chemical registration system. *J. Cheminform.*, 4(1):11, 2012.
- [122] J.J. Irwin, T. Sterling, M.M. Mysinger, E.S. Bolstad, and R.G. Coleman. Zinc: A free tool to discover chemistry for biology. *J. Chem. Inf. Model.*, 52(7):1757–1768, 2012.
- [123] K.T. Schomburg, L. Wetzer, and M. Rarey. Interactive design of generic chemical patterns. *Drug. Discov. Today*, 18(13-14):651–658, 2013.
- [124] F. Milletti, L. Storchi, G. Sforza, S. Cross, and G. Cruciani. Tautomer enumeration and stability prediction for virtual screening on large chemical databases. *J. Chem. Inf. Model.*, 49(1):68–75, 2009.
- [125] T. ten Brink and T.E. Exner. Influence of protonation, tautomeric, and stereoisomeric states on protein-ligand docking results. *J. Chem. Inf. Model.*, 49(6):1535–1546, 2009.

BIBLIOGRAPHY

- [126] T. Kalliokoski, H.S. Salo, M. Lahtela-Kakkonen, and A. Poso. The effect of ligand-based tautomer and protomer prediction on structure-based virtual screening. *J. Chem. Inf. Model.*, 49(12):2742–2748, 2009.
- [127] Pospisil P., Ballmer P., Scapozza L., and Folkers G. Tautomerism in computer-aided drug design. *J. Recept. Signal. Transduct. Res.*, 23(4):361–371, 2003.
- [128] F. Oellien, J. Cramer, C. Beyer, W.-D. Ihlenfeldt, and P. M. Selzer. The impact of tautomer forms on pharmacophore-based virtual screening. *J. Chem. Inf. Model.*, 46(6):2342–2354, 2006.
- [129] T. Clark. Tautomers and reference 3d-structures: The orphans of in silico drug design. *J. Comput.-Aided Mol. Des.*, 24(6-7):605–611, 2010.
- [130] P. Kenny and J. Sadowski. Structure modification in chemical databases. In T. Oprea, editor, *Chemoinformatics in Drug Discovery*, pages 271–285. Wiley-VCH Verlag GmbH & Co. KGaA, 2005.
- [131] M. Haranczyk and M. Gutowski. Quantum mechanical energy-based screening of combinatorially generated library of tautomers. tautgen: A tautomer generator program. *J. Chem. Inf. Model.*, 47(2):686–694, 2007.
- [132] T. Thalheim, A. Vollmer, R.-U. Ebert, R. Kühne, and G. Schüürmann. Tautomer identification and tautomer structure generation based on the inchi code. *J. Chem. Inf. Model.*, 50(7):1223–1232, 2010.
- [133] N. T. Kochev, V. H. Paskaleva, and N. Jeliaskova. Ambit-tautomer: An open source tool for tautomer generation. *Mol. Inf.*, 32(5-6):481–504, 2013.
- [134] T. Will, M. C. Hutter, J. Jauch, and V. Helms. Batch tautomer generation with moltpc. *J. Comput. Chem.*, 34(28):2485–2492, 2013.
- [135] R.A. Sayle. So you think you understand tautomerism? *J. Comput.-Aided Mol. Des.*, 24(6-7):485–496, 2010.
- [136] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The protein data bank. *Nucleic Acids Res.*, 28(1):235–242, 2000.
- [137] A.M. Davis, S.A. St-Gallay, and G.J. Kleywegt. Limitations and lessons in the use of x-ray structural information in drug design. *Drug. Discov. Today*, 13(19-20):831–841, 2008.
- [138] N.J. Fuller, J.C. Burgoyne and R.M. Jackson. Predicting druggable binding sites at the protein-protein interface. *Drug. Discov. Today*, 27(3-4):155–161, 2009.
- [139] C. Barillari, J. Taylor, R. Viner, and J.W. Essex. Classification of water molecules in protein binding sites. *J. Am. Chem. Soc.*, 129(9):2577–2587, 2007.

- [140] J.C. Cole, O. Korb, T.S.G. Olsson, and J. Liebeschuetz. The basis for target-based virtual screening: Protein structures. In C. Sotriffer, editor, *Virtual Screening - Principles, Challenges, and Practical Guidelines*, volume 48 of *Methods and Principles in Medicinal Chemistry*, pages 87–114. Wiley-VCH, Weinheim, 2011.
- [141] J.C. Cole, O. Korb, T.S.G. Olsson, and J. Liebeschuetz. The basis for target-based virtual screening: Protein structures. In C. Sotriffer, editor, *Virtual Screening - Principles, Challenges, and Practical Guidelines*, volume 48 of *Methods and Principles in Medicinal Chemistry*, pages 92–94. Wiley-VCH, Weinheim, 2011.
- [142] J.W.M. Nissink, C. Murray, M. Hartshorn, V.L. Verdonk, J.C. Cole, and R. Taylor. A new test set for validating predictions of protein-ligand interaction. *Proteins*, 49(4):457–471, 2002.
- [143] Allen. F.H. The cambridge structural database: A quarter of a million crystal structures and rising. *Acta Cryst.*, B58(1):380–388, 2002.
- [144] Pdb file formats, . http://www.pdb.org/pdb/static.do?p=file_formats/index.jsp. (accessed February 25, 2014).
- [145] E.C. Meng and R.A. Lewis. Determination of molecular topology and atomic hybridization states from heavy atom coordinates. *J. Comput. Chem.*, 12(7):891–898, 1991.
- [146] J.C. Baber and E.E. Hodgkin. Automatic assignment of chemical connectivity to organic molecules in the cambridge structural database. *J. Chem. Inf. Comput. Sci.*, 32(5):401–406, 1992.
- [147] M. Hendlich, F. Rippmann, and G. Barnickel. Bali: Automatic assignment of bond and atom types for protein ligands in the brookhaven protein databank. *J. Chem. Inf. Comput. Sci.*, 37(4):774–778, 1997.
- [148] M. Froeyen and P. Herdewijn. Correct bond order assignment in a molecular framework using integer linear programming with application to molecules where only non-hydrogen atom coordinates are available. *J. Chem. Inf. Model.*, 45(5):1267–1274, 2005.
- [149] P. Labute. On the perception of molecules from 3d atomic coordinates. *J. Chem. Inf. Model.*, 45(2):215–221, 2005.
- [150] Y. Zhao, T. Cheng, and R. Wang. Automatic perception of organic molecules based on essential structural information. *J. Chem. Inf. Model.*, 47(4):1379–1385, 2007.
- [151] L.R. Forrest and B. Honig. An assessment of the accuracy of methods for predicting hydrogen positions in protein structures. *Proteins*, 61(2):296–309, 2005.
- [152] P. Labute. Protonate3d: Assignment of ionization states and hydrogen coordinates to macromolecular ctructures. *Proteins*, 75(1):187–205, 2009.

BIBLIOGRAPHY

- [153] E. Krieger, R.L. Jr Dunbrack, R.W.W. Hooft, and B. Krieger. Assignment of protonation states in proteins and ligands: Combining pka prediction with hydrogen bonding network optimization. In R. Baron, editor, *Computational Drug Discovery and Design*, pages 405–421. Springer, New York, 2012.
- [154] T. Lippert and M. Rarey. Fast automated placement of polar hydrogen atoms in protein-ligand complexes. *J. Cheminform.*, 1(1):13, 2009.
- [155] P. Ferrara, H. Gohlke, D.J. Price, G. Klebe, and C.L. 3rd. Brooks. Assessing scoring functions for protein-ligand interactions. *J. Med. Chem.*, 47(12):3032–3047, 2004.
- [156] G.L. Warren, C.W. Andrews, A.M. Capelli, B. Clarke, J. LaLonde, M.H. Lambert, M. Lindvall, N. Nevins, SF. Semus, S. Senger, G. Tedesco, I.D. Wall, J.M. Woolven, C.E. Peishoff, and M.S. Head. A critical assessment of docking programs and scoring functions. *J. Med. Chem.*, 49(20):5912–5931, 2006.
- [157] T. Cheng, X. Li, Y. Li, Z. Liu, and R. Wang. Comparative assessment of scoring functions on a diverse test set. *J. Chem. Inf. Model.*, 49(4):1079–1093, 2009.
- [158] E. Kellenberger, J. Rodrigo, P. Muller, and D. Rognan. Comparison of automated docking programs as virtual screening tools. *Proteins*, 57(2):225–242, 2004.
- [159] M.D. Cummings, R.L. DesJarlais, A.C. Gibbs, V. Mohan, and E.P. Jaeger. Comparison of automated docking programs as virtual screening tools. *J. Med. Chem.*, 48(4):962–976, 2005.
- [160] B. Waszkowycz, D. E. Clark, and E. Gancia. Outstanding challenges in protein-ligand docking and structure-based virtual screening. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 1(2):229–259, 2011.
- [161] C. Bissantz, G. Folkers, and D. Rognan. Protein-based virtual screening of chemical databases. 1. evaluation of different docking/scoring combinations. *J. Med. Chem.*, 43(25):4759–4767, 2000.
- [162] M. Stahl and H.-J. Böhm. Development of filter functions for protein-ligand docking. *J. Mol. Graph. Model.*, 16(3):121–132, 1998.
- [163] R.D. Taylor, P.J. Jewsbury, and J.W. Essex. Fds: Flexible ligand and receptor docking with a continuum solvent model and soft-core energy function. *J. Comput. Chem.*, 24(13):1637–1656, 2003.
- [164] D. Muthas, Y.A. Sabnis, M. Lundborg, and A. Karlen. Is it possible to increase hit rates in structure-based virtual screening by pharmacophore filtering? an investigation of the advantages and pitfalls of post-filtering. *J. Mol. Graph. Model.*, 26(8):1237–1251, 2008.

- [165] C. Sotriffer, editor. *Virtual Screening - Principles, Challenges, and Practical Guidelines*, volume 48 of *Methods and Principles in Medicinal Chemistry*. Wiley-VCH, Weinheim, 2011.
- [166] H.-J. Böhm and M. Stahl. The use of scoring functions in drug discovery applications. In K.B. Lipkowitz and D.B. Boyd, editors, *Reviews in Computational Chemistry*, volume 18, pages 41–87. John Wiley & Sons, Inc., New Jersey, 2003.
- [167] C. Sotriffer and H. Matter. The challenge of affinity prediction: Scoring functions for structure-based virtual screening. In C. Sotriffer, editor, *Virtual Screening - Principles, Challenges, and Practical Guidelines*, volume 48 of *Methods and Principles in Medicinal Chemistry*, pages 179–222. Wiley-VCH, Weinheim, 2011.
- [168] Appendix a: Software overview. In C. Sotriffer, editor, *Virtual Screening - Principles, Challenges, and Practical Guidelines*, volume 48 of *Methods and Principles in Medicinal Chemistry*, pages 491–500. Wiley-VCH, Weinheim, 2011.
- [169] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.*, 261(3):470–489, 1996.
- [170] A. M. Henzler and M. Rarey. In pursuit of fully flexible protein-ligand docking: Modeling the bilateral mechanism of binding. *Mol. Inf.*, 29(3):164–173, 2010.
- [171] A. R. Leach, B. K. Shoichet, and C. E. Peishoff. Prediction of protein-ligand interactions. docking and scoring: Successes and gaps. *J. Med. Chem.*, 49(20):5851–5855, 2006.
- [172] N. Moitessier, P. Englebienne, D. Lee, J. Lawandi, and C. R. Corbeil. Towards the development of universal, fast and highly accurate docking/scoring methods: A long way to go. *Br. J. Pharmacol.*, 153 Suppl 1:7–26, 2008.
- [173] T. J. Ewing, S. Makino, A. G. Skillman, and I. D. Kuntz. Dock 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Des.*, 15(5):411–428, 2001.
- [174] A.N. Jain. Surfex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.*, 46(4):499–511, 2003.
- [175] Z. Zsoldos, D. Reid, A. Simon, B.S. Sadjad, and A.P. Johnson. ehits: An innovative approach to the docking and scoring function problems. *Curr. Protein Pept. Sci.*, 7(5): 421–435, 2006.
- [176] M. Totrov and Abagyan R. Flexible protein-ligand docking by global energy optimization in internal coordinates. *Proteins: Struct., Funct., Genet.*, 29(1):215–220, 1997.
- [177] G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.*, 267(3):727–748, 1997.

BIBLIOGRAPHY

- [178] G. M. Morris, D. S. Goodsell, R.S. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson. Automated docking using a lamarkian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.*, 19(14):1639–1662, 1998.
- [179] M.R. McGann, H.R. Almond, A. Nicholls, J.A. Grant, and F.K. Brown. Gaussian docking functions. *Biopolymers*, 68(1):76–90, 2003.
- [180] R.A. Friesner, J.L. Banks, R.B. Murphy, T.A. Halgren, J.J. Klicic, D.T. Mainz, M.P. Repasky, E.H. Knoll, M. Shelley, J.K. Perry, D.E. Shaw, P. Francis, and P.S. Shenkin. Glide: A new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *J. Med. Chem.*, 47(7):1739–1749, 2004.
- [181] I. Schellhammer and M. Rarey. Trixx: Structure-based molecule indexing for large-scale virtual screening in sublinear time. *J. Comput.-Aided Mol. Des.*, 21(5):223–238, 2007.
- [182] J. Schlosser and M. Rarey. Beyond the virtual screening paradigm: Structure-based searching for new lead compounds. *J. Chem. Inf. Model.*, 49(4):800–809, 2009.
- [183] C.G. Wermuth, C.R. Ganellin, P. Lindberg, and L.A. Mitscher. Glossary of terms used in medicinal chemistry (iupac recommendations 1998). *Pure Appl. Chem.*, 70(5):1129–1143, 1998.
- [184] A.R. Leach, V.J. Gillet, R.A. Lewis, and R. Taylor. Three-dimensional pharmacophore methods in drug discovery. *J. Med. Chem.*, 53(2):539–558, 2010.
- [185] G. Wolber, T. Seidel, F. Bendix, and T. Langer. Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug Discov. Today*, 13(1-2): 23–29, 2008.
- [186] M.L. Peach and M. C. Nicklaus. Combining docking with pharmacophore filtering for improved virtual screening. *J. Cheminform.*, 1(1):6, 2009.
- [187] S.A. Hindle, M. Rarey, C. Buning, and T. Lengauer. Flexible docking under pharmacophore type constraints. *J. Comput.-Aided Mol. Des.*, 16(2):129–149, 2002.
- [188] J.R. Greenwood, D. Calkins, A.P. Sullivan, and J.C. Shelley. Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. *J. Comput.-Aided Mol. Des.*, 24(6-7):591–604, 2010.
- [189] C. Scharfer, T. Schulz-Gasch, J. Hert, L. Heinzerling, B. Schulz, T. Inhester, M. Stahl, and M. Rarey. Confect: Conformations from an expert collection of torsion patterns. *ChemMedChem*, 8(10):1690–1700, 2013.
- [190] D.J. Huggins, W. Sherman, and B. Tidor. Rational approaches to improving selectivity in drug design. *J. Med. Chem.*, 55(4):1424–1444, 2012.

BIBLIOGRAPHY

- [191] T.T. Ashburn and K.B. Thor. Drug repositioning: Identifying and developing new uses for existing drugs. *Nat. Rev. Drug. Discov.*, 3(8):673–683, 2004.
- [192] J.L. Medina-Franco, M.A. Giulianotti, G.S. Welmaker, and R.A. Houghten. Shifting from the single to the multitarget paradigm in drug discovery. *Drug Discov. Today*, 18(9-10):495–501, 2013.
- [193] N. Schneider, S. Hindle, G. Lange, R. Klein, J. Albrecht, H. Briem, K. Beyer, H. Claussen, M. Gastreich, C. Lemmen, and M. Rarey. Substantial improvements in large-scale redocking and screening using the novel hyde scoring function. *J. Comput.-Aided Mol. Des.*, 26(6):701–723, 2012.
- [194] Qt - a cross-platform application and ui framework. <http://qt-project.org/>. (accessed February 25, 2014).
- [195] Jenkins - an extendable open source continuous integration server. <http://jenkins-ci.org/>. (accessed February 25, 2014).

Bibliography of this Dissertation's Publications

- [D1] S. Urbaczek, A. Kolodzik, J.R. Fischer, T. Lippert, S. Heuser, I. Groth, T. Schulz-Gasch, and M. Rarey. Naomi: On the almost trivial task of reading molecules from different file formats. *J. Chem. Inf. Model.*, 51(12):3199–3207, 2011.
- [D2] A. Kolodzik, S. Urbaczek, and M. Rarey. Unique ring families: A chemically meaningful description of molecular ring topologies. *J. Chem. Inf. Model.*, 52(8):2013–2021, 2012.
- [D3] M. Hilbig, S. Urbaczek, I. Groth, S. Heuser, and M. Rarey. Mona - interactive manipulation of molecule collections. *J. Cheminform.*, 5(1):38, 2013.
- [D4] S. Urbaczek, A. Kolodzik, and M. Rarey. The valence state combination model - a generic framework for handling tautomers and protonation states. *J. Chem. Inf. Model.*, 54(3):756–766, 2014.
- [D5] S. Urbaczek, A. Kolodzik, S. Heuser, I. Groth, and M. Rarey. Reading pdb: Perception of molecules from 3d atomic coordinates. *J. Chem. Inf. Model.*, 53(1):76–87, 2013.
- [D6] S. Bietz, S. Urbaczek, B. Schulz, and M. Rarey. Protoss: A holistic approach to predict tautomers and protonation states in protein-ligand complexes. *J. Cheminform.*, 6(1):12, 2014.
- [D7] A. M. Henzler, S. Urbaczek, M. Hilbig, and M. Rarey. An integrated approach to knowledge-driven structure-based virtual screening. *J. Comput.-Aided Mol. Des.*, 28(9):927–939, 2014.
- [D8] A. M. Henzler, S. Urbaczek, and M. Rarey. Consistent handling of tautomers and protonation states in virtual screening. *in preparation*.
- [D9] K. Schomburg, S. Bietz, H. Briem, A. M. Henzler, S. Urbaczek, and Matthias Rarey. Facing the challenges of structure-based target prediction by inverse virtual screening. *J. Chem. Inf. Model.*, 54(6):1676–1686, 2014.

BIBLIOGRAPHY OF THIS DISSERTATION'S PUBLICATIONS

- [D10] M. v. Behren, A. Volkamer, A. M. Henzler, K. T. Schomburg, S. Urbaczek, and M. Rarey. Fast protein binding site comparison via an index-based screening technology. *J. Chem. Inf. Model.*, 53(2):411–422, 2013.

Appendix A

Publications and conference contributions

A.1 Publications in scientific journals

This section summarizes the author's publications in scientific journals and specifies the author's contributions.

- D1 **S. Urbaczek**, A. Kolodzik, J.R. Fischer, T. Lippert, S. Heuser, I. Groth, T. Schulz-Gasch and M. Rarey. NAOMI: On the Almost Trivial Task of Reading Molecules from Different File Formats. *Journal of Chemical Information and Modeling*, 51(12):3199-3207, 2011.

Based on preliminary work of S. Wefing, S. Urbaczek developed the concepts of the chemical model and the computational representation of molecules in the NAOMI framework as well as the procedures for the assignment of the necessary data (initialization procedures). S. Urbaczek, assisted by A. Kolodzik, designed and implemented the workflows for the interpretation of molecules from chemical file formats. The NAOMI software library was developed and implemented in a joint effort by S. Urbaczek, A. Kolodzik, J.R. Fischer, and T. Lippert. I. Groth, S. Heuser, and T. Schulz-Gasch provided general support. M. Rarey supervised the work.

- D2 A. Kolodzik, **S. Urbaczek**, and M. Rarey. Unique Ring Families: A Chemically Meaningful Description of Molecular Ring Topologies. *Journal of Chemical Information and Modeling*, 52(8):2013-2021, 2012.

A. PUBLICATIONS AND CONFERENCE CONTRIBUTIONS

The general idea for the URF was developed by A. Kolodzik and S. Urbaczek in a joint effort. A. Kolodzik both established the theoretical foundation for the description of the URF and developed the algorithms for their calculation. A. Kolodzik integrated the method into the NAOMI framework. Together, A. Kolodzik and S. Urbaczek designed the validation procedures which were carried out by A. Kolodzik. M. Rarey supervised the work.

- D3 M. Hilbig, **S. Urbaczek**, S. Heuser, I. Groth, and M. Rarey. MONA - Interactive Manipulation of Molecule Collections. *Journal of Cheminformatics*, 5(1):38, 2013.

A preliminary version of MONA, which provided the basis for the work presented in the publication, was developed as a student project under the supervision of M. Hilbig and was refined and improved by M. Hilbig and S. Urbaczek. M. Hilbig and S. Urbaczek developed the concepts for the design of the MolDB in a joint effort. This includes the distinction between molecules and instances as well as the idea of handling compound collections as sets of molecules rather than sets of instances. S. Urbaczek developed the algorithms, both canonicalization and string generation, for the generation of MolStrings and the procedures for the rebuilding of molecules from this description. Additionally, S. Urbaczek designed the methods for the calculation of molecular properties based on the NAOMI framework and also developed the necessary functionality for element and functional group filters. MONA was designed and implemented by M. Hilbig. S. Urbaczek assisted with design of the user interface and software testing. I. Groth and S. Heuser provided general support. M. Rarey supervised the work.

- D4 **S. Urbaczek**, A. Kolodzik, and M. Rarey. The Valence State Combination Model: A Generic Framework for Handling Tautomers and Protonation States. *Journal of Chemical Information and Modeling*, 54(3):756-766, 2014.

The general concept for the generation of tautomers and protonation states was developed by S. Urbaczek and A. Kolodzik in a joint effort. S. Urbaczek devised the VSC model which provides the theoretical framework of the associated methods. S. Urbaczek developed the algorithms for the selection of valence states, the generation of valid valence bonds structures, the scoring of solutions and the enumeration of the remaining results. A. Kolodzik developed the functionality for the partitioning of molecules into zones. S. Urbaczek developed the algorithmic

strategies for the canonicalization, normalization and generation workflows and integrated them into the NAOMI system. S. Urbaczek and A. Kolodzik devised and performed the evaluation in a joint effort. M. Rarey supervised the work.

- D5 **S. Urbaczek**, A. Kolodzik, S. Heuser, I. Groth and M. Rarey. Reading PDB: Perception of Molecules from 3D Atomic Coordinates. *Journal of Chemical Information and Modeling*, 53(1):76-87, 2013.

The general concept for the perception of molecules from 3D coordinates was developed by S. Urbaczek and A. Kolodzik in a joint effort. S. Urbaczek developed the algorithms and methods for the individual steps of the workflow and derived the parameters for the scoring of individual valence state assignments. S. Urbaczek integrated the methods in the NAOMI framework. Together, S. Urbaczek and A. Kolodzik designed the validation procedures which were carried out by S. Urbaczek. M. Rarey supervised the work.

- D6 S. Bietz, **S. Urbaczek**, B. Schulz, and M. Rarey. Protoss: A Holistic Approach to Predict Tautomers and Protonation States in Protein-Ligand Complexes. *Journal of Cheminformatics*, 6(1):12, 2014.

Precursors of the presented Protoss version were developed in two individual student projects. During his master thesis and a following project, S. Bietz created a prototype for the integration of tautomers and protonation states into Protoss which was based on a tautomer and protonation state generation module developed by S. Urbaczek. During the diploma thesis of B. Schulz, B. Schulz and S. Urbaczek, who supervised this project, refined and reimplemented this concept on the basis of the NAOMI model. Eventually, S. Bietz and S. Urbaczek completed this second prototype for the final version presented in the publication. S. Urbaczek devised the computational representation of proteins in the NAOMI framework and developed the procedures for the assignment of the necessary data (initialization procedures). This includes both the perception of proteins from PDB files and the calculation of initial hydrogen positions. S. Bietz contributed to the evaluation and the stabilization of the protein initialization procedure. Based on the valence state combination model, S. Urbaczek implemented the handling of tautomers and protonation states for both ligand and protein residues. S. Urbaczek also designed the methods for the identification of variable mode regions and the conversion of different tautomeric and protonation

A. PUBLICATIONS AND CONFERENCE CONTRIBUTIONS

states into interaction modes. S.Bietz developed the enhanced scoring scheme for the evaluation of hydrogen bonds, metal interactions, and unfavorable polar interactions. The integration of the chemical stabilities into the Protoss objective function was performed by S.Bietz and S. Urbaczek in a joint effort. S. Bietz also adapted and improved the network optimization algorithm. B. Schulz, in addition to his work during the development of the second prototype, contributed substantially to the improved stability of the respective software tool. The validation procedures were designed and carried out by S. Bietz. This includes the selection and preprocessing of the datasets, the implementation of the automated evaluation experiments, and the result analyses. M. Rarey supervised the work.

- D7 A.M. Henzler, **S. Urbaczek**, M.Hilbig, and M. Rarey. An Integrated Approach to Knowledge-Driven Structure-Based Virtual Screening. *Journal of Computer-Aided Molecular Design*, 28(9):927-939, 2014.

Although the RAISE approach is still based on the general ideas of the TrixX screening tool previously developed by I. Schellhammer and J. Schlosser, it is a modified reimplemention on the basis of the NAOMI framework. The RAISE platform has been implemented by S. Urbaczek and A.M. Henzler in a joint effort. S. Urbaczek developed the concepts for the representation of directed interactions in the NAOMI framework as well as the procedures for their generation in the context of both molecules and proteins. Based on this representation, S. Urbaczek developed a generic approach for the generation of RAISE descriptors. The strategy for the identification of apolar points in the binding pockets has been designed by A.M. Henzler. In a joint effort, S. Urbaczek and A.M. Henzler devised the general concepts, structures, and procedures for descriptor matching in the RAISE framework. The respective implementation and subsequent refinement was conducted by A.M. Henzler. A.M. Henzler developed all algorithms and strategies necessary for both the generation and assessment of ligand poses and the creation and evaluation of pharmacophores and combined these components into a working screening pipeline. M. Hilbig assisted with the integration of the MolDB into the screening procedure. A.M. Henzler designed and performed the evaluation studies. M. Rarey supervised the work.

- D8 A.M. Henzler, **S. Urbaczek**, S. Bietz, and M. Rarey. Consistent Handling of Tautomers and Protonation States in Virtual Screening. *Journal of Computer-Aided Molecular Design*, in preparation.

The general concept for the integration of protomers in the RAISE workflow was developed by A.M. Henzler and S. Urbaczek in a joint effort. Based on the VSC method and the Protoss approach, S. Urbaczek developed the workflow for the default protomer generation for both proteins and molecules. Furthermore, S. Urbaczek developed the concepts for the handling of directed interactions by a unified representation in the context of protomers. The transformation into RAISE multi-state descriptors was developed by A.M. Henzler and S. Urbaczek in a joint effort. A.M. Henzler adapted the methods for descriptor matching and pose generation. The final screening pipeline was also implemented by A.M. Henzler. The evaluation studies were designed by S. Urbaczek and A.M. Henzler in a joint effort and were performed by A.M. Henzler.

- D9 K. Schomburg, S. Bietz, H. Briem, A.M. Henzler, **S. Urbaczek**, and M. Rarey. Facing the Challenges of Structure-based Target Prediction by Inverse Virtual Screening. *Journal of Chemical Information and Modeling*, 54(6):1676-1686, 2014.

K.T. Schomburg established the general concept for the inverse screening procedure and implemented the associated workflow. Fundamental RAISE functionality was contributed by A. M. Henzler and S. Urbaczek. K.T. Schomburg designed the ProteinDB and implemented it with the assistance of S. Urbaczek and S. Bietz. In a joint effort, S. Urbaczek and S. Bietz developed the methods for the generation of unique string identifiers for the storage of proteins in the ProteinDB and the procedures to recreate proteins from this data. K.T. Schomburg designed and performed the evaluation studies. M. Rarey supervised the work.

- D10 M. v. Behren, **A. Volkamer**, A. M. Henzler, K. T. Schomburg, S. Urbaczek, and M. Rarey. Fast protein binding site comparison via an index-based screening technology. *Journal of Chemical Information and Modeling*, 53(2):411-422, 2013.

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

A. PUBLICATIONS AND CONFERENCE CONTRIBUTIONS

A.2 Conferences

This section lists the author's presentations at international conferences.

Talk **S. Urbaczek**, S. Bietz, M. Rarey, Automated Prediction of Tautomeric States in Protein-Ligand Complexes, 240th ACS National Meeting, 2010, Boston, USA

Poster **S. Urbaczek**, A. Kolodzik, R. Fischer, T. Lippert, M. Rarey, The File IO Round Robin Game: On the Development of a Consistent Chemical Representation, 9th International Conference on Chemical Structures, 2010, Noordwijkerhout, NL.

Poster **S. Urbaczek**, A. Kolodzik, S. Heuser, I. Groth, M. Rarey, NAOMI: On the Almost Trivial Task of Reading Molecules from Different File Formats, Gordon Research Conference, 2011, Tilton, USA

Appendix B

Additional Data

In this chapter, additional data which was not included in the publications comprising this thesis is presented. This includes a list of valence states currently supported by the NAOMI framework, a table containing the corresponding pairs of valence states needed for the selection step as well as the substructure patterns underlying the scoring step of the VSC method. Additionally, a more detailed description of the associated methods is provided.

B.1 Valence States

The construction of the valence state layer is the internal valence check of the NAOMI framework. Molecules are rejected if this stage could not be successfully passed. The causes for rejection are undefined element identities, failed assignment of valence states to atoms, and failed localization of aromatic bond orders. The process starts with a list containing all valence states associated with the atom’s element (see Table B.1 for a complete list). Each valence state is checked for compatibility with the atom and a list of accordant states is compiled. A valence state is compatible if its formal charge is identical to the atom’s and the following conditions for the numbers of single (N_{SB}), double (N_{DB}), and triple bonds (N_{TB}) hold true:

$$N_{SB}^{Atom} \leq N_{SB}^{State} \quad N_{DB}^{Atom} = N_{DB}^{State} \quad N_{TB}^{Atom} = N_{TB}^{State}$$

Aromatic bond orders (N_{AB}^{Atom}) are converted to additional single bonds (N_{SB}^{Add}) and double bonds (N_{DB}^{Add}) prior to each individual comparison in the following way:

$$N_{DB}^{Add} = \max(N_{DB}^{State} - N_{DB}^{Atom}, 0) \quad N_{SB}^{Add} = N_{AB}^{Atom} - N_{DB}^{Add}$$

B. ADDITIONAL DATA

Table B.1: Valence states of elements with covalent bonds. Valence states are represented as element symbol followed by the number of single, double, triple bonds and the formal charge. The most relevant valence states of each element are underlined. For all elements not listed generic valence states without bond orders are used.

Element	Valence states						
Hydrogen	<u>H100</u>	H000+	H000-				
Boron	<u>B300</u>	B400-					
Carbon	<u>C400</u>	<u>C210</u>	<u>C101</u>	C020	C300+	C300-	C001-
	C110-						
Nitrogen	<u>N300</u>	<u>N110</u>	<u>N210+</u>	<u>N400+</u>	N020+	N101+	N001
	N010-		N200-				
Oxygen	<u>O200</u>	<u>O010</u>	<u>O100-</u>	O110+	O300+	O001+	O000-2
Fluorine	<u>F100</u>	F000-					
Silicon	<u>Si400</u>						
Phosphorus	<u>P310</u>	P300	P400+	P600-	P500	P110	P001
Sulfur	<u>S220</u>	<u>S200</u>	<u>S010</u>	<u>S210</u>	S300+	S100-	S110+
	S400	S310+	S600	S020	S030	S410	S000-2
Chlorine	<u>Cl100</u>	Cl000-	Cl130	Cl120	Cl110	Cl020	Cl110
	Cl200	Cl020+	Cl300	Cl500			
Germanium	<u>Ge400</u>						
Arsenic	<u>As300</u>	As500	As310	As400+			
Selenium	<u>Se200</u>	Se220	Se210				
Bromine	<u>Br100</u>	Br000-	Br130	Br120	Br110	Br300	Br500
Iodine	<u>I100</u>	I000-	I200-	I120	I300	I500	I700

B.2 Corresponding Pairs of Valence States

As explained in [D4], the selection step in the VSC workflow is based on pairs of valence states corresponding to particular types of transformations. These are listed in Table B.2 for all cases which are considered in the current implementation of the protomer generation routines. This restriction is, however, not necessary and the method is able to handle every pair adhering to the rules for corresponding valence states shown in B.3. The protonation type pairs are implemented for all valence states and are not explicitly shown here. These can be easily derived from Table B.1 by application of the rules in B.3.

B.2 Corresponding Pairs of Valence States

Table B.2: Valence states and corresponding donors and acceptors used for the state selection step of the VSC method.

Element	Valence State	Protonation		Tautomer		Resonance	
		Donor	Acceptor	Donor	Acceptor	Donor	Acceptor
Carbon	C400	-	C300-	-	C210	-	-
	C210	-	C110-	C400	-	-	-
Nitrogen	N300	N400+	N200-	-	N110	-	N210+
	N110	N210+	N010-	N300	-	N200-	-
	N210+	-	N110	-	-	N300	-
	N200-	N300	-	-	-	-	N110
Oxygen	O200	O300+	O100-	-	O010	-	O110+
	O010	O110+	-	O200	-	O100-	-
	O100-	O200	O000-2	-	-	-	O200
	O110+	-	O010	-	-	O200	-
Sulfur	S200	S300+	S100-	-	S010	-	S110+
	S010	S110+	-	S200	-	S100-	-
	S100-	S200	S000-2	-	-	-	S200
	S110+	-	S010	-	-	S200	-

Table B.3: Rules for the identification of the different types of corresponding valence states.

Type	Single Bonds	Double Bonds	Formal Charge
Tautomer Donor	+2	-1	=
Tautomer Acceptor	-2	+1	=
Resonance Donor	+1	-1	-1
Resonance Acceptor	-1	+1	+1
Protonation Donor	+1	=	+1
Protonation Acceptor	-1	=	-1

B. ADDITIONAL DATA

B.3 Substructure Patterns

As explained in [D4], the scoring scheme of the VSC method is based on the identification of predefined structural fragments corresponding to either rings or functional groups. The resulting scores are calculated using the following equations:

$$S_{VSC} = \sum S_{ring} + \sum S_{group} \quad (\text{B.1})$$

$$S_{ring} = \sum cycle + \sum S_{sub} \quad (\text{B.2})$$

$$S_{group} = \sum S_{subgroup} \quad (\text{B.3})$$

The patterns for cycles (see Table B.5), substituents (see Table B.6) and functional groups (see Table B.4) are provided using SMILES-like identifiers. The scoring procedure is performed in two steps, the calculation of the bond order including tautomers and resonance forms and the subsequent calculation of the protonation score involving only the addition or removal of hydrogen atoms. For this reason two distinct score values are provided. It must be noted that the omission of particular patterns corresponds to a penalty as the generic scores are always smaller than the arbitrary value of the reference system (see [D4]).

Table B.4: Patterns and associated scores for functional groups used in the scoring step of the VSC method. The first value of the score pair represents the score of the bond order arrangement, the second value the score of the ionization state.

Pattern	Scores	Pattern	Scores
ON	(100,100)	O=N	(50,100)
NN	(100,100)	N=N	(100,100)
O	(100,100)	N	(100,100)
[NH+]	(100,111)	[NH2+]	(100,111)
[NH3+]	(100,111)	[NH4+]	(100,111)
O=CN	(100,100)	OC=N	(50,100)
O=S(=O)(N)NC=N	(100,100)	O=S(=O)(N)N=CN	(100,100)
N=CN	(100,100)	[NH+]=CN	(80,127)
[NH2+]=CN	(80,127)	N=[N+]=N	(100,0)
N=[N+]=[N-]	(100,100)	[N-]=[N+]=[N-]	(100,140)
C=NNC=O	(100,100)	O=CO	(100,100)
O=C[O-]	(100,111)	O=C(O)O	(100,100)
O=C(O)[O-]	(100,111)	O=C([O-])[O-]	(100,112)
N=C(N)N	(100,100)	[N+]=C(N)N	(80,100)

B.3 Substructure Patterns

$[\text{NH}_+] = \text{C}(\text{N})\text{N}$	(80,132)	$[\text{NH}_2+] = \text{C}(\text{N})\text{N}$	(80,132)
$\text{N} = \text{C}(\text{N})\text{NC}(=\text{N})\text{N}$	(100,100)	$\text{N} = \text{C}(\text{N})\text{N} = \text{C}(\text{N})\text{N}$	(100,100)
$[\text{NH}_2+] = \text{C}(\text{N})\text{NC}(=\text{N})\text{N}$	(100,132)	$[\text{NH}_2+] = \text{C}(\text{N})\text{N} = \text{C}(\text{N})\text{N}$	(100,132)
$\text{N} = \text{C}(\text{N})[\text{NH}_+] = \text{C}(\text{N})\text{N}$	(100,132)	$\text{ONC}(=\text{N})\text{N}$	(100,100)
$\text{ON} = \text{C}(\text{N})\text{N}$	(110,100)	$\text{NNC}(=\text{O})\text{N}$	(100,100)
$\text{NNC}(=\text{S})\text{N}$	(100,100)	$\text{SNC} = \text{O}$	(100,100)
$\text{SNC}(=\text{O})\text{N}$	(100,100)	$\text{SC} = \text{O}$	(100,100)
$\text{SC}(=\text{O})\text{O}$	(100,100)	$\text{N} = \text{S} = \text{N}$	(100,100)
$\text{O} = \text{C}(\text{N})\text{N}$	(100,100)	$\text{OC}(=\text{N})\text{N}$	(80,100)
$\text{O} = \text{C}$	(100,100)	$\text{OC} = \text{C}$	(50,100)
$\text{S} = \text{C}(\text{N})\text{N}$	(100,100)	$\text{SC}(=\text{N})\text{N}$	(80,100)
$\text{ONC}(\text{S}) = \text{O}$	(100,100)	$\text{ON} = \text{C}$	(100,100)
$\text{NN} = \text{C}$	(100,100)	$\text{O} = [\text{N}^+]\text{O}$	(100, 0)
$\text{O} = [\text{N}^+][\text{O}^-]$	(100,100)	$\text{O} = [\text{N}^+](\text{O})\text{O}$	(100, 0)
$\text{O} = [\text{N}^+](\text{O})[\text{O}^-]$	(100,200)	$\text{NNC} = \text{O}$	(100,100)
$\text{NNC}(=\text{N})\text{N}$	(100,100)	$[\text{N}^+]\text{NC}(=\text{N})\text{N}$	(100,100)
$\text{O} = \text{S}(=\text{O})\text{N}$	(100,100)	$\text{O} = \text{S}(=\text{O})[\text{NH}^-]$	(100, 85)
$\text{O} = \text{S}(=\text{O})[\text{N}^-]$	(100, 85)	$\text{O} = \text{S}(=\text{O})(\text{N})\text{N}$	(100,100)
$\text{O} = \text{S}(=\text{O})\text{NC}(=\text{O})\text{N}$	(100,100)	$\text{O} = \text{S}(=\text{O})[\text{N}^-]\text{C}(=\text{O})\text{N}$	(100,105)
$\text{SC}(=\text{O})\text{N}$	(100,100)	$\text{S} = \text{CN}$	(100,100)
$\text{SC} = \text{N}$	(80,100)	$\text{O} = \text{C}(\text{O})\text{N}$	(100,100)
$\text{ONC}(=\text{O})\text{N}$	(100,100)	$\text{ONC} = \text{O}$	(100,100)
$[\text{O}^-]\text{NC} = \text{O}$	(100,93)	$\text{ON} = \text{CN}$	(100,100)
$\text{O} = \text{S} = \text{O}$	(100,100)	$\text{O} = \text{PO}$	(100,100)
$\text{O} = \text{P}[\text{O}^-]$	(100,120)	$\text{O} = \text{P}(\text{O})\text{O}$	(100,100)
$\text{O} = \text{P}(\text{O})[\text{O}^-]$	(100,123)	$\text{O} = \text{P}([\text{O}^-])[\text{O}^-]$	(100,125)
$\text{O} = \text{P}(\text{N})(\text{N})\text{N}$	(100,100)	$\text{O} = \text{P}(\text{O})(\text{O})\text{O}$	(100,100)
$\text{O} = \text{P}(\text{O})(\text{O})[\text{O}^-]$	(100,123)	$\text{O} = \text{P}(\text{O})([\text{O}^-])[\text{O}^-]$	(100,124)
$\text{O} = \text{P}([\text{O}^-])([\text{O}^-])[\text{O}^-]$	(100,125)	$\text{S} = \text{P}(\text{O})(\text{O})\text{O}$	(120,100)
$\text{S} = \text{P}(\text{O})(\text{O})[\text{O}^-]$	(120,120)	$\text{S} = \text{P}(\text{O})([\text{O}^-])[\text{O}^-]$	(120,121)
$\text{S} = \text{P}([\text{O}^-])([\text{O}^-])[\text{O}^-]$	(120,122)	$\text{SP}(=\text{O})(\text{O})\text{O}$	(100,100)
$\text{SP}(=\text{O})(\text{O})[\text{O}^-]$	(100,121)	$\text{SP}(=\text{O})([\text{O}^-])[\text{O}^-]$	(100,122)
$\text{S} = \text{P}(\text{S})(\text{O})\text{O}$	(100,100)	$\text{SP}(=\text{O})(\text{O})\text{N}$	(100,100)
$\text{SP}(=\text{O})([\text{O}^-])\text{N}$	(100,120)	$\text{O} = \text{P}(\text{O})(\text{O})\text{N}$	(100,100)
$\text{O} = \text{P}(\text{O})([\text{O}^-])\text{N}$	(100,120)	$\text{O} = \text{P}([\text{O}^-])([\text{O}^-])\text{N}$	(100,125)
$\text{O} = \text{P}(\text{O})(\text{O})[\text{N}^-]$	(100, 90)	$\text{O} = \text{P}(\text{O})([\text{O}^-])[\text{N}^-]$	(100,110)
$\text{O} = \text{P}([\text{O}^-])([\text{O}^-])[\text{N}^-]$	(100,115)	$\text{O} = \text{P}(\text{O})\text{N}$	(100,100)
$\text{O} = \text{P}([\text{O}^-])\text{N}$	(100,120)	$\text{O} = \text{P}(\text{O})[\text{N}^-]$	(100, 90)
$\text{O} = \text{P}([\text{O}^-])[\text{N}^-]$	(100,110)	$\text{O} = \text{P}(\text{O})(\text{N})\text{N}$	(100,100)
$\text{O} = \text{P}([\text{O}^-])(\text{N})\text{N}$	(100,120)	$\text{O} = \text{S}(=\text{O})(\text{O})\text{N}$	(100,100)
$\text{O} = \text{S}(=\text{O})([\text{O}^-])\text{N}$	(100,180)	$\text{O} = \text{S}(=\text{O})\text{O}$	(100,100)
$\text{O} = \text{S}(=\text{O})[\text{O}^-]$	(100,200)	$\text{O} = \text{S}(=\text{O})(\text{O})\text{O}$	(100,100)
$\text{O} = \text{S}(=\text{O})(\text{O})[\text{O}^-]$	(100,200)	$\text{O} = \text{S}(=\text{O})([\text{O}^-])[\text{O}^-]$	(100,220)

B. ADDITIONAL DATA

O=S(=O)S	(100,100)	O=[N+](O)N	(100, 0)
O=[N+](O)N	(100,100)	N#C	(100,100)
[O-][N+]=C	(100,100)	O[N+]=C	(100, 0)
S	(100,100)	[S-]	(100, 80)
[S+]	(100,100)	[F-]	(100,100)
[Cl-]	(100,100)	[Br-]	(100,100)
[I-]	(100,100)	[O+]#[C-]	(100,100)

Table B.5: Patterns and associated scores for cycles used in the scoring step of the VSC method. In this particular case the SMILES-like identifiers differ from their usual meaning. The symbol [C] represents carbon atoms with an sp³ hybridization, the symbol C without an adjacent = represents carbon atoms with an sp² hybridization in which the double bond is not part of the cycle. The first value of the score pair represents the score of the bond order arrangement in case of isolated rings, the second represents the score of the bond order arrangement in case of rings which are fused to another conjugated rings, the third value the score of the ionization state.

Pattern	Scores	Pattern	Scores
c1ccccc1	(100,100,100)	C1C=CCCC=1	(95, 40,100)
C1=CCC=CC1	(100, 70,100)	n1ccccc1	(100,100,100)
[n+]1ccccc1	(80,100,100)	[nH+]1ccccc1	(100,100, 93)
N1CCC=CC=1	(99,100,100)	N1CC=CCC1	(99,100,100)
N1C=CCC=C1	(99,100,100)	N1C=CC=CC1	(99,100,100)
[N-]1C=CC=CC1	(99,100, 93)	[N-]1C=CCC=C1	(99,100, 93)
n1ncccc1	(100,100,100)	N1N=CC=CC1	(100,100,100)
N1N=CCC=C1	(100,100,100)	N1NCC=CC1	(100,100,100)
n1cnccc1	(100,100,100)	[n+]1cnccc1	(80,100,100)
[nH+]1cnccc1	(100,100, 93)	N1C=CCNC1	(99,100,100)
[N-]1CNC=CC1	(99,100, 93)	[N-]1C=CCNC1	(99,100, 93)
N1C=CC=NC1	(99,100,100)	N1CN=CCC1	(99,100,100)
N1C=NC=CC1	(99,100,100)	[N-]1C=NC=CC1	(99,100, 93)
[NH+]1=CNC=CC1	(99,100, 93)	N1C=NCC=C1	(99,100,100)
[N-]1C=NCC=C1	(99,100, 93)	N1CN=CC=C1	(99,100,100)
N1CNC[C]C1	(60,100,100)	N1CN=C[C]C1	(60,100,100)
N1C=NC[C]C=1	(60,100,100)	[NH+]1C=NC[C]C=1	(60,100,110)
[NH+]1=CN=C[C]C1	(60,100,110)	N1[C]C=CNC1	(80,100,100)
n1ccncc1	(100,100,100)	[nH+]1ccncc1	(100,100, 93)
N1C=CNC=C1	(80,100, 80)	N1C=CN=CC1	(99,100,100)
N1C=NCNC1	(99,100,100)	N1C=CNCC1	(99,100,100)
[N-]1C=CNCC1	(99,100, 80)	n1cnenc1	(100,100,100)
N1C=NC=NC1	(100,100,100)	N1C=NCN=C1	(100,100,100)
N1CNCNC1	(100,100,100)	[N-]1CNCNC1	(100,100, 80)

B.3 Substructure Patterns

n1ncnc1	(100,100,100)	N1N=CCN=C1	(100,100,100)
N1N=CN=CC1	(100,100,100)	N1N=CC=NC1	(100,100,100)
N1C=CN=NC1	(100,100,100)	N1N=CNC=C1	(80,100, 80)
N1NC=CN=C1	(80,100, 80)	N1N=CCNC1	(100,100,100)
[N-]1CNN=CC1	(100,100, 80)	N1C=NN=CC1	(100,100,100)
[N-]1C=NN=CC1	(100,100, 80)	O1C=CN=CC1	(80,0,80)
O1CCNC=C1	(80,0,80)	N1C=CC=C1	(100,100,100)
N1C=CCC1	(100,100,100)	N1=CC=CC1	(90,100,100)
N1CC=CC1	(100,100,100)	N1C=C=CC1	(100,100,100)
N1[C]C=CC1	(90,100,100)	N1C=C[C]C1	(90,100,100)
S1C=CC=C1	(100,100,100)	O1C=CC=C1	(100,100,100)
N1N=CC=C1	(100,100,100)	N1N=CCC1	(100, 90,100)
N1=NCC=C1	(90,100, 80)	N1NCC=C1	(90,100, 80)
N1C=NC=C1	(100,100,100)	[N+]1C=NC=C1	(100,100,100)
[NH+]1=CNC=C1	(100,100, 95)	[N-]1C=NC=C1	(100,100, 60)
N1C=NCC1	(100,100,100)	N1CC=NC=1	(100,100,100)
N1CCN=C1	(100,100,100)	N1C=CNC1	(100,100,100)
N1[C]CNC1	(95,100,100)	N1C=N[C]C1	(100,100,100)
[NH+]1[C]CNC=1	(100,100,90)	O1N=CC=C1	(100,100,100)
O1NCC=C1	(90,100,100)	O1C=NC=C1	(100,100,100)
N1N=NC=C1	(100,100,100)	N1N=CC=N1	(90, 0, 90)
N1N=CN=C1	(100,100,100)	N1C=NN=C1	(100,100,100)
N1N=CNC1	(100,100,100)	N1NC=NC1	(100,100,100)
N1NC=NC(=O)1	(100,100,100)	N1N=CNC(=O)1	(100,100,100)
N1N=NN=C1	(100,100,100)	N1N=NC=N1	(100,100,100)
[N-]1N=NN=C1	(100,100,105)	[N-]1N=NC=N1	(100,100,105)
S1N=CC=C1	(100,100,100)	S1C=NC=C1	(100,100,100)
S1C=CNC1	(100,100,100)	S1C=[N+]C=C1	(80,100,100)
O1N=CC=N1	(100,100,100)	O1NCC=N1	(80,100, 80)
O1N=CN=C1	(100,100,100)	O1C=NN=C1	(100,100,100)
S1N=NC=C1	(100,100,100)	S1N=CN=C1	(100,100,100)
S1C=NN=C1	(100,100,100)	S1N=CC=N1	(100,100,100)
S1NCC=N1	(80,100,100)		
[N-]1C=CC=NC=CC[N-]CC=CN=CC=C1			(200,100,360)
N1C=CC=NC=CCNCC=CN=CC=C1			(200,100,200)

B. ADDITIONAL DATA

Table B.6: Patterns and associated scores for ring substituents used in the scoring step of the VSC method. The first value of the score pair represents the score of the bond order arrangement, the second value the score of the ionization state.

Pattern	Scores	Pattern	Scores
CN	(100, 100)	C=N	(60, 60)
C=[NH+]	(60, 50)	C=[NH2+]	(60, 50)
CO	(100, 100)	C[O-]	(100, 90)
C=O	(115, 115)	CS	(100, 100)
C[S-]	(100, 90)	C=S	(115, 115)
C=NO	(110, 110)	CN=O	(100, 100)
C=CN	(110, 110)	CC=N	(100, 100)
C=NN	(120, 120)	CN=N	(100, 100)
NO	(100, 100)	[N+]O	(140, 0)
[N+][O-]	(140, 160)	S=O	(150, 150)

Appendix C

Software Architecture

In this chapter, the software architecture of the NAOMI framework is presented including a description of the underlying software components, the application programs based on them and the automated test framework used to ensure that these programs are reliable and have a high degree of internal consistency. The aim is to provide some insight into the design decisions made during the development of the NAOMI framework considering its potential application scope and to give an idea about the effort put into the internal consistency and reliability with respect to the developed algorithms and software tools. The chapter is organized into three sections. First, the general concepts, central classes, and mutual dependencies of the internal libraries are presented and discussed. Then, an overview about the numerous application programs is given including tools developed by the author and additional software based on the library. The third section gives a quick overview about the automated test framework which is used to ensure stable development and consistent results.

C.1 Software Libraries

The NAOMI framework follows a library-centric software design principle with the aim to provide wide-ranging and highly reusable components for the rapid development of both new scientific methods and application programs in the context of CADD. The individual libraries encapsulate data structures and associated functionality with a clearly defined scope of application. The purpose of this section is not to provide a comprehensive and detailed description of each component including its classes and functions but to give a general insight into the system's structure and to relate it to the methods presented in the previous chapters. For a better overview, the libraries are classified into different categories (see figure C.1). **Support libraries** provide basic functionality

C. SOFTWARE ARCHITECTURE

which is essentially required in many applications in the context of cheminformatics. As they do not depend on other libraries of the system they can be considered as its lowest hierarchical level. **Core libraries** focus on those chemical objects which play the most central role in typical cheminformatics and screening applications. These are small molecules, proteins, protein-ligand complexes, and molecular interactions. As has been described previously, a lot of effort has been put into the appropriate and consistent modeling of each of these structures. **Database libraries** implement the different database schemes, while **Application libraries** incorporate the data structures and algorithms for specific applications. In the following six libraries will be discussed in order to present the core structure of the NAOMI framework.

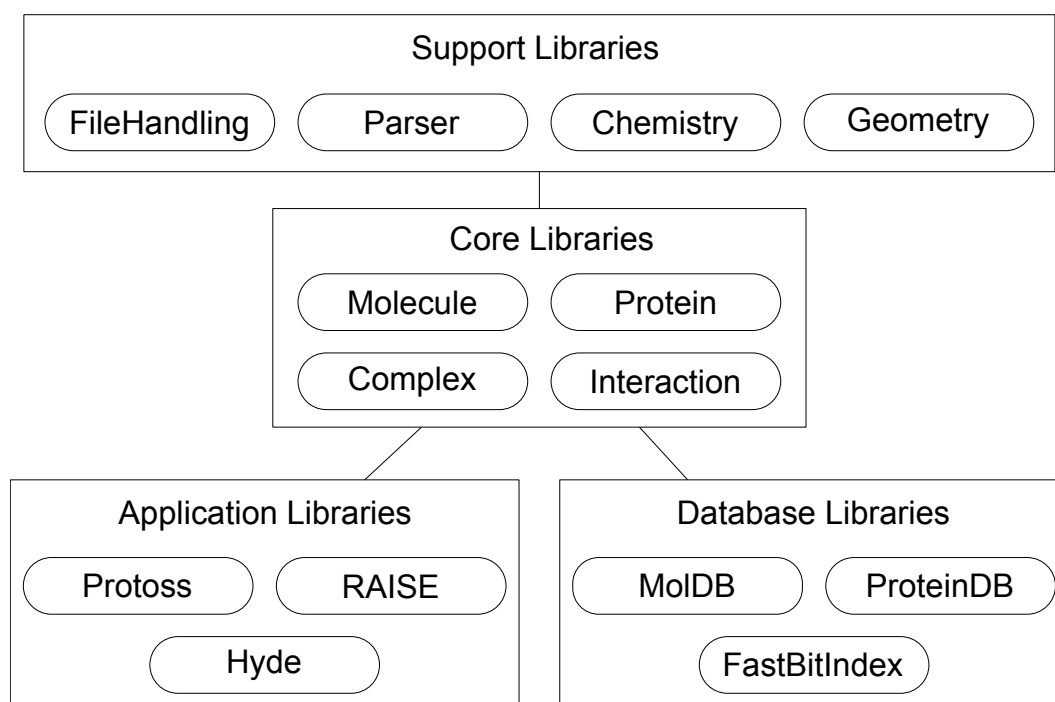


Figure C.1: General structure of the libraries in the NAOMI framework.

File Handling - Support Library

Specialized file formats are the most common source of chemical structure data in the context of cheminformatics. The corresponding files are structured in such a way that they comprise multiple entries corresponding to different chemical entities. The main purpose of the File Handling library is to process the supported types of chemical data files (SDF, MOL2, Smiles, PDB) and to provide access to their individual entries.

This is the first step of the input procedure shown in Figure C.4. The most central classes are `InputSplitter`, `IODevice`, and `MultiDeviceHandler`. The `InputSplitter` is an abstract base class with the purpose to split an input file into its individual entries. As the exact way to do this depends on the respective chemical file format, different `InputSplitters` have to be implemented for each of these cases. `IODevice` is a container for exactly one input file and provides access to its individual entries in textual form. `MultiDeviceHandler` essentially offers a consecutive index for the unified handling of multiple input devices.

Parser - Support Library

Chemical file formats are based on different concepts for the description of molecules, e.g., string identifiers or connection tables. Each of these has its own specification and needs to be interpreted accordingly. The main purpose of the Parser library is to provide parsers for each supported chemical file format which are used to both extract the included data and at the same time check the adherence to format specifications. The parsed information is stored in specialized data objects which can be passed on to other libraries. This is the second step of the input procedure shown in Figure C.4. Currently, parsers for the following formats are included: PDB, MOL, MOL2, SDF, SMILES.

Chemistry - Support Library

The Chemistry library incorporates the data associated with different layers of the internal chemical model presented in Section 2.1. This includes both classes representing the basic objects and their properties (`Element`, `ValenceState`, `AtomType`) and global containers reflecting their respective connections (see figure C.2). This information is needed in the third step of the of the input procedure shown in figure C.4.

Molecule - Core Library

The Molecule library is the most extensive of the core libraries and is at the heart of the NAOMI framework. Its central class is the `Molecule` which is implemented as an undirected graph based on an adjacency-list representation. Its nodes and edges correspond to `Atoms` and `Bonds` respectively. The latter structures play a prominent role throughout the core libraries as they incorporate both the internal chemical description as well as information about the molecule's topology (see figure C.3).

As `Molecules` are complex objects, they are created by specialized builder classes which use the format-specific data provided by the Parser library in order to build

C. SOFTWARE ARCHITECTURE

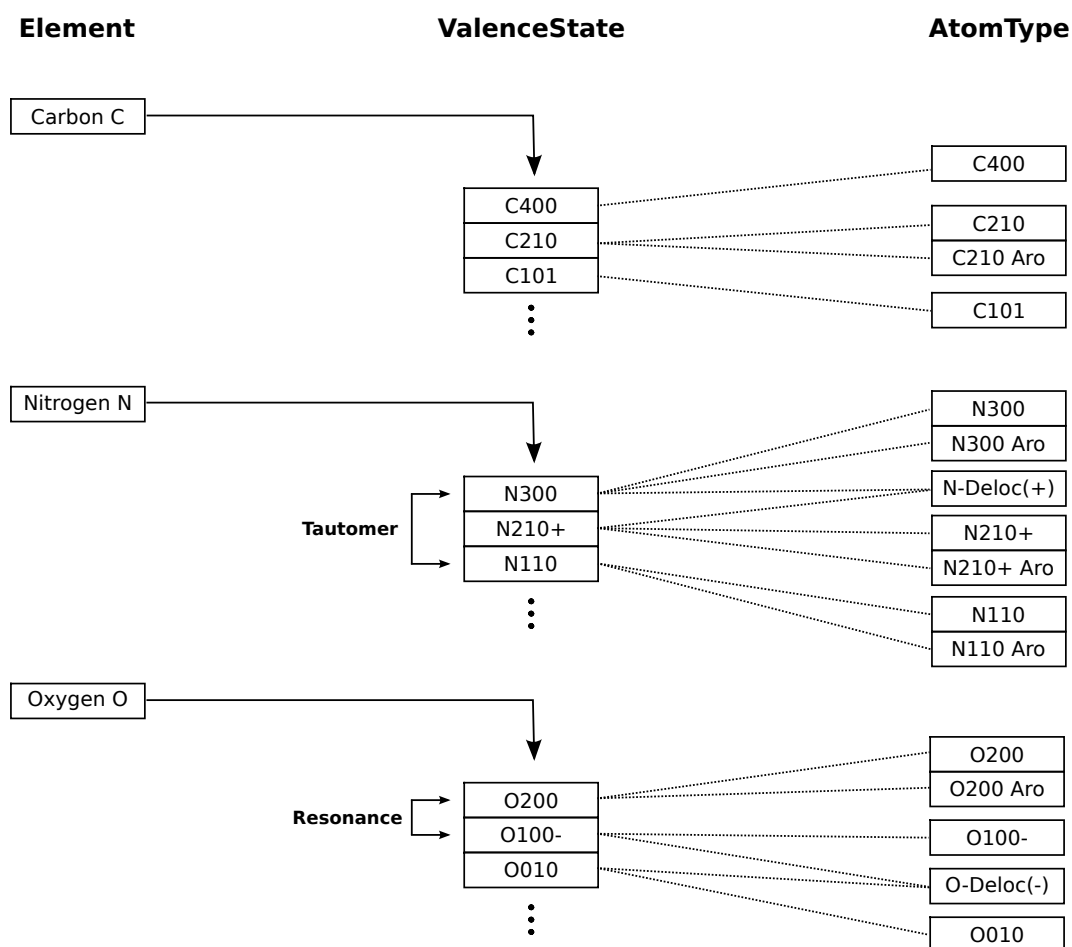


Figure C.2: Internal structure of the Chemistry library.

the internal representation according to the procedures described in section 2.3. This workflow generally consists of two parts, a format-dependent build-up step in which the complete graph structure is created and a format-independent initialization procedure during which additional information, e.g., atom types, functional groups, and rings, is calculated (see figure C.4).

The Molecule library additionally includes a number of submodules providing functionality for canonicalization, superposition, export to different file formats, and the calculation of physicochemical properties (see table C.1). Each submodule incorporates its own data structures which can be used by other libraries.

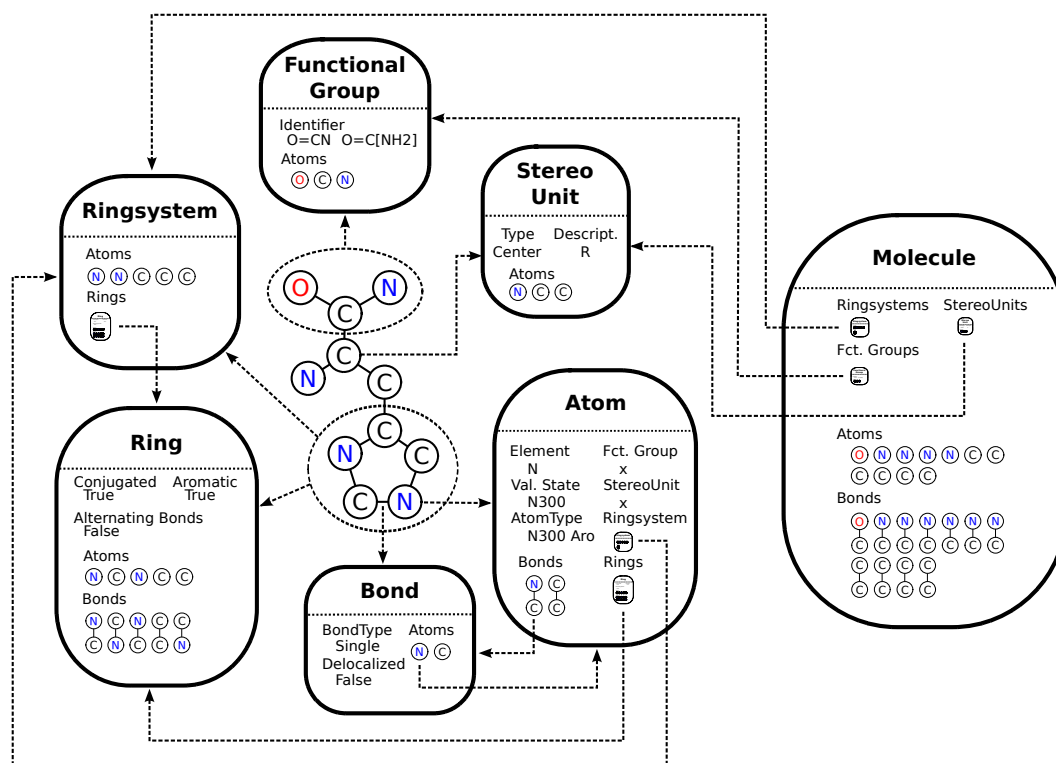


Figure C.3: Internal representation of molecules in the NAOMI framework.

Protein - Core Library

As was explained in Section 3.1, the internal representation of proteins in the NAOMI framework is largely based on the same models developed for the description of molecules. This is directly reflected on the software level as shown in figure C.5. Proteins are composed of molecules, which represent their connected molecular components, and individual Residues. Additionally, they include data structures which allow the mapping of particular atoms to their respective residues. Not storing the information about residues in the Atom class was a conscious decision in order to avoid mutual dependencies between the libraries.

Proteins are created by specialized builder classes based on the same general procedure shown in figure C.4. As they consist of the same data structures as molecules, functionality from the submodules of the Molecule library can be easily reused. The protein library also has a similar structure as the latter and thus includes additional submodules (see table C.2). Some functionality such as the superimposition of protein structures is implemented in a library of its own due to the inherent complexity of the task.

C. SOFTWARE ARCHITECTURE

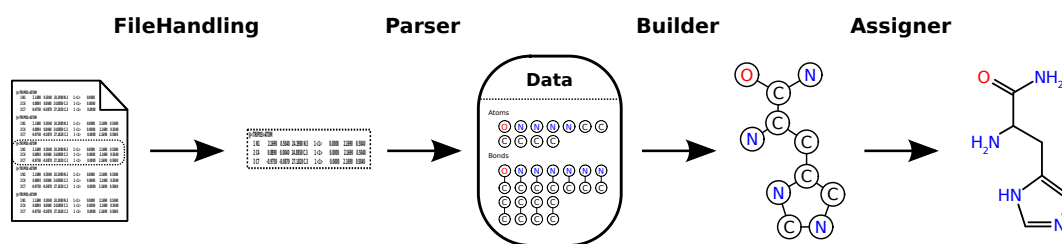


Figure C.4: General procedure for the creation of Molecules in the NAOMI framework.

Table C.1: Submodules of the Molecule library.

Submodule	Description
Assigner	Builders for the internal chemical description
Builder	MoleculeBuilders for different file formats
Canonizer	Canonicalization of molecules and atoms
MultiState	Handling of protomers
Properties	Calculation of physicochemical properties
Superposer	Superposition of molecules
Writer	Export of molecules to different file formats

Complex - Core Library

Complexes in the NAOMI framework are essentially a combination of proteins and molecules, so that the Complex library directly builds on the functionality associated with the respective objects. Apart from the Complex class, the library introduces the ActiveSite as an additional data structure. A Complex is a composition of a protein, waters, metals, and all remaining molecules. The ActiveSite class is structured similarly, but does include a list of residues rather than the complete protein (see figure C.6).

Table C.2: Submodules of the Protein library.

Submodule	Description
Builder	ProteinBuilders for different file formats
Chain	Handling of protein chains
MultiState	Handling of protomers in residues
Writer	Export of proteins to different file formats

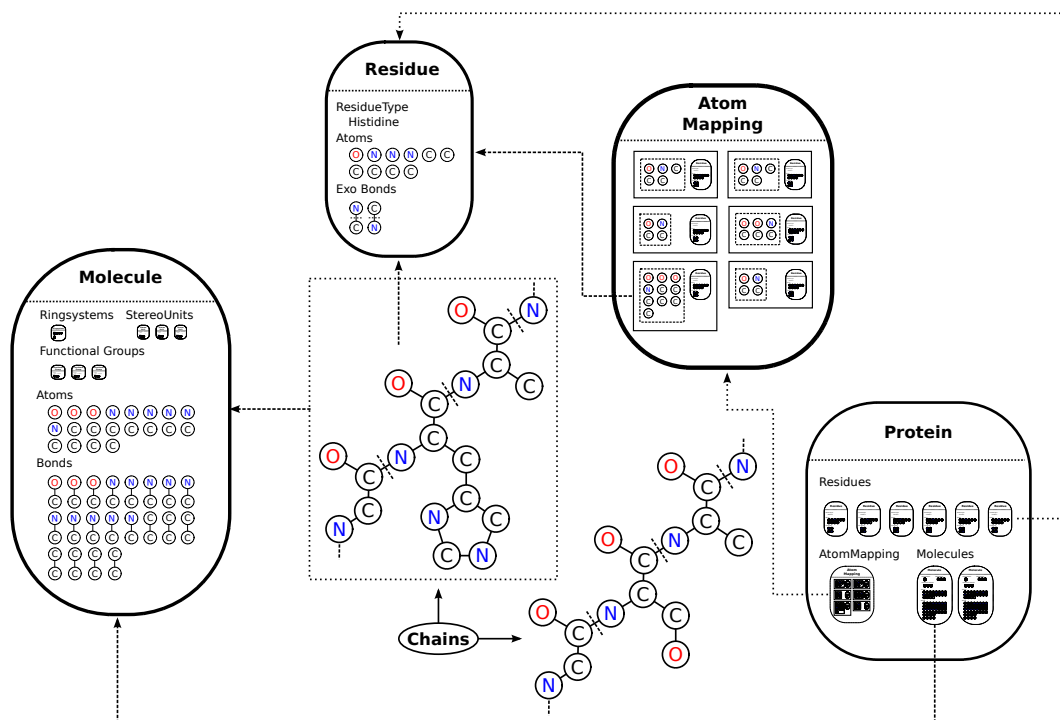


Figure C.5: Internal representation of Proteins in the NAOMI framework.

C.2 Application Software

The NAOMI software system includes a large number of application programs which have been implemented using the functionality provided by the software libraries presented in the previous section. These range from small command-line tools to complex GUI-based programs covering a broad scope of different application scenarios.

C.3 Software Testing

In order to ensure both the constant scientific quality of the included methods and algorithms as well as the technical reliability of its individual components, the NAOMI framework employs a number of completely automated software tests.

Unit-Tests

The unit tests are based on the QtTest framework[194] and can be compiled and executed locally. They exist for each of the internal libraries and usually cover most of their basic functionality.

C. SOFTWARE ARCHITECTURE

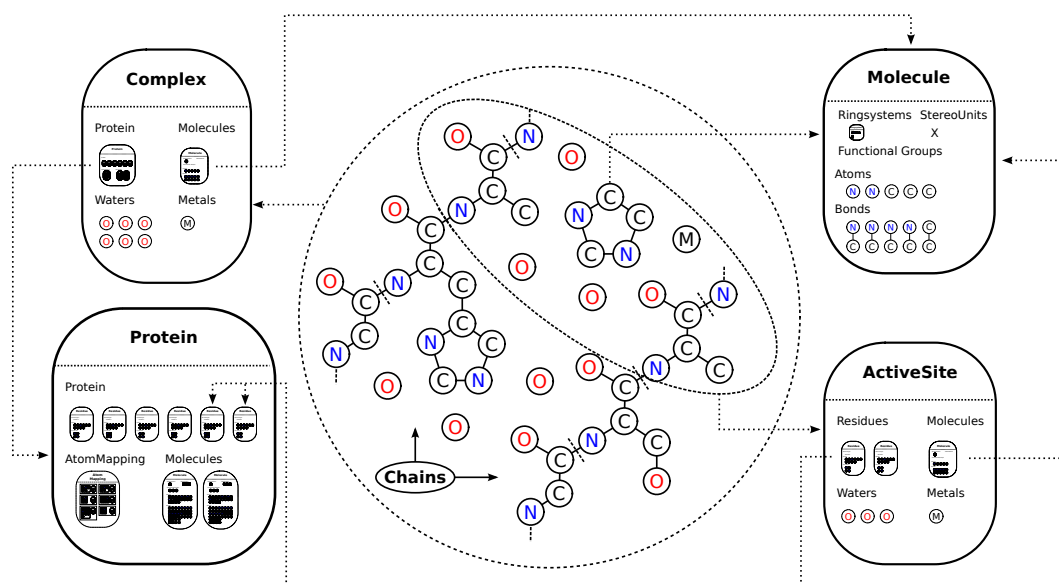


Figure C.6: Internal representation of Complexes and ActiveSites in the NAOMI framework.

System-Tests

The internal system tests are based on the continuous integration tool jenkins[195] and which is extended by a python-based automated test framework developed and maintained by the BioSolveIT. On the one hand, this approach ensures that both libraries and application software can be compiled and executed on different platforms. On the other, it allows to perform tests encompassing the complete system in an automated fashion.

Table C.3: Software applications using the NAOMI framework.

Tool	Type	Availability
LeadIT ¹	GUI application	www.biosolveit.de/LeadIT
MONA	GUI application	www.zbh.uni-hamburg.de/Mona
NAOMI	Command-line tool	www.zbh.uni-hamburg.de/NAOMI
Protoss	Web server	www.zbh.uni-hamburg.de/protoss
SMARTSviewer	Web server	smartsview.zbh.uni-hamburg.de
SMARTSeditor	GUI application	www.zbh.uni-hamburg.de/smartseditor
SeeSAR	GUI application	www.biosolveit.de/SeeSAR
Torsion Analyzer	GUI application	www.biosolveit.de/TorsionAnalyzer

¹The handling of small-molecules is based on NAOMI

Appendix D

Software Tools

NAOMI Converter

The NAOMI converter is a command-line tool for the conversion of the most common chemical file formats (MOL2, SDF, SMILES, PDB). It is completely based on the NAOMI software system and incorporates both the method for the consistent conversion of molecules described in [D1] and the method the perception of molecules from three-dimensional coordinates presented in [D5]. As a batch processing tool, the NAOMI converter does not need any kind of manual intervention and all relevant program options can be specified using predefined command-line parameters.

NAOMI
Molecule file converter (2.0)

Authored by: S.Urbaczek, A.Kolodzik, R.Fischer and T.Lippert
Many thanks to: Prof. M.Rarey, Dr. H.Claussen (BioSolveIT),
and R.Kraus (BioSolveIT)

Supported file formats:

Input:

*.mol *.mol2 *.pdb *.sdf *.smi *.smiles

Output:

*.mol *.mol2 *.sdf *.smi *.smiles

General options:

-h [--help] Prints help message
-v [--verbosity] arg Set verbosity level (0 = Quiet ,

D. SOFTWARE TOOLS

1 = Errors , 2 = Warnings , 3 = Info)

Input options :

-i [--input] arg Input file(s), suffix is required.
Several input files can be
given separated by spaces ,
e.g. -i a.mol2 b.sdf

Output options :

-o [--output] arg Output file , suffix is required.

Configuration :

--all Convert all components from entry
(Largest Component is converted
by default).

Although being first and foremost a conversion tool, the NAOMI converter can also be applied in several other scenarios depending on the input format and the chosen options.

- Concatenation of Datasets - The NAOMI can accept multiple input files and combines their respective entries in a single output file
- Concatenation of Datasets - The NAOMI can accept multiple input files and combines their respective entries in a single output file
- Cleanup of Datasets - The NAOMI converter does not convert neither invalid file entries nor invalid molecules. This feature can be used to eliminate such entries from datasets.
- Calculation of three-dimensional hydrogen coordinates - The NAOMI adds missing hydrogens to molecules and calculates three-dimensional coordinates if possible. If no conversion is performed, meaning that input and output file have the same format, this feature can be used to eliminate invalid entries.
- Generation of unique SMILES - The NAOMI converter generates unique SMILES by default.
- Extraction of small molecules from pdb data

Appendix E

Journal articles

NAOMI: On the Almost Trivial Task of Reading Molecules from Different File Formats

[D1] S. Urbaczek, A. Kolodzik, J.R. Fischer, T. Lippert, S. Heuser, I. Groth, T. Schulz-Gasch and M. Rarey. NAOMI: On the Almost Trivial Task of Reading Molecules from Different File Formats. *Journal of Chemical Information and Modeling*, 51(12):3199-3207, 2011.

<http://pubs.acs.org/articlesonrequest/AOR-hRTTf9abf9PGggQX9ztR>

Reproduced with permission from S. Urbaczek, A. Kolodzik, J.R. Fischer, T. Lippert, S. Heuser, I. Groth, T. Schulz-Gasch and M. Rarey. NAOMI: On the Almost Trivial Task of Reading Molecules from Different File Formats. *Journal of Chemical Information and Modeling*, 51(12):3199-3207, 2011. Copyright 2011 American Chemical Society.

NAOMI: On the Almost Trivial Task of Reading Molecules from Different File formats

Sascha Urbaczek,[†] Adrian Kolodzik,[†] J. Robert Fischer,[†] Tobias Lippert,[†] Stefan Heuser,^{†,||} Inken Groth,[‡] Tanja Schulz-Gasch,[§] and Matthias Rarey^{*,†}

[†]Center for Bioinformatics (ZBH), Bundesstrasse 43, 20146 Hamburg, Germany

[‡]Research Active Ingredients, Beiersdorf AG, Tropelwitzstrasse 15, 22529 Hamburg, Germany

[§]Pharmaceutical Division, F. Hoffmann-La Roche Ltd., CH-4070 Basel, Switzerland

 Supporting Information

ABSTRACT: In most cheminformatics workflows, chemical information is stored in files which provide the necessary data for subsequent calculations. The correct interpretation of the file formats is an important prerequisite to obtain meaningful results. Consistent reading of molecules from files, however, is not an easy task. Each file format implicitly represents an underlying chemical model, which has to be taken into consideration when the input data is processed. Additionally, many data sources contain invalid molecules. These have to be identified and either corrected or discarded. We present the chemical file format converter NAOMI, which provides efficient procedures for reliable handling of molecules from the common chemical file formats SDF,¹ MOL2,² and SMILES.³ These procedures are based on a consistent chemical model which has been designed for the appropriate representation of molecules relevant in the context of drug discovery. NAOMI's functionality is tested by round robin file IO exercises with public data sets, which we believe should become a standard test for every cheminformatics tool.



INTRODUCTION

Chemical file formats provide the necessary data for application programs and offer a means to share results with other scientists in a computer readable form. For small molecules, the most commonly used formats are Symyx SDF V2000 (formerly MDL SDF),¹ Tripos MOL2,² and Daylight SMILES.³ Virtually all public databases provide molecular files of at least one of these types.

Unfortunately, many programs do not accept all formats as input or generate only some of them as output. Hence, file format converters are needed to exchange data between these tools. This becomes especially important if several of these tools are combined in a workflow. The consistent conversion of molecules is crucial at this point, since even minor alterations might result in errors in subsequent calculations.

The conversion process is difficult and error prone. File formats implicitly represent an underlying chemical model which has to be taken into account. Hence, the file format conversion is actually a conversion between different chemical representations. Furthermore, some programs generate files that do not follow format specifications or contain errors. Converters must thus be able to identify errors and ambiguities in input data and resolve them consistently or discard the corresponding molecule.

Since chemical file formats play such a central role in cheminformatics, every tool and software package must be able to read and write molecular files. Hence, every tool that supports more than one file format can be used as a converter. However, there are tools which have specifically been designed for file format conversion, such as the free software OpenBabel⁴ and, more

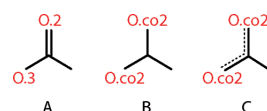


Figure 1. Different representations of carboxylates as observed in MOL2 files.

recently, fconv⁵ or the commercial tools MOL2Mol,⁶ MN. Convert,⁷ and Babel.⁸ Furthermore, there is a large number of programming libraries for cheminformatics, both open source and proprietary, which provide the necessary functionality to read and write molecules. Evidently, these can be used to implement converter tools. Examples of such libraries are OpenBabel,⁹ CDK,¹⁰ CACTVS,¹¹ JOELib,¹² PerlMol,¹³ OEChem,¹⁴ and RDKit.¹⁵ Additionally, some tools are routinely used for file format conversions, although that is not their specific purpose. Typical examples are programs for the generation of 3D coordinates, such as CORINA,¹⁶ LigPrep,¹⁷ and CONCORD.¹⁸

We have implemented a new tool for the consistent conversion of chemical file formats called NAOMI. This converter is based on a robust chemical model which is designed to appropriately describe organic molecules relevant in the context of drug discovery. It provides a reliable and accurate internal representation which allows for a consistent interconversion of the widely used

Received: July 14, 2011

Published: November 08, 2011

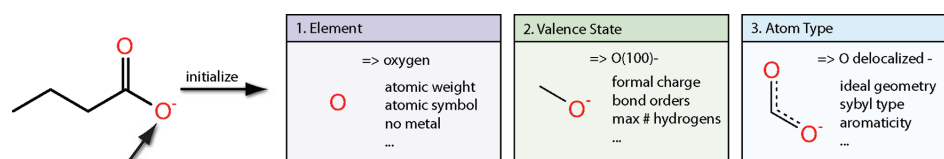


Figure 2. Annotation of the three levels of chemical information for an oxygen of a carboxylate.

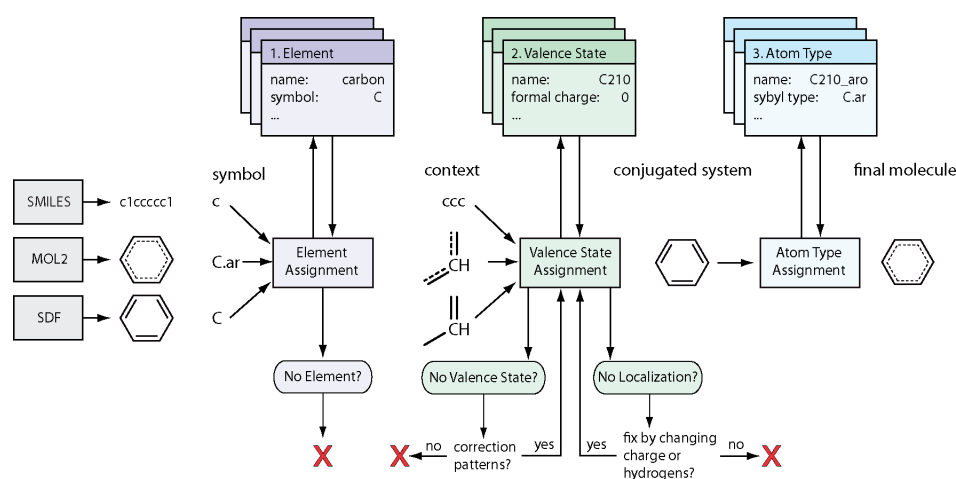


Figure 3. Schematic view of the three steps of molecule initialization.

molecular file formats SDF V2000,¹ MOL2,² and SMILES.³ NAOMI also supports reading and writing SDF V3000 files but does currently not implement all associated features, e.g., self-contained sequence representation. NAOMI checks the chemical validity of molecules and calculates molecular descriptors independent of input file formats.

Although file IO is a task all cheminformatics tools have to perform, not very much is known about the methodologies applied to address the problems related to file conversion. We assume that many tools use approaches very similar to NAOMI, but unfortunately these are mostly not published. Furthermore, file IO and conversion is rarely tested and validated exhaustively. The aim of this paper is to explicitly put the focus on these tasks to demonstrate the complexity and typical pitfalls. We present a round robin test for cheminformatics tools able to read and write different file formats and advocate the use of such tests routinely.

File Format Conversion. The conversion of file formats involves two steps: First, the information provided by the input format is interpreted to build an internal representation of the molecule. Second, all relevant data for the target format is derived from this representation. Due to the different underlying chemical models of the file formats, the conversion usually involves switching from one chemical description to another. Thus, it is important to consider the requirements and limitations of these descriptions.

The Symyx SDF format¹ represents molecules by a single valence bond structure, also called Lewis structure.¹⁹ Hydrogens are frequently omitted to save disk space, while the file format specification ensures the presence of formal charges. The valence bond description has limitations concerning kekulé and resonance

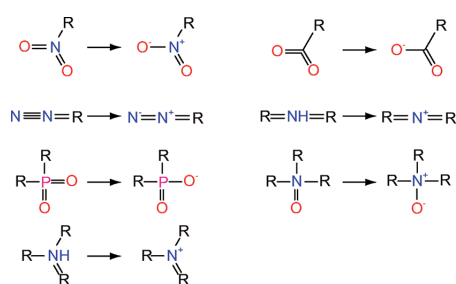


Figure 4. If no valence state can be identified for an atom, then a set of simple correction patterns is applied.

structures, since multiple equivalent valence bond forms of the same molecule may exist.

SMILES²⁰ can represent molecules by a single valence bond structure, whereas hydrogens are virtually always omitted. The format also implements the concept of aromatic atoms and bonds, which allows to represent aromatic systems with different equivalent kekulé forms by a single delocalized description. According to the Daylight theory manual,²⁰ aromaticity in SMILES is however not intended to model physicochemical properties (Daylight theory manual, page 14). Nevertheless, aromatic atoms and bonds are commonly used to describe molecules which are aromatic in a chemical sense, although a single valence bond structure would be sufficient to characterize these molecules unambiguously.

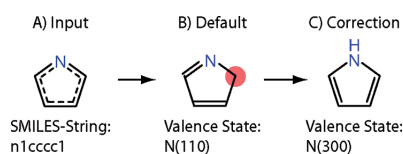


Figure 5. If an input file annotates aromatic atoms and bonds (A), default valence states are assigned in a first step (B). If this attempt is not successful, alternative valence states are considered (C) to correct the input.

The TRIPOS MOL2 format implements the concept of aromatic atoms and bonds, too. Furthermore, the format offers the possibility to describe equivalent resonance forms of common functional groups, such as carboxylates and guanidinium groups, with a delocalized representation. This is realized using specific atom types, called sybyl types, which include information about the atom's hybridization. Usually, MOL2 files do not provide formal charges, but hydrogens are specified. Unfortunately, there is no exact documentation on how the sybyl types must be assigned. This leads to considerable differences between MOL2 files written by different tools. As shown in Figure 1 there are many ways to combine sybyl types, bond orders, and charges to describe the same functional group.

METHODOLOGY

Chemical Model. A consistent chemical model is the keystone for an appropriate internal representation of molecules in cheminformatics application. It also provides the framework for the identification and correction of erroneous input molecules.

The atom-centered chemical model of NAOMI comprises three different levels of chemical information which are assigned to each atom during an initialization procedure. Each level extends the environment that is considered and provides a more detailed description of the atom.

The element is the first and most basic level of description. It provides properties which depend only on the atom's chemical element. These properties comprise the element symbol, the element name, the atomic number, the atomic weight, the van der Waals radius, the number of valence electrons, the covalent radius, and whether the element is considered a metal.

The valence state is the second level of chemical information and extends the scope of the chemical element by taking bonds and formal charges into account. Each valence state represents a valid bond pattern of an atom in a valence bond structure of the molecule. Valence states contain topological information which include formal charge, number of bonds, bond orders, number of free electrons, and whether the corresponding atom can be part of a conjugated or aromatic system.

The atom type extends the valence state to model effects, such as aromaticity and the existence of equivalent resonance forms. This is needed to compensate for the shortcomings of a localized molecular description.

To determine an atom type, the atom and all atoms in its conjugated system (if applicable) are considered. Atom types provide an ideal geometry, a corresponding sybyl type, mark atoms as conjugated or aromatic, and contain information about delocalized electrons. Additionally, an atom type marks the corresponding atom as a hydrogen-bond acceptor or as a potential hydrogen-bond donor.

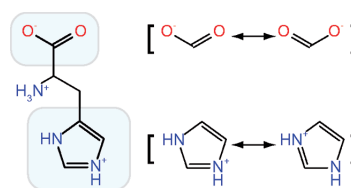


Figure 6. Molecules are partitioned into zones of conjugated atoms. The two oxygen atoms of the carboxylate group and the two nitrogen atoms of the imidazole ring have different valence states but identical atom types. Therefore, the valence states describe a localized structure with a defined formal charge, and the atom types describe a delocalized structure, with a delocalized charge.

Each atom is assigned a corresponding element, valence state and atom type (see Figure 2). Valence states ensure that each molecule has a valid valence bond structure, while atom types allow easy access to a delocalized description.

The basic assumption of the chemical model is that organic molecules which are relevant in the drug discovery context can always be represented by at least one valence bond structure. If that is not the case, then the molecule will either be corrected or discarded. Since there are no strict valence rules for metallic elements, only monatomic ions are accepted. Molecules containing covalently bound metals are currently not supported by the model.

Molecule Initialization. *Overview.* During the molecule initialization data from input files is used to build the internal representation of the molecule. This task is carried out in three separate steps (see Figure 3).

Element Assignment. First, the molecular graph is built from the connectivity data provided by the input file. During this process, the element for each atom is determined, and initial bond types are assigned. The perception of elements, bond types, and connectivity from the different file formats is implemented according to their respective specifications. All elements of the periodic table and bonds of type single, double, triple, and aromatic are supported. Molecules which have atoms or bonds with undefined types are discarded at this point, since this information is required in the subsequent steps.

The initial data are used to generate a valid valence bond form of the molecule. A valence bond form is valid if valence states can be assigned to all atoms and the aromatic bonds can be localized. If no valence bond form can be generated and no correction is possible, the molecule is discarded.

Valence State Assignment. They are selected on basis of the formal charge and bond orders of the atom. Hence, molecules with formal charges, hydrogens, and localized bond orders are the optimal input for this procedure. In this case, the assignment is straightforward and unambiguous. The omission of hydrogens or the use of aromatic bonds, which basically corresponds to the omission of bond orders, also poses no problem, since the remaining properties are still sufficient to reach an unambiguous assignment. If charges or multiple properties are missing, then additional data from the input format is necessary to resolve ambiguities.

Each file format makes use of a different molecular representation and applies certain strategies to omit redundant information. Hence, individual assignment procedures are needed for each file format.

Molecules from SDF are supplied in a valence bond form, which allows a direct comparison to valence states. If hydrogens are

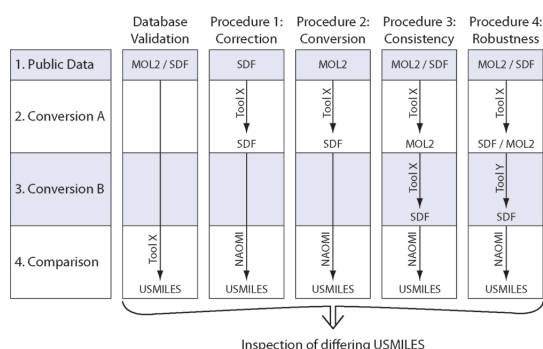


Figure 7. Various procedures test different aspects of file format conversions. During these procedures, molecules are converted by different combinations of tools. USMILES are used for the comparison of the resulting molecules.

Table 1. Options Used for Computing Time Benchmarks

tool/options	explanation
CORINA	
-d wh	write hydrogens to output file
-d no 3d	disable generation of 3D coordinates
-t n	do not write trace file
MOE	
-SVL script	(see Supporting Information)
NAOMI	
-v 0	do not print messages to shell
Open Babel	
-o can	generate USMILES (only for SMILES as output)

omitted, formal charges and multiple bonds are sufficient to unambiguously identify the correct valence states.

Molecules from SMILES may provide information on the bond orders explicitly, whereas hydrogens are virtually always omitted. If this is the case, the assignment works the same way as for SDF. Additionally, SMILES implements the concept of aromatic bonds. This means that bond orders and hydrogens can be missing, and hence ambiguities arise for certain types of atoms. The most prominent example is the pyrrole-like aromatic nitrogen (see Figure 5) which has to be provided with explicit hydrogens for an unambiguous assignment.

Molecules from MOL2 usually have all hydrogens attached but lack the specification of formal charges. If they also contain aromatic bonds, two properties are missing. These ambiguities can only be resolved by using sybyl type information. Additionally, some resonance forms of common functional groups are indicated by specific sybyl types. Their bond types and valence states are adapted accordingly in a postprocessing step.

If no valence state could be found for an atom, the atom's environment is checked by using simple patterns representing common valence errors (see Figure 4). If a pattern matches, then a valence state is assigned, and the bond orders and valence states of the environment are adapted. Otherwise the molecule is discarded.

Afterward, the bonds marked as aromatic in the input file are localized to ensure a valid valence bond form. Information about the localized bond orders for each atom is provided by its

Table 2. Validation of Input Data Sets by NAOMI

data set	no. molecules	no. rejected molecules	corrected molecules	no. diffs MOL2 ↔ SDF
DUD ligands ²³	3961	0	10	0
DUD decoys ²³	124 413	1	13	0

corresponding valence state. The information is used in a recursive algorithm to assign defined bond orders to all bonds.

If the assignment of bond orders was not successful using the default valence states, all atoms of a molecule are checked for an alternative valence state assignment using rule sets specific to the respective file formats (see Figure 5). All combinations of these alternatives are enumerated, and the most probable solution is picked by a simple scoring scheme. The score is calculated as the sum of atoms which have the same valence states with respect to the initial structure. Thus, the procedure assures a minimum deviation from the default assignment. If there are multiple solutions with equal scores, a canonical solution is picked. If no solution could be found, then the molecule is discarded.

Atom Type Assignment. At this point, a valid valence bond form of the molecule is available and can be accessed during subsequent calculations. Since all necessary information can now be derived from the internal representation, the following steps are independent of the input file format.

The next step is the generation of a delocalized description for the molecule. The description allows to overcome the limitations of the valence bond representation concerning kekule and resonance structures. Although these aspects are handled by separate procedures, both need information about the molecule's rings. These are calculated using the relevant cycles algorithm as described by Vismara.²¹

Since equivalent kekule structures can only occur in cyclic systems, this information is stored directly in the molecule's rings. A ring is marked as delocalized if it has alternating single and double bonds and the number of delocalized electrons does fulfill Hueckel's rule. Bonds from rings which are already marked are considered both single and double during the check of neighboring rings. To ensure that the assignment for all rings is independent from the initial valence bond form, the assignment procedure is repeated until the total number of marked rings does not change anymore.

For the identification of equivalent resonance forms, the molecule is partitioned into zones which correspond to its conjugated systems (see Figure 6). This is done by using the information provided by the valence states in combination with the molecule's rings. Each zone is checked for pairs of atoms for which a formal charge can be exchanged. These atoms can be identified by comparison of their corresponding valence states. Then all possible resonance forms are enumerated, and all atoms with delocalized charges are marked. Finally, suitable atom types are selected from a list provided by the valence state using the information about the conjugated system and the delocalization of the atom.

After the initialization procedure, the molecule is represented by a valence bond description (valence states and bond orders) and a delocalized description (atom types and delocalization flags). Both descriptions can be used in subsequent steps.

Validation. To evaluate the quality of file format conversions, a method for comparing input and converted molecules is required. Unfortunately, there is no direct way to determine if two molecular representations are identical. This is especially true if they are stored in different file formats.

The comparison of unique SMILES (USMILES)²² is an easy and verifiable way to identify differences between molecules. Two things have to be taken into consideration with this approach: First, USMILES generated with different tools are often not identical. This means that the method will only be reliable if the USMILES come from the same source. Second, some file format specific information will be lost during the conversion. Therefore, USMILES should be obtained from SDF files, since it provides an unambiguous valence bond structure.

The public DUD ligand and the DUD decoy²³ data sets are used in all validation procedures. To establish a reference for the comparison, both were converted from SDF and MOL2 to USMILES (see Figure 7). These USMILES serve as a basis to determine whether molecules change during conversion steps.

To investigate a tool's ability to convert file formats, four validation procedures are used as shown in Figure 7. In the first

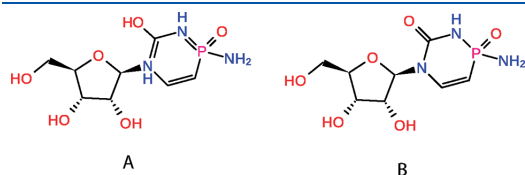


Figure 8. Molecule ZINC0153034: (A) Rejected by NAOMI in DUD decoy data set and (B) in current ZINC database.

Table 3. Data Sets Converted To USMILES by MOE and Open Babel^a

tool	data set	no. rejected molecules	MOL2 ↔ SDF	
			no. diffs	% of data
MOE	DUD ligands	0	1598	40%
	DUD decoys	0	67 042	54%
Open Babel	DUD ligands	0	1875	47%
	DUD decoys	0	46 987	38%

^a Shown are the differences between the generated USMILES originating from MOL2 and SDF.

Table 4. Investigation of Correction Functionality

tool	DUD decoys		DUD ligands	
	no. rejected	no. corrected	no. rejected	no. corrected
CORINA	0	0	0	0
MOE	0	13	0	8
NAOMI	1	13	0	10
Open Babel	0	0	0	0

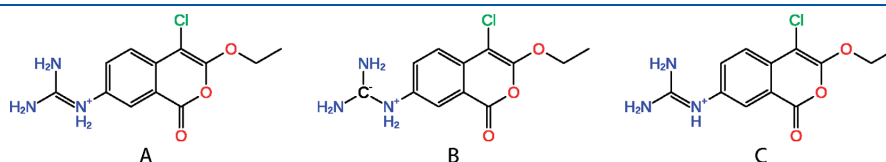


Figure 9. (A) Molecule from DUD ligand data set. (B) Corrected molecule from MOE. (C) Corrected molecule from NAOMI.

procedure, the internal error correction of the tools is analyzed by conversion of molecules from SDF to SDF. The ability to convert molecules from one format into another is investigated in the second procedure by converting molecules from MOL2 to SDF. The third procedure focuses on a tool's internal consistency by converting back and forth using the same tool twice. Finally, the robustness is checked by using different tools subsequently in a pipeline.

All validation procedures are performed with CORINA,²⁴ MOE,²⁵ Open Babel,⁴ and NAOMI. CORINA is commonly used for generating 3D coordinates and for molecular file format conversion and is considered the gold standard. MOE is used for a variety of applications in drug design and supports preparation of ligands for subsequent calculations. This includes the generation of protonation states and tautomers as well as filtering according to molecular descriptors. Correct and consistent reading and writing of molecules forms the basis for these applications. An open source alternative to these tools is Open Babel. Open Babel supports a variety of molecular file formats and is designed to be used as a file format converter.

Computing Time Benchmarks. Although the consistency and the quality of the converted molecules are of superior importance, computing times play a significant role due to the increasing sizes of current data sets. Hence, the runtime behavior is analyzed in order to assess their applicability in large setups.

To investigate NAOMI's performance, the ZINC-everything data set is converted from and to MOL2, SDF, and USMILES. Measured computing times are compared to the commonly used tools CORINA, Open Babel, and MOE. For an unbiased comparison, optional settings of these tools are selected to yield similar results compared to NAOMI. Therefore, generation of USMILES and writing of hydrogens are enforced, and output of additional information is minimized (see Table 1). Conversion from SMILES to MOL2 and SDF using CORINA is omitted since CORINA automatically generates 3D coordinates upon conversion. Furthermore, SMILES is not supported as an output format by CORINA. Although, NAOMI is able to conduct its calculations in parallel, this option is disabled for an easier comparison. All file format conversions are performed on a Linux PC with two Intel Xeon CPUs (2.53 GHz) and 32 GB of main memory.

RESULTS

Data Set Validation. Results of the validation of the DUD ligand and DUD decoy data sets²³ are shown in Table 2. NAOMI successfully converts all molecules except one from MOL2 and SDF to USMILES. A small number of incorrectly protonated nitrogens are corrected. One molecule (ZINC1583034) is rejected, as it contains invalid phosphorus and nitrogen atoms (see Figure 8) which cannot be corrected and localized. Since USMILES

generated by NAOMI are identical for both file formats, they can serve as a reference for the following validation procedures.

Both data sets could also be successfully converted to USMILES by MOE and Open Babel. The molecule which was rejected by NAOMI is neither discarded nor corrected by both tools.

Table 5. Investigation of Conversion Functionality

tool	DUD decoys		DUD ligands	
	no. diffs	% of data	no. diffs	% of data
CORINA	5522	4%	439	11%
MOE	4287	3%	181	5%
NAOMI	0	0%	0	0%
Open Babel	13 469	11%	966	24%

Table 6. Investigation of tool consistency

tool	starting file format	DUD decoys		DUD ligands	
		no. diffs	% of data	no. diffs	% of data
CORINA	MOL2	5522	4%	439	11%
	SDF	4174	3%	235	6%
MOE	MOL2	5770	5%	457	12%
	SDF	5683	5%	453	11%
NAOMI	MOL2	0	0%	0	0%
	SDF	0	0%	0	0%
Open Babel	MOL2	17 351	14%	1168	29%
	SDF	17 364	14%	1168	29%

Table 7. Investigation of Tool Robustness

tool X	tool Y	starting file format	DUD decoys		DUD ligands	
			no. diffs	% of data	no. diffs	% of data
CORINA	MOE	MOL2	4265	3%	176	4%
		SDF	5931	5%	449	11%
		NAOMI	MOL2	58	0%	0
	Open Babel	SDF	4149	3%	235	6%
		MOL2	5522	4%	439	11%
		SDF	19 192	15%	1371	35%
MOE	CORINA	MOL2	6755	5%	504	13%
		SDF	4656	4%	245	6%
		NAOMI	MOL2	3159	3%	167
	Open Babel	SDF	4585	4%	239	6%
		MOL2	4483	4%	174	4%
		SDF	19 311	16%	1374	35%
NAOMI	CORINA	MOL2	0	0%	0	0%
		SDF	643	1%	17	0%
		MOE	MOL2	176	0%	0
	Open Babel	SDF	1217	1%	221	6%
		MOL2	0	0%	0	0%
		SDF	14 172	11%	1164	29%
Open Babel	CORINA	MOL2	29 896	24%	1887	48%
		SDF	10 047	8%	289	7%
		MOE	MOL2	13 693	11%	973
	NAOMI	SDF	43 285	35%	1703	43%
		MOL2	13 469	11%	966	24%
		SDF	1790	1%	24	1%

USMILES originating from MOL2 and SDF, however, differ significantly (see Table 3).

Tool Validation 1: Correction. As mentioned above, the DUD data sets contain 24 invalid molecules in total of which one has been rejected and 23 could be corrected. CORINA and Open Babel convert those without performing any error correction (Table 4). MOE and NAOMI correct the nitrogens with invalid protonation states with differing results (see Figure 9 for an example). Additionally, NAOMI corrects invalid phosphate groups.

Tool Validation 2: Conversion. Results of the investigation of the conversion functionality (see Figure 7) are shown in Table 5. By inspection of the differing molecules, we were able to identify a small number of error classes that will be discussed for every tool:

CORINA places positive charges on carbon atoms of guanidinium- and amidinium-like groups. This error also occurs in five-membered aromatic rings containing this substructure.

MOE places positive charges on carbon atoms of guanidinium- and amidinium-like groups in five-membered aromatic rings. Depending on the substituents, the carbon atom is either charged twice or a carbon atom next to it is negatively charged.

Open Babel's most prominent class of errors is the incorrect conversion of aromatic systems containing charged nitrogen atoms. All bonds in these systems are converted to single bonds in the resulting SDF file. The second kind of error concerns protonation states. Open Babel does not consider input hydrogens to determine formal charges. Therefore, many atoms are neutralized during the conversion process. Since MOL2 entries often do not provide formal charges, this may lead to unexpected results.

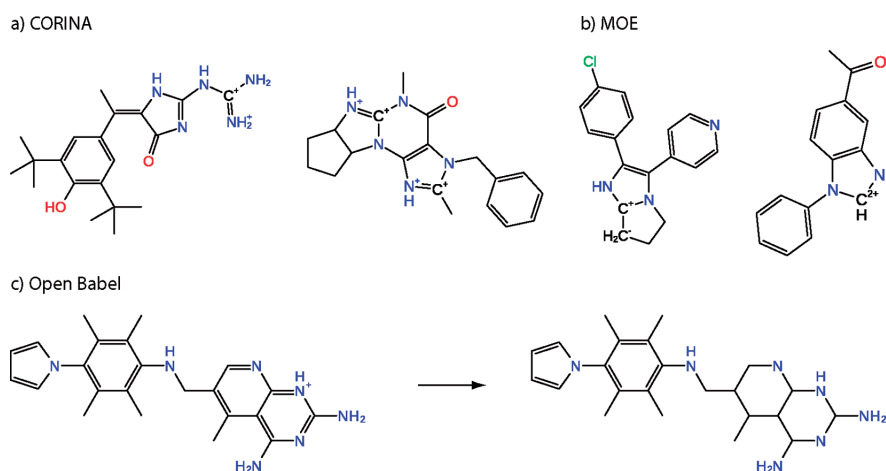


Figure 10. Examples of conversion problems with CORINA, MOE, and Open Babel.

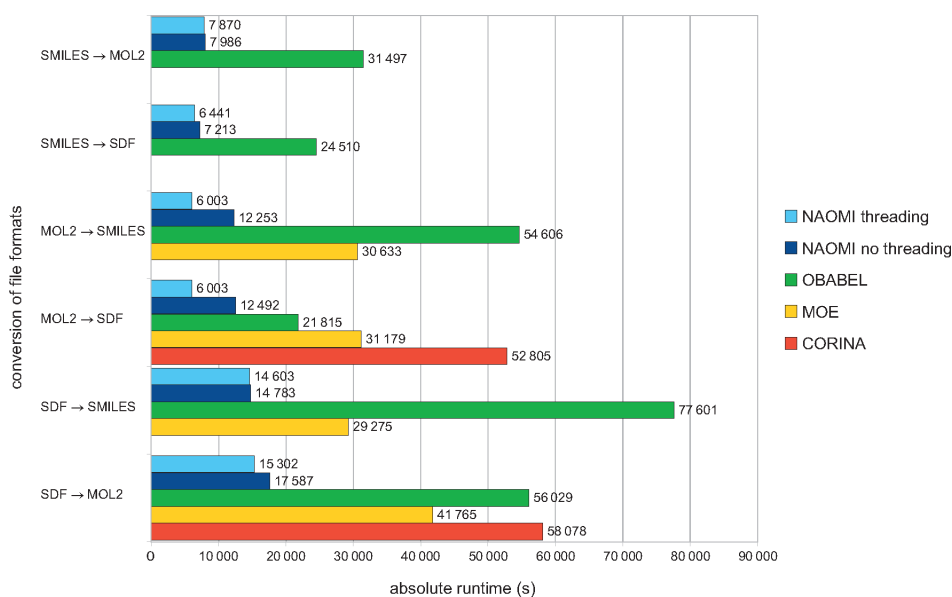


Figure 11. Computing times (wall clock time) for file format conversion of the ZINC-everything data set. For CORINA and MOE, only the computation from SDF and MOL2 are comparable, since the conversion from SMILES includes 3D coordinate generation which is not the case for NAOMI and OpenBabel. Furthermore, CORINA does support SMILES as output format.

Tool Validation 3: Consistency. Results of the investigation of consistency (see Figure 7) are shown in Table 6. Starting from MOL2, the numbers of differences should be identical to those of validation procedure 2 (see Table 5), since no additional file format conversion is performed. A higher number of errors indicates inconsistencies in reading and writing from and to MOL2. Starting from SDF, no differences at all should occur.

CORINA and NAOMI convert molecules consistently in both cases. The differences which were observed for CORINA when the first input was provided from SDF are introduced by switching from a delocalized to a localized description. Nevertheless, they

only represent different valid resonance forms of the original data and are therefore not considered conversion errors. MOE and Open Babel show inconsistencies in both cases.

Tool Validation 4: Robustness. The robustness of the investigated tools is analyzed by combining two different tools in a pipeline. Since tools tend to interpret input from file formats differently, the molecules can change with each additional tool included in the workflow. Table 7 indicates that inconsistencies during file format conversion are not uncommon and depend both on the kind of tools used and on the order in which they are combined.

Furthermore, the success of the conversion strongly depends on the source of the input data. The experiment clearly shows that all tools benefit significantly from preprocessing data sources with NAOMI toward consistency and high quality (see Figure 10).

Computing Time Benchmarks. Figure 11 summarizes the computing times for conversion of the ZINC-everything data set. Since NAOMI is designed for large scale cheminformatics applications, it is not surprising that it is substantially faster than the modeling platform MOE. NAOMI supports multithreading resulting in a speed-up by another factor of 1.4. For SDF and SMILES, file IO is usually the rate-determining step. Therefore, threading does not lead to an improvement of runtimes. The MOL2 format however needs a more advanced initialization procedure, thus leading to gains in runtimes when threading is enabled.

In summary, NAOMI achieves a conversion speed of up to 2841 molecules/second on a PC with two Intel Xeon CPUs (2.53 GHz) and 32 GB of main memory.

CONCLUSION

Handling chemical structures is and remains a complex task. File formats contain chemical descriptions at different levels of detail and are therefore not easy to convert. Since the description of file formats are sometimes ambiguous when it comes to details, software tools tend to interpret them differently. This in turn causes errors in data sets and misinterpretations in tools. For the cheminformatics community, it would be a great benefit to build clear standards for file formats and to certify software with respect to these standards.

Meanwhile, it is important that software tools are at least self-consistent when reading and writing file formats. Evidently, errors in reading molecules from files usually have a substantial impact on downstream algorithms and methods. NAOMI will most certainly have flaws of its own, and in order to find them, consistency checks as those presented are needed. We urge that more of these tests should be published and that the existing ones become a standard validation procedure for all cheminformatics applications.

The command-line converter NAOMI has been implemented in C++ and can be downloaded at <http://www.zbh.uni-hamburg.de/naomi>. It will be available free of charge for academic use. A convenient graphical user interface for NAOMI's functionality will soon be provided by the chemical library preprocessor MONA (see <http://www.zbh.uni-hamburg.de/mona>).

ASSOCIATED CONTENT

Supporting Information. Original and corrected structures for both DUD data sets are provided. The corrected structures are supplied in the same file format as the respective input files (SDF or MOL2). Furthermore, a text file containing the SVL commands used for the computations with MOE is supplied. This material is available free of charge via the Internet at <http://pubs.acs.org/>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: rarey@zbh.uni-hamburg.de.

Present Address

[†]Current address: Georg Simon Ohm University of Applied Sciences, Nuremberg, Germany.

ACKNOWLEDGMENT

The authors thank Stefan Wefing for his initial ideas concerning the chemical model, Dr. Holger Claußen for testing, Rene Kraus for IO support, and Matthias Hilbig for supplying a graphical interface.

REFERENCES

- (1) *Symyx CTfile Formats*; <http://www.symyx.com/downloads/public/ctfile/ctfile.jsp>, (accessed January 27, 2011).
- (2) *TRIPOS Mol2 File Format*; <http://tripos.com/data/support/mol2.pdf>, (accessed January 27, 2011).
- (3) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
- (4) *The Open Babel Package*, version 2.3.0; <http://openbabel.org>, (accessed January 18, 2011).
- (5) Neudert, G.; Klebe, G. fconv: format conversion, manipulation and feature computation of molecular data. *Bioinformatics* **2011**, *27*, 1021–1022.
- (6) *Mol2Mol*; <http://www.gunda.hu/mol2mol/index.html>, (accessed January 27, 2011).
- (7) *MN.Convert*; Molecular Networks GmbH - Computerchemie: Erlangen, Germany; <http://www.molecular-networks.com/products/convert>, (accessed January 27, 2011).
- (8) *Babel*; OpenEye Scientific Software, Inc.: Santa Fe, NM; <http://www.eyesopen.com/docs/babel/current/pdf/BABEL.pdf>, (accessed January 27, 2011).
- (9) Guha, R.; Howard, M.; Hutchison, G.; Murray-Rust, P.; Rzepa, H.; Steinbeck, C.; Wegner, J.; Willighagen, E. The Blue Obelisk-Interoperability in Chemical Informatics. *J. Chem. Inf. Model.* **2006**, *46*, 991–998.
- (10) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493–500.
- (11) Ihlenfeldt, W.; Takahashi, Y.; Abe, H.; Sasaki, S. Computation and Management of Chemical Properties in CACTVS: An Extensible Networked Approach toward Modularity and Compatibility. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 109–116.
- (12) *JOELib/JOELib2*; <http://sourceforge.net/projects/joelib/>, (accessed January 27, 2011).
- (13) *PerlMol*; <http://www.perlmo.org/>, (accessed January 27, 2011).
- (14) *OEChem*; OpenEye Scientific Software, Inc.: Santa Fe, NM; <http://www.eyesopen.com/oechem-tk>, (accessed January 27, 2011).
- (15) *RDKit*; <http://rdkit.org/>, (accessed Jan 27, 2011).
- (16) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000–1008.
- (17) *LigPrep*; Schrödinger, LLC: Cambridge, MA; <http://www.schrodinger.com/products/14/10/>, (accessed January 27, 2011).
- (18) *Concord*; Tripos: St. Louis, MO; http://tripos.com/data/SYBYL/Concord_072505.pdf, (accessed January 27, 2011).
- (19) Lewis, G. N. The Atom and the Molecule. *J. Am. Chem. Soc.* **1916**, *38*, 762–785.
- (20) *Daylight theory manual*, Daylight version 4.9; Daylight Chemical Information Systems, Inc.: Aliso Viejo, CA; <http://www.daylight.com/dayhtml/doc/theory/index.pdf>, (accessed March 8, 2011).
- (21) Vismara, P. Union of all the minimum cycle bases of a graph. *Electron. J. Comb.* **1997**, *4*, 1–15.
- (22) Weininger, D.; Weininger, A.; Weininger, J. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97–101.
- (23) Huang, N.; Shoichet, B.; Irwin, J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789–6801 Data sets (SDF and Mol2) downloaded April 12, 2011.

(24) *CORINA - Fast Generation of High-Quality 3D Molecular Models*, version 3.48; Molecular Networks GmbH - Computerchemie : Erlangen, Germany; <http://www.molecular-networks.com/products/corina>, (accessed January 18, 2011).

(25) *MOE*, version 2010.10; Chemical Computing Group: Montreal, Quebec, Canada; <http://www.chemcomp.com/software.htm>, (accessed January 18, 2011).

Unique Ring Families: A Chemically Meaningful Description of Molecular Ring Topologies

[D2] A. Kolodzik, **S. Urbaczek**, and M. Rarey. Unique Ring Families: A Chemically Meaningful Description of Molecular Ring Topologies. *Journal of Chemical Information and Modeling*, 52(8):2013-2021, 2012.


<http://pubs.acs.org/articlesonrequest/AOR-WBYvgICi3nrRd4gG9DNN>

Reproduced with permission from A. Kolodzik, S. Urbaczek, and M. Rarey. Unique Ring Families: A Chemically Meaningful Description of Molecular Ring Topologies. *Journal of Chemical Information and Modeling*, 52(8):2013-2021, 2012. Copyright 2012 American Chemical Society.

Unique Ring Families: A Chemically Meaningful Description of Molecular Ring Topologies

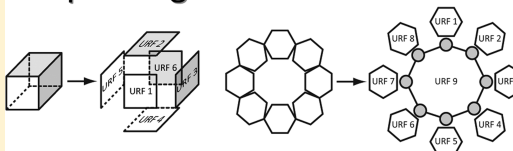
Adrian Kolodzik,^{†,‡} Sascha Urbaczek,[†] and Matthias Rarey^{*,†}

[†]Center for Bioinformatics (ZBH), University of Hamburg, Bundesstr. 43, 20146 Hamburg, Germany

 Supporting Information

ABSTRACT: The perception of a set of rings forms the basis for a number of cheminformatics applications, e.g. the systematic naming of compounds, the calculation of molecular descriptors, the matching of SMARTS expressions, and the generation of atomic coordinates. We introduce the concept of unique ring families (URFs) as an extension of the concept of relevant cycles (RCs).^{1,2} URFs are consistent for different atom orders and represent an intuitive description of the rings of a molecular graph. Furthermore, in contrast to RCs, URFs are polynomial in number. We provide an algorithm to efficiently calculate URFs in polynomial time and demonstrate their suitability for real-time applications by providing computing time benchmarks for the PubChem Database.³ URFs combine three important properties of chemical ring descriptions, for the first time, namely being unique, chemically meaningful, and efficient to compute. Therefore, URFs are a valuable alternative to the commonly used concept of the smallest set of smallest rings (SSSR) and would be suited to become the standard measure for ring topologies of small molecules.

Unique Ring Families



INTRODUCTION

Ring perception is a crucial step in many cheminformatics applications, including the calculation of molecular descriptors, the matching of SMARTS expressions, and the generation of two- and three-dimensional atomic coordinates. In order to obtain consistent results, a set of rings has to be unique in the sense that it depends only on the molecule's topology. Efficient algorithms and ring perception concepts that lead to a limited number of cycles provide the means for interactive applications. Chemically meaningful rings allow for an easy analysis and interpretation of the resulting set of rings. Due to their high relevance in chemistry, several computational methods for automatic ring perception have been developed over the past 35 years.⁴ Each of these methods has deficiencies in being either not unique or not polynomial in number or not chemically meaningful. The paper of Berger et al.⁵ impressively demonstrates this for a number of ring perception concepts including the widely used SSSR.⁴

A molecule can be interpreted as a simple, connected, unweighted and undirected graph $G = (V, E)$ where the atoms are interpreted as a set of vertices V and bonds are considered a set of edges E . A cycle is a subgraph of G such that any vertex degree is exactly two. A connected cycle is called elementary. Since elementary cycles meet our expectation of rings in a molecular graphs we will use the terms elementary cycle and ring synonymously. $E(v_1, v_2)$ is the edge connecting the vertices v_1 and v_2 . For the set of vertices or edges of a cycle (or a general subgraph) C , we will write $V(C)$ and $E(C)$, respectively. A cycle C containing the edges $E(C)$ has a length of $|C|$ which is equal to its number of edges $|E(C)|$. It can be described by the incidence vector of its edges. A cycle with n edges is called n -cycle.

A connected n -cycle is called n -ring. A chord is an edge e connecting two vertices of a ring C with $e \notin E(C)$. A ring is chord-less if it has no chord. Cycles can be combined to larger ones by forming the symmetric difference of their edges; this operation is considered the "addition" of cycles. In order to describe the addition of cycles, we utilize the xor operator \oplus in agreement with the nomenclature used by Berger et al.⁵ Thus, the addition of two cycles C_A and C_B that forms the cycle C_C will be written as $C_A \oplus C_B = C_C$. All cycles of G form the cycle space $S(G)$. A cycle base $B(G)$ is a subset of $S(G)$ that allows to construct all cycles of $S(G)$ by the addition operation. The length of $B(G)$ is equal to the sum of the lengths of its cycles. All cycles of a cycle base are elementary.

In the following, we will discuss common concepts of ring perception in order to motivate our new approach. The set of all rings⁶ (Ω) includes all elementary rings of a molecular graph and efficient algorithms for its calculation have been developed. The number of rings and the computational run-times, however, grow dramatically for complex ringsystems. Additionally, not all resulting rings are meaningful in a chemical context, and Ω is, thus, a unique description that is neither chemically meaningful nor polynomial in size.

The most frequently applied strategy of ring perception is the calculation of the smallest set of smallest rings⁴ (SSSR) which is a subset of Ω . An SSSR represents a minimum cycle base (MCB). It contains a polynomial number of rings and can be calculated in polynomial time.⁷ If a molecular graph contains only a single MCB, the SSSR is unique and intuitive. If this is

Received: December 31, 2011

Published: July 10, 2012

not the case, the resulting SSSR is arbitrary and depends on the specific algorithm used for its construction. Furthermore, the selected SSSR often depends on the input atom order.⁸

The problems arising from nonunique ring descriptions can be exemplified with SMARTS pattern matching. According to page 20 of the Daylight Theory Manual,⁹ the SMARTS pattern [R3] describes an atom which is part of three SSSR rings. The matching of this SMARTS pattern on the highly symmetric molecule cubane (SMILES = C12C3C4C1C5C2C3C45) using the Daylight web service¹⁰ illustrates the problems arising from the SSSR's lack of uniqueness (see Figure 1). Any combination

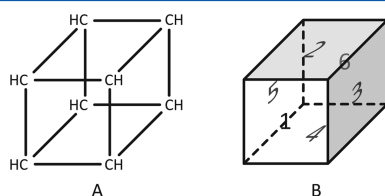


Figure 1. Cubane contains six alternative MCBs. Each combination of five of the 4-rings forms an SSSR.

of five of the shown 4-rings forms a valid SSSR. The sixth ring can be constructed by adding the rings of the SSSR.¹¹ Consequently, the SMARTS pattern [R3] only matches four of the eight equivalent carbon atoms depending on the selected SSSR.

The **essential set of essential rings** (ESER)¹² and the approaches published by Corey¹³ and Wipke¹⁴ try to perceive chemically meaningful rings by calculating an MCB and adding rings up to a certain size or rings including certain elements. Due to their heuristic nature, these approaches lack a mathematical foundation and are not suitable for all kinds of molecular graphs.⁵

In addition, there is a number of graph theoretical ring perception concepts which are limited to planar graphs. The **minimum planar cycle base** and the **extended set of smallest rings**¹⁵ are examples of such concepts. Since molecular graphs are not necessarily planar, these ring perception concepts are of limited use for general applications in cheminformatics.

The **set of β -rings**¹⁶ is defined on a plane embedding of a molecular graph. The chord-less faces of the embedding are processed by increasing size. The set of β -rings includes all faces representing 3-rings or 4-rings. Additionally, it contains all faces which are linearly independent of three or less shorter faces already contained in the set. Berger et al.⁵ suggested to use the **β^* -rings** instead. These rings are calculated on all chord-less rings of a graph instead of the chord-less faces of a specific plane embedding. In contrast to the set of β -rings, the set of β^* -rings is unique but contains an exponential number of rings.

An additional set of rings which is defined for general graphs is the **set of smallest cycles at edges** (SSCE).¹⁷ The SSCE is calculated on the basis of Ω by recursively deleting all edges included in more than one ring. The SSCE does, however, not necessarily contain a cycle base. Consequently, it does not provide a complete description of the rings of a molecular graph.

Relevant cycles^{1,2} (RCs) are defined as the union of all MCBs. They comprise a unique set of rings and an intuitive description of most molecular graphs. Some molecules, however, contain an exponential number of RCs. Examples are

cyclophane-like structures which will be discussed in more detail in the following sections.

To tackle the exponential number of rings, Gleiss et al.¹¹ suggested to classify RCs into **interchangeability classes** (ICs). ICs are calculated by dividing RCs into essential and interchangeable rings. An essential ring is included in all MCBs. Rings which are not essential are called interchangeable. An IC contains either a single essential ring or all interchangeable rings which can be constructed from a subset of the IC and shorter cycles. While treating the rings of an interchangeability class as a union can be suitable for the prediction of RNA secondary structures, this concept is not generally applicable in cheminformatics. For example, the description of the six RCs of cubane or the 6-rings of fullerene as single ICs is too coarse for most applications and, especially in the case of fullerene, it is not chemically meaningful.

Relevant cycle families (RCFs)¹ are conceptually similar to ICs. An RCF contains all RCs generated on the basis of a single relevant cycle prototype (RCP). RCPs are not unique and their number depends on the order of the molecule's atoms. Since each RCP results in an RCF, the RCFs are also not unique and their number can vary for a molecule.

None of the mentioned concepts of ring perception efficiently calculate a complete and polynomial set of unique and chemically intuitive rings for molecular graphs. We introduce the concept of **unique ring families** (URFs), which meets all of these requirements.

UNIQUE RING FAMILIES

Generation of Relevant Cycles. Since unique ring families (URFs) are defined on the basis of RCs, we provide a short outline of Vismara's RC detection algorithm.¹ The perception of RCs involves five consecutive steps which are explained below (see Figure 2):

1. Calculate all 2-connected components of the molecular graph G .
2. For each 2-connected component, calculate the shortest paths from each vertex r to each other vertex, only passing through vertices which follow r in an arbitrary but fixed order π .
3. Calculate RCPs by combining pairs of shortest paths of identical size starting from the same vertex r .
4. Eliminate RCPs which linearly depend on strictly smaller cycles with respect to cycle addition.
5. Calculate RCs by a backtracking procedure on the basis of the RCPs.

2-Connected components of the molecular graph can be calculated using the algorithm published by Tarjan.¹⁸ The 2-connected components will be called ringsystems in the following sections. An order π of the vertices is established by sorting them according to their degree in descending order. Vertices of identical degree are ordered arbitrarily. This ordering guarantees polynomial runtime complexity for the calculation of RCPs. In the second step, a breadth-first-search is used to calculate a single shortest path $P(r,t)$ from each vertex r to each other vertex through vertices following r in the ordering π . Thus, only paths through vertices of equal or lower degree are considered. If two shortest paths $P(r,p)$ and $P(r,q)$ of identical size solely share the vertex r , and if furthermore p and q are directly connected by an edge, an uneven ring is identified. If p and q are both directly connected to a vertex z which is neither a member of $P(r,p)$ nor a member of $P(r,q)$, an

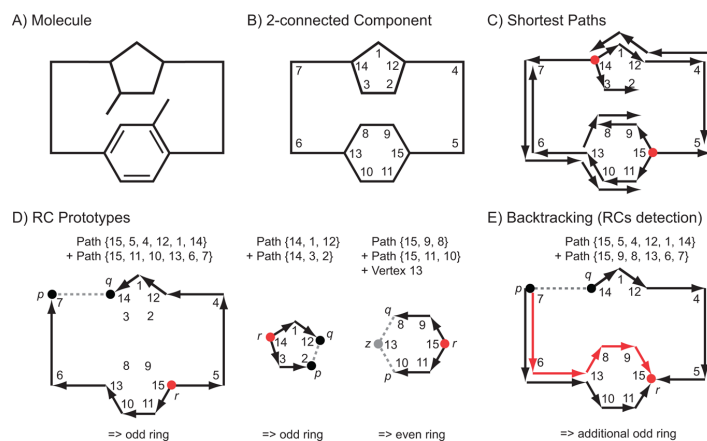


Figure 2. Process to identify the RCs of a molecular graph. (A) At first, 2-connected components are calculated (B) and vertices are ordered according to their degree. Vertices of higher degree are labeled with higher numbers than vertices of lower degree. (C) Shortest paths only passing through vertices following r in the order π are calculated from each vertex r to each other vertex of the graph (shown for vertices 14 and 15). (D) The polynomial number of RCPs are calculated on the basis of the identified shortest paths. Two shortest paths of equal lengths which only share the vertex r form an uneven RCP if their end points (p, q) are adjacent. If they share an adjacent vertex z , they form an even RCP. (E) RCs are enumerated on the basis of RCPs by combining alternative shortest paths (red arrows) connecting p or q to r .

even RCP is identified. The length of the shortest paths used to identify an RCP of size n is therefore given by the following equation:

$$|E(P(r, p))| = |E(P(r, q))| = \begin{cases} \frac{n-1}{2} & \text{if } n \text{ is odd} \\ \frac{n}{2} - 1 & \text{if } n \text{ is even} \end{cases} \quad (1)$$

As described above, only a single shortest path is considered for each pair of vertices. Multiple shortest paths connecting two vertices may exist, however. Thus, the polynomial number of RCPs represent only a subset of the exponential number of RCs. To identify all RCs on the basis of the RCPs, Vismara's algorithm uses a backtracking procedure. The set of RCs calculated during backtracking on the basis of a single RCP is defined as an RCF. This backtracking procedure includes the following steps:

First, for each RCP the set S_p of all shortest paths from p to r and the set S_q of all shortest paths from q to r are calculated. If an RCP is uneven, each combination of $P(r, p) \in S_p$ and $P(r, q) \in S_q$ forms an uneven RC with the edge $E(p, q)$ (see, for example, the 11-ring in Figure 2E). If an RCP is even, each combination of $P(r, p) \in S_p$ and $P(r, q) \in S_q$ forms an even RC with the edges $E(p, z)$ and $E(q, z)$.

Note that all RCs of an RCF have the same size. If their size is uneven, they share at least the vertices r, p , and q and the edge $E(p, q)$. Otherwise, they share at least the vertices r, p, q , and z and the edges $E(p, z)$ and $E(q, z)$. All RCFs of a molecular graph are disjoint with respect to their rings and their union forms the set of all RCs of a graph. In the following, the RCF of a ring C_x will be called RCF_x . Furthermore, we will write $E(\text{RCF}_x)$ and $V(\text{RCF}_x)$ to denote the union of the edges or vertices of all rings of an RCF_x , respectively.

Introduction of Unique Ring Families. On the basis of the RCs of a molecular graph, we define the terms URF-pair-related and URF-related as follows:

Definition 1. Let C_1 and C_2 be two RCs of a graph G , then C_1 and C_2 are URF-pair-related if and only if all of the following conditions hold:

1. $|C_1| = |C_2|$
2. $E(C_1) \cap E(C_2) \neq \emptyset$
3. It exists a set S of strictly smaller rings in G such that $C_1 \oplus (\bigoplus_{c \in S} c) = C_2$

Definition 2. The URF-relation is defined as the transitive closure of the URF-pair-relation. A URF is defined as the set of URF-related RCs and hence represents an equivalence class. The length $|URF|$ is defined as the length of each of its RCs. The number of URFs of a graph is called URF-number.

For an efficient calculation of molecular ring topologies in case of complex ringsystems, a description of rings should be at most polynomial in number with respect to the size of the graph. In the following, we estimate the URF-number of a molecular graph by comparing it to the polynomial number of RCFs.

Theorem 1. Any two rings of an RCF are URF-related.

Due to the construction of RCFs as described above, any two RCs of an RCF have identical lengths and share at least either an edge $E(p, q)$ or the edges $E(p, z)$ and $E(q, z)$. Thus, all rings of an RCF meet conditions 1 and 2 of definition 1. Furthermore, the RCs of an RCF differ only by alternative shortest paths replacing $P(r, p)$ or $P(r, q)$. As a consequence of eq 1, the following equation describes the length of two shortest paths used to construct an RCP of size n :

$$|P(r, p) \in S_p| = |P(r, q) \in S_q| < \frac{n}{2} \quad (2)$$

Since $P(r, p)$ contains less than half of the edges of the RCP, the symmetric difference of any two alternative shortest paths of S_p forms a set of rings which are smaller than n . Since the same is true for any two alternative paths of S_q , each two rings of an RCF can be constructed by cycle addition of each other and a set of smaller rings. Hence, all rings of an RCF meet condition 3 of definition 1. Consequently, any two RCs of an

RCF are URF-related and the URF-number is less or equal to the number of RCFs. Since the number of RCPs and RCFs is polynomial according to Theorem 4 of Vismara's paper,¹ the URF-number is at most polynomial, too.

Calculation of URFs. In the following, we provide an algorithm to calculate the polynomial number of URFs in polynomial time on the basis of the RCPs. The algorithm uses the described properties of RCPs as well as their linear dependency with respect to cycle addition in order to describe URFs by their edges sets.

Lemma 1. Let C_A and C_B be two URF-related RCs, then C_A and C_B linearly depend on each other and a set of smaller rings with respect to cycle addition.

According to condition 3 of definition 1, two URF-pair-related RCs linearly depend on each other and a set of smaller rings with respect to cycle addition. Since a URF consists of the transitive closure of the URF-pair-relation, any two URF-related RCs linearly depend on each other and a set of smaller rings. Thus, URFs can be calculated in three steps.

1. Calculate RCPs according to Vismara's algorithm.
2. Let $B_z(G)$ be a subset of a minimum cycle basis B with $B_z(G) = \{C \in B \mid |C| < |C_A| = |C_B|\}$. Identify all 2-pairs of RCPs (C_A, C_B) with

$$C_A \oplus \left(\bigoplus_{c \in B_z(G)} c \right) = C_B \quad (3)$$

Note that this operation is already performed during the calculation of RCPs. In Vismara's ring construction algorithm, a Gaussian elimination is used to eliminate rings which depend on smaller rings. Any ring C_A which depends on smaller and equal sized rings is marked as relevant. If the set of equal sized rings on which C_A depends on, only consists of a single ring C_B , C_A and C_B are marked as potentially URF-related. Furthermore, please note that any two rings of $\text{RCF}_A \cup \text{RCF}_B$ meet conditions 1 and 3 for being URF-pair-related.

3. If any two rings of RCF_A and RCF_B share an edge, these two rings are URF-pair-related. Since the URF-relation is an equivalence relation, C_A and C_B are URF-related if

$$E(\text{RCF}_A) \cap E(\text{RCF}_B) \neq \emptyset \quad (4)$$

In order to calculate RCPs according to Vismara's algorithm, rings which linearly depend on strictly smaller rings are eliminated. If a ring depends linearly on rings of the same size and strictly smaller rings, it is marked as relevant. All RCs of identical size which are identified in this step to be linearly dependent on each other and a set of smaller rings form pairs of possibly URF-related RCPs. For each RCP, all edges and vertices belonging to the same RCF can be identified using a simple breadth first search starting from r followed by a backtracking procedure involving the following steps:

1. Starting from r each vertex v is labeled according to its distance d_v to r using a breadth-first-search.
2. E_{cur} and V_{cur} represent the vertices and edges currently identified as belonging to E_{RCF} and V_{RCF} , respectively. V_{cur} is initialized with $V_{\text{cur}} \leftarrow \{p, q, z\}$ if C_A has even size and $V_{\text{cur}} \leftarrow \{p, q\}$ if C_A has uneven size. E_{cur} is initialized with $E_{\text{cur}} \leftarrow \{E(p, z), E(q, z)\}$ if C_A has even size and $E_{\text{cur}} \leftarrow \{E(p, q)\}$ if C_A has uneven size. A list Q of vertices is initialized with $Q \leftarrow \{p, q\}$.
3. For a vertex $v_{\text{cur}} \in Q$ identify each directly connected vertex v_{adj} . If $d_{v_{\text{cur}}} - 1 = d_{v_{\text{adj}}}$ then

- $E_{\text{cur}} \leftarrow E_{\text{cur}} \cup E(v_{\text{cur}}v_{\text{adj}})$
- $V_{\text{cur}} \leftarrow V_{\text{cur}} \cup (v_{\text{adj}})$
- $Q \leftarrow Q \cup v_{\text{adj}}$ if $v_{\text{adj}} \notin V_{\text{cur}}$

$$4. Q \leftarrow Q \setminus v_{\text{cur}}$$

5. If $Q = \emptyset$, then $E_{\text{cur}} = E(\text{RCF}_A)$ and $V_{\text{cur}} = V(\text{RCF}_A)$. Otherwise, continue with step 3.

For a connected graph containing $|E|$ edges and $|V|$ vertices, RCPs can be calculated in $\mathcal{O}(Z|E|^3)$ with $Z = |E| - |V| + 1$ being the cyclomatic number of G .¹ A Gaussian elimination to identify RCPs of identical size, which depend on each other and strictly smaller rings, can be performed in $\mathcal{O}(|E|R^2)$ operations with R being the number of RCPs. The sets of edges belonging to each RCF are calculated in $\mathcal{O}(|E|R)$. Finally, the edge set intersections of all 2-pairs of RCFs can be calculated in $\mathcal{O}(|E|R^2)$. According to Vismara¹ the number of RCPs (R) is limited by the following relation:

$$R \leq 2|E|^2 + Z|V| \Rightarrow R \leq 2|E|^2 + |E||V| \quad (5)$$

Consequently, the Gaussian elimination and the calculation of the edge intersection of 2-pairs of RCFs are the speed-limiting steps and URFs can be perceived in $\mathcal{O}(|E|^5 + |V|^2)$. Thus, URFs represent a polynomial description of the ring topologies of a molecular graph and can be calculated in polynomial time.

Interpretation of URFs. From a chemical perspective, URFs can be best understood by calculating the union of the edges of all URF-related rings. Since a URF can contain smaller URFs, it can be illustrated by merging these smaller URFs to single nodes. This illustration represents a quotient graph of the partition of smaller URFs. Examples of molecular graphs and their corresponding RCs, RCPs and URFs are shown in Figures 3 and 4.

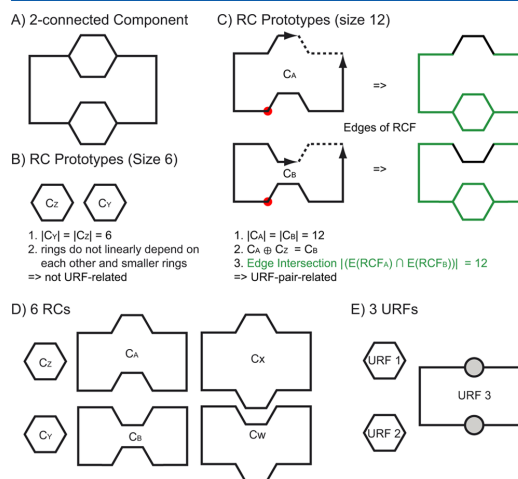


Figure 3. Ring system (A) containing 2 RCPs of size 6 (B) and 2 RCPs of size 12 (C). The two small rings form individual URFs (E). The two 12-rings belong to the same URF since they have the same size, share edges, and are linearly dependent on each other and one of the 6-rings. The molecular graph contains a total of six RCs (D) and three URFs (E). The URFs are illustrated as a quotient graph with the smaller URFs merged to individual nodes.

Compared to common strategies of ring perception, URFs have the major advantages that they are unique, intuitive, polynomial in number and provide a complete description of

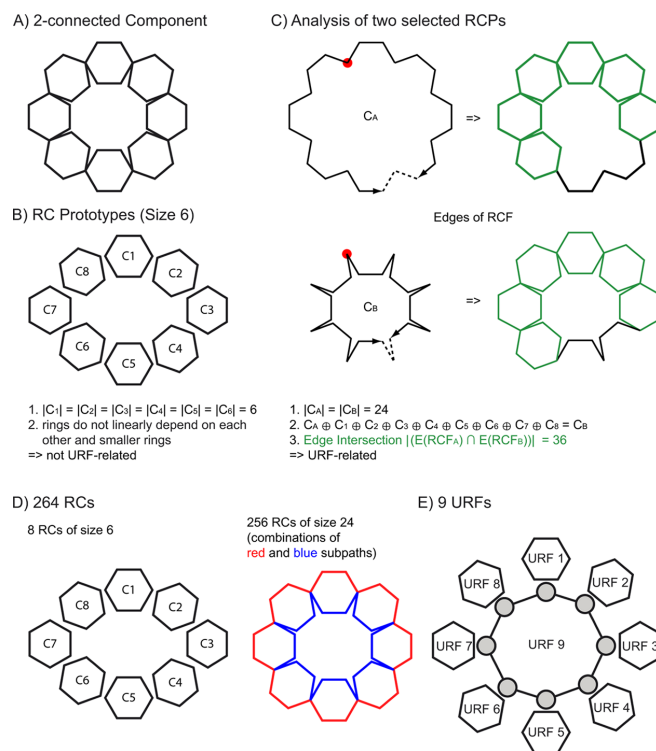


Figure 4. Ring system (A) consisting of 8 RCPs of size 6 (B) and 4 RCPs of size 24 of which 2 are illustrated (C). While the large RCPs have the same size and are linearly dependent according to condition 3 of definition 1, they do not share any edge. Their RCPs, however, share 36 edges. Note that this demonstrates, that two URF-related rings are not necessarily URF-pair-related. (D) The molecular graph contains 8 RCs of size 6 and 256 RCs of size 24. The set of all 264 RCs can be represented by 9 URFs. Eight URFs each contain a single 6-ring. One URF represents a macrocycle including the small URFs. This URF is illustrated as a quotient graph of the partition of smaller URFs. Note that the number of RCs increases exponentially with the number of para-bridged 6-rings, while the number of URFs increases linearly and stays intuitive.

the ring topology of a molecular graph. Macrocycles with para-substituted rings are a well-known problem (see Figures 3 and 4). The molecular structure shown in Figure 4 contains 264 RCs and 256 different possible SSSR cycle bases. The 256 large RCs belong to the same URF, resulting in 9 URFs. Thereby, URFs model the intuitive description of the molecule as a macrocycle containing eight smaller rings.

A frequently found specification in chemical patterns is the number of rings an atom is involved in. In the pattern language SMARTS, this is modeled with the R-feature. As discussed in the introduction, the R-feature is based on an SSSR which causes problems due to nonuniqueness. So far, no alternative approach resulting in a unique and polynomial number of ring representatives was available. Describing atoms by the number of URFs they are involved in represents an easy to implement solution to this problem.

Figure 5A shows the number of rings that contain the atoms A1 and A2. Using SSSRs, the result depends on the selected cycle base. In contrast, the number of RCs is large and chemically nonintuitive. Similar problems occur for symmetric cyclic structures like cubane (see Figure 5B). The calculation of URFs results in a consistent and chemically meaningful value for each atom. Furthermore, if an application requires the construction of an MCB, this can be easily achieved by selecting a

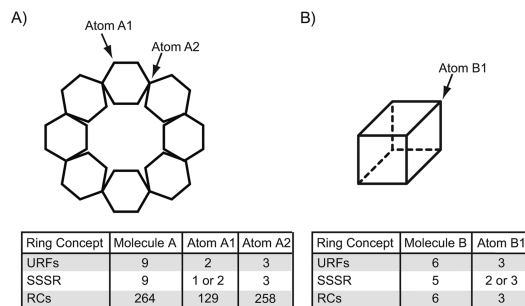


Figure 5. Two complex ring systems with their number of SSSR-rings, relevant cycles, and unique ring families. Additionally, ring memberships for the marked atoms are listed.

single arbitrary RCP of each URF followed by a Gaussian elimination of the resulting set of rings. Since the number of URFs is greater than or equal to the number of cycles of an MCB and smaller than or equal to the number of RCPs, the URF-number can be estimated by the following equation:

$$(E - V + 1) \leq \text{URF-number} \leq (2E^2 + EV) \quad (6)$$

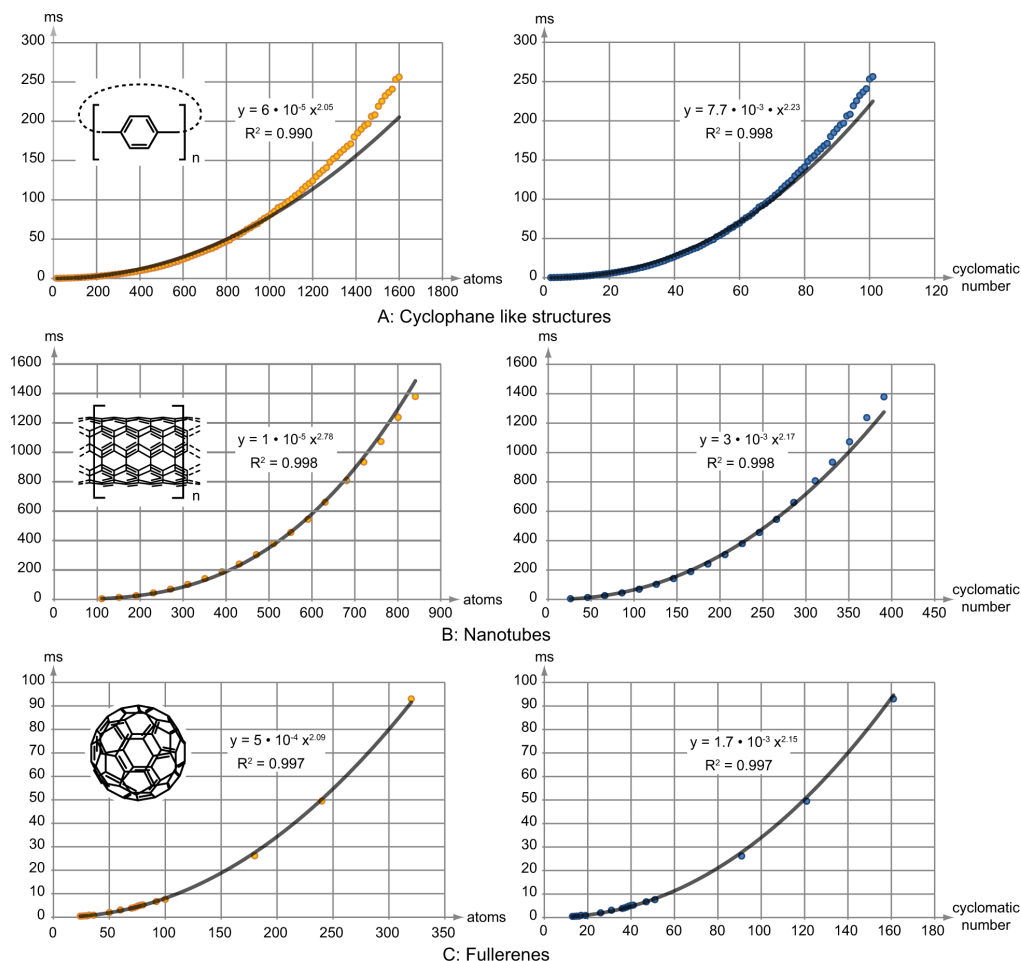


Figure 6. Required runtimes for the calculation of URFs depending on the number of atoms (left) and the cyclomatic number (right) for cyclophane-like structures (A), nanotubes (B), and fullerenes (C).

COMPUTING TIME BENCHMARKS

Ring perception is an important step in almost all cheminformatics tasks. Applications which process large data sets thus require a fast method to identify the rings of molecular graphs. To check the large-scale applicability of the described method to calculate URFs, we measured the runtimes for the perception of URFs for a number of test sets. Time measurements were performed in a single thread on a PC with an Intel Core2 Quad Q9550 CPU (2.83 GHz) and 4 GB of main memory. For each molecule of the data set, the runtime for 100 iterations of ring perception was measured and on the basis of this measurement, the average runtime for a single ring perception was calculated. For file-I/O we used the NAOMI framework.¹⁹ Measured runtimes shown in Figure 7 do not include file I/O and molecular preprocessing. The data structures of the NAOMI framework are not specifically optimized for the detection of URFs but focus on the correct chemical modeling of small molecules. The listed runtimes thus provide an

estimate of URF detection in the context of a common cheminformatics application.

To investigate the maximum runtime for the perception of URFs, we generated a number of molecules containing highly complex ring systems. First, we generated cyclophane-like structures that contain a large macrocycle with n para-bridged 6-rings. The generated molecules have a cyclomatic number Z_n of $Z_n = n + 1$, contain $n^2 + n$ RCs and $n + 1$ URFs. The runtime for the calculation of the URFs of these molecules is shown in Figure 6A. The required runtime for molecules containing $|V|$ atoms and a cyclomatic number of Z increases approximately with $|V|^2$ and Z^2 .

As a second type of molecules that contain complex rings, single walled nanotubes were generated using ConTub.²⁰ While the parameters i and k were set to 5 nm, the length of the nanotube was increased in steps of 5 nm starting with a length of 10 nm up to a maximum of 100 nm. Both V and Z increase linearly with the length of the nanotube. As shown in Figure 6B,

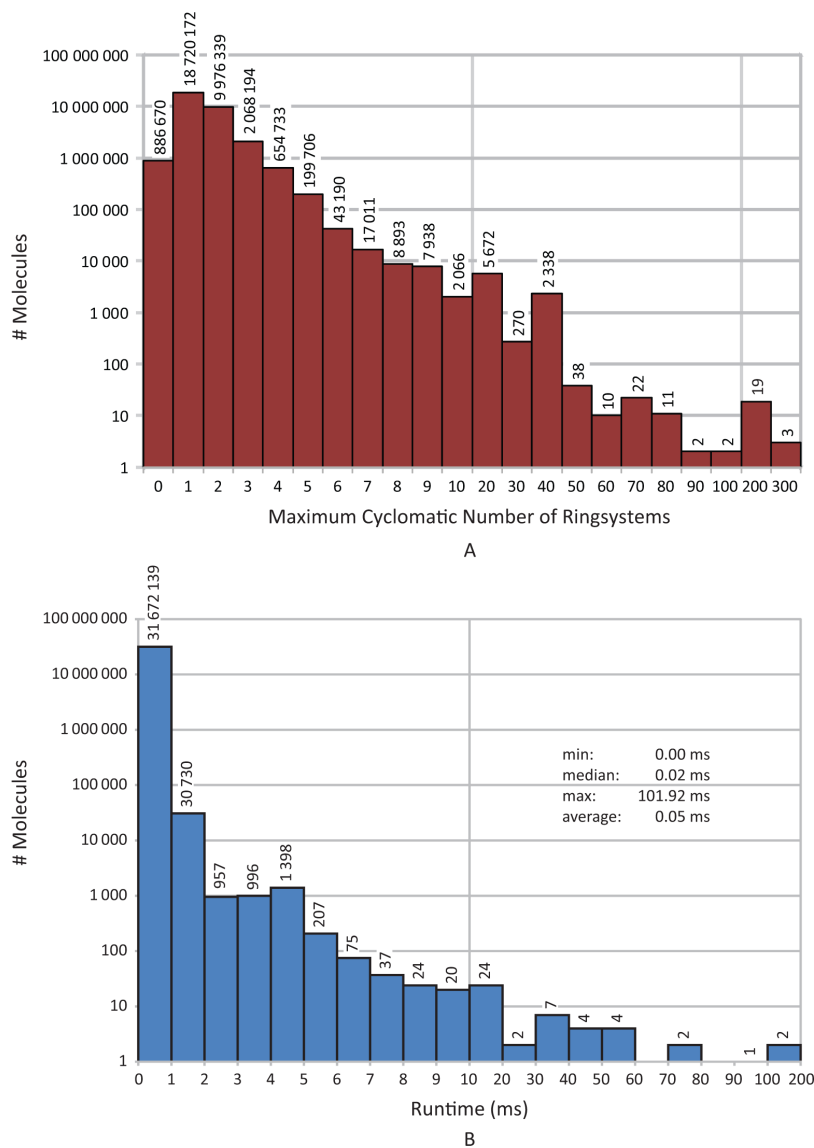


Figure 7. (A) Maximum cyclomatic number of the ringsystems of the molecules of the Pubchem-2D data set. (B) Benchmarks for URF perception for those molecules of the PubChem-2D data set having a cyclomatic number of at least one.

the runtime for the calculation of URFs increases slower than V^3 or Z^3 .

As a third set of complex molecules, a number of fullerenes ranging from C₂₄ to C₃₂₀ were generated. Coordinates of these molecules were taken from a Fortran program specialized in the generation of fullerenes.²¹ The runtime requirement again increased approximately with V^2 as well as with Z^2 (see Figure 6C).

Finally, to investigate the runtime which is required to perceive rings of commonly used molecules, we perceived URFs for the PubChem Compound 2D data set.³ The data set was

downloaded on March 27th, 2011 from <ftp://ftp.ncbi.nlm.nih.gov/pubchem/Compound/CURRENT-Full/> and contains 32 593 299 molecular structures. These include a number of molecules of high complexity not present in the respective 3D data set. Figure 7A illustrates the complexity of the data set by showing the maximum cyclomatic number for the ringsystems of each molecule.

Shown runtimes represent the required runtime for 100 iterations of ring perception. Nevertheless, these runtimes are close to zero for most common molecules. The median for the perception of URFs for a molecule of the Pubchem Data set is

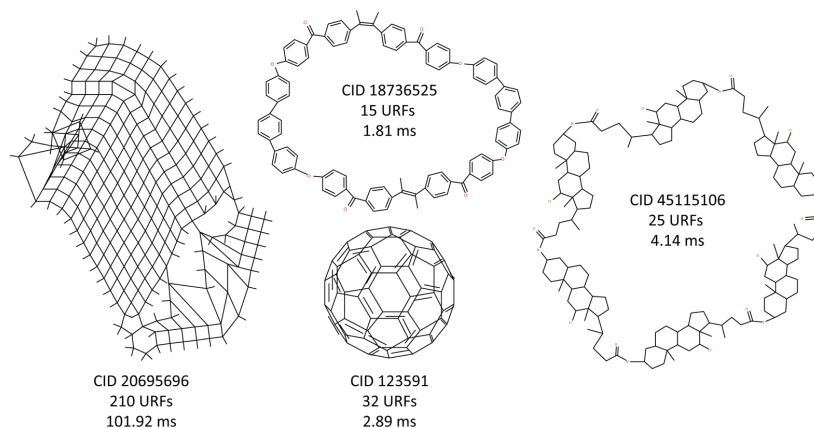


Figure 8. Runtimes and compound IDs for a selection of molecules of the Pubchem-2D data set.

0.02 ms, the average runtime is 0.05 ms, and the maximum runtime is 102 ms. This demonstrates that URFs can be calculated on the fly even for interactive applications and large databases. Only 34 490 molecules (0.11% of the database) show runtimes of more than 1 ms for the calculation of URFs. A list of those 100 molecules which require the highest runtimes for the calculation of URFs is added to this paper as Supporting Information. Some representative examples are shown in Figure 8.

A common molecular file format conversion, tested with Open Babel for the ZINC-everything data set, requires approximately 2 ms.¹⁹ Due to the low runtime for calculating URFs of about 0.02 ms for commonly used molecules, the perception of URFs is suitable for high throughput cheminformatics applications. Even for an artificially complex cyclophane-like structure containing $100 + 2^{100}$ RCs, the URFs can be calculated in less than 2 s.

CONCLUSION

We have introduced the concept of unique ring families (URFs). In contrast to common ring perception approaches, URFs are polynomial in number, unique, and provide a complete description of the rings of a molecular graph. Furthermore, we have described an efficient method to calculate URFs in polynomial time. We demonstrated its applicability on large scale by showing computing time benchmarks for the Pubchem 2D data set. For these reasons, URFs represent a valuable alternative to common ring perception concepts and are worthwhile to be considered as a standard description for ring topologies in molecular graphs.

ASSOCIATED CONTENT

Supporting Information

100 molecular structures of the PubChem Database which require the highest runtimes for the perception of URFs. This material is available free of charge via the Internet at <http://pubs.acs.org/>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: rarey@zbh.uni-hamburg.de.

Present Address

[‡]Evotec AG, Essener Bogen 7, 22419 Hamburg. Phone: 0049 40 56081 230. Email: Adrian.Kolodzik@evotec.com.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors would like to thank Christian Ehrlich and J. Robert Fischer for helpful comments and proofreading of the manuscript. Furthermore, the authors thank J. Robert Fischer and Tobias Lippert for their work on the NAOMI framework, which was used for reading the molecules of the Pubchem data set.

REFERENCES

- (1) Vismara, P. Union of all the minimum cycle bases of a graph. *Electron. J. Comb.* **1997**, *4*, 1–15.
- (2) Plotkin, M. Mathematical Basis of Ring-Finding Algorithms in CIDS. *J. Chem. Doc.* **1971**, *11*, 60–63.
- (3) Wang, Y.; Xiao, J.; Suzek, T.; Zhang, J.; Wang, J.; Bryant, S. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **2009**, *37*, 623–33.
- (4) Zamora, A. An Algorithm for Finding the Smallest Set of Smallest Rings. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 40–43.
- (5) Berger, F.; Flamm, C.; Gleiss, P.; Leydold, J.; Stadler, P. Counterexamples in Chemical Ring Perception. *J. Chem. Inf. Model.* **2004**, *44*, 323–331.
- (6) Hanser, T.; Jauffret, P.; Kaufmann, G. A New Algorithm for Exhaustive Ring Perception in a Molecular Graph. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1146–1152.
- (7) Balducci, R.; Pearlman, R. S. Efficient exact solution of the ring perception problem. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 822–831.
- (8) Carta, G.; Onnis, V.; Knox, A.; Fayne, D.; Lloyd, D. Permuting input for more effective sampling of 3D conformer space. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 179–190.
- (9) *Daylight Theory Manual 4.9*. <http://www.daylight.com/dayhtml/doc/theory/index.pdf> (accessed June 9th, 2012).
- (10) *Daylight Depictmatch*. http://www.daylight.com/daycgi_tutorials/depictmatch.cgi (accessed June 9th, 2012).
- (11) Petra M. Gleiss, J. L.; Stadler, P. F. Interchangeability of Relevant Cycles in Graphs. *Electron. J. Comb.* **2000**, 1–16.
- (12) Fujita, S. A new algorithm for selection of synthetically important rings. The essential set of essential rings for organic structures. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 78–82.

- (13) Corey, E.; Perersson, G. Algorithm for machine perception of synthetically significant rings in complex cyclic organic structures. *J. Am. Chem. Soc.* **1972**, *94*, 460–465.
- (14) Wipke, W.; Dyott, T. Use of Ring Assemblies in Ring Perception Algorithm. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 140–147.
- (15) Downs, G.; Gillet, V.; Holliday, J.; Lynch, M. Theoretical aspects of ring perception and development of the extended set of smallest rings concept. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 187–206.
- (16) Nickelsen, H. Ringbegriffe in der Chemie-Dokumentation. *Nachr. Dok.* **1971**, *3*, 121–123.
- (17) Dury, L.; Latour, T.; Leherte, L.; Barberis, F.; Vercauteren, D. A new graph descriptor for molecules containing cycles. Application as screening criterion for searching molecular structures within large databases of organic compounds. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1437–1445.
- (18) Tarjan, R.; Vishkin, U. An Efficient Parallel Biconnectivity Algorithm. *SIAM J. Comput.* **1985**, *14*, 862–874.
- (19) Urbaczek, S.; Kolodzik, A.; Fischer, J. R.; Lippert, T.; Heuser, S.; Groth, I.; Schulz-Gasch, T.; Rarey, M. NAOMI - On the almost trivial task of reading molecules from different file formats. *J. Chem. Inf. Model.* **2011**, *51*, 3199–3207.
- (20) Melchor, S.; Martin-Martinez, F. J.; Dobado, J. A. CoNTub v2.0 - Algorithms for Constructing C₃-Symmetric Models of Three-Nanotube Junctions. *J. Chem. Inf. Model.* **2011**, *51*, 1492–1505.
- (21) Schwerdtfeger, P. *Topological Analysis of Fullerenes - A Fortran Program*. <http://ctcp.massey.ac.nz/index.php?group=page=fullerenes&menu=fulleren> (accessed March 11th, 2012).

MONA - Interactive Manipulation of Molecule Collections

[D3] M. Hilbig, **S. Urbaczek**, S. Heuser, I. Groth, and M. Rarey. MONA - Interactive Manipulation of Molecule Collections. *Journal of Cheminformatics*, 5(1):38, 2013.

<http://dx.doi.org/10.1186/1758-2946-5-38>

RESEARCH ARTICLE

Open Access

MONA – Interactive manipulation of molecule collections

Matthias Hilbig¹, Sascha Urbaczek¹, Inken Groth², Stefan Heuser³ and Matthias Rarey^{1*}

Abstract

Working with small-molecule datasets is a routine task for cheminformaticians and chemists. The analysis and comparison of vendor catalogues and the compilation of promising candidates as starting points for screening campaigns are but a few very common applications. The workflows applied for this purpose usually consist of multiple basic cheminformatics tasks such as checking for duplicates or filtering by physico-chemical properties. Pipelining tools allow to create and change such workflows without much effort, but usually do not support interventions once the pipeline has been started. In many contexts, however, the best suited workflow is not known in advance, thus making it necessary to take the results of the previous steps into consideration before proceeding. To support intuition-driven processing of compound collections, we developed MONA, an interactive tool that has been designed to prepare and visualize large small-molecule datasets. Using an SQL database common cheminformatics tasks such as analysis and filtering can be performed interactively with various methods for visual support. Great care was taken in creating a simple, intuitive user interface which can be instantly used without any setup steps. MONA combines the interactivity of molecule database systems with the simplicity of pipelining tools, thus enabling the case-to-case application of chemistry expert knowledge. The current version is available free of charge for academic use and can be downloaded at <http://www.zbh.uni-hamburg.de/mona>.

Background

The compilation and preparation of small-molecule datasets forms the core of virtually all cheminformatics applications. The careful selection of relevant compounds and the thorough processing of the associated data are essential in order to obtain meaningful results. Although the necessary steps for this process strongly depend on the respective context, there are nevertheless a number of common and recurring tasks. These include, among others, the removal of duplicates, filtering by physico-chemical properties or substructure matching and the visual inspection of the respective compounds.

Workflow or pipelining tools support this recurrence by providing components or nodes corresponding to such common tasks. These nodes can be individually parameterized and combined in a pipeline, thus enabling the generation of a variety of customized workflows. The specification of these workflows is usually facilitated by a

graphical interface. The most commonly used programs in the context of cheminformatics are Pipeline Pilot [1] and the open-source alternative Knime [2] which have been compared in a recent review [3]. There are numerous further examples of scientific workflow systems described in the literature [4]. All these programs contain a certain number of predefined components and are extensible by allowing users to program their own modules. In addition to the flexibility concerning the specification of workflows, pipelining tools have the advantage that the processes are completely automated. This makes workflow processing the method of choice when all steps are known in advance and no intervention is necessary. Furthermore, there are usually only short setup times compared to the laborious installation and initialization of a server-based molecular database system. Molecular databases, on the other hand, make it possible to compile datasets in a more interactive manner. Data needed for common cheminformatics tasks can be calculated in advance and stored in the database, resulting in noticeably reduced run times for data access. For most common database systems chemical cartridges exist which provide

*Correspondence: rarey@zbh.uni-hamburg.de

¹Center for Bioinformatics (ZBH), University of Hamburg, Bundesstrasse 43, 20146 Hamburg, Germany

Full list of author information is available at the end of the article

the functionality to import chemical data. Molecules are typically written to SQL tables in the form of line notations such as (U)SMILES [5] or InChI [6]. These unique topological identifiers are used to ensure the uniqueness of molecules or to rapidly find particular molecules in the database. It is possible to reduce run times for substructure searches by annotating common substructures in molecules and for similarity searches by using pre-calculated fingerprints. Physico-chemical properties can be stored in databases using indices to boost the run times of filter operations. Depending on the number and kind of pre-calculated molecular descriptors, run times for setting up the databases can be quite large. Additionally, database systems often need to be installed on the respective operating system.

Here, we present MONA, a software tool aiming at combining the advantages of both approaches. In this way, the software enables a more interactive and intuitive approach to deal with large compound collections. In different validation procedures we show the internal consistency of all provided operations. Additionally we provide benchmarks showing that all provided operations are sufficiently fast for interactive use.

Methods

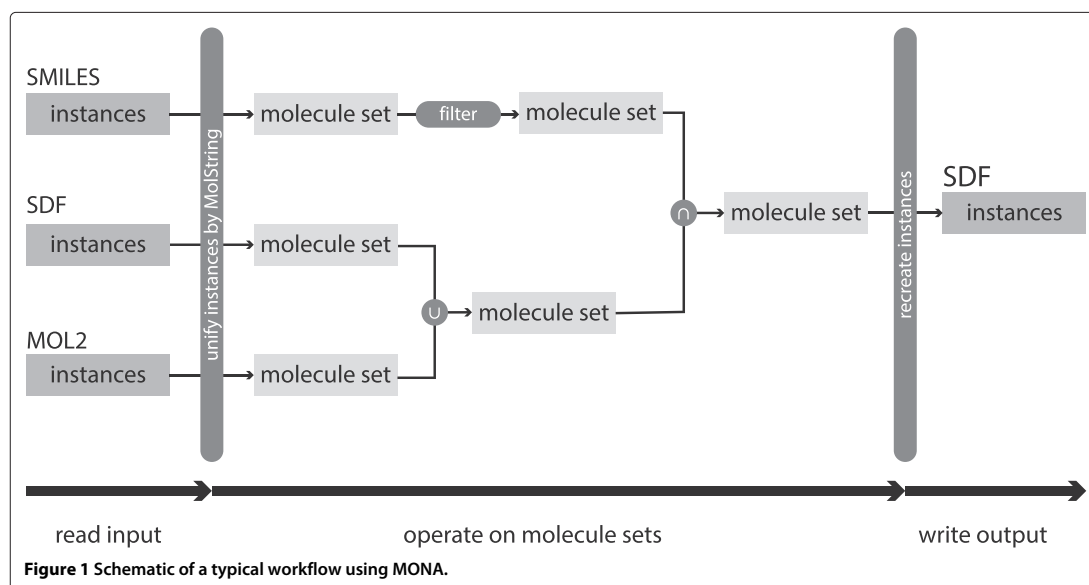
Based upon the NAOMI framework [7], MONA allows to interactively prepare, inspect and convert small-molecule datasets. The most important aspect of MONA is that the primary objects handled are molecules, not their occurrences in a particular dataset. During the import procedure, molecules are converted into a unique

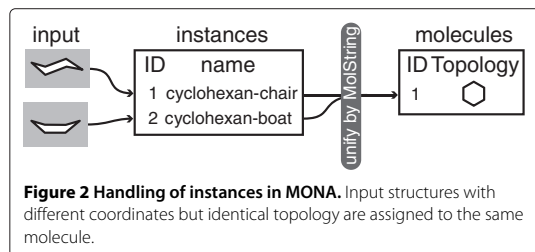
topological description, duplicates are automatically detected and stored as so-called instances. A typical MONA workflow scheme is shown in Figure 1. To ensure high efficiency, MONA employs a relational SQL database for all operations on datasets. Furthermore, MONA's architecture allows an efficient handling of molecule sets including their instant creation as well as classical set operations like union, intersection and difference.

The following sections describe the concepts behind MONA. This includes molecular representation and management by a relational database, performing operations on molecule sets, and rapid visualization of large compound collections.

Molecules and instances

In the context of MONA the terms molecule and instance are used to distinguish between the actual compound and its occurrence in a dataset (see Figure 2). There can be multiple instances of the same molecule originating from different entries of input files. Depending on the context these instances can be interpreted as either conformations or duplicate entries. In order to reliably assign instances to their corresponding molecules, a canonical topological description is needed. MONA uses an internal string representation called MolString which serves two purposes. First and foremost it is used to efficiently rebuild the molecule as this is needed for particular operations as explained in the following sections. Furthermore, it is used as unique topological descriptor for the assignment of instances to molecules during registration. Molecules are serialized to and from the database, where each molecule





and each instance is identified internally by a unique id called *Molecule Key* and *Instance Key* respectively.

Instances can be imported from common chemical file formats (SMILES, SDF, MOL2) using the NAOMI framework. The procedures for the consistent handling of these formats have been described in detail in [7]. If an entry consists of multiple disconnected components, currently solely the largest component is kept. Furthermore, it is possible to import small molecules from PDB files using the method described in [8]. In this case all components of the entry are imported. Additional data from SDF files is stored for each entry and can be recreated during export. Since the identification of molecules is based on a topological description, different tautomeric forms and protonation states are generally handled as separate entities. The same also applies to molecules with and without explicit specification of stereo descriptors. In order to customize the way molecules are assigned to instances, MONA offers different rules for the import of molecules. Depending on the context, molecules can be imported in a neutralized form, as canonized tautomer and without stereochemistry.

Molecule sets

MONA allows to organize compounds in molecule sets. Molecule sets are collections of pair-wise different molecules (not instances) which are used for all operations in MONA. As has been mentioned above, molecules are considered equal if and only if their canonical MolString representation is identical. We believe that this concept of molecular identity follows the basic understanding of chemists. Additionally, there are various technical reasons why sets of molecules are used rather than sets of instances. All available operations, such as filtering, manual selection and visualization, are based on molecular topology, so that there would not be any benefit from using sets of instances. Furthermore, some operations are based on the equality of the sets' elements. Due to the additional data from the input format equality of instances is ambiguous at best, whereas it is well defined for molecules on the topological level. In the end, working with molecule sets is more efficient and the results from set operations can be intuitively understood.

Molecule sets are stored internally as lists of *Molecule Keys*. MONA is able to handle an arbitrary number by keeping these lists in a relational database. When exporting molecule sets to chemical file formats, molecules must be converted back to instances. As instances for a given molecule may come from different input files, it is necessary to choose which source should be used for output generation. For that purpose, a list of original molecule sources is kept in the database. Data associated with a molecule, such as names and coordinates, are then either taken from the first found instance or from all instances in the chosen data sources and eventually exported to the output file.

Visualization of molecule sets

The analysis of the distribution of different physico-chemical properties is a simple way to get a first impression of a molecule set. For that purpose MONA offers customizable histograms for a number of common physico-chemical properties. It is also possible to include multiple sets in one histogram, which allows to compare their properties at a quick glance.

For further analysis, MONA offers a fast visualization of molecule sets using two-dimensional structure diagrams. This provides a means to visually inspect large molecule collections and manually select molecules for the creation of smaller sets. MONA does not offer any type of three-dimensional visualization which would only be needed to show differences between instances such as conformational variability. The necessary two-dimensional coordinates are generated by a built-in layout algorithm on the fly. In order to browse large molecule sets, the results of such calculations for the molecules must be available instantly. Even with a fast layout algorithm the pre-calculation of coordinates for all molecules in a set would take a prohibitively long time. Fortunately, coordinates for all molecules are really never needed. By using a model-view architecture and lazily calculating coordinates only when they are needed, browsing of molecule sets with hundred thousands of molecules becomes instantaneous. On modern hardware depictions of the few molecules a user can capture simultaneously on the computer screen appear without much latency. By intelligent multi-threading, including the cancellation of coordinate calculations for molecules that are no longer visible, fast scrolling of large sets does not lead to congested threads.

Operations on molecule sets

In general, MONA operates on molecule sets and creates new sets as results (see Figure 3). All sets can be used in further operations resulting in a high degree of flexibility. The intention of the set concept is to enable the typical workflow of interactive processing, namely to

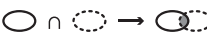
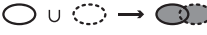
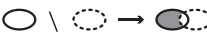
operations	signature	description	
	$\text{intersect}(S_1, S_2, \dots, S_n) \rightarrow S_r$	set S_r contains molecules that are in all original sets	
	$\text{union}(S_1, S_2, \dots, S_n) \rightarrow S_r$	set with all molecules contained in any of the original sets	
	$\text{difference}(S_1, S_2) \rightarrow S_r$	contains molecules from S_1 that are not in S_2	
filter set	$\text{filter}(S) \rightarrow S_r$	available filters	example
		physico chemical property	molecular weight > 200
		chemical element	contains oxygen, no nitrogen
		functional group	contains Pyridol
		smarts	contains c1cccc1
visual select subset	$\text{select}(S) \rightarrow S_r$	select subset of S by picking structure diagrams	
split subset	$\text{split}(S, n) \rightarrow (S_1, S_2, \dots, S_n)$	split set S into n equally sized parts	

Figure 3 Supported operations in MONA.

browse, select, and store data iteratively. The common mathematical set operations (union, intersection and difference) work on multiple input sets and produce a single set as result. Since these operations are solely based on the evaluation of identities of the contained molecules, they can be realized directly by the database using SQL statements. Because molecule sets are internally handled as lists of *Molecule Keys* the respective operations can be carried out efficiently. Mathematical set operations produce results instantaneously even for large datasets, which makes them suitable for interactive use. For the same reasons, the splitting of molecule sets by various criteria is interactively possible.

Filtering and visual selection

Both filtering and visual selection are operations on a single molecule set which generate a subset by excluding particular elements. The criterion for the exclusion is either a combination of molecular properties or manual selection. Filter chains for molecular properties are specified as a logical conjunction of elementary filters. Four elementary filter types are currently supported: (a) physico-chemical properties, (b) chemical elements, (c) functional groups, and (d) SMARTS patterns.

The physico-chemical properties comprise mostly topological descriptors such as the number of rings, molecular weight, and the topological surface area. This is extended by properties which can be derived from the chemical structure such as LogP [9]. Property filters always include or exclude a range of values the molecules must conform to. In contrast to that, substructure filters only ensure the presence or absence of a specific substructure in the molecules of the set. Chemical element filters are the most

basic type of substructure filters. They are typically used to remove large classes of molecules such as halogenated compounds. Functional group filters allow the exclusion or inclusion of a set of common functional groups including both aromatic rings and acyclic structures. The number of groups and their types are currently predefined in MONA. If these should not be sufficient, SMARTS expressions can be used to handle any type of chemical patterns. Additionally, MONA allows to upload collections of SMARTS patterns and use them in a single query. The efficiency of the filtering operation strongly depends on the selected filter types. Property filters are fast since the values for molecules are pre-calculated and stored in the database. These filters can therefore be realized by directly using database functionality. The same holds true for element and functional group filters. Both resort to pre-calculated bitfields saved in the database. These are slower than the property filter as SQL databases do not support bitfield matches. SMARTS filters are the computationally most demanding types, since all molecules have to be rebuilt from their MolString and tested against the SMARTS expression.

Elementary filters can be combined into complex queries which can be applied to any molecule set. In order to make filtering with criteria such as the Rule-of-Five for orally bioavailable molecules [10] possible, a tolerance can optionally be specified for a filter chain. This means that not all elementary filters need to match but only m of n filters, where $m \leq n$ can be arbitrarily chosen. Using tolerances has an impact on the speed of filtering operations. If $m < n$ the filter process becomes slower, since the filter chain needs to be transformed into multiple database queries instead of one.

MONA as application

MONA is a cross-platform application, which can be started without prior installation as no setup of an external database system is required. Currently SQLite is used as underlying database backend for its simplicity in setup and administration. SQLite is connected via a regular SQL API such that any other relational database system could be used instead.

The user interface consists of three different areas reflecting the functionality described in the previous sections. Imported molecule files are contained in the molecule sources view, from where molecule sets can be created at any time. The current molecule sets are shown in the list on the left side. They can be visualized in the respective views either as histograms or as a sortable table of structure diagrams. Operations for sets as described above are available in the toolbar or via the context menu. Filter chains can easily be built in the filter view (see Figure 4) using particular GUI elements for each type of elementary filter. Physico-chemical property filters are created with the help of a histogram that shows the distribution of the selected property in the currently chosen set. Chemical elements in the element filter can be selected in a periodic table, and functional groups are specified using structure diagrams. SMARTS expressions are entered in text form, the syntax is checked while typing and wrong expressions are highlighted.

All operations run in separate threads, which is the basis of this responsive user interface. It maintains its performance even if more demanding tasks are running in the background. Created molecule sets can be saved persistently in the database and restored when opening the database again. Molecule sets can eventually be exported to one of the supported chemical file formats from the context menu.

Results and discussion

The main focus of MONA are interactive scenarios where large molecule files need to be handled. To illustrate this further, three different workflows are described:

Scenario 1: Preparing a molecule dataset for screening

The compilation of a set of molecules for a virtual or experimental screening is a very common task in cheminformatics. Starting with a large collection of compounds the preparation mainly consists of selecting a subset of molecules with suitable properties for the target to be addressed (see Figure 5). For this purpose various filters can be iteratively created and tested. A few common filters, e.g., the Rule-of-Five, are already predefined in MONA and can be used directly. In addition to the use of filters, molecules can also be selected manually using visual selection. The manual selection can often be facilitated by sorting the molecules according to a specific property. If the results of different filter runs are kept as sets, they can be compared to each other using set operations. Set operations can also be used to eliminate particular molecules (rather than substructures) from molecule sets. One can simply load a file containing unwanted compounds and subtract them from the current set. All steps can be iteratively applied after visual inspection of the remaining and the rejected molecules. For example, bounds related to physico-chemical properties can be adapted on a case-to-case basis depending on the size of the remaining library. After finding the right combination of filters the final candidate set can be exported into an appropriate file format and used by another program. All data including 3D coordinates from instances previously read into the database are retained in this step.

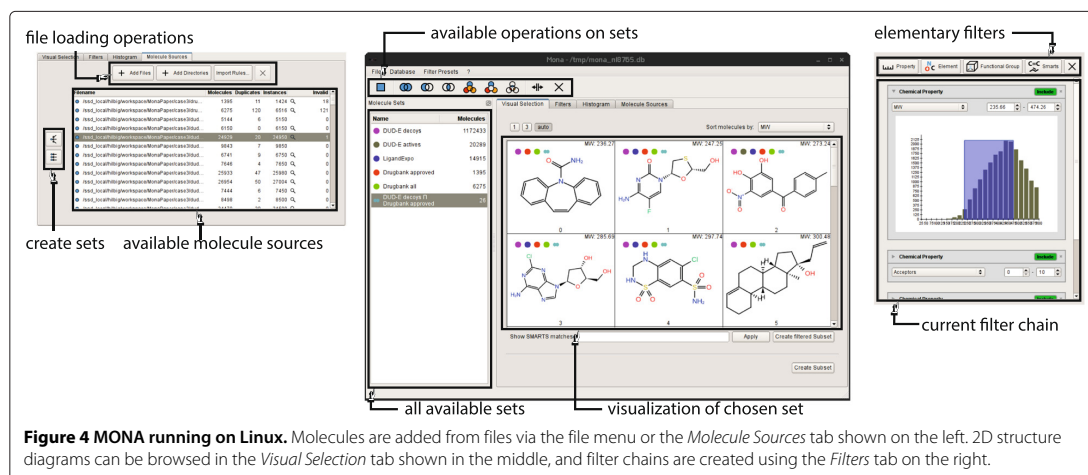
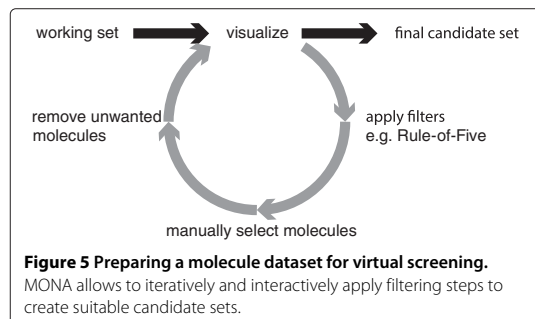
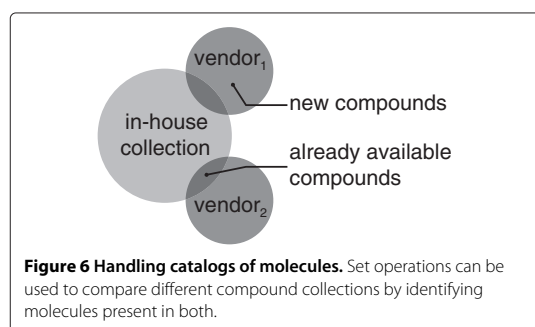


Figure 4 MONA running on Linux. Molecules are added from files via the file menu or the *Molecule Sources* tab shown on the left. 2D structure diagrams can be browsed in the *Visual Selection* tab shown in the middle, and filter chains are created using the *Filters* tab on the right.



Scenario 2: Handling catalogs of molecules

The second scenario is taken from the field of compound management. Many vendors offer their compound catalogs in the form of chemical data files. These files can be used to compare the compound portfolio of the different vendors with each other or with an in-house library (see Figure 6). This task is usually complicated by the fact that each vendor uses different standards for the representation of the respective compounds. When loading vendor catalogs as sets within MONA, different file formats and molecules across different vendors are automatically unified. Optionally, the user can decide to unify additional properties like the tautomeric state or the protonation. The resulting individual sets can be intersected with each other for comparison and evaluation. In this way either compounds offered by various vendors or substances that are uniquely supplied by one vendor can be easily identified. Furthermore, the sets can also be intersected with a current in-house collection, so that potential additions may be identified. Vendor catalogs usually contain price information and order numbers for each compound. Exporting all instances for molecule sets preserves this information and allows to compare prices for all molecules in the exported set.



Scenario 3: Verifying existing molecular databases

Databases like DUD-E [11,12] are widely used to test and evaluate the performance of docking algorithms. The functionality provided by MONA can be used to simplify verification tasks that are tedious to do manually. In order to validate the new DUD-E database, we tried to answer the following three questions (see Figure 7):

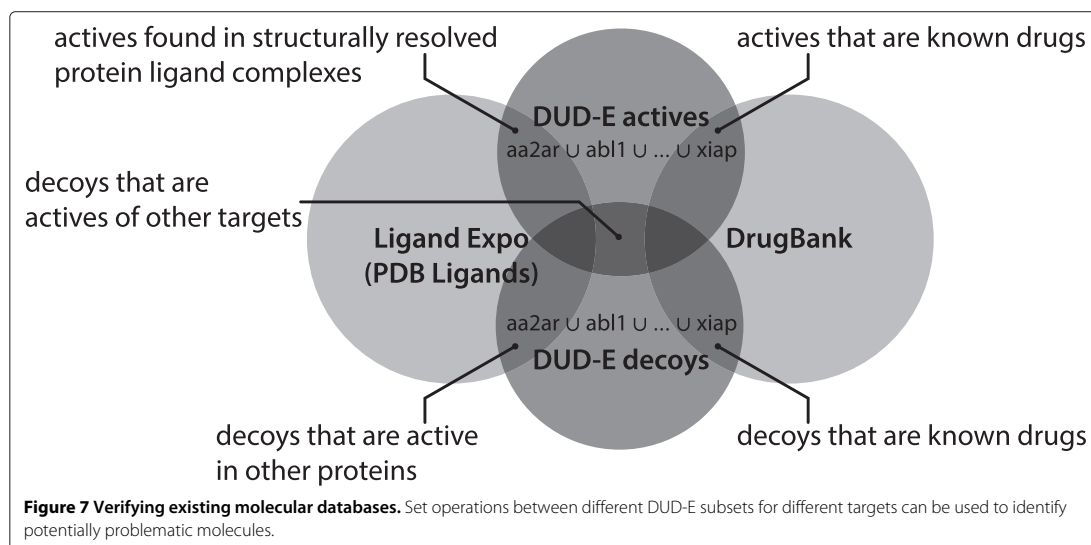
- Are any of the actives decoys for other targets?
- Are any of the decoy molecules ligands found in structurally resolved protein-ligand complexes?
- Are any of the decoy molecules already known drugs?

In order to investigate the first question, one molecule set with actives and one set with decoys was created from the respective files for each individual target. Then, all active sets were united into one set *A* and all decoys were united into one set *D*. The intersection of both sets directly provides the answer to the first question. The resulting set contains 123 molecules (provided in Additional file 1).

To answer the second question, the decoy set *D* has to be intersected with a set containing known ligands from protein-ligand complexes. The necessary data is provided by LigandExpo [13,14] which offers a SMILES file containing all small molecules from crystal structures in the Protein Data Bank (PDB) [15]. The resulting intersection contains 141 decoys which are ligands of at least one protein in the PDB (provided in Additional file 2).

The third question can be answered in the same way. This time, a substance set of approved drugs from Drugbank [16,17] was used as reference. Drugbank currently lists 1395 molecules registered as drugs. The intersection of these molecules with *D* contains 26 molecules (provided in Additional file 3) each of which is approved as a drug. Most interestingly, the resulting set contains the compound cladribine (see Figure 8), which is known to interact to deoxycytidine kinase and considered as a decoy molecule of mitogen-activated protein kinase 1. The compound nandrolone phenpropionate is a known substrate to cytochrome P450 19A1 and considered decoy for cytochrome P450 3A4. Although these two molecules might in fact be inactive against their decoy targets, this analysis at least points to critical cases where the decoy status should be further clarified.

Furthermore, it is possible to quickly exploit the data sources like the PDB for seeking alternative targets for all the actives in the DUD-E dataset. Let A_i be the set of active compounds for each target *i*. The intersections between each A_i and the LigandExpo set results in one set per target containing all compounds for which complex structures are deposited in the PDB. Exporting these sets with all instances taken from LigandExpo results in one file for each target containing other proteins in the PDB with



the same ligand. As an example the active flavopiridol for cdk2 was found which also inhibits glycogen phosphorylase (PDB code 1e1y). Note that searching for flavopiridol in the PDB easily gives the same result but with MONA, this search process was performed with all 20289 active molecules of DUD-E simultaneously without the need for scripting.

It took seven minutes to import all 1.2 million molecules necessary for this scenario into the database and one minute to create all sets in the GUI on an Intel Core i7-2600 CPU with 3.4 GHz and 8 GB of memory. All individual set operations ran in less than 10 seconds.

Correctness

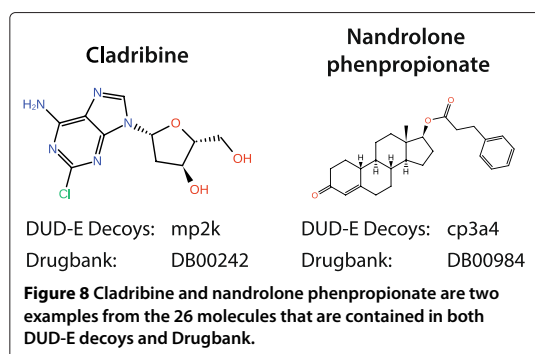
All operations provided by MONA depend on the consistent internal representation of molecules and their respec-

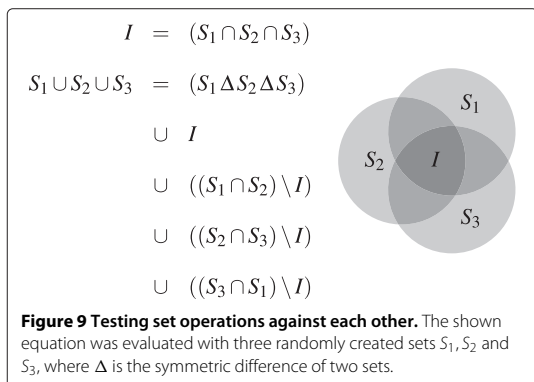
tive properties. This applies to both the internal chemical model and the operations performed by the underlying database. The consistency of the chemical model concerning the handling of different chemical file formats has already been validated in [7]. Therefore, the validation of MONA was focused on the correctness of the database functionality. This was done by ensuring the following invariants:

1. Molecules stored in the database are restored exactly as before.
2. Molecule sets can be created and combined with set operations.
3. Different types of filters can be correctly applied to molecule sets.

Storage of molecules in the database is tested by comparing a molecule restored from the database with the original molecule. The order of atoms and bonds may change, but if any valence states or atom coordinates differ the test fails. All molecules passing NAOMI initialization from PubChem Substance (100 M molecules) [18,19] and from emolecules (5 M molecules) [20] can be correctly restored from the database.

Operations on sets of molecules were tested against each other by verifying that the general equation in Figure 9 holds. Sets S_1 , S_2 and S_3 are created by randomly distributing molecules of a test set to one, two or all three sets. Then the union of S_1 , S_2 and S_3 must be the same as the union of the symmetric difference ($S_1 \Delta S_2 \Delta S_3$), the intersection of all three sets and all pair-wise intersections of two sets.



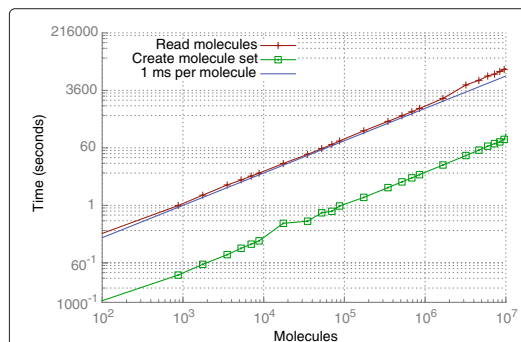


Confirming filter operations was done by comparing results returned by the database against the results retrieved by linearly applying each filter against every molecule in turn.

Computing time

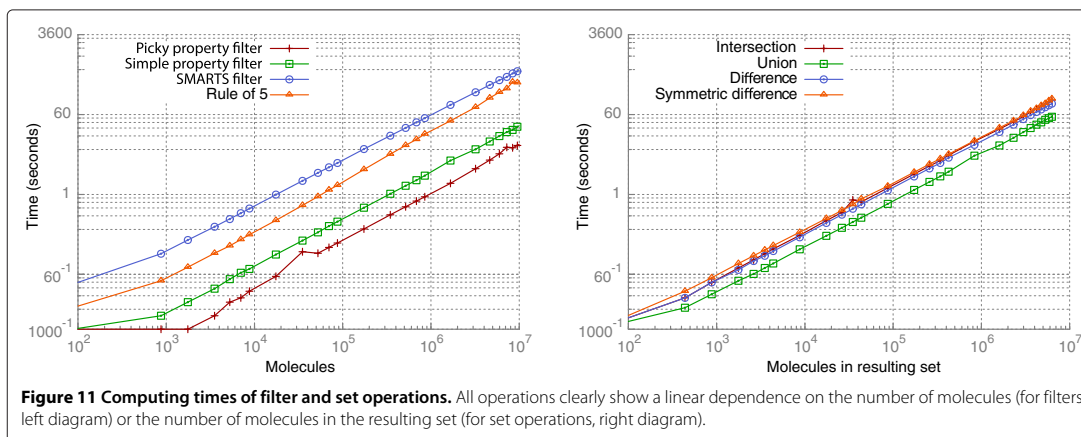
In order to assess the computing time requirements of MONA, scaling tests for important operations on the database were performed. As most of the operations only consist of database queries the results are highly dependent upon the used database backend. Here, SQLite was used with a page cache of 1 GB. This value was chosen as the best compromise for modern workstations.

All benchmarks were done on a workstation with an Intel Xeon E5630 CPU running at 2.53 GHz and 64 GB of available main memory. A subset of molecules from the PubChem Substances database was used as benchmark set. The molecules in this set were randomly chosen with uniform probability from the whole PubChem Substance database.



Naturally, the size of the database depends linearly on the size of the input. In our case the size of the database corresponds roughly to the size of a compressed SD file of the same compound set. All in all it takes approximately 1000 seconds to read 1 million molecules from SDF (see Figure 10), resulting in a database of size 1 GB, which is much smaller than the respective uncompressed MOL2 or SDF files.

The relative order of run times for different types of filters (see Figure 11) has been discussed in Section "Filtering and visual selection". Additionally, all filters and set operations do not only depend linearly upon the size of the input set but also on the size of the resulting set. This can be seen when comparing the picky property filter to the simple property filter from Figure 11 as the picky filter has to write considerable less results into a new subset in the database.



In summary, we conclude that MONA is efficient enough to handle sets with up to one million molecules interactively on a current workstation with at least 2 GB of main memory. Therefore, it can be used as a desktop application for most cheminformatics tasks.

Conclusion

MONA is an intuitive, interactive tool for processing large small-molecule datasets. It offers functionality to perform many common cheminformatics tasks such as combining datasets, filtering by molecular properties, and visualization using a built-in 2D engine. Since MONA is based on a robust cheminformatics framework, molecules from common file formats (SMILES, SDF, MOL2) can be handled consistently. The low setup time despite the use of a database makes MONA a reasonable compromise between pipelining tools and molecule database systems. More importantly, MONA offers a different way of working with molecule datasets. Compared to pipelining tools, it supports an interactive and case-driven process. While chemical databases and pipelining tools are mostly in the hands of cheminformaticians, MONA's lightweight interface offers chemists an easy way to deal with large compound collections.

We have provided three prototypical scenarios from different fields of applications which emphasize the great versatility of MONA. Various validation procedures show that MONA is internally consistent concerning both the representation of molecules and the database operations. Furthermore, the run times for dataset operations from the benchmarks are sufficient for interactive use in most situations with up to one million molecules.

Since working with datasets is such a central task in cheminformatics there are a lot of potential additional features which could be included in future versions of MONA. We are confident, that MONA's functionality will be substantially extended over the next year. The main focus will be on the introduction of new types of visualizations for molecular sets with respect to molecular similarity and molecular scaffolds. The current version can be downloaded at <http://www.zbh.uni-hamburg.de/mona>. It is available free of charge for academic use.

Additional file

Additional file 1: The file contains all molecules from the intersection between a set containing all DUD-E decoys and a set containing all DUD-E actives (123 molecules).

Additional file 2: The file contains all molecules from the intersection between a set containing all DUD-E decoys and a set containing all LigandExpo molecules (141 molecules).

Additional file 3: The file contains all molecules from the intersection between a set containing all DUD-E decoys and a set containing all DrugBank molecules (23 molecules).

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

M.H. and S.U. developed the algorithmic concepts behind MONA, implemented the software and tested it. I.G. and S.H. participated in the user interface design and performed initial tests. M.R. initiated the development and supervised the project. All authors read and approved the final manuscript.

Acknowledgements

Thanks to Mathias v. Behren, Andreas Heumeier and Thomas Otto for the first version of MONA and demonstrating that sets of molecules are a worthwhile idea. We further thank Thomas Lemcke (University of Hamburg) for his pharmaceutical advice and Marcus Gastreich and Christian Lemmen (BioSolveIT GmbH) for critically reviewing the usability of MONA.

Author details

¹Center for Bioinformatics (ZBH), University of Hamburg, Bundesstrasse 43, 20146 Hamburg, Germany. ²Beiersdorf AG, Research Active Ingredients, Tropelwitzstrasse 15, 22529 Hamburg, Germany. ³Nuremberg Institute of Technology Georg Simon Ohm, Kesslerplatz 12, 90121 Nuremberg, Germany.

Received: 12 June 2013 Accepted: 31 July 2013

Published: 28 August 2013

References

1. Accelrys Software Inc: **Pipeline Pilot 8.5**. 2012.
2. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, Ohl P, Sieb C, Thiel K, Wiswedel B: *KNIME: The Konstanz information miner, Studies in Classification, Data Analysis, and Knowledge Organization*. Berlin Heidelberg: Springer; 2008.
3. Warr W: **Scientific workflow systems: Pipeline pilot and KNIME**. *J Comput-Aided Mol Des* 2012, **26**(7):801–804.
4. Kappeler M: **Software for rapid prototyping in the pharmaceutical and biotechnology industries**. *Curr Opin Drug Discov Dev* 2008, **11**(3):389–392.
5. Weininger D, Weininger A, Weininger J: **SMILES. 2. Algorithm for generation of unique SMILES notation**. *J Chem Inf Comput Sci* 1989, **29**(2):97–101.
6. Heller S, McNaught A, Stein S, Tchekhovskoi D, Pletnev I: **InChI - the worldwide chemical structure identifier standard**. *J Cheminform* 2013, **5**:7.
7. Urbaczek S, Kolodzik A, Fischer R, Lippert T, Heuser S, Groth I, Schulz-Gasch T, Rarey M: **NAOMI - On the almost trivial task of reading molecules from different file formats**. *J Chem Inf Model* 2011, **51**(12):3199–3207.
8. Urbaczek S, Kolodzik A, Groth I, Heuser S, Rarey M: **Reading PDB Perception of molecules from 3D atomic coordinates**. *J Chem Inf Model* 2013, **53**(1):76–87.
9. Wildman SA, Crippen GM: **Prediction of physicochemical parameters by atomic contributions**. *J Chem Inf Comput Sci* 1999, **39**(5):868–873.
10. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ: **Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings**. *Adv Drug Deliv Rev* 2001, **46**(1–3):3–26.
11. Mysinger M, Carchia M, Irwin J, Shoichet B: **Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better Benchmarking**. *J Med Chem* 2012, **55**(14):6582–6594.
12. **DUD-E**. [<http://dude.docking.org/>] [Data set as SDF downloaded on 2013-02-01].
13. Feng Z, Chen L, Maddula H, Akcan O, Oughtred R, Berman H, Westbrook J: **Ligand Depot: a data warehouse for ligands bound to macromolecules**. *Bioinformatics* 2004, **20**(13):2153–2155.
14. **Ligand Expo**. [<http://ligand-expor.rcsb.org/>] [Data set as SMILES (CACTVS with stereo) last accessed on 2013-02-01].
15. Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov I, Bourne P: **The protein data bank**. *Nucleic Acids Res* 2000, **28**(1):235–242.
16. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djoumbou Y, Eisner R, Guo A, Wishart D: **DrugBank 3.0: a comprehensive resource for 'omics' research on drugs**. *Nucleic Acids Res* 2011, **39**(suppl 1):D1035–D1041.

17. **DrugBank**. [<http://www.drugbank.ca/>] [Data set with approved drugs as SDF downloaded on 2013-02-01].
18. Bolton E, Wang Y, Thiessen P, Bryant S, Elsevier: **Chapter 12 PubChem: Integrated platform of small molecules and biological activities**. *Annu Rep Comput Chem* 2008, **4**:217–241.
19. **PubChem Substance**. [<http://www.ncbi.nlm.nih.gov/pcsubstance>] [Data set as SDF downloaded on 2012-20-09].
20. **eMolecules**. [<http://www.emolecules.com/>] [Data set as SDF downloaded on 2012-20-09].

doi:10.1186/1758-2946-5-38

Cite this article as: Hilbig *et al.*: MONA – Interactive manipulation of molecule collections. *Journal of Cheminformatics* 2013 **5**:38.

Publish with **ChemistryCentral** and every scientist can read your work free of charge

“Open access provides opportunities to our colleagues in other parts of the globe, by allowing anyone to view the content free of charge.”

W. Jeffery Hurst, The Hershey Company.

- available free of charge to the entire scientific community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
<http://www.chemistrycentral.com/manuscript/>



ChemistryCentral

The Valence State Combination Model: A Generic Framework for Handling Tautomers and Protonation States

[D4] S. Urbaczek, A. Kolodzik, and M. Rarey. The Valence State Combination Model: A Generic Framework for Handling Tautomers and Protonation States. *Journal of Chemical Information and Modeling*, 54(3):756-766, 2014.


<http://pubs.acs.org/articlesonrequest/AOR-aNbSPU9D97tS5tadunet>

Reproduced with permission from S. Urbaczek, A. Kolodzik, and M. Rarey. The Valence State Combination Model: A Generic Framework for Handling Tautomers and Protonation States. *Journal of Chemical Information and Modeling*, 54(3):756-766, 2014. Copyright 2014 American Chemical Society.

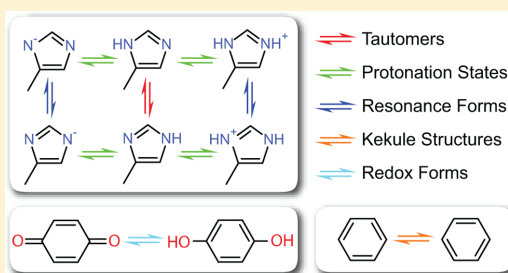
The Valence State Combination Model: A Generic Framework for Handling Tautomers and Protonation States

Sascha Urbaczek,[†] Adrian Kolodzik,[†] and Matthias Rarey*[‡]

University of Hamburg, Center for Bioinformatics (ZBH), Bundesstrasse 43, 20146 Hamburg, Germany

 Supporting Information

ABSTRACT: The consistent handling of molecules is probably the most basic and important requirement in the field of cheminformatics. Reliable results can only be obtained if the underlying calculations are independent of the specific way molecules are represented in the input data. However, ensuring consistency is a complex task with many pitfalls, an important one being the fact that the same molecule can be represented by different valence bond structures. In order to achieve reliability, a cheminformatics system needs to solve two fundamental problems. First, different choices of valence bond structures must be identified as the same molecule. Second, for each molecule all valence bond structures relevant to the context must be taken into consideration. The latter is especially important with regard to tautomers and protonation states, as these have considerable influence on physicochemical properties of molecules. We present a comprehensive method for the rapid and consistent generation of reasonable tautomers and protonation states for molecules relevant in the context of drug design. This method is based on a generic scheme, the Valence State Combination Model, which has been designed for the enumeration and scoring of valence bond structures in large data sets. In order to ensure our method's consistency, we have developed procedures which can serve as a general validation scheme for similar approaches. The analysis of both the average number of generated structures and the associated runtimes shows that our method is perfectly suited for typical cheminformatics applications. By comparison with frequently used and curated public data sets, we can demonstrate that the tautomers and protonation state produced by our method are chemically reasonable.



INTRODUCTION

One of the most fundamental requirements in cheminformatics is the consistent handling of molecules from different sources. There is always the implicit assumption that the results of cheminformatics software applications are only dependent on the actual compounds and not on the way these are provided in the input data. Yet, apart from problems arising from the interpretation of data from chemical file formats, there are certain ambiguities in the way molecules are represented which considerably complicate this task. Virtually all modern cheminformatics systems are based on a description of molecules by valence bond structures (Lewis structures). The inherent limitations of this molecular representation and their implications on tautomer generation have been recently discussed in detail by Sayle.¹ In the following, we will largely follow the nomenclature used in his publication and refer back to particular aspects mentioned therein.

The main problem with respect to consistency is the fact that different valence bond structures can represent the same molecule. Some of these correspond to distinct chemical entities, e.g., tautomers and protonation states, whereas others are artifacts of valence theory, i.e., resonance forms and Kekule structures. In some contexts even oxidation states may be

interpreted as alternative forms of the same molecule (see Figure 1 for examples).

From a formal point of view, each of these valence bond structures could be chosen as a representation for a particular compound. In practice, not all members of this set of alternatives are equally likely to be encountered due to automated normalization procedures and manual curation. However, despite all these efforts, a certain degree of ambiguity cannot be entirely avoided. The resulting implications for cheminformatics systems in general² and large compound databases in particular³ have been thoroughly investigated in the literature. In his publication, Sayle¹ has identified five specific tasks associated with the ambiguities of molecular representations. With respect to consistency; these are comparison (#1) and, more importantly, canonicalization (#2). A cheminformatics system must be able to reliably identify and treat alternative valence bond structures as the same molecule. This is usually done by conversion to a canonical form which serves as input for subsequent methods.

Received: December 6, 2013

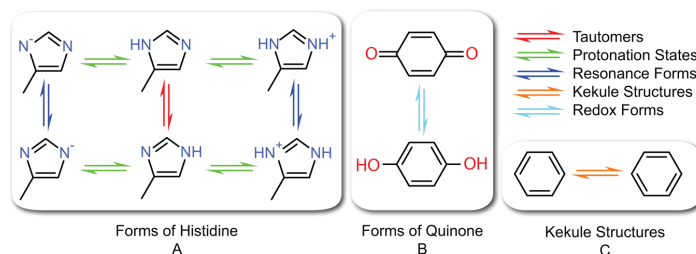


Figure 1. (A) Different valence bond structures of the imidazole ring of histidine including prototropic tautomers, protonation states, and resonance forms. (B) Two oxidation forms (quinone and hydroquinone) may in some context be considered as the same molecule. (C) Kekule structures are valence bond structures of aromatic rings with alternating single and double bonds.

The generation of unique identifiers, e.g., InChI,^{4,5} is a typical application scenario for canonicalization procedures.

Another quite opposite problem arises with regard to the general reliability of cheminformatics calculations. In many cases, it is necessary to consider multiple valence bond structures to sufficiently represent a molecule. The most prominent examples are certainly tautomers and protonation states, which will be summarized under the term protomers in the following. Since these correspond to actual physical entities, their respective ratios can have significant influence on a compound's observed physicochemical properties.^{6–11} The problem is, however, not exclusive to this scenario, as different resonance forms also play a role during the calculation of partial charges.¹² The respective tasks identified by Sayle¹ are (complete) enumeration (#3) and selection (#4). Both refer to the generation of valence bond structures, the difference being that selection (#4) restricts the results to a subset containing only relevant, e.g., energetically stable, solutions. Virtual screening techniques such as molecular docking are applications in which selection (#4) plays an important role. Relying on only one valence bond structure can lead to false-negative results as particular protomers may interact differently with target proteins. On the other hand, a large number of (possibly energetically unfavorable) alternatives can result in an increased false-positive rate and unnecessarily high runtimes. The general implications on structure-based and ligand-based screening methods have been investigated in several publications.^{13–15} The final task mentioned by Sayle¹ is prediction (#5), which extends selection by additionally ranking the relevant solutions by their respective energy.

The basic problem associated with the interconversion of valence bond structures is to transform groups of atoms according to specific rules with respect to bond orders and atomic properties (formal charges, bound hydrogens). As has been proposed by Sayle,¹ the methods developed for that purpose can be roughly divided into two categories: (1) Local approaches rely on pattern matching to identify relevant groups of atoms. These patterns are associated with rules describing the respective changes in the molecule. Pattern-based methods thus only use transformations that were anticipated in advance, thereby reducing the risk of generating unexpected and probably unwanted results. On the other hand, there is always the possibility of omitting relevant structures due to missing patterns. This can occur even if rules of a similar type are already included in the pattern library. Transformations covering long bond paths are a typical example for that problem. There are multiple publications describing local methodologies in the literature.^{13,16–18} (2) Global approaches

predefine substructures in a molecule, identify atoms with variable states within, and subsequently enumerate valid valence bond structures. This is usually done in a more generic manner than matching specific patterns, so that the results can easily contain completely artificial, i.e. chemically unreasonable, results. These either have to be omitted directly during or removed after the enumeration procedure. The omission of transformations in more complex structures, however, is generally not a problem. Global approaches have also been described in the literature^{19–21} and other sources.²² It must be noted that the previous differentiation between the two types of methods has been introduced mainly for classification purposes. Local approaches, for instance, often include a number of long-range patterns which, in combination with the underlying transformation engine, makes them suitable for the handling of the vast majority of molecules relevant in the field of drug design.

Here, we present the valence state combination model, a new concept for the description and classification of valence bond structures based on the NAOMI²³ framework. Using this model, we have developed, based on similar ideas as the ones presented by Sayle et al.,²² an extended and significantly improved method for the generation of valence bond structures which falls into the general category of global approaches. By application of a generic scoring scheme, this method combines the inherent consistency of the global strategy with the high reliability generally attained by local approaches. In contrast to previously published global methods, our approach consistently deals with all aspects relevant for the generation of protomers, including resonance forms and ionization states. Our method has been used to solve three common cheminformatics tasks, namely the generation of a canonical form (canonicalization), the generation of a preferential representation (normalization), and the generation of a set of reasonable protomers (generation). We have tested each application with respect to consistency using a general and comprehensible validation scheme. Furthermore, we have assessed the general suitability of our approach for common cheminformatics applications on the basis of these three operations. The criteria for the evaluation comprise runtime, the average number of generated structures, and the quality of the resulting protomers.

METHODOLOGY

Valence State Combination Model. Valence bond structures of molecules are generally represented as graphs in which nodes correspond to atoms and edges correspond to bonds. Each atom is associated with an element and a formal charge and each bond with a localized bond order (single,

double, or triple). In the NAOMI model,²³ this description is extended by an atom-based valence state descriptor. A valence state is a chemically valid combination of bond orders and formal charge for a particular element (see Figure 2). This

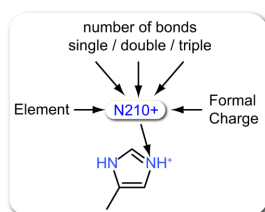


Figure 2. Example of a valence state descriptor for a nitrogen atom. The descriptor comprises the atom's element, bond order distribution, and formal charge.

additional descriptor is used to ensure the chemical validity of a molecule. A valence bond structure is valid if a valence state with the given bond orders and formal charge exists for each atom. Furthermore, valence states provide the means to systematically classify and generate different valence bond structures of molecules as explained below.

A set of valence states for all atoms of the molecule is called a valence state combination (VSC). A VSC is valid if a distribution of bond orders compatible with these valence states exists. Valid VSCs thus correspond to valence bond structures associated with a particular heavy atom skeleton. Note that bond orders are not part of the VSC representation; they are used for validation purposes only. Relations between valence bond structures can be determined by comparison of their corresponding VSCs (see Figure 3).

The description of these relations is based on atoms with different valence states, considering both their number and their types. Depending on the changed properties, substitutions of valence states for atoms are classified as protonation type, tautomer type, and resonance type as shown in Table 1. The involved states are called donors (higher number of single bonds) or acceptors (lower number of single bonds). The respective numbers of substitutions in VSCs are denoted as $\Delta_{\text{type}}(D \rightarrow A)$ and $\Delta_{\text{type}}(A \rightarrow D)$.

Table 2 lists the six basic relation types together with their conditions. Distinct valence bond structures with identical VSCs correspond to Kekule forms. They differ only in their respective bond order distribution. If all substitutions between two VSCs are of the protonation type, two cases need to be distinguished. When changing a donor to an acceptor or vice versa, the formal charge of the respective atom changes due to the addition or removal of hydrogen atoms. If the number of substitutions of donors and acceptors is not equal, the total charge of the molecule is altered, resulting in a different ionization state. Otherwise, the net charge of the molecule is identical, meaning that protons are merely occupying different locations. Tautomers and mesomers contain only changes of the tautomer-type and the resonance-type, respectively. Additionally, the number of donors and acceptor substitutions must be equal. Otherwise, the VSCs represent different redox forms of the molecule.

Substitution types can also occur in mixed constellations, and the resulting relations are best described as combinations of the just presented basic types. The 1-hydroxy-2-pyridone men-

tioned by Sayle¹ is an interesting example. The valence bond structures shown in Figure 4 can be best characterized as different resonance forms, a zwitterionic and a neutral one, with different proton positions.

The algorithms presented in the following chapters are based on the VSC representation of molecules. One of its major advantages is the fact that all of the potentially relevant molecule states can be consistently generated by considering different types of valence state substitutions. By explicitly handling all the different cases described in this section, a high degree of generality can be achieved.

Overview. The complete workflow for the generation of valence bond structures is shown in Figure 5. In the first step, the molecule is subdivided into multiple nonoverlapping substructures which are then treated independently. This partitioning reduces the computational costs for both the generation and the subsequent scoring of VSCs. A partition is considered valid if the independent enumeration of VSCs of each part and a subsequent combination of these lead to the same VSCs as if the enumeration would have been performed on the whole molecule. A partition is optimal if it is valid and has the smallest possible substructures. In the following sections, two partitioning schemes (generic and heuristic) are presented. Both are applied for the solution of different cheminformatics tasks described in later sections.

After partitioning, the atoms of each substructure are checked for alternative valence state assignments. Which valence states are included strongly depends on the context and will be explained in more detail later. As well as partitioning, valence state selection has a strong influence on the computational costs of the subsequent steps. The more alternatives are selected, the more VSCs must be generated and potentially scored. An optimal selection scheme thus only selects valence states for atoms that actually need to be modified. Again, two selection schemes (generic and heuristic) for different applications will be presented.

In the next step, VSCs are generated for each substructure using the alternative states selected in the previous step. Each of these VSCs is checked for validity by attempting to calculate a bond order distribution. VSCs for which this is not possible are invalid and therefore rejected. During the calculation, additional boundary conditions, e.g., the oxidation state of the initial molecule, are preserved.

The resulting VSCs are all chemically valid but may still contain undesired valence bond structures. These include unstable tautomers, unlikely protonation states, unreasonable resonance forms, or unusual representations of functional groups. In order to identify and eventually remove these VSCs, a pattern-based scoring scheme is applied. The resulting score expresses how well a particular substructure of the molecule is represented by the respective VSC. It must be stressed that the scoring scheme has not been designed to accurately predict the ratios between different molecular species. Its two main purposes are the elimination of completely artificial representations, i.e., energetically inaccessible states, and the coarse categorization of the remaining VSCs into stability classes. After eliminating all undesired VSCs, the final valence bond structures are completely enumerated by combining the VSC of the different substructures.

Partitioning of Molecules. The partitioning algorithm is based on the exclusion of atoms and bonds from the molecular graph and the subsequent identification of the remaining connected components. These will be referred to as Multi State

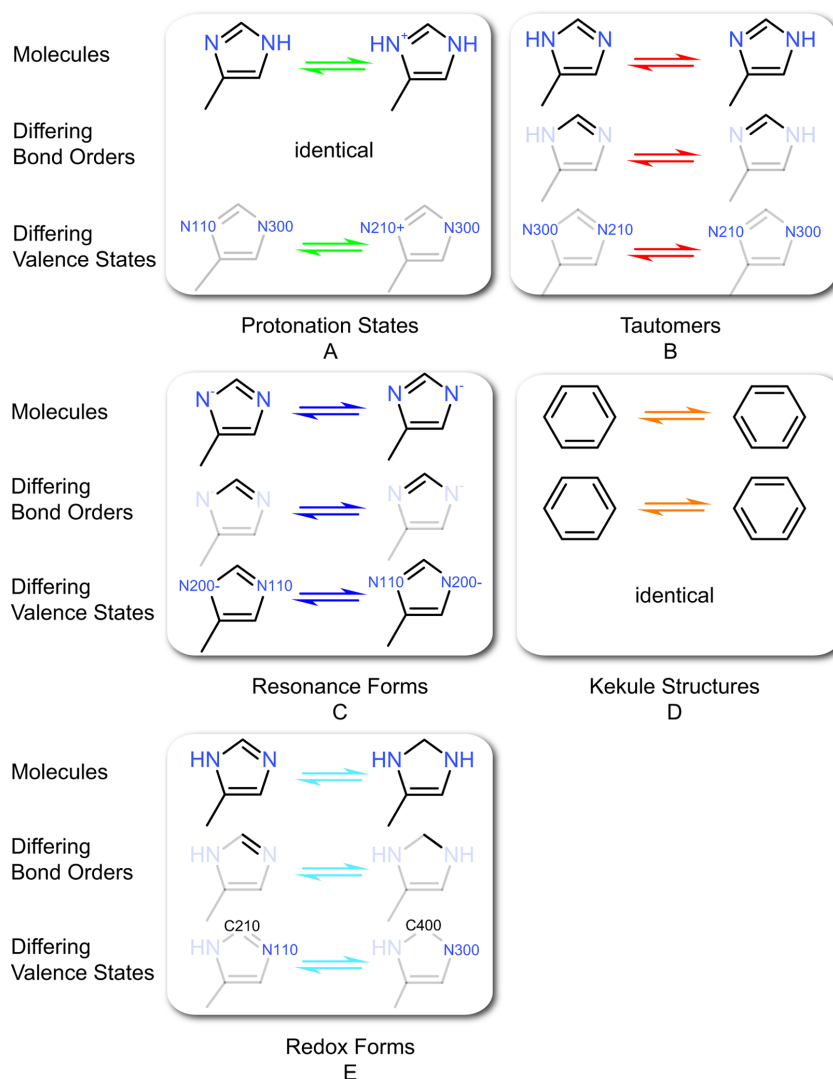


Figure 3. Differences between protonation states (A), tautomers (B), resonance forms (C), Kekule structures (D), and redox forms (E).

Table 1. Substitution Types for Valence States Including the Affected Properties^a

type	double bonds	# bonds	charge	examples	
				donor	acceptor
protonation	0	±	±	O200	O100-
resonance	±	0	±	O100-	O010
tautomer	±	±	0	O200	O010

^aChanged properties are marked with a ± and unchanged properties with 0. The pairs of valence states on the right side of the table represent common substitutions for oxygen atoms.

Table 2. Relations between Valence Bond Structures on the Basis of Valence State Substitution

relation	substitution type	condition
kekule	none	
ionization	protonation	$\Delta(D \rightarrow A) \neq \Delta(A \rightarrow D)$
protonation	protonation	$\Delta(D \rightarrow A) = \Delta(A \rightarrow D)$
mesomer	resonance	$\Delta(D \rightarrow A) = \Delta(A \rightarrow D)$
tautomer	tautomer	$\Delta(D \rightarrow A) = \Delta(A \rightarrow D)$
redox	resonance	$\Delta(D \rightarrow A) \neq \Delta(A \rightarrow D)$
	tautomer	$\Delta(D \rightarrow A) \neq \Delta(A \rightarrow D)$

Partitions (MSP) in the following discussion. The generic partitioning scheme only involves the exclusion of sp³-hybridized carbon atoms (corresponds to valence state

C400). There are only two particular cases in which atoms with valence state C400 are included in MSPs: first, if the atom is bound to an atom with valence state C210, which in turn has

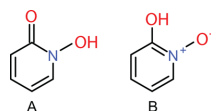


Figure 4. Example for the combination of valence state substitutions. The relation between the pyridone form (A) and the pyridine form (B) cannot be described by one of the basic types from Table 2.

at least one neighbor with the element nitrogen, oxygen or sulfur, and second, if the atom is part of a ring and is the only atom with valence state C400 in this ring. Bonds are excluded if one of the connected atoms is excluded.

The MSPs resulting from the generic partitioning scheme are usually large, and it is often possible to further reduce their size. This is achieved by removing bonds within the MSPs with the goal to effectively split them into smaller substructures. The exclusion of a bond is only valid if its bond order in the current structure is identical in all relevant VSCs. Since the final bond orders are not known at this point, the decision that a bond will keep its current type must be in accordance with the subsequent scoring procedure. This means that VSCs with a different bond order would be rejected in the following steps in any case.

The heuristic partitioning scheme builds on the results from the generic scheme and uses a set of rules to identify additional bonds for exclusion. These rules are based on the classification of each MSP into conjugated rings, conjugated chains, and functional groups. Rings are considered conjugated if all of their atoms are part of the respective MSP. Conjugated chains consist only of carbon atoms which have a multiple bond and are bound only to other carbon atoms. The remaining connected components represent functional groups. In a first step, bonds connecting functional groups with conjugated rings or conjugated chains are investigated. A bond is excluded if it is a single bond and the atom from the functional groups does not fulfill one of the following two criteria: (1) It has a valence state of type N300. (2) It has a valence state of type O200 or S200 and only one non-hydrogen bond. In these cases, a change in bond order is not unlikely, as is shown for two examples in Figure 6.

Since conjugated chains consist of only carbon atoms, they are merely bridges between the other two types of

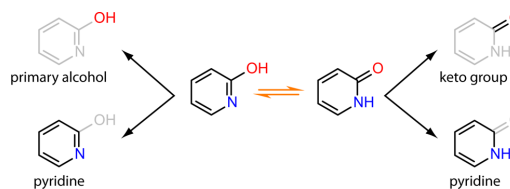


Figure 6. Two examples for which functional groups and conjugated rings have to be treated as a union to avoid missing VSCs.

substructures. Therefore, if a conjugated chain has only one bond to another structure (ring or functional group), this bond can be safely excluded. This is also done if the chain has multiple bonds which were previously excluded by the functional group rule. The complete partitioning of the NAD⁺ molecule is shown as an example in Figure 7.

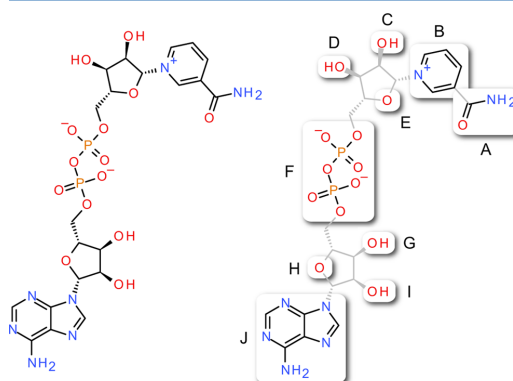


Figure 7. Partitioning of NAD⁺ into functional groups and conjugated rings. The amide group and the pyridine ring have been separated, whereas the amino group remains connected to the purine.

Selection of Valence States. The selection of valence states is based on the substitution types introduced above (see Table 1). Each substitution corresponds to a pair of valence states which are known in advance and can be retrieved starting

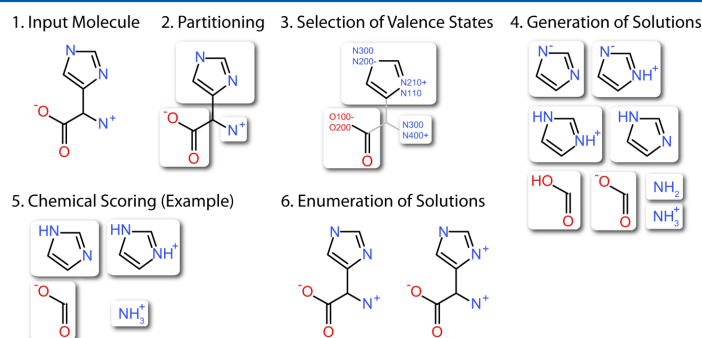


Figure 5. Overview of the generation of protonation states for an input molecule (1). In a first step, the molecule is partitioned into substructures (2) which are handled separately. In the next step, alternative valence states are selected (3). Afterward, valid VSCs are generated (4). These are scored (5), and only the best solutions for each zone are retained. The final list of valence bond structures results from the combination of all remaining VSCs.

from any valence state. A list of alternatives for an atom can thus be easily obtained by consecutively and uniquely adding the respective members of the pairs for each of the relevant substitution types. In order to select an alternative assignment, the compatibility with the atom's topology must be ensured. This means that the number of bonds of the valence state must be larger than or equal to the atom's number of non-hydrogen bonds. Otherwise, the assignment would correspond to the removal of non-hydrogen bonds. Although this may be interesting with respect to transformations such as ring-chain tautomerism, it will not be further considered here.

For the sake of generality, the **generic** selection procedure includes all possible valence states for each atom in a MSP. This usually results in potentially many more alternatives than are actually needed. The **heuristic** selection scheme aims at reducing this number by explicitly excluding valence states for particular atoms. The problem at this point is similar to the one discussed in the previous section. The final VSCs are not yet known, and the decisions must be in accordance with the subsequent scoring procedure in order to avoid missing VSCs.

The exclusion of particular valence states in the **heuristic** selection scheme is based solely on an analysis of the atom's environment. For atoms in functional groups, this includes their direct neighbors from the same functional group. These are transformed into a SMILES-like identifier which reflects the valence bond structure of the input molecule. This identifier is looked up in a list of predefined structures. If the identifier is present, information concerning the exclusion of particular substitution types is retrieved. In this way, groups that already have a preferred representation in the initial valence bond structure need not be modified. The information provided from the patterns is described in Figure 8.

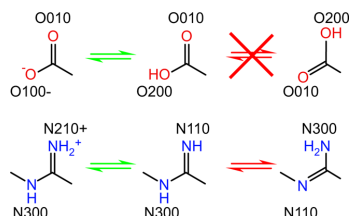


Figure 8. Selection of valence states for carboxylic acid and amidine groups. Due to the group's symmetry, different tautomers of carboxylic acids are not considered.

For the generation of tautomers, carboxylic acids are irrelevant. Due to the symmetry of the group the transfer of the hydrogen from one oxygen to the other would only result in a different rotamer. In this case both oxygen atoms are excluded from tautomer substitution. With respect to protonation, both the charged and the neutral form need to be included. This means that both oxygens are not excluded from protonation substitution. The same procedure is applied to atoms in conjugated rings with the ring constituting the atom's environment. If the identifier is not included, the **generic** scheme is used to identify alternative states for the atom.

Generation of Valid VSCs. Prior to the generation of VSCs, each MSP is analyzed to ensure that the generation of additional states is at all possible. MSPs can be ignored if no atom with alternative valence states could be found. For

tautomers and mesomers, i.e. if new bond order distributions are to be generated, MSPs can also be omitted if only either donors or acceptors are present. In this case, no substitution of valence states is possible (see Figure 9 for examples). Changing the number of donors and acceptors corresponds to changing the oxidation state of the molecule, which is not desired in most contexts.

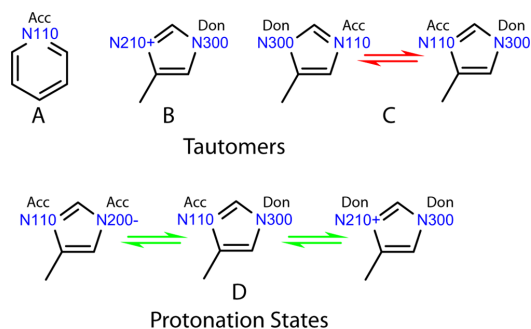


Figure 9. Criteria for the generation of additional states. The generation of tautomers requires at least one tautomer acceptor and one tautomer donor in a zone. Pyridine (A) has only a single tautomer acceptor, and the imidazolium ion (B) only has one tautomer donor. No tautomers can be generated in such cases. Imidazole (C) contains a tautomer acceptor as well as a tautomer donor and can tautomerize. Protonation states (D) can also be generated if a molecule only contains either protonation-type donors or acceptors.

The algorithm for the generation of valid VSCs is based on a backtracking procedure with pruning. The atoms of the MSP are processed in a specific order which is established prior to the actual assignment procedure. The algorithm starts with terminal atoms, i.e. atoms with only one bond in the MSP, followed by internal atoms with at least one terminal neighbor. The remaining atoms are processed last. The order of the atoms inside the three classes is arbitrary and does not affect the result. As a combinatorial problem, the procedure can be represented by a tree, where each node corresponds to the assignment of a valence state to an atom. Inner nodes thus represent partial VSCs while the tree's leaves correspond to complete VSCs. For each node, the chemical validity of the corresponding VSC is verified. In most cases, this can be performed without actually generating bond orders for the bonds of the MSP. The checks are based on the compatibility between valence states of different atoms with respect to the expected bond types as well as their oxidation states: (1) For atoms with only one bond in the MSP, the assignment of a valence state is equivalent to the assignment of a bond order to the corresponding bond. The compatibility with the atom's neighbors can be easily checked by ensuring that the count of this particular bond type is not exceeded. This check is always performed when an atom with terminal neighbors is encountered. (2) When reaching a leaf, the valence states with an uneven number of multiple bonds are counted. If this number is uneven, no valid bond order distribution exists, and the VSC can be further ignored. (3) The number of donors in the initial valence bond structures is counted in order to retain the molecule's oxidation state. VSCs differing in the number of donors compared to the initial valence bond structures can be discarded. Note that since information about being a donor or

acceptor is also stored in the valence states, VSCs not fulfilling this boundary condition can be easily identified. (4) Eventually, for each VSC passing all previous checks, a recursive bond localization routine is used which assigns bond orders to all bonds in the MSP. If this routine is successful, the solution represents a valid valence bond structure and is stored.

Scoring of VSCs. Scores for each VSC are calculated under consideration of the bond order distribution generated in the previous step. The scoring procedure is mainly based on the recognition of predefined structural fragments contained within particular substructures, i.e., conjugated rings and functional groups, of the molecule. The final score of the VSC (S_{VSC}) is calculated as the sum of the individual scores obtained for each of these substructures (see eq 1). Please note that due to changes in bond orders and valence states, the scores have to be recalculated for each VSC.

$$S_{\text{VSC}} = \sum S_{\text{ring}} + \sum S_{\text{group}} \quad (1)$$

$$S_{\text{ring}} = \sum \text{cycle} + \sum S_{\text{sub}} \quad (2)$$

$$S_{\text{group}} = \sum S_{\text{subgroup}} \quad (3)$$

The structural fragments in the substructures are identified using canonical SMILES-like identifiers. These are generated on the basis of the bond types and valence states of the respective VSC. The predefined data are stored in multiple databases which can be queried with the identifiers in order to retrieve the score associated with a fragment.

In case of conjugated rings, the score S_{ring} comprises two types of contributions, one from the ring itself, S_{cycle} , and one from its substituents, S_{sub} (see eq 2). The reference point for S_{cycle} is the isolated aromatic system without exocyclic double bonds, e.g., pyrrole for a five-membered ring with one nitrogen atom. In case there are multiple structures fulfilling this requirement, e.g., the 1H and 2H tautomers of 1,2,3-triazole, one is arbitrarily selected. The score of the reference system is set to an arbitrary value of 100. If a ring with an identical heavy atom connectivity does contain a structural deviation from the reference, e.g., an sp^3 hybridized carbon atom, the associated fragment has an individual score. This can be higher or lower depending on the stability assigned to this particular arrangement. The substructures representing ring substituents comprise the ring atom, the exocyclic atom, and the exocyclic atom's direct neighbors. The associated scores have fixed values and are independent from the concrete ring system they are connected to. Again, one particular representation of the substituent, the one with an exocyclic single bond and without charges, receives an arbitrary reference score of 100. Functional groups are first treated as a whole; i.e., an identifier for the complete group is generated. If the pattern was present, the associated score is directly set as the score of the substructure. Otherwise, the group is partitioned into smaller pieces which serve as starting points for further queries. In this case, the score for the group is composed of the scores of the smaller fragments (see eq 3). The reference system for a subgroup is preferably neutral and corresponds to the most stable tautomeric form where possible.

If no predefined data are available in any of the three cases, a generic score is calculated according to eq 4:

$$S_{\text{generic}} = \max(0, 80 - \sum P) \quad (4)$$

This is done by subtracting various penalties (P) which are summarized together with the respective conditions in Table 3.

Table 3. Classification and Conditions for the Penalties used during the Calculation of Generic Scores

substructure	type	penalty	condition
ring	aromaticity	20	nonaromatic ring (Hueckel's rule)
ring	charge	20	single charge in ring
ring	charge	80	multiple charges in ring and substituents
ring	stability	80	three consecutive donors ^a in the ring
substituent	bond order	20	substituent has exocyclic double bond
substituent	charge	20	single charge in substituent
substituent	charge	80	multiple charges in substituent
group	charge	80	multiple positive charges in group

^aDonors are atoms with the following valence states: O200, N300, S200.

Since S_{generic} is used only as a fallback, the respective maximal score is deliberately set lower than that of the reference system. If the sum of the penalties (P) exceeds 80, the score of the substructure is set to zero.

The relative differences between the scores of rings, substituents, and functional groups have been derived from multiple pairs of tautomers and ionization states for which the major form was known from either experiments or theoretical calculations.²⁴ The databases currently contain 252 entries in total (113 in cycles, 121 in subgroups, 18 in substituents). Examples for ring and functional groups patterns are shown in Figure 10.

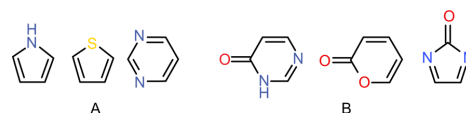


Figure 10. Examples for ring and functional groups patterns. (A) A reference score of 100 is assigned to isolated aromatic rings without exocyclic double bonds. (B) The score for rings with exocyclic double bonds comprises one contribution from the ring and another from the carbonyl substituent.

VSC MODEL APPLICATIONS

In the following applications, we will consider resonance forms, prototropic tautomers, and protonation states as instances of the same molecule, whereas oxidation forms are interpreted as distinct chemical species. The method is, however, not restricted to this assumption in general and can be easily modified so that different types of valence bond structures are perceived as identical.

Canonicalization. The generation of a canonical representation is the first workflow in which our method is applied. Canonical representations are mainly used to determine whether two valence bond structures represent the same molecule. In this context, it does not matter if the result corresponds to the most stable form or even a chemically reasonable one.

The workflow starts with the partitioning of the molecule into MSPs and the selection of alternative valence states as described above. The atoms of each MSP are then sorted in a

canonical way using a variant of the Morgan extended sums algorithm.²⁵ The backtracking algorithm in the generation step processes the atoms in this exact order until the first valid VSC has been found. This VSC serves as the canonical form of the respective substructure. Since no additional scoring is needed, the canonical VSCs of each substructure can be directly combined to yield the canonical representation for the complete molecule.

For the canonicalization to work correctly, the results must be identical for each possible valence bond structure of the molecule provided as input. This can only be achieved if the substructures generated in the partitioning step and the lists of valence states identified in the selection step are both identical in each case. The heuristic algorithms for partitioning and selection are therefore inappropriate, and the generic variants are applied. Since only a single valid VSC must be generated in the end, the size of the substructures and the number of alternative states are of less importance for the resulting compute time. Nevertheless, in order to further accelerate the process, all charged valence states are transformed into their neutral states where possible (considering the number of hydrogens) using the protonation-type substitution. Consequently, only tautomer-type and resonance-type substitutions need to be considered in the next steps.

The canonicalization procedure applied to the atoms of each substructure differs only in one aspect from the CANON algorithm used for the generation of USMILES.²⁶ In the CANON algorithm, the atomic invariants correspond to the atom's valence state in combination with the number of attached hydrogens. This means that the initial ranks of atoms can normally be deduced by comparing valence states and hydrogens. In case of a yet unknown valence bond structure, the final valence state of an atom is, however, not defined. Instead, a list of valence states is used to describe the topology of each atom and provide the initial ranks. Furthermore, the number of non-hydrogen bonds serves as a replacement for the number of hydrogens.

Normalization. The aim of normalization is the generation of a canonical valence bond structure which additionally adheres to common conventions for the representation of molecules. This task seems, at least at first glance, quite similar to the previously described canonicalization. The main difference results from the necessity of a scoring step in order to determine the best suited choice for the molecule. This implies that multiple VSCs have to be generated and compared with each other. Here, we have chosen a neutral form as normalized representation, meaning that all atoms are neutralized when possible (considering bound hydrogens). The only exception to this rule is functional groups which are represented in a zwitterionic form by convention, e.g., nitro groups and *n*-oxides. The method is, however, not restricted to this preference and can be easily modified so that, for instance, the preferred ionization state is generated.

Again, the workflow starts with the partitioning of the molecule into substructures and the selection of alternative valence states. Due to the enumeration of VSCs in the later steps, the size of the substructures and the number of states are relevant factors. Therefore, the heuristic strategies for both partitioning and state selection are used. In contrast to canonicalization, the initial substructures and alternative valence states do not have to be identical for each starting structure. The additional scoring step ensures that the results are consistent.

In the next two steps, valid VSCs are generated and scores are assigned as explained in the sections above. For each substructure, only those solutions with the highest score are retained. If there is only one VSC left for a substructure, it can be directly assigned, and no further steps are necessary. Otherwise, a canonical solution has to be picked from the VSCs with the highest score. This is done using the canonicalization method described in the previous section. However, since this method only works correctly in case of identical MSPs and lists of valence states, a preprocessing step is required. The respective MSP is repartitioned by exclusion of bonds having the same bond type in all VSCs. Additionally, all valence states which could not be found in one of the remaining VSCs are removed from the lists of alternatives. This eventually creates the necessary conditions for the canonicalization procedure.

Generation. The last application of our method is the generation of a set of reasonable tautomers and protonation states of a molecule. The resulting molecules can be used as input for methods that rely on the positions of hydrogen atoms such as docking. They can also serve as a starting point for the determination of the energetically most stable form of a molecule under consideration of the molecule's local environment, e.g., bound to a protein. The inclusion of multiple resonance structures, although possible with our method, is not considered useful in this context.

The initial steps of the workflow are identical to those described for normalization. But instead of canonically selecting one of the remaining VSCs of each zone, the combinations are enumerated in order to generate a set of molecules. One major difference from the previously presented approaches is the possibility of generating duplicates due to molecular symmetry. This is avoided by removing VSCs from each zone that would lead to identical valence bond structures in the resulting molecules. For this purpose, automorphism classes for atoms are calculated using the Morgan algorithm, which is also used for the canonicalization. In combination with the respective valence state of an atom in a VSC, these classes can be used to generate a string representation of each VSC in a zone, which are used to identify and remove duplicates.

For molecules containing more than one ionizable group, it is usually not desirable to enumerate all combinations of VSCs from the respective zones. To avoid chemically unreasonable species with a high number of charges, the maximum number of charges in the complete molecule is restricted by three simple rules: (1) The number of charged groups must be smaller than four, (2) the number of pairs of oppositely charged groups is smaller than two, and (3) the maximum number of positive charges in a ring system is restricted to one.

RESULTS AND DISCUSSION

The three applications presented in the previous sections are the basis for the evaluation of our method in terms of consistency, quality, and performance. Throughout these studies, the following commonly used public data sets served as input: (1) ZINC clean leads^{27,28} (ZINC-CL), (2) LigandExpo component dictionary^{29,30} (LEXPO-CD), (3) Drugbank^{31,32} (DRUGBANK), and (4) ChEMBL.^{33,34} All calculations and runtime measurements were performed on a PC with an Intel Core i5-3570 CPU (3.40 GHz) and 8 GB of main memory.

Consistency. Independence from the initial valence bond structure of a molecule is a fundamental requirement of the presented method and has been thoroughly investigated for

Table 4. Runtimes for the Three Workflows with Different Data Sets

	ZINC-CL	LEXPO-CD	DRUGBANK	ChEMBL
# total molecules	5735035	17310	6583	1318187
runtime canonicalization [ms/cmpd]	0.28	0.41	0.45	0.71
runtime normalization [ms/cmpd]	0.31	0.50	0.6	0.73
runtime generation [ms/cmpd]	0.45	0.75	0.75	1.37

each of the three applications. Consistency can be verified by a simple and straightforward procedure. The starting point is a set containing different representations of the same molecule, e.g., as different molecule entries in a file. After applying the respective workflow to each representation, the resulting molecules are converted to USMILES for comparison. If the method is consistent, all resulting USMILES are identical. In case of enumeration, lists of USMILES must be compared.

The best way to ensure consistency would be to test all possible valence bond structures of the molecule with the procedure described above. This is, however, not feasible in many cases due to the prohibitively large number of resulting molecular states. We therefore decided to reduce the set by exclusion of protonation and ionization states (see Table 2 for our definition), since the main complexity of the task results from valence bond structures with different bond order distributions.

The input structures needed for the assessment of our method's consistency were generated using a workflow corresponding to the one described for the canonicalization of molecules. But instead of selecting a canonical form, we generated valid VSCs without any scoring step and enumerated all possible combinations. Identical results could be achieved for all three workflows, canonicalization, normalization, and generation, with all four data sets mentioned above.

Runtimes. Table 4 lists the runtimes for the three workflows with the above-mentioned data sets. The results for canonicalization and normalization are comparable in both cases, whereas the time needed for the generation of a set of states is higher. This is not surprising since the workflow involves the enumeration of multiple molecule states and the built-in elimination of duplicates based on automorphism classes. In all cases, an average runtime lower than 1.5 ms per molecule is measured, thus showing that our method is suitable for processing large data sets. The similarity of results for canonicalization and normalization are most probably a consequence of the normalization procedures used during the curation of the used data sets. As has been explained above, the runtimes for normalization are highly dependent on the input form of the molecule, and the process is accelerated by reasonable initial representations.

Normalization. The main purpose of normalization is to transform different input forms of the same molecule into an identical and at the same time chemically reasonable representation. We have already shown that our normalization workflow is consistent for the four data sets used in this study. Here, we will focus on the second aspect. We believe that the best way to investigate if results are chemically reasonable is to compare the resulting valence bond structure with those found in frequently used and curated public data sets.

The procedure applied for this purpose is again based on the comparison of USMILES. Directly using the input molecule and the normalized version is, unfortunately, not suitable in many cases. As has been explained above, a canonicalization step at the end of the workflow is used to arbitrarily select one

of multiple equally acceptable solutions. This makes the comparison to a reference structure, which has most likely been normalized by a different procedure, pointless. We therefore decided to enumerate all combinations of VSCs with the highest score and to check if the input structure is contained within the obtained set (best). A negative result does, however, not necessarily mean that our method generated an unreasonable result. The representation in the data set could simply correspond to a VSC which received a lower score based on our scoring scheme. For that reason, we additionally enumerated all VSCs with a score of at least 75% of the best score and also searched in this extended set (extended). The results of this validation procedure are summarized in Table 5.

Table 5. Classification of the Input Structures from Three Data Sets into Mutually Exclusive Categories for the Generation of Tautomers

	LEXPO-CD	DRUGBANK	ChEMBL
# total molecules	17310	6583	1318187
# molecules (best)	16837	6431	1252408
# molecules (extended)	364	118	52491
# molecules (not found)	135	48	11433

The differences encountered during the process can be subdivided into two classes. First, there are input structures which are not found in the best set, but in the extended set. These correspond in many cases to keto and enol tautomers of aromatic heterocycles, which are ranked differently by our scoring method (see 138 in Figure 11). Second, there are input structures which are not present in either of both sets. After visual inspection, we think that the results generated by our method are in general at least equally acceptable and in some cases even better than the representations found in the data set. The latter especially applies to charged structures for which a reasonable neutral form can be formulated (see 3MC in Figure 11). The normalized molecules generated by our method are provided as Supporting Information for all entries of LEXPO-CD and DRUGBANK which were not included in either of the two sets.

Finding the input structure in a set of equally scored alternatives is, however, only one aspect of the method's performance. Additionally, one has to make sure that the success is not simply based on the enumeration of an unreasonably large number of representations. For that reason, the sizes of the respective sets are also an important performance indicator and are shown in Table 6.

The average number of generated states is considerably lower than the result of an exhaustive enumeration. Only for a small percentage of molecules (less than 0.5%) does the number of equivalent structures actually exceed a size of five. This is in all cases caused by the combination of states from independent zones, e.g., molecules having multiple imidazole rings.

Generation. The aim of our generation workflow is to generate a set of chemically reasonable protomers of a molecule

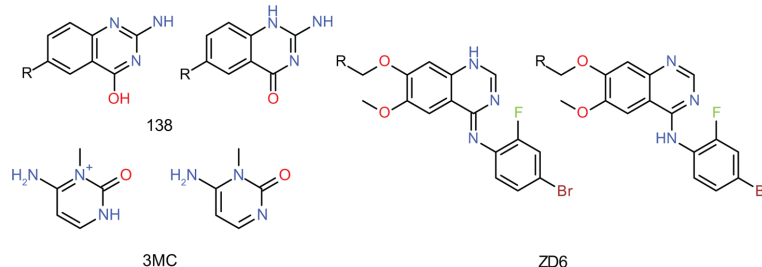


Figure 11. Examples of differences between the normalized forms generated by our method (right side) and those found in the Ligand Expo data set (left side).

Table 6. Number of Molecules with More than One and More than Five Tautomers in the Best Set^a

	ZINC-CL	LEXPO-CD	DRUGBANK	ChEMBL
# total molecules	5735035	17310	6583	1318187
# tautomers >1	207207	1483	520	93430
# tautomers >5	671	5	5	4699
# average	2.27	2.21	2.37	8.3

^aThe provided average refers only to cases with more than one tautomer.

for typical cheminformatics applications, e.g., docking calculations. Considering this context, the resulting set should only contain states which are realistically expected to be stable in a protein–ligand complex. In order to assess the quality of our results, we used ZINC-CL as a reference set since it was generated for the exact same scenario. The procedure is identical to the one described for the evaluation of the normalization workflow. The input structure is searched in two sets, one containing the states with the highest score (best) and one containing states with a score of at least 75% of the highest score (extended). The results of the procedure are summarized in Table 7.

Table 7. Classification of the Input Structures from the ZINC-CL Data Set into Mutually Exclusive Categories for the Generation of Protomers

	ZINC-CL
# total molecules	5735035
# molecules (best)	4764463
# molecules (not best)	914921
# molecules (not found)	55651

As has already been discussed above, one important parameter for the evaluation of the method's performance certainly is the number of generated states. The results for all four data sets are summarized in Table 8.

CONCLUSION

The simple fact that the same molecule can be represented by different valence bond structures constitutes a complex challenge for cheminformatics applications. It complicates the determination of molecular identity and makes the results of cheminformatics calculations prone to inconsistencies. Furthermore, it imposes the task of selecting the best suited structure or structures for the respective context of application. The identification, description, and consistent handling of these

Table 8. Number of Molecules with More than One and More than Five Protomers in the Best Set^a

	ZINC-CL	LEXPO-CD	DRUGBANK	ChEMBL
# total molecules	5735035	17310	6583	1318187
# protomers >1	1007976	2221	770	159663
# protomers >5	9240	183	78	13231
# average	2.54	3.14	3.20	4.40

^aThe provided average refers only to cases with more than one protomer.

different molecular representations is thus a fundamental requirement in the field of cheminformatics.

To cope with these problems, we have introduced a formalism which describes different valence bond structures of a molecule on the basis of the recently published NAOMI model. Using this description, we developed a general method for their fast and consistent enumeration and presented three exemplary applications. In our validation, we have shown that the devised methodology can be successfully applied to relevant tasks in cheminformatics in a consistent manner. We have also demonstrated the low runtime of our approach which makes it suitable for processing large data sets.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information includes the normalized structures of all entries from the Ligand Expo Component Dictionary and Drugbank whose input form was not included in the results of our method. These are provided as separate SMILES files. This material is available free of charge via the Internet at <http://pubs.acs.org/>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: rarey@zbh.uni-hamburg.de.

Author Contributions

[†]Equal contribution.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank one of the reviewers for bringing the existence of the freely available source code of the method developed by Sayle and Delaney³⁵ to their attention.

REFERENCES

- (1) Sayle, R. So you think you understand tautomerism? *J. Comput.-Aided Mol. Des.* **2010**, *24*, 485–496.
- (2) Warr, W. A. Tautomerism in chemical information management systems. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 497–520.
- (3) Sitzmann, M.; Ihlenfeldt, W.-D.; Nicklaus, M. Tautomerism in large databases. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 521–551.
- (4) Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. InChI - the worldwide chemical structure identifier standard. *J. Cheminform.* **2013**, *5*, 7.
- (5) InChI version 1, software version 1.04 (2011)-Technical Manual. http://www.inchi-trust.org/fileadmin/user_upload/software/inchi-v1.04/INChI_TechMan.pdf (last accessed Dec 06, 2013).
- (6) Milletti, F.; Storch, L.; Sforna, G.; Cruciani, G. New and Original pK_a Prediction Method Using Grid Molecular Interaction Fields. *J. Chem. Inf. Model.* **2007**, *47*, 2172–2181.
- (7) Shelley, J.; Chollet, A.; Frye, L.; Greenwood, J.; Timlin, M.; Uchimaya, M. Epik: a software program for pK_a prediction and protonation state generation for drug-like molecules. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 681–691.
- (8) Martin, Y. Let's not forget tautomers. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 693–704.
- (9) Clark, T. Tautomers and reference 3D-structures: the orphans of in silico drug design. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 605–611.
- (10) Cramer, R. Tautomers and topomers: challenging the uncertainties of direct physicochemical modeling. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 617–620.
- (11) Greenwood, J.; Calkins, D.; Sullivan, A.; Shelley, J. Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 591–604.
- (12) Gilson, M.; Gilson, H.; Potter, M. Fast assignment of accurate partial atomic charges: an electronegativity equalization method that accounts for alternate resonance forms. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1982–1997.
- (13) Oellien, F.; Cramer, J.; Beyer, C.; Ihlenfeldt, W.-D.; Selzer, P. M. The Impact of Tautomer Forms on Pharmacophore-Based Virtual Screening. *J. Chem. Inf. Model.* **2006**, *46*, 2342–2354.
- (14) ten Brink, T.; Exner, T. Influence of Protonation, Tautomeric, and Stereoisomeric States on Protein-Ligand Docking Results. *J. Chem. Inf. Model.* **2009**, *49*, 1535–1546.
- (15) Kalliokoski, T.; Salo, H.; Lahtela-Kakkonen, M.; Poso, A. The effect of ligand-based tautomer and protomer prediction on structure-based virtual screening. *J. Chem. Inf. Model.* **2009**, *49*, 2742–2748.
- (16) Kenny, P.; Sadowski, J. In *Chemoinformatics in Drug Discovery*; Oprea, T., Ed.; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2005; pp 271–285.
- (17) Milletti, F.; Storch, L.; Sforna, G.; Cross, S.; Cruciani, G. Tautomer enumeration and stability prediction for virtual screening on large chemical databases. *J. Chem. Inf. Model.* **2009**, *49*, 68–75.
- (18) Kochev, N. T.; Paskaleva, V. H.; Jeliazkova, N. Ambit-Tautomer: An Open Source Tool for Tautomer Generation. *Mol. Inf.* **2013**, *32*, 481–504.
- (19) Haranczyk, M.; Gutowski, M. Quantum Mechanical Energy-Based Screening of Combinatorially Generated Library of Tautomers. TauTGen: A Tautomer Generator Program. *J. Chem. Inf. Model.* **2007**, *47*, 686–694.
- (20) Thalheim, T.; Vollmer, A.; Ebert, R.-U.; Kühne, R.; Schüürmann, G. Tautomer Identification and Tautomer Structure Generation Based on the InChI Code. *J. Chem. Inf. Model.* **2010**, *50*, 1223–1232.
- (21) Will, T.; Hutter, M. C.; Jauch, J.; Helms, V. Batch tautomer generation with MolTPC. *J. Comput. Chem.* **2013**, *34*, 2485–2492.
- (22) Sayle, R.; Delany, J. In *Innovative Computational Applications: the Interface of Library Design, Bioinformatics, Structure Based Drug Design and Virtual Screening*; IIRG publishers: San Francisco, CA, 1999. http://www.daylight.com/meetings/emug99/Delany/taut_html/sld001.htm (accessed Jan 30, 2014).
- (23) Urbaczek, S.; Kolodzik, A.; Fischer, J. R.; Lippert, T.; Heuser, S.; Groth, I.; Schulz-Gasch, T.; Rarey, M. NAOMI - On the almost trivial task of reading molecules from different file formats. *J. Chem. Inf. Model.* **2011**, *51*, 3199–3207.
- (24) Raczynska, E.; Kosinska, W.; Osmialowski, B.; Gawinecki, R. Tautomeric Equilibria in Relation to Pi-Electron Delocalization. *Chem. Rev.* **2005**, *105*, 3561–3612.
- (25) Morgan, H. L. The generation of a unique machine description for chemical structures - a technique developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–113.
- (26) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97–101.
- (27) Irwin, J.; Shoichet, B. ZINC - A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
- (28) ZINC Database - Version 12. <https://zinc.docking.org/> (Clean Leads Reference as SMILES downloaded Dec 03, 2013).
- (29) Feng, Z.; Chen, L.; Maddula, H.; Akcan, O.; Oughtred, R.; Berman, H.; Westbrook, J. Ligand Depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics* **2004**, *20*, 2153–2155.
- (30) Ligand Expo; RCSB PDB. <http://ligand-expo.rcsb.org/> (chemical component dictionary as SMILES (OpenEye with stereo) downloaded Jul 10, 2012).
- (31) Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A.; Wishart, D. DrugBank 3.0: a comprehensive resource for "Omics" research on drugs. *Nucleic Acids Res.* **2011**, *39*, D1035–D1041.
- (32) DrugBank 3.0. <http://www.drugbank.ca/> (all drugs as SDF downloaded Dec 03, 2013).
- (33) Bellis, L. J.; et al. Collation and data-mining of literature bioactivity data for drug discovery. *Biochem. Soc. Trans.* **2011**, *39*, 1365–1370.
- (34) ChEMBLdb - Version 17. <https://www.ebi.ac.uk/> (ChEMBLdb including an SDF file downloaded Dec 03, 2013).
- (35) Source code of the tautomer generation method by Sayle and Delany. <http://www.daylight.com/meetings/emug99/Delany/tautomers/> (accessed Jan 30, 2014).

Reading PDB: Perception of Molecules from 3D Atomic Coordinates

[D5] **S. Urbaczek**, A. Kolodzik, S. Heuser, I. Groth and M. Rarey. Reading PDB: Perception of Molecules from 3D Atomic Coordinates. *Journal of Chemical Information and Modeling*, 53(1):76-87, 2013.

<http://pubs.acs.org/articlesonrequest/AOR-RKZNmYgpW6rVhdXcQ8Gr>

Reproduced with permission from S. Urbaczek, A. Kolodzik, S. Heuser, I. Groth and M. Rarey. Reading PDB: Perception of Molecules from 3D Atomic Coordinates. *Journal of Chemical Information and Modeling*, 53(1):76-87, 2013. Copyright 2013 American Chemical Society.

Reading PDB: Perception of Molecules from 3D Atomic Coordinates

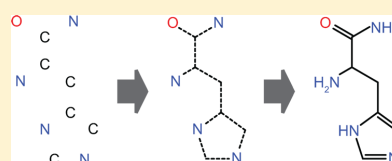
Sascha Urbaczek,[†] Adrian Kolodzik,^{†,||} Inken Groth,[‡] Stefan Heuser,^{‡,§} and Matthias Rarey^{*,†}

[†]Center for Bioinformatics (ZBH), University of Hamburg, Bundesstrasse 43, 20146 Hamburg, Germany

[‡]Research Active Ingredients, Beiersdorf AG, Tropelwitzstrasse 15, 22529 Hamburg, Germany

S Supporting Information

ABSTRACT: The analysis of small molecule crystal structures is a common way to gather valuable information for drug development. The necessary structural data is usually provided in specific file formats containing only element identities and three-dimensional atomic coordinates as reliable chemical information. Consequently, the automated perception of molecular structures from atomic coordinates has become a standard task in cheminformatics. The molecules generated by such methods must be both chemically valid and reasonable to provide a reliable basis for subsequent calculations. This can be a difficult task since the provided coordinates may deviate from ideal molecular geometries due to experimental uncertainties or low resolution. Additionally, the quality of the input data often differs significantly thus making it difficult to distinguish between actual structural features and mere geometric distortions. We present a method for the generation of molecular structures from atomic coordinates based on the recently published NAOMI model. By making use of this consistent chemical description, our method is able to generate reliable results even with input data of low quality. Molecules from 363 Protein Data Bank (PDB) entries could be perceived with a success rate of 98%, a result which could not be achieved with previously described methods. The robustness of our approach has been assessed by processing all small molecules from the PDB and comparing them to reference structures. The complete data set can be processed in less than 3 min, thus showing that our approach is suitable for large scale applications.



INTRODUCTION

Crystal structures of protein–ligand complexes provide valuable insights into the interactions between proteins and small molecules. The statistical analysis of these structures has become an important tool in many different areas of research in the life sciences. Because of the large number of entries, the Protein Data Bank (PDB)¹ is the most important resource for experimentally determined structures of protein–ligand complexes. The structural data in the PDB is made available via different chemical file formats (PDB, mmCIF, PDBML/XML),² of which the PDB format³ is the most common. PDB files contain element identities, three-dimensional coordinates, and connectivities for all atoms. However, unlike many other chemical file formats, this format does neither provide information about bond orders, formal charges, and aromaticity nor any kind of atom typing. Many cheminformatics methods and tools, however, depend on those and similar properties. Hence, when PDB files are supported as input, those properties have to be derived from the information provided by the file format. Although many current software packages include functionality to perceive molecular structures from three-dimensional coordinates, only a small number of these approaches has been published.^{4–10}

The initial steps of all methods are similar to a certain extent. First, covalent bonds between atoms are identified by either using distance criteria or by simply relying on the connectivity data (CONNECT entries) provided by the PDB format. In some approaches, this step is followed by a valence check during which spurious bonds arising from distorted geometries are

removed. Subsequently, possible hybridizations for atoms are determined by analyzing bond lengths and bond angles. In the next step bond orders and atom types are assigned. Depending on the way these assignments are handled, the methods can be divided into two classes. Approaches from the first class determine bond orders independently of hybridization states, either by using the bond lengths directly or by matching of functional group patterns. This is often followed by an additional step during which inconsistencies in the assignments are handled. In methods from the second class, bond orders are derived directly from previously determined hybridization states using different bond localization routines.

We present a new method for the perception of molecular structures from three-dimensional atomic coordinates, which is based on the recently published NAOMI model.¹¹ Using its robust chemical description, the molecules are constructed in a hierarchical scoring approach. The first steps are based on the local geometry of each individual atom, whereas later steps include larger parts of the atom's environment to generate a correct chemical representation. This bottom-up approach has the advantage that it does not rely on definite assignments at early stages, for example, by assigning bond orders by torsion angles, or by matching of functional group patterns. In contrast to previously published methods, the final solution is selected from a list of potential candidate structures which are ranked using both confidence values for the atoms' geometry and

Received: July 30, 2012

Published: November 25, 2012

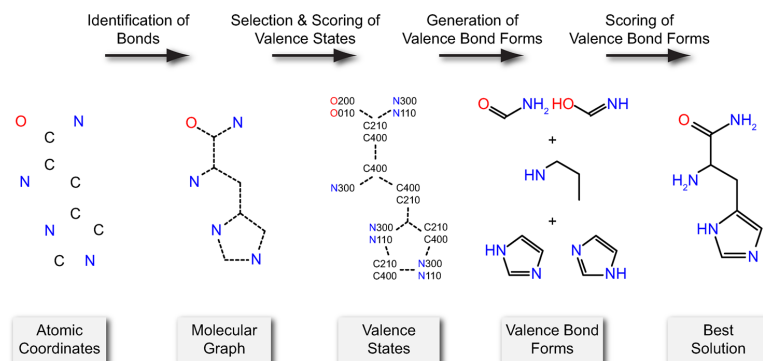


Figure 1. Schematic view of the workflow for the generation of molecules from three-dimensional coordinates.

chemical knowledge. This combination is the key to circumvent the shortcomings of other approaches, which either put too much focus on the provided coordinates or simply ignore them by using pattern-matching. The method's robustness and reliability are validated in different procedures by comparing reference molecules to the generated molecular structures. Furthermore, benchmark studies show its suitability for large scale applications.

METHODOLOGY

Overview. The aim of the presented method is the generation of both chemically valid and reasonable molecular structures from element identities and three-dimensional atomic coordinates. A molecule is considered chemically valid if a valence bond structure (Lewis structure) can be found, in which the valences of the atoms' elements are not violated. Not every possible valid valence bond form, however, provides a reasonable description of the molecule. On the one hand, geometric features, for example, interatomic distances and planar groups, must be reflected in the assigned bond orders. On the other hand, common standards for the representation of particular functional groups and resonance forms should be met. The last point is especially important since resonance forms and, depending on the quality of the provided coordinates, even tautomeric forms can not be deduced from geometry alone. For this purpose, we make use of the NAOMI model,¹¹ which has been successfully applied to the consistent conversion of chemical file formats. In this model, atoms are represented by three chemical descriptors, namely element, valence state, and atom type, which are assigned in three consecutive steps. Valence states represent valid bond order distributions for atoms in valence bond structures. They are defined by an element identity, the number of associated single, double, and triple bonds and a formal charge (e.g., N400+ for quaternary nitrogen atoms). As will be explained below, valence states can be used to generate valence bond forms if the atoms' connectivities are known. Atom types are derived from valence states and are thus independent of the input file format.

The perception of molecular structures from atomic coordinates is performed in four steps (see Figure 1). At first, covalent bonds are identified on the basis of interatomic distances. The second step comprises identification of possible valence states for each atom and scoring according to the atom's local environment. In the third step valence bond forms of the molecule are generated by enumerating valid

combinations of valence states and their associated bond orders. These combinations are scored in the final step to determine the most appropriate valence bond representation of the molecule. The strategy adopted in our method is based on the opinion that the best possible compatibility between the perceived molecules and the provided coordinates should be sought. We believe, that the best way to do so is to build the molecular structure based on the atom's local geometries and use chemical knowledge only when either inconsistencies are encountered or ambiguities need to be resolved.

Identification of Bonds. To determine if a covalent bond exists between two atoms, the distance criterion originally proposed by Meng⁴ is applied. A bond is created if

$$\delta_{\text{bond}} = r_{ij} - (R_i + R_j) < 0.4 \text{ \AA} \quad (1)$$

where r_{ij} is the distance between the atoms i and j and R_i and R_j denote the covalent radii¹² of the atoms' corresponding elements. The high tolerance value of 0.4 Å in eq 1 ensures that no potential covalent bond is missed during the identification process. The softness of the criterion can, however, lead to an erroneous bond perception in case of distorted geometries. The resulting superfluous bonds give rise to two different types of chemical errors, which can readily occur at the same time. On the one hand, the atom's number of bonds may exceed the maximum valence of its associated element. On the other hand, distorted geometries may lead to the formation of incorrect cyclic structures (usually rings of size three or four). To deal with these errors, the bond perception is performed in several consecutive steps: (1) identification of bonds between non-hydrogen atoms, (2) valence check for all atoms and removal of superfluous bonds, (3) perception of the molecule's rings, (4) length check for all ring bonds and removal of superfluous bonds, and (5) identification of hydrogen bonds.

After the perception of all non-hydrogen bonds, each atom is checked for violations of its valence. This is done by comparing the number of identified bonds to the number of allowed bonds for its element. If a violation is encountered, long bonds ($\delta_{\text{bond}} > 0.1 \text{ \AA}$) are removed in order of their lengths until either the valence is restored or all long bonds are eliminated. In case of short non-hydrogen bonds ($r_{ij} < 0.5 \cdot (R_i + R_j)$), the coordinates are considered incorrect and the molecule cannot be constructed. After the ring perception each ring is checked for long bonds ($\delta_{\text{bond}} > 0.1 \text{ \AA}$). If such a ring is encountered, its longest bond is removed and the molecule's rings are

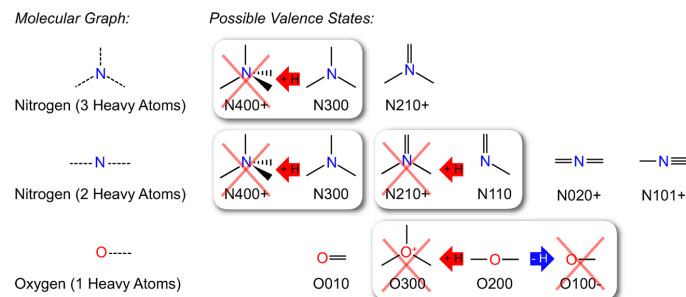


Figure 2. Examples for the selection of valence states. The crossed-out states are not selected since they can be deduced from the corresponding neutral states shown in the same box.

Table 1. Most Common Candidate Valence States for Typical Elements in Organic Molecules^a

element	valence states						
hydrogen	H100						
carbon	C400	C210	C101	C020			
oxygen	O200	O010	O110+	O300+	O001+		
nitrogen	N300	N110	N210+	N400+	N020+	N101+	N001
phosphorous	P310	P300	P400+				
sulfur	S220	S210	S300+	S200	S110+	S010	S001+

^aValence states are represented as element symbol followed by the number of single, double, and triple bonds and the formal charge.

recalculated. This process is repeated until all long bonds in rings are eliminated. In contrast to non-hydrogen atoms, hydrogens are only allowed to have one bond and only the closest heavy atom needs to be identified. The hydrogen bond is created if the resulting bond is not short ($\Rightarrow r_{ij} \geq 0.5 \cdot (R_i + R_j)$) and the heavy atom's valence is not violated. Otherwise the hydrogen atom is discarded.

Selection of Valence States. In the next step, suitable valence states are selected from a list of allowed states for the respective element for each atom. Since bond orders have not been assigned at this point and formal charges are usually not provided, the number of bonds from the previous step is the only criterion for this selection. Valence states are selected in two cases. First, if the valence state and the atom have an identical number of bonds. Second, if the atom's bond count is smaller, but the missing bonds can be saturated by hydrogens. Charged valence states are only considered if no corresponding neutral state exists or a formal charge has been specified for the atom. Examples for this identification procedure are shown in Figure 2.

In many cases, this results in an ambiguous assignment since multiple valence states may be compatible with a particular number of bonds. To deal with this ambiguity, all selected valence states are scored to determine the most appropriate choice as explained below. This score reflects the state's compatibility with the atom's local environment, which is characterized by the spatial distribution of the atom's neighbors and their respective element identities. The use of a predefined list of valid valence states is an important aspect of ensuring a molecule's chemical validity. Atoms with an invalid number of bonds can be easily identified by the fact that no candidate valence state has been found. This evidently applies to all cases, where the number of bonds exceeds the maximum allowed number for the respective element. In addition to that, it is also possible to identify atoms with unusual bond counts in case of higher row elements such as sulfur or phosphorus. A typical

example would be a phosphate group that is missing two of its terminal oxygen atoms thus leaving the central phosphorus with only two covalent bonds. This constellation is rather unlikely in organic molecules and simply saturating the atom's valences by addition of hydrogens seems questionable in a chemical sense. If no candidate valence state for an atom can be found, the molecule is considered incorrect and cannot be constructed. The most common candidate valence states for typical elements in organic molecules are shown in Table 1.

Evaluation of Geometrical Parameters. The compatibility of valence states is mainly assessed on the basis of the atom's local geometry. For that purpose, several geometrical parameters g are evaluated and used to derive scores $G_p(g)$ for different chemical properties p , for example, bond orders. These scores are calculated according to the following scheme. For each property, a minimum and a maximum value are defined, which correspond to the scores of 0.0 and 1.0, respectively (see Table 2). Between the minimum and the maximum values a linear function is used.

The absolute value of the scalar **triple product** π of the normalized bond vectors connecting an atom and its neighbors

Table 2. Parameters for the Calculation of Scores $G_p(g)$ for Different Properties p ^a

property	parameter	minimum (0.0)	maximum (1.0)
$G_{\text{planar}}(\pi)$	π	≥ 0.6	≤ 0.15
$G_{\text{linear}}(\alpha)$	$\alpha[^\circ]$	≤ 150	≥ 170
$G_{\text{sp}^2}(\alpha)$	$\alpha[^\circ]$	≤ 114	≥ 118
$G_{\text{single}}(\delta)$	$\delta[\text{\AA}]$	≤ -0.1	≥ -0.04
$G_{\text{double}}(\delta)$	$\delta[\text{\AA}]$	≥ -0.04	≤ -0.1
$G_{\text{triple}}(\delta)$	$\delta[\text{\AA}]$	≥ -0.15	≤ -0.25
$G_{\text{planar}}(\tau)$	$\tau[^\circ]$	≥ 40	≤ 10

^aBetween the minimum and the maximum values a linear function is used.

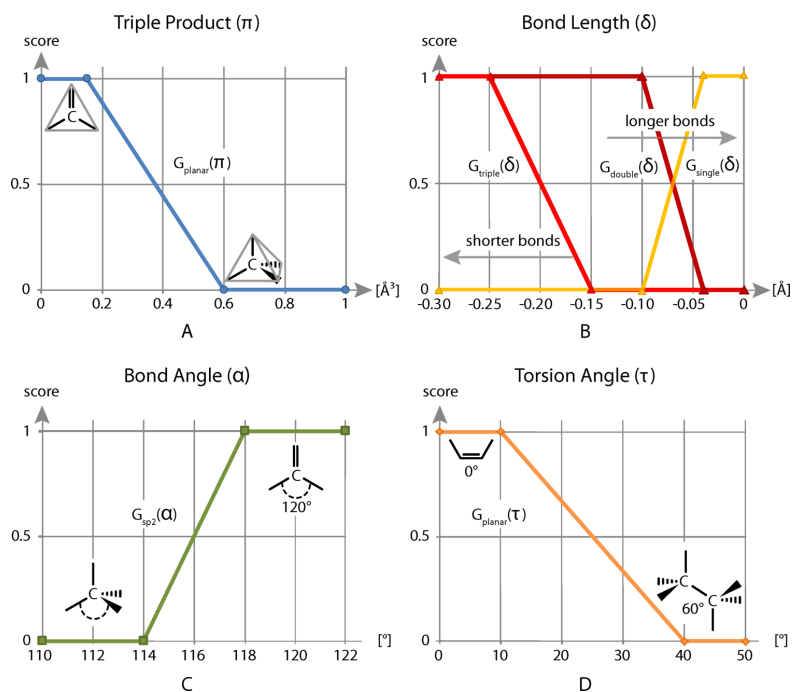


Figure 3. A: Score for the planarity of an atom using the triple product. B: Score for bond orders using the bond length. C: Score for an sp^2 hybridization on the basis of an atom's bond angle. D: Score for planarity using the largest torsion angle.

is a direct measure for its planarity ($G_{\text{planar}}(\pi)$) and can thus be used to distinguish sp^2 from sp^3 hybridizations. A triple product smaller than 0.15 indicates planarity, whereas a value larger than 0.6 (the triple product of an ideal tetrahedron is approximately 0.7) indicates the opposite. **Bond angles** α are used to determine the hybridization of an atom. They are especially important for the identification of linear geometries ($G_{\text{linear}}(\alpha)$), for example, in the presence of triple bonds. Because of the large difference to the bond angles of other hybridizations, sp hybridization can be easily distinguished. The smaller difference between the angles associated with sp^2 and sp^3 hybridizations makes the distinction between these cases rather difficult ($G_{\text{sp}^2}(\alpha)$). Scores for particular bond orders ($G_{\text{single}}(\delta)$, $G_{\text{double}}(\delta)$, $G_{\text{triple}}(\delta)$) are determined using the **bond length** δ which is calculated as described in eq 1. In the case of double bonds, the largest **torsion angle** τ at the respective bond is taken into consideration ($G_{\text{planar}}(\tau)$). Torsion angles can be used to check if the atoms surrounding the bond partners are coplanar, which is a precondition for double bonds. By taking torsion angles into account, invalid double bond assignments due to shortened interatomic distances can be avoided. Single bonds joining an aromatic ring with either an alkyl substituent or another aromatic ring are typical examples for this case. Although the bond length might be shortened, the torsion angle often clearly contradicts the double bond order. The torsion bond probability $G_{\text{double}}(\delta, \tau)$ is the product of $G_{\text{double}}(\delta)$ and $G_{\text{planar}}(\tau)$. For atoms in rings, **torsion angles** τ can be used to determine the planarity of the ring. In this case, only bonds in the same ring are included during the calculation of the largest torsion angle.

Probabilities of Hybridization States. The scores $G_p(g)$ are the basis for the calculation of probabilities for different hybridization states P_{hyb} . Since the number and kind of parameters used strongly depends on the atom's topology, each case is discussed separately.

For atoms with one bond, the bond length is the only available geometrical parameter.

$$P_{\text{sp}} = G_{\text{triple}}(\delta) \quad (2)$$

$$P_{\text{sp}^2} = G_{\text{double}}(\delta) - G_{\text{triple}}(\delta) \quad (3)$$

$$P_{\text{sp}^3} = G_{\text{single}}(\delta) \quad (4)$$

In this case the probabilities for the hybridization states correspond to the scores for the respective bond orders as described in eqs 2–4. Since sp hybridization is always associated with a linear geometry, the number of bonds at the atom's neighbor is checked. If the neighbor has more than two bonds this condition cannot be fulfilled and the value of P_{sp} is added to P_{sp^2} and then set to 0.0.

For atoms with two bonds, one bond angle and two bond lengths are available. The score for the presence of a double bond at the atom A_{double} is calculated as the sum of the torsion bond scores $G_{\text{double}}(\delta, \tau)$ of the atom's bonds, whereas its maximum value is limited to 1.0.

$$A_{\text{double}} = \min(1.0, \sum G_{\text{double}}(\delta, \tau)) \quad (5)$$

The sum in eq 5 is used to account for the limitations of valence bond structures. In delocalized systems, for example, aromatic rings, bonds can have lengths between the expected

values of single and double bonds. In this case, the score for the presence of a double bond might be underestimated if only the larger of both values is considered. Because of the geometric restraints in small rings, we then distinguish two cases. For atoms in an acyclic environment or in large rings (at least eight atoms), the following probability scheme is used:

$$P_{sp} = 2/3 \cdot (G_{\text{linear}}(\alpha) + 0.5 \cdot A_{\text{double}}) \quad (6)$$

$$P_{sp^2} = \begin{cases} 1/2 \cdot (1.0 - P_{sp}) & \text{if } P_{sp} > 0.0 \\ 2/3 \cdot (A_{\text{double}} + 0.5 \cdot G_{sp^2}(\alpha)) & \text{else} \end{cases} \quad (7)$$

$$P_{sp^3} = 1.0 - (P_{sp^2} + P_{sp}) \quad (8)$$

Since only the sp hybridization is compatible with a linear geometry, the bond angle has a higher weighting factor in the calculation of the associated probability in eq 6. For the probability of an sp² hybridization in eq 7 it is considered less reliable due to the small difference to the ideal value of the sp³ hybridization. If the atom is part of a small ring (less than eight atoms), ring torsion angles can be used as an additional parameter to assess the planarity of the respective ring. Furthermore, a linear arrangement is extremely unlikely in these cases, so that only sp² and sp³ hybridizations need to be considered. The probabilities and scores are adapted in the following way:

$$P_{sp^2} = 2/5 \cdot (A_{\text{double}} + A_{\text{planar}} + 0.5 \cdot G_{sp^2}(\alpha)) \quad (9)$$

$$P_{sp^3} = 1.0 - P_{sp^2} \quad (10)$$

Since bond angles in rings with a size smaller than six are strongly influenced by the strain of the cyclic arrangement, they are not a reliable measure for the atom's hybridization. In this case, the score is automatically set to 0.5 to indicate that no decision can be made. The planarity score A_{planar} in eq 9 for an atom is the minimum of the $G_{\text{planar}}(\tau)$ (see Figure 3D) scores of each bond.

For atoms with three bonds, three bond angles, three bond lengths, and one triple product can be calculated. Since sp hybridization is not possible in this case, a decision between sp² and a sp³ hybridization has to be made. For the calculation of the atom's angle score $A_{sp^2}(\alpha)$ the mean bond angle $\bar{\alpha}$ is used.

$$P_{sp^2} = 1/6 \cdot (3 \cdot G_{\text{planar}}(\bar{\alpha}) + 2 \cdot A_{\text{double}} + A_{sp^2}(\alpha)) \quad (11)$$

$$P_{sp^3} = 1.0 - P_{sp^2} \quad (12)$$

Again, the geometrical parameters are not considered equally reliable which is reflected in the different weighting factors in eq 11. The scoring of valence states for atoms with four or more bonds is solely based on scores for bond orders, and no probabilities for hybridizations need to be calculated for these cases.

Scoring of Valence States. The probabilities P_{hyb} from the previous step are used to calculate integer-based scores for all selected valence states of each atom. This score reflects the compatibility between the atom's local environment and the respective valence state and is used to identify the best suited state for an individual atom. Additionally, the absolute value of the score also provides a measure of confidence, which can be used to compare possible valence state assignments for different atoms. The scoring procedure makes use of the fact that valence states are not compatible with all hybridization states.

In case of compatibility, the score S_{VS} is calculated using the probability P_{hyb} according to the following scheme:

$$S_{\text{VS}} = \begin{cases} 1 & \text{if } P_{\text{hyb}} < 0.6 \\ \lfloor P_{\text{hyb}} \cdot c + 0.5 \rfloor & \text{else} \end{cases} \quad (13)$$

The confidence factor c in eq 13 determines the maximum value of the score and depends on the topology of the respective atom (see Table 3). The values are based on the

Table 3. Confidence Values for Different Topologies

topology	confidence c
1 bond	2.0
2 bonds(acyclic)	3.0
2 bonds(cyclic)	4.0
≥ 3 bonds	5.0

number of geometrical parameters available for the calculation of the probabilities P_{hyb} . A single bond length, for example, is not well suited to reliably distinguish between hybridizations, since even small geometrical distortions may cause the bond order perception to fail. This lack of reliability is reflected in a small confidence factor of 2.0 for atoms with one bond. The integer-based scheme ensures that only those valence states which are clearly favored by the atom's local geometry receive scores larger than one. This prevents the elimination of valence states based on small geometrical differences.

If the compatibility between the selected valence states and their associated hybridization states is mutually exclusive, the scoring procedure is straightforward. Because of the limitation of valence bond structures, this is, however, not always the case (see Figure 4 for examples). On the one hand, there are atoms which are represented by the same valence state but have different hybridizations, such as nitrogens in amines and amides. These cases are handled by assigning the largest score obtained for all compatible hybridizations to the respective valence state. On the other hand, some atoms are not sufficiently represented by a single valence state such as oxygens in a carboxylate group. In this case both compatible valence states receive identical scores. Examples for the scoring procedure are shown in Figure 5.

The calculation of scores for atoms with four or more bonds can in most cases be avoided due to the fact that there is only one suitable valence state. If this is not the case, the multiple bond score A_{double} introduced in eq 5 is used in place of P_{hyb} to calculate the score for all selected valence states. This is always sufficient to distinguish between the alternatives.

In some cases, it is beneficial to remove valence states from the list of candidates if their associated hybridization is not compatible with the atom's local geometry ($P_{\text{hyb}} = 0.0$). These valence states will not be considered during the generation of valence state forms, which in turn reduces the complexity of the next steps. Since distorted geometries could easily lead to the premature exclusion of relevant valence states, this is only done in two rather unambiguous cases. First, if the corresponding valence state is only compatible with an sp hybridization and second if the atom has three bonds.

Distorted geometries can also result in incorrect scores which will eventually lead to undesired valence bond structures. This is especially true if atoms with only one bond are involved since the resulting assignment cannot be corrected by the valence states of the surrounding atoms. To avoid these errors, valence

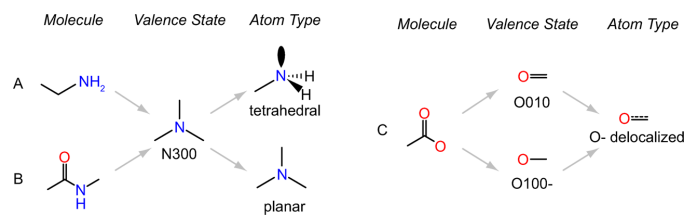


Figure 4. Limitations of valence bond structures: (a) Nitrogens in amides and amines have the same valence states but different geometries. (b) Oxygens in carboxylates have different valence states but have the same bond length.

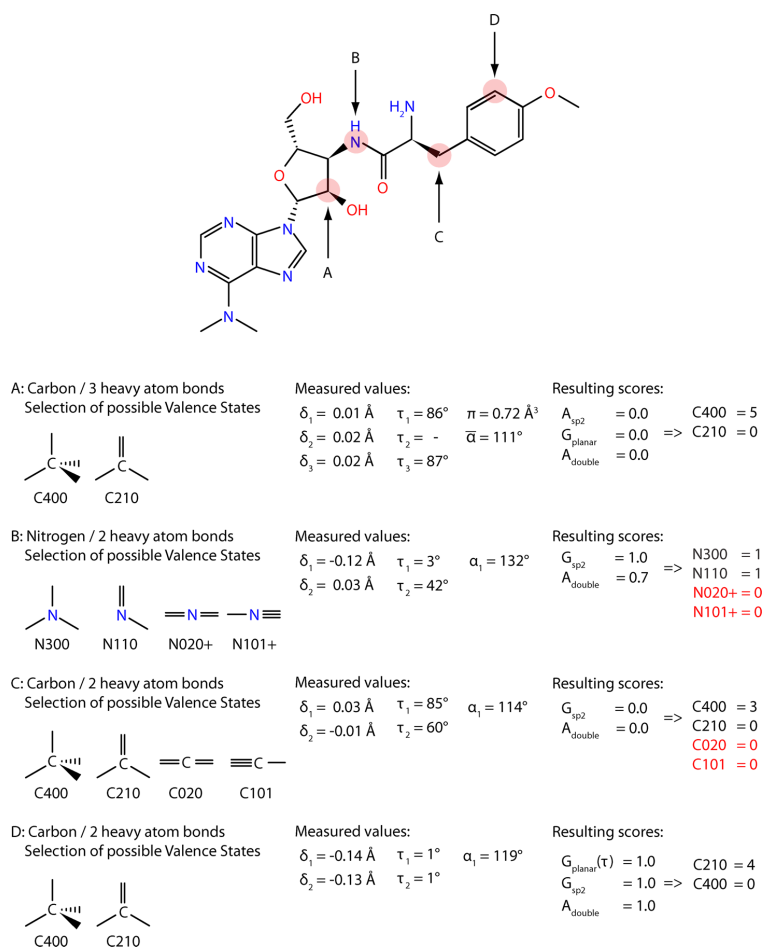


Figure 5. Examples for the valence state scoring procedure. Relevant geometrical parameters are triple products, bond lengths, bond angles and torsion angles. If bond angles and bond lengths do not indicate a linear geometry, valence states associated with a linear geometry are excluded (marked in red). For the atoms A, C, and D the geometrical parameters clearly support one of the valence states. In the case of atom B, there is no clear preference and two valence states are equally probable.

state scores of atoms that are part of specific substructures are increased by +2 (see Figure 6). These resulting scores are, however, not high enough to change the assignment in case of a perfect geometric compatibility.

The purpose of the described procedure is to provide reliable scores which can be used to identify the best valence state

representation of the molecule. Due to the differing quality of available input data, this must also apply if the provided coordinates are of poor quality. As mentioned above, the scores are not only used to find the best choice for an individual atom but also to compare assignments between atoms. This means that valence states with higher scores have a stronger influence

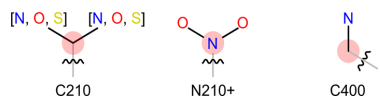


Figure 6. Additional scores of +2 are assigned to valence states for atoms (marked with red spheres) in specific substructures. The number of bonds at the atoms corresponds to the number of bonds identified during the bond perception.

on the resulting valence state form. The score does not only depend on the number of available parameters at the respective atom, but also on their consistency. This is assessed by the individual evaluation of the geometrical parameters during the calculation of the probabilities P_{hyb} . A value of 1.0 is only possible if all geometrical parameters are consistent, which in turn results in low scores for atoms with inconsistent local geometries.

Generation of Valence Bond Forms. In the next step, chemically valid valence bond representations of the molecule are generated by assigning valence states to all atoms and bond orders to all bonds. A combination of valence states is valid if a bond order distribution can be generated which is in accordance with the valence states of the atoms. The score of such a combination is calculated as the sum of the scores of the valence states from the previous step. The best combination can be identified by enumerating all valid combinations with a maximum score. For the enumeration a branch and bound algorithm with a depth-first search strategy is used. Prior to the enumeration, the list of valence states for each atom is checked for cases where only one valence state is remaining. This state is assigned directly and the orders of the adjacent bonds are adapted accordingly. Afterward, the molecule is partitioned into zones containing atoms connected by bonds with unassigned bond orders. The individual processing of each zone further decreases the number of possible combinations. If a single best scored combination for a zone exists, it is selected. Otherwise, combinations with equal scores are ranked using additional geometrical and chemical criteria as described below.

Scoring of Valence Bond Forms. Each combination of valence states generated in the last step is a valid valence bond form (in the sense that no valences are violated) and is also compatible with the local geometry of the atoms. This does, however, not necessarily imply that each form provides a reasonable description of the molecule. On the one hand, discrepancies between the assigned bond orders and the actual bond lengths might exist, which could not be resolved during the atom based valence state scoring procedure. On the other hand, the combination might contain unusual representations of functional groups or conjugated systems, which could not be excluded using geometrical parameters alone. Hence, an additional scoring scheme, which makes explicit use of the assigned bond orders, is applied to distinguish reasonable from undesired valence bond forms. In contrast to the previous steps, where geometrical parameters had a high priority, this step focuses mainly on chemical aspects.

Prior to the scoring procedure, valence states and bond orders are assigned if they are identical in all generated valence bond forms. Afterward, the molecule is again partitioned into zones containing atoms connected by bonds with unassigned bond orders. Then, substructures (see Figure 7) including at least one of the unassigned atoms are identified in each of the remaining valence bond forms. These substructures correspond to preferred representations of functional groups and for each

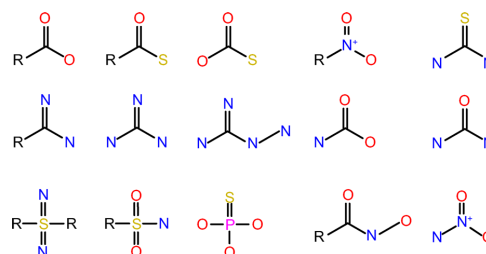


Figure 7. Substructures representing favored representations of particular functional groups in valence bond structures. The R represents both carbon and hydrogen.

match a score of +1 is assigned to the respective form. If an unassigned atom is part of a ring with a size smaller than eight, Hueckel's rule is applied to assess its aromaticity. Valence bond forms receive a score of +1 for each ring, where the rule is fulfilled. It must be stressed that our approach does not favor particular functional groups or aromatic rings in general but only if the geometrical parameters were not sufficient to resolve the structure.

If a bond with an unassigned bond order is part of a substructure or ring which has been scored in the previous step, no further scoring is performed. Otherwise, $G_{\text{order}}(\delta)$ from Table 2 is used to determine if the current bond order is compatible with the calculated bond length. In the case of double bonds, $G_{\text{double}}(\delta, \tau)$ is used. If the respective value exceeds a threshold of 0.7 a score of +1 is assigned to the valence bond form. Hence, solutions in which bond lengths do not correspond to the assigned bond orders receive lower scores. If the bond is also part of a ring with less than eight atoms, $G_{\text{planar}}(\tau)$ is used as an additional parameter to assess the bond's planarity. A score of +1 is assigned if either $G_{\text{planar}}(\tau)$ is smaller than 0.3 (planar geometry) for a double bond or $G_{\text{planar}}(\tau)$ is larger than 0.7 (ring is not planar) for a single bond.

Again, only the solutions with the largest scores are kept. If there are still multiple solutions left, they are considered equivalent and a canonization scheme is used to choose a unique form for each zone. Since a detailed explanation of the canonization algorithm extends the scope of this publication, only a brief description of the general idea will be given. The atoms of the respective zone are ordered in a procedure similar to the CANON algorithm¹³ used for the generation of USMILES. The zone is then processed atom by atom according to this newly generated order. At each step the respective solutions are sorted by the valence states (using ids as sorting criterion) of the particular atom and all solutions with lower ranks are eliminated. This process is repeated until only one solution remains. Obviously, it is also possible to omit the canonization and use the solutions for each zone to enumerate all equivalent valence bond forms of the molecule.

RESULTS AND DISCUSSION

Validation with Curated Structures. In a first validation procedure we tested if our method was able to generate the expected valence bond structures for small molecules from different PDB entries. The success was verified by comparison of the resulting molecules to manually curated reference structures provided as USMILES.¹³ Small molecules were extracted from PDB entries used in the studies of Hendlich⁶ and Labute.⁷ Because of its importance in the field of

cheminformatics, we also included the ligands from the PDB entries of the Astex Diverse Set.¹⁴ The complete validation set consists of 563 molecules from 363 PDB entries. Both PDB entries and SMILES files for the respective compounds are provided as Supporting Information. Table 4 lists the PDB

Table 4. PDB IDs and Component Names of All Molecules for Which Our Method Did Not Generate the Expected Structure

Labute ⁷	Hendlich ⁶	Astex ¹⁴
2R04 (W71)	1MIO (CFM)	1G9V (HEM)
3FX2 (FMN)	1PMP (OLA)	1Q4G (HEM)
5TLN (BAN)	6RSA (UVC)	
8XIA (XLS)		

entries and the component names of the ligands for which our method failed to generate the expected structure. Five of these examples are shown together with the reference structures taken from the respective publications in Figure 8.

The dihydro-oxazol ring of ligand W71 from 2R04 (see Figure 8A) is perceived as oxazol. Because of a short bond of C4A to the nitrogen atom and the planarity of the five-membered ring, the valence state C210 (which is compatible with an sp^2 hybridization) receives a higher score. This eventually leads to a structure including an aromatic ring. One of the hydroxy groups of the flavin mononucleotide ligand FMN from 3FX2 (see Figure 8B) is interpreted as a carbonyl group. In this case the valence state C210 is favored due to the trigonal planar geometry of C2'. The same also applies to the α carbon CA2 in BAN from 5TLN (see Figure 8E). The carbonyl group of the molecule XLS from 8XIA (see Figure 8C) is interpreted as a hydroxy group because of the tetrahedral geometry at C2. The double bonds of the olefinic moiety of OLA (1PMP) (see Figure 8D) and of one of the vinylic groups

in HEM (1G9V, 1Q4G) are perceived as single bonds due to the bond lengths and associated bond angles.

Our method was able to generate the correct structure in 98% of the cases. All observed differences were caused by strong deviations from the expected molecular geometries. The valence bond forms generated by our method are, however, equally reasonable in a chemical sense and also in agreement with the supplied atomic coordinates. Only in the case of BAN the generated structure does not correspond to the tautomeric form which would be expected for the isolated compound with respect to the hydroxamic acid group. The molecular geometry may, however, be influenced by the interactions with a metal atom in the protein–ligand complex. The PDB entry CFM contains an Fe–Mo–S cluster, for which our method does not produce a valence bond form but isolated atoms. Since valence bond forms are not well suited to describe metal clusters, we do not consider this a perception error, but think it should be mentioned at this point. The same is true for the vanadate in 6RSA in which no bonds between the oxygens and the vanadium atom are formed. The uridine molecule, however, is perceived correctly.

Comparison with Other Methods. To compare our results with those of other existing methods, we used the tools I-interpret,¹⁵ fconv,¹⁶ and MOE¹⁷ to generate molecules for the above-mentioned 363 PDB entries. This was done by first converting the entries from PDB to SDF (since fconv does not support sdf as output format, mol2 was chosen in this case) and then using the converted file as input for the comparison to the reference structures. The results are summarized in Table 5. Since our method will be part of the NAOMI converter, it is referred to as NAOMI in the table. The comparison to the reference structures was done using the NAOMI framework. Since all files (PDB input, SDF/MOL2 files from different tools, SMILES for comparison) are supplied as Supporting

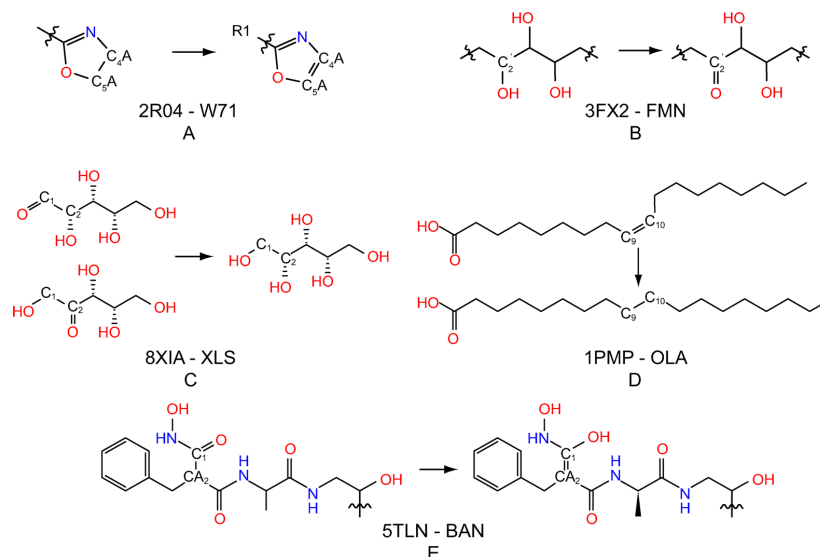


Figure 8. Five of the nine molecules for which our method did not generate the expected structure. The expected results are shown on the left side of the arrow, the results of our method on the right. The names from the PDB files are listed for all atoms for which incorrect valence states were identified.

Table 5. Results of the Generation of Molecules from the 363 PDB Entries Using Different Tools^a

PDB-Code	NAOMI	fconv	I-interpret	MOE	PDB-Code	NAOMI	fconv	I-interpret	MOE	PDB-Code	NAOMI	fconv	I-interpret	MOE
1AAQ (PSI)					1ABE (ARA)					1ABE (ARB)				
1A1A (PMP)					1A1C (PMP)					1AMR (PMP)				
1APT (CHAIN I)					1APU (CHAIN I)					1AQB (RTL)				
1BAP (ARA)					1BAP (ARB)					1CHM (CMS)				
1CPS (CPM)					1CRP (GDP)					1CRR (GDP)				
1DHF (FOL)					1DR1 (HBI)					1EFG (GDP)				
1ERB (ETR)					1FEM (REA)					1G9V (HEM)				
1G1A (GSP)					1GKC (NFH)					1GM8 (SOX)				
1GPK (HUP)					1HFC (PLH)					1HQ2 (PH2)				
1HSN (BME)			X		1HVR (XK2)					1HVV (D16)				
1IA1 (NDP)					1IA1 (TQ3)					1IG3 (V1B)				
1J31 (CP6)					1J31 (NDP)					1JLA (TNK)				
1KE5 (LS1)					1L91 (BME)					1LH7 (NBE)				
1MBI (HEM)					1MBI (IMD)		X			1MEH (MOA)				
1MIO (CFM)					1MLN (HEM)		X			1MMV (H4B)				
1MNC (PLH)					1MYJ (HEM)					1N1M (A3M)				
1N21 (PAF)					1NNB (DAN)					1OF6 (DTY)				
1OPB (RET)					1OWE (675)					1P62 (GEO)				
1P2Y (HEM)					1PBF (FAD)					1PHE (HEM)				
1PHF (HEM)					1PMN (984)					1PMP (OLA)				
1POE (GEL)					1Q41 (IXM)					1Q4G (HEM)				
1R58 (AO5)					1R90 (FLP)					1R90 (HEM)				
1RBP (RTL)					1S3V (TQD)					1SQ5 (PAU)				
1SQN (NDR)					1T9B (1CS)					1T9B (FAD)				
1TRP (2GP)					1TT1 (KAI)					1U4D (DBQ)				
1UNL (RRC)					1UOU (CMU)					1V48 (HA1)				
1XOZ (CIA)					1Y6B (AAX)					1YST (U10)				
1YV3 (BIT)					1YWR (LJ9)					2BR1 (PPF)				
2DRI (RIP)					2FKE (FK5)					2R04 (W71)				
2RNT (GPG)					2SNS (THP)					2TDD (UFP)				
2XIM (XYL)					2XIS (XYL)					2YPI (PGA)				
3CSC (ACO)					3DFR (MTX)					3DFR (NDP)				
3DRC (MTX)					3ER3 (0EL)					3FX2 (FMN)				
3POR (C8E)					4AT1 (ATP)					4CP4 (CAM)				
4DFR (MTX)					4FAB (FDS)					4FBP (AMP)				
4GR1 (RGS)					5CPP (HEM)					5LDH (LNC)				
5XIA (XYL)					6ABP (ARA)					6ABP (ARB)				
6RSA (UVC)					7CAT (NDP)					7HVP (CHAIN I)				
7TLN (INC)					8CAT (NDP)					8XIA (XLS)				

^aThe colors represent the quality of the resulting structures. Green cells: Correct structure. Yellow cells: Suboptimal structure. Red cells: Structure substantially differing from reference. X: No structure generated.

Information, the comparison can be carried out using other tools with the same functionality. The differences between the generated molecules and the references can be divided into two categories. First, there are molecules for which hybridization states or bond orders have been differently assigned. All of these differences are caused by deviations from the expected geometries and are thus directly linked to the quality of the respective coordinates. Second, there are molecules with unusual or even chemically unreasonable resonance or tautomeric forms. Although these differences are not wrong considering the molecule's geometry, they deviate from conventions concerning the representation of particular substructures. Depending on the gravity of these deviations, the solutions are either considered invalid or simply not optimal. Examples for both cases are shown in Figure 9.

Table 5 shows that many differences appearing with other tools are avoided by our method. Incorrect perceptions because of geometrical distortions are often prevented by considering all aspects of an atom's environment. The confidence values for valence states are derived from multiple geometrical parameters so that the assignment has a certain stability against small geometrical distortions. This is a considerable advantage over methods which rely on definite assignments based on particular geometrical parameters. By considering the confidence values of surrounding atoms during the generation of valence bond

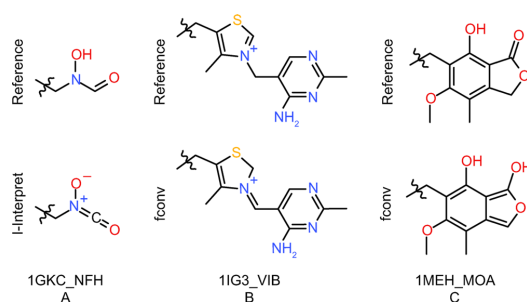


Figure 9. Comparison of reference structures and perceived structures generated by other tools. The structures A and B are classified as errors, whereas structure C is classified as not optimal.

structures even strong distortions can be compensated in some cases. The explicit inclusion of chemical knowledge in the last step of the workflow helps to reliably resolve the remaining ambiguities. Errors concerning the representation of molecules typically occur with methods that put too much emphasis on the evaluation of the geometrical parameters during the generation of valence bond forms. One has to keep in mind that localized bond orders are only an approximation and do

not have to strictly adhere to molecular geometries. By scoring multiple alternative structures using a combination of chemical and geometrical criteria, our method is able to generate molecules that are both in agreement with the atomic coordinates and chemically reasonable.

Validation with Complete Ligand Expo Data Set. The main purpose of our method is the automatic generation of reasonable molecular representations for large data sets. To show that our method is both, efficient and robust, we applied it to all entries of the Ligand Expo data set¹⁸ in PDB format and analyzed the results in terms of runtime and quality. The generated structures were compared to the respective molecules in the SDF format, which are also provided on the Ligand Expo Web site.¹⁹ Again, USMILES served as a basis for the comparison. Since the NAOMI model does not support covalently bound metal atoms, all metal bonds were ignored and only the largest resulting component was used. Additionally, monatomic entries were skipped, since the ionization state of single atoms can not be deduced without knowledge of the environment. Empty entries and entries with multiple disconnected components were also ignored, since this usually indicates missing atoms. Some entries were rejected due to unusually small distances between atoms (coordinate errors). The results of this procedure are summarized in Table 6.

Table 6. Results of the Analysis of the 602704 Entries in the Ligand Expo Data Set for Both SDF and PDB

	SDF	PDB
no. total	602704	602704
no. format errors	0	3015
no. empty entries	7688	7678
no. monatomic entries	241002	239452
no. disconnected entries	10254	10193
no. coordinate errors	499	939
no. converted entries	343261	341427
no. compared entries	334121	

Both data sets initially contained 602704 entries, of which 334121 (55.4%) were eventually used for comparison. To avoid inconsistencies concerning ionization states, all molecules were neutralized in advance (see Figure 10). In 91.7% (306341) of the cases identical valence bond structures were found. The reasons for the observed 27780 differences are quite diverse, as shown shown in Table 7.

In 10012 (36.0%) of the cases, a different tautomeric form of the molecule was generated. Tautomeric forms can often not be distinguished on the basis of the provided coordinates and multiple solutions are equally acceptable. As described above these cases are handled by a canonization procedure, so that different tautomeric forms do not indicate perception errors but rather different default representations. Typical examples for substructures with equivalent tautomeric states are substituted imidazoles, pyrimidones, and guanidinium groups. 810 (2.9%)

Table 7. Analysis of the Reasons for Different Valence Bond Structures for the 334121 Compared Entries of the Ligand Expo Data Set^a

	entries	% of data set	% of differences
no. different valence bond form	27780	8.3	100
no. different tautomeric form	10012	3.0	36.0
no. different oxidation state	810	0.2	2.9
no. different bond order	10349	3.1	37.3
no. different terminal bond order	6063	1.8	21.8
no. small molecule	3523	1.1	12.7

^aMolecules are considered small if they have less than 8 heavy atoms.

of the differences were due to different oxidation states of particular heterocyclic compounds such as NAD/NADP. As with tautomers, these states can not be reliably distinguished on the basis of atomic coordinates, especially in entries with low resolution. Therefore, these cases are also not considered perception errors meaning that 94.9% of the results are essentially identical.

The remaining 16958 entries were further investigated in order to determine the reason for the incorrect perception. These entries correspond to 2341 different components, of which the 20 with the highest counts are shown in Table 8. Evidently, 22.8% (3864) of the differences are caused by only 1% of the components. These entries will be used for the discussion of specific problems encountered with the LigandExpo data set.

The errors associated with HEM are almost exclusively caused by the vinylic double bonds. As discussed above, the number of available geometrical parameters for the determination of bond orders for terminal bonds is small and makes the perception less stable with respect to deviations from ideal geometries. PGV, BCR, PEK, PEV, and OLC are molecules with long aliphatic chains and a specific number of double bonds. In many entries there is a considerable disagreement between our method and the LigandExpo references concerning both the presence and position of these double bonds. We have encountered numerous examples where we did not even find a single shortened bond length in the molecule although a double bond was present in the LigandExpo structure. Many of the incorrect perceptions concerning FAD, NAD, and UMP are caused by strong geometrical distortions of the respective aromatic rings. In some cases torsion angles that reach up to 40° are encountered in these usually completely planar structures. In case of CYC, BLA, and MDO exocyclic carbon-carbon double bonds at five-membered aromatic heterocyclic are interpreted as single bonds. These assignments were in all cases a result of an unambiguous single bond length at the respective bond. The difference from the entries ACB, MLE, and MYR are caused by the specific way covalently bound compounds are handled in the PDB format. If a molecule is bound to a residue of a protein or nucleic acid, the atom involved in this bond is usually assigned to the residue.

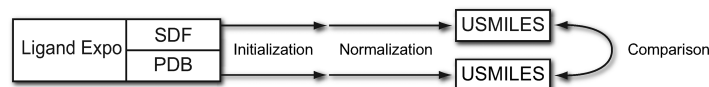


Figure 10. Scheme for the comparison of molecules from the Ligand Expo data set. Generated molecules from the PDB format are compared to the respective structures from the SDF format.

Table 8. PDB Component Names and Numbers of Errors for Those Molecules for Which the Most Errors Occurred

name	no. errors	name	no. errors	name	no. errors	name	no. errors	name	no. errors
HEM	1794	PGV	194	CYC	187	BCR	182	LLP	164
ACB	124	FAD	124	PEK	120	PEV	102	MLE	84
PSO	124	BLA	83	IMA	81	MYR	80	7MG	80
OLC	79	MDO	77	NAD	77	PDU	76	UMP	73

This means that the compound in the entry does not represent an isolated molecule and that necessary information is missing. These errors can often be avoided when the complete PDB entry including the protein environment is used. The reasons for the differences encountered for PSO are quite similar. The psoralen is also covalently bound to a nucleotide but in this case no atoms from the initial component are missing. This connection is, nevertheless, reflected in the coordinates by a change of hybridization geometries for the carbon atoms in the five-membered ring. Since the molecule contained in the LigandExpo data set is an isolated psoralen, the different perception is not surprising. In case of LLP, PDU, and 7MG the structures provided by the LigandExpo data set seem to be wrong (see Figure 11).

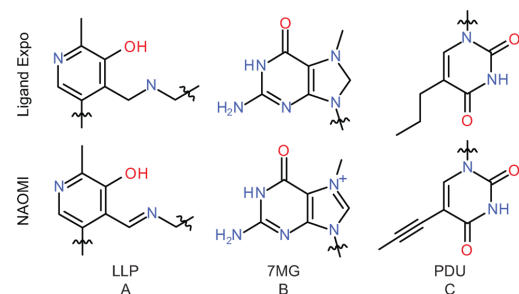


Figure 11. Comparison of inconsistent structures from the LigandExpo data set to those generated by NAOMI.

The compound LLP represents a lysine residue covalently linked to a pyridoxal phosphate via an imine group. This double bond is not present in any of the structures from LigandExpo although it is reflected by a short bond length in the coordinates. The name for the compound PDU on the LigandExpo Web site is 5(1-propynyl)-2'-deoxyuridine-5'-monophosphate which indicates the presence of a triple bond. This triple bond is, however, not present in the reference structure. 7MG is supposed to be 7*N*-methyl-guanosine-5'-monophosphate, a molecule with a charged five-membered heterocycle which is generated by our method. The carbon found in the LigandExpo data set, however, has a carbon atom with an sp^3 hybridization in the five-membered ring.

We think that these examples are sufficient to provide a general overview of the reasons for the observed differences. A special case worth mentioning are molecules with fewer than eight heavy atoms, such as solvents and auxiliary agents. Because of the extreme deviations from ideal geometries, these entries can often not be handled on the basis of atomic coordinates alone. We believe that in some cases these molecules were of minor interest to the researchers and less care was taken during the structure determination process.

When interpreting the results of the comparison one has to keep in mind that our method solely relies on the atomic coordinates provided by the file format. The reference molecules in the Ligand Expo data set are, however, derived from various inputs. In particular, this includes information about the components provided by the crystallographers. This means that the provided coordinates are not necessarily in perfect agreement with the structures present in the data set. In the end 10349 (61.0%) of the 16958 remaining entries differ by only one bond order and the respective bond is terminal in 6063 (35.8%) of these cases. This shows that the generated structures, even if they are not identical, are generally in good agreement for the larger part of the molecules.

Runtimes. The runtimes for the conversion from both the PDB and the SDF format to USMILES are shown in Table 9.

Table 9. Runtimes for the Conversion of the Ligand Expo Data Set from PDB and SDF to USMILES

data set	entries	runtime (s)
PDB (all)	602704	147
SDF (all)		79
PDB (>7 atoms)	204797	110
SDF (>7 atoms)		64

The conversion from SDF provides a point of reference for the performance of our method, since the steps after the generation of the valence bond structure are identical for both formats. Due to the numerous monatomic and small molecules (e.g., solvent molecules) in the data set, we also used a subset where all entries with less than eight atoms have been excluded. This data set provides a more realistic picture of the average runtimes per molecule. The molecule entries in the PDB format were only supplied as single files in a tar archive, which can cause large IO overhead. To avoid this, we concatenated all files into one large file which is a common procedure for other formats such as SDF.

Time measurements were performed on a PC with an Intel Core2 Quad Q9550 CPU (2.83 GHz) and 4 GB of main memory. The average runtime for the conversion of a single molecule from the PDB format is approximately 1 ms. The comparison to the value obtained for the SDF format (0.4 ms/molecule) shows, that the runtimes lie well in the range of conventional file format conversions. Our method can hence be used even in large scale applications.

CONCLUSION

We have presented a novel method for the perception of molecular structures from atomic coordinates. This method is based on the recently published NAOMI model,¹¹ which has been developed for the appropriate representation of organic molecules. The robustness of our approach has been assessed by processing the Ligand Expo data set in PDB format and comparing the resulting molecules to the structures from the corresponding SDF files. The results are correct in more than

95% of the cases showing that our method is able to produce reasonable results even when working with coordinates of varying quality. The method's accuracy has been demonstrated by comparison to manually curated molecules from previously published benchmarking sets. Our method was successful in 98% of the cases and was able to generate reasonable molecular representations even from structures with distorted geometries. A direct comparison to the tools fconv, I-interpret, and MOE shows that the combination of geometrical and chemical criteria used in our method is the key to avoid many assignment problems. Due to the average runtime of less than 1 ms per molecule the method is perfectly suitable for large scale applications.

Since the method is based on the NAOMI model, it is currently limited to organic molecules which can be represented by valence bond structures. This limitation does, however, only exclude a small number of molecules in the PDB and is thus considered acceptable. Because of missing hydrogen atoms and low resolution of most PDB entries the appropriate tautomeric form can usually not be deduced from the atomic coordinates alone. This would require a more advanced analysis of the ligand's energy or the explicit consideration of the molecule's environment, for example, the binding pocket of the protein, neither of which are in the scope of our method. The method is included in the current version of the NAOMI-converter which can be downloaded at <http://www.zbh.uni-hamburg.de/naomi>. It is available free of charge for academic use.

■ ASSOCIATED CONTENT

📄 Supporting Information

PDB files of the 563 molecules used in the validation studies, the corresponding USMILES of the reference structures, and the converted molecules for which the perception was considered incorrect are provided. This material is available free of charge via the Internet at <http://pubs.acs.org/>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: rarey@zbh.uni-hamburg.de.

Present Addresses

[§]Georg Simon Ohm University of Applied Sciences, Kesslerplatz 12, 90121 Nuremberg, Germany.

^{||}Evotec AG, Essener Bogen 7, 22419 Hamburg.

Notes

The authors declare no competing financial interest.

■ REFERENCES

- (1) Berman, H.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.; Weissig, H.; Shindyalov, I.; Bourne, P. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
- (2) PDB File Formats. http://www.pdb.org/pdb/static.do?p=file_formats/index.jsp (accessed Oct 19, 2011).
- (3) PDB Format, version 3.3. <http://www.wwpdb.org/documentation/format33/v3.3.html> (accessed Oct 19, 2011).
- (4) Meng, E.; Lewis, R. Determination of Molecular Topology and Atomic Hybridization States from Heavy Atom Coordinates. *J. Comput. Chem.* **1991**, *12*, 891–898.
- (5) Baber, J.; Hodgkin, E. Automatic Assignment of Chemical Connectivity to Organic Molecules in the Cambridge Structural Database. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 401–406.
- (6) Hendlich, M.; Rippmann, F.; Barnickel, G. BALL: Automatic Assignment of Bond and Atom Types for Protein Ligands in the Brookhaven Protein Databank. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 774–778.
- (7) Labute, P. On the Perception of Molecules from 3D Atomic Coordinates. *J. Chem. Inf. Model.* **2005**, *45*, 215–221.
- (8) Froeyen, M.; Herdewijn, P. Correct Bond Order Assignment in a Molecular Framework Using Integer Linear Programming with Application to Molecules Where Only Non-Hydrogen Atom Coordinates Are Available. *J. Chem. Inf. Model.* **2005**, *45*, 1267–1274.
- (9) Zhao, Y.; Cheng, T.; Wang, R. Automatic Perception of Organic Molecules Based on Essential Structural Information. *J. Chem. Inf. Model.* **2007**, *47*, 1379–1385.
- (10) Sayle, R. PDB: Cruft to Content (Perception of Molecular Connectivity from 3D Coordinates). Daylight Chemical Information Systems Inc. MUG'01 Presentation, 2001. <http://www.daylight.com/meetings/mug01/Sayle/m4xbondage.html> (accessed Oct 18, 2011).
- (11) Urbaczek, S.; Kolodzik, A.; Fischer, R.; Lippert, T.; Heuser, S.; Groth, I.; Schulz-Gasch, T.; Rarey, M. NAOMI - On the Almost Trivial Task of Reading Molecules from Different File Formats. *J. Chem. Inf. Model.* **2011**, *51*, 3199–3207.
- (12) Cordero, B.; Gomez, V.; Platero-Prats, A. E.; Reyes, M.; Echeverria, J.; Cremades, E.; Barragan, F.; Alvarez, S. Covalent radii revisited. *Dalton Trans.* **2008**, 2832–2838.
- (13) Weininger, D.; Weininger, A.; Weininger, J. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97–101.
- (14) Hartshorn, M.; Verdonk, M.; Chessari, G.; Brewerton, S.; Mooij, W.; Mortenson, P.; Murray, C. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* **2007**, *50*, 726–741.
- (15) I-Interpret, version 1.0, Shanghai Institute of Organic Chemistry. <http://www.sioc-ccbg.ac.cn/?p=42> software=i-interpret (accessed Oct 4, 2012).
- (16) fconv—A tool not only for file conversion, version 1.24, Gerd Neudert, University of Marburg. http://pc1664.pharmazie.uni-marburg.de/drugscore/fconv_download.php (accessed Oct 4, 2012).
- (17) Molecular Operating Environment (MOE), version 2011.10, Chemical Computing Group Inc. <http://www.chemcomp.com/software.htm>, (accessed Oct 4, 2012).
- (18) Feng, Z.; Chen, L.; Maddula, H.; Akcan, O.; Oughtred, R.; Berman, H.; Westbrook, J. Ligand Depot: A Data Warehouse for Ligands Bound to Macromolecules. *Bioinformatics* **2004**, *20*, 2153–2155.
- (19) Ligand Expo, RCSB PDB. <http://ligand-expo.rcsb.org/> (SDF and PDB dataset downloaded Jul 10, 2012).

Protoss: A Holistic Approach to Predict Tautomers and Protonation States in Protein-Ligand Complexes

[D6] S. Bietz, **S. Urbaczek**, B. Schulz, and M. Rarey. Protoss: A Holistic Approach to Predict Tautomers and Protonation States in Protein-Ligand Complexes. *Journal of Cheminformatics*, 6(1):12, 2014.

<http://dx.doi.org/10.1186/1758-2946-6-12>

RESEARCH ARTICLE

Open Access

Protoss: a holistic approach to predict tautomers and protonation states in protein-ligand complexes

Stefan Bietz^{1†}, Sascha Urbaczek^{1†}, Benjamin Schulz^{1,2} and Matthias Rarey^{1*}

Abstract

The calculation of hydrogen positions is a common preprocessing step when working with crystal structures of protein-ligand complexes. An explicit description of hydrogen atoms is generally needed in order to analyze the binding mode of particular ligands or to calculate the associated binding energies. Due to the large number of degrees of freedom resulting from different chemical moieties and the high degree of mutual dependence this problem is anything but trivial. In addition to an efficient algorithm to take care of the complexity resulting from complicated hydrogen bonding networks, a robust chemical model is needed to describe effects such as tautomerism and ionization consistently. We present a novel method for the placement of hydrogen coordinates in protein-ligand complexes which takes tautomers and protonation states of both protein and ligand into account. Our method generates the most probable hydrogen positions on the basis of an optimal hydrogen bonding network using an empirical scoring function. The high quality of our results could be verified by comparison to the manually adjusted Astex diverse set and a remarkably low rate of undesirable hydrogen contacts compared to other tools.

Keywords: Protein-ligand complex, Tautomers, Protonation states, Hydrogen placement

Background

Crystal structures of protein-ligand complexes play an important role in the drug development process. They provide valuable insights into where and how molecules interact with their respective target proteins and thus are the basis for further optimization strategies. They also serve as starting point for numerous structure-based in-silico techniques such as molecular docking or pharmacophore generation. Furthermore, the statistical analysis of large collections of crystal structures is a common means to gain general knowledge about molecular interactions and geometry. These results are often used to derive parameters for various computational methods. All of the above-mentioned applications depend on information about the interactions between protein and molecules with hydrogen bonds being one of the most important types. Due to insufficient resolution, the vast majority of

the entries in the Protein Data Bank (PDB) [1] only contain coordinates of non-hydrogen atoms. In order to be able to work with these entries, automated procedures for the placement of hydrogen atoms are needed. Considering its importance, it is not surprising that a large number of different methodologies have been developed to tackle this task. A thorough review of these different approaches has been given by Forrest and Honig [2].

While many of these applications show substantial differences concerning their subjective function or their underlying optimization algorithms, most of them share the degrees of freedom which are used to tackle the uncertainties of structure determination [3-8]. Typically, these comprise rotatable hydrogens, tautomers and protonation states of particular amino acids, alternative water orientations, and terminal side chain flips. Indeed, this covers the most important ambiguities of protein structures, but neglects crucial aspects of ligand molecules. Different tautomers and protonation states can lead to substantially different interaction patterns. Hence, considering alternative ligand states has a high impact on the quality of hydrogen bonding networks, especially for applications dealing

*Correspondence: rarey@zbh.uni-hamburg.de

†Equal contributors

¹Center for Bioinformatics(ZBH), Universität Hamburg, Bundesstr. 43, 20146 Hamburg, Germany

Full list of author information is available at the end of the article



with ligand binding. Neglecting these degrees of freedom might easily lead to erroneous predictions, including the omission of relevant hydrogen bonds and the generation of hydrogen clashes. Nevertheless, targeting this problem has not drawn much attention in the literature yet. This might be reasoned in a deviating focus of most hydrogen prediction tools, which concentrate rather on the whole protein than on single binding sites, but it might also reflect the difficulty of properly modeling complex phenomena like tautomerism and ionization of arbitrary organic molecules. However, some of the more recently developed methods consider these effects at least to some extent.

Protonate 3D [9,10] has been developed for the prediction of hydrogen coordinates as a preprocessing step to structure-based computational applications, e.g. protein-ligand docking or molecular dynamics. Beside well-established degrees of freedom for protein side chains, it is also capable of considering selected alternative states of other chemical groups. This is technically realized by a SMARTS [11]/SMILES [12]-based template collection stored in a predefined parameter file which must explicitly contain all tautomeric and protonation states that should be considered for a specific chemical group. Furthermore, Protonate 3D uses a prioritizing branch-and-bound algorithm in combination with a preceding dead-end elimination to handle the state space optimization problem and a force field based energy model including additional terms for tautomerism and ionization effects.

The modeling and simulation suite YASARA [13,14] provides a sub-module for the prediction of hydrogen coordinates which is able to consider alternative protonation states and tautomers of non-protein-like chemical substructures. Similarly to Protonate 3D, a configuration file contains template definitions for different potential states of these substructures represented as SMILES strings. Its default collection of considered substructures is a little more comprehensive, but its generality is still limited by the fact that all molecular states have to be explicitly defined. The optimization problem is tackled with an algorithm, originally developed for side chain prediction, which combines a dead-end elimination, a branch-and-bound backtracking, and a graph decomposition approach [15]. Interestingly, the underlying empirical scoring model, in contrast to most other hydrogen prediction tools, targets a minimization of the amount of unsaturated hydrogen bond donors or acceptors instead of a maximization of the number of attractive interactions.

We present a novel method for the placement of hydrogen coordinates in protein-ligand complexes. By using the consistent chemical description provided by the NAOMI model [16], tautomeric and protonation states of both protein and ligand are handled consistently. The method is a substantial extension of Protoss [17] which has been

developed earlier. The optimal hydrogen bonding network is determined on the basis of the quality of all possible hydrogen bonds in combination with the stability of the involved chemical groups. There is to the best of our knowledge no other method described in the literature which is able to handle the degrees of freedom for protein and ligand in a comparable generality.

Methods

The purpose of the presented method is the generation of the most probable hydrogen placement for a given protein-ligand complex. The underlying optimization procedure is based on an empirical scoring scheme designed to identify an optimal hydrogen bonding network. This scheme takes both the quality of hydrogen bond interactions and the relative stability of different chemical species into account. The procedure is performed in separate steps which will be explained in detail in the following sections. Due to the exceptional importance of the PDB as source for input structures, we have added a subsection in which the necessary preprocessing steps for working with PDB files are discussed.

Input from PDB files

In contrast to most other chemical file formats, the PDB format [18] does not include any information about bond orders or atom types so that these properties must be derived directly from the provided atomic coordinates. In case of biological macromolecules, e.g., proteins, this process can be considerably facilitated by using structural templates for standard residues. The necessary data for both the subdivision of proteins into residues and the identification of particular atoms is provided in the coordinate section of the PDB format. In case of incomplete residues, the missing atoms are topologically added in order to ensure an accurate description. They will, however, not have valid coordinates and are thus ignored during the calculation of interactions. For the large and steadily growing number of different small molecules in the PDB, predefined structural templates are generally not a viable option. In this case a generic method for the construction of molecules from three-dimensional coordinates is needed. This evidently also applies to non-standard residues for which no predefined template is available. We use a method based on the NAOMI model for that purpose [19]. Both strategies eventually result in isolated components which have to be connected in order to build the complete protein structure. The connection of standard residues with peptidic bonds is again handled with recourse to predefined templates. All other types are based on a procedure similar to that used for the generic construction from three-dimensional coordinates. The only difference is that the method is applied to a substructure rather than the complete molecule. In this

way the consistent description of molecules can be used to reliably handle the integration of residues into proteins. The description of both proteins and molecules is based on the NAOMI model, meaning that consistent atom type and bond order information is available throughout the next steps.

Initial hydrogen positions

Initial hydrogen coordinates are calculated on the basis of idealized geometries provided by the atom types of the NAOMI model. These geometries reflect the hybridization states of the respective atoms and are based on the general concepts of VSEPR theory [20]. In combination with the coordinates of the covalently-bound non-hydrogen atoms, knowledge about the atom's hybridization state can be used to calculate reasonable positions for hydrogens. The concrete orientation of the respective hydrogen bonds is in many cases unambiguously determined by the constraints imposed by the atom's local geometry. In case of an sp^3 hybridized carbon atom with three non-hydrogen bonds, for instance, the direction of the bond coincides with the connection line to the unoccupied vertex position of the underlying tetrahedron. There are, however, a few cases for which multiple acceptable orientations exist. The most prominent examples in protein-ligand complexes are isolated atoms (e.g. water), terminal atoms (e.g. alcohols, acyclic amines) and particular types of ring atoms (e.g. cyclic secondary amines). In these cases the orientation of hydrogens cannot be unambiguously derived from the heavy atom skeleton of the respective molecule. The final decision can only be made under consideration of all chemical moieties in close vicinity so that only preliminary positions can be calculated at this point. Another type of ambiguity arises with respect to the initial tautomeric and ionization states of both residues and ligands as these will obviously influence the corresponding hydrogen positions. For this purpose the normalization procedures described in [21] are applied prior to the generation of initial hydrogen coordinates. Free amino and acid groups of residues resulting from chain breaks are a special case. If the PDB file does not indicate that these residues are in fact terminal, they will be treated internally as incomplete parts of an amide bond and thus kept in their neutral state. At the end of the procedure, each hydrogen atom in the protein will have three-dimensional coordinates which are in accordance with the hybridization states of the respective bond partners. In case of multiple alternatives, these preliminary positions, however, are just needed for technical reasons and will be adapted in later steps.

Enumeration of alternative hydrogen positions

Based on the initial assignment of hydrogen positions, tautomers and protonation states, substructures with

variable hydrogen positions in both protein and ligand are identified. The considered types of variability are rotations of terminal hydrogen atoms, potential side-chain flips for specific residues, alternative tautomeric forms, different protonation states, and alternative orientations of water molecules. For each substructure, all different placements of hydrogen atoms, called alternative modes in the following, are enumerated (see Figure 1 for examples). In contrast to the previously published Protoss version, tautomeric and protonation states for small molecules and non-standard residues are also taken into account. These are generated using the valence state combination model presented in a separate publication [21]. Since the details of these calculations are beyond the scope of the presented method, we will only give a short overview with focus on those aspects relevant in the current context. The workflow starts with the partitioning of the molecule into non-overlapping substructures which correspond for the most part to conjugated ringsystems and functional groups. In some cases substituents, e.g., alcohols and amines, are considered as part of a ringsystem as they are necessary for the consistent generation of tautomers. The partitioning is retained throughout the following steps as it reflects the dependency between the hydrogen positions for the atoms in the substructures. These will be referred to as Variable Mode Regions (VMR) in the following. Protonation states and tautomers are enumerated for each VMR individually and stored in form of a list containing the alternative modes together with an integer-based score. These scores provide an order of preference which is crucial when deciding if the default mode of a VMR should be changed in order to optimize the hydrogen

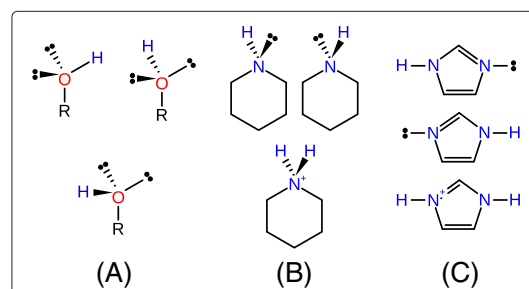


Figure 1 Three examples for VMRs with alternative hydrogen positions and free electron pairs. Primary alcohols (A) are considered as rotatable and the associated hydrogen atom can occupy any position on the orbit around the oxygen atom. Three exemplary orientations are shown. Cyclic secondary amines (B) can either be protonated and positively charged or neutral. In the latter case the hydrogen atom can occupy two distinct positions. The imidazole ring (C) can either occur as one of the two different tautomeric species or in its ionized form. In contrast to the other examples the VMR contains multiple atoms in this case.

bonding network. The underlying scoring scheme is based on the identification of predefined structural fragments in the respective modes of the VMRs. Each fragment corresponds to a different tautomeric form or protonation state and is associated with a partial score. The total score of the mode is calculated as the sum of these individual contributions. In case of ringsystems the score comprises contributions from each ring and its respective substituents. Scores for functional groups are either generated by matching the whole group directly, which is the usual case, or by partitioning the group into subgroups and adding the scores of the smaller subgroups. The values for the individual contributions of the respective substructures have been derived from different pairs of tautomers for which the preference was experimentally known and from pK_a tables.

Hydrogen bond interactions

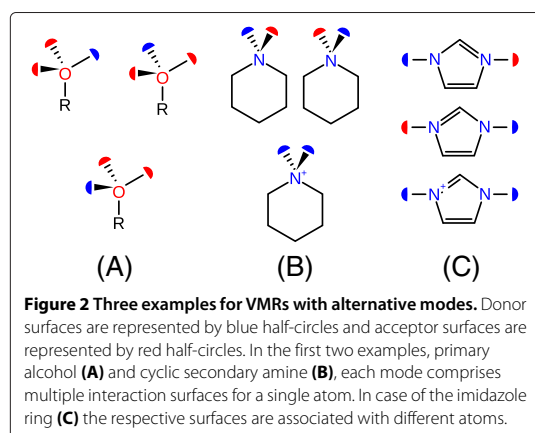
Since Protoss is designed to identify the best hydrogen bonding network, it requires structural information for the evaluation of potential polar interactions. Therefore, each mode is internally represented as a set of interaction surfaces, originally developed in the context of molecular docking [22]. This is shown in Figure 2 (A) for the straightforward case of a rotatable hydroxyl group. Each mode includes one interaction surface associated with the orientation of the hydrogen atom (donor surface) and two additional interaction surfaces associated with the atom's free electron pairs (acceptor surface). The modes for a secondary amine are shown in Figure 2 (B) in order to exemplify the handling of protonation states. In this case the number of donor and acceptor surfaces of each mode is not necessarily identical. Modes for tautomeric states introduce a higher complexity since they involve hydrogen positions at multiple atoms simultaneously. The corresponding modes for an

imidazole moiety are shown in Figure 2 (C). In this case, only specific combinations of interaction surfaces are considered reflecting the different tautomeric states of the molecule. These combinations are derived from the alternative modes for the VMRs generated in the previous step.

The objective function for the evaluation of the hydrogen bonding network comprises, as in the previous Protoss version, the analysis of hydrogen bonds as well as metal interactions. In order to prevent the generation of undesirable contacts of polar groups in the protein-ligand interface, the scoring function has been extended by an additional term for the assessment of repulsive contributions such as donor-donor, donor-metal, or acceptor-acceptor contacts. The interaction quality is for both cases, attractive and repulsive interactions, determined by a geometric criterion which measures the relative orientation of two interaction surfaces (see [22]). However, in contrast to hydrogen bonds and metal interactions, repulsions have naturally a destabilizing influence on the total energy of the hydrogen bonding network.

Optimization procedure

The optimization procedure is based on two main aspects, namely the scoring of hydrogen bond interactions and the resolution of dependencies in the hydrogen bonding network. The latter is represented by a graph structure in which each node corresponds to a single VMR with all its associated alternative modes. Edges between nodes are formed if there exists at least one relevant interaction between the atoms of the respective VMRs. This is determined by a geometric criterion. In the first step, the scoring phase, the alternative modes of each node are assigned a base score which is composed of an intrinsic stability contribution reflecting the preference of the respective tautomeric form or protonation state it represents and a term for the interaction energies with all non-variable parts of the complex. The value of the stability contribution is derived from the score calculated by the generic scoring scheme described above. Each edge contains a matrix that stores an interaction score for each combination of modes of its two incident nodes. In the second step, the optimization phase, a combination of a cycle decomposition and a dynamic programming algorithm is used to find an optimal hydrogen bonding network by minimizing the total score and selecting a distinct mode for every VMR. For a set of selected modes M , the total score is therefore calculated by Equation 1.



$$\begin{aligned} totalScore(M) = & \sum_{m \in M} baseScore(m) \\ & + \sum_{m, n \in M} (interactions(m, n) + repulsions(m, n)) \end{aligned} \quad (1)$$

Further details about the optimization procedure can be found in a previous publication [17]. Finally, the optimized coordinates of all variable hydrogen atoms are generated by transferring the structural information of the individual modes back onto the protein-ligand complex.

Results and discussion

Tautomeric frequencies

Most hydrogen prediction tools for protein-ligand complexes only handle tautomerism for moieties from proteinogenic amino acids or by explicit lists of substructure transformation rules. In order to demonstrate the insufficiencies of this approach, we counted all substructures contained in the Ligand Expo database (accessed Jan 3, 2014) [23], for which we were able to identify sensible alternative tautomers or protonation states. Furthermore, we split the set into two groups. First, the set of functional groups which also appear in protein side chains, namely carboxylates, primary amines, and imidazoles (classical VMRs). Second, all other functional groups and conjugated substructures for which more than one sensible state could be created (advanced VMRs). Rotational degrees of freedom were neglected for this analysis.

We found that only 19% of the Ligand Expo database molecules did not show any VMR with alternative tautomers or protonation states. Furthermore, 17% of all molecules only contain substructures from the classical VMRs set. For all other molecules, at least one advanced VMR was observed.

Overall, we found 1802 structurally different, canonical VMR types. In order to analyze the relevance of these different substructures, we first sorted the list of VMRs according to the portion of molecules containing the respective VMR and then plotted the amount of molecules whose variability with respect to tautomerism and protonation can be completely described by a set of the k most frequent VMRs (see Figure 3). The results show that, e.g.,

a set of approximately 430 substructures is required to consider the full variability for 90% of all molecules in the Ligand Expo database. In general, the curve progression clearly illustrates the strong dependency of low prediction error rates on the consideration of a wide range of chemical substructures.

Figure 4 additionally depicts the absolute amount of different VMRs for various chemical classes. This classification demonstrates that the high amount of different VMRs is mostly reasoned in the diversity of aromatic substructures. The difficulty of correctly treating more complicated substructures, such as annulated aromatic ringsystems, motivates a generic approach for handling tautomerism.

Undesirable contacts

One of the primary requirements on hydrogen placement is to avoid the generation of undesirable contacts such as close donor-donor, donor-metal or acceptor-acceptor interactions. In order to evaluate the effect of considering alternative protonation and tautomeric states on this issue, we analyzed the occurrence of undesirable contacts in the results of the hydrogen prediction tools Protonate 3D (as implemented in MOE 2013.08 [10]), YASARA (version 13.9.8 [14]), and Protoss. The latter was used in two alternative versions, with and without an analysis of alternative tautomers and protonation states. Apart from that, all tools were applied with default settings. The sc-PDB database v.2012 [24] served as basis for this test, as it constitutes a comprehensive and diverse database of pharmacological relevant protein-ligand complexes. However, as the protein files provided by the sc-PDB do not contain water molecules, we used the corresponding original files from the PDB instead. The sc-PDB v.2012 consists of all in all 8077 protein-ligand complexes. Nine of them were not available in the PDB anymore (November 2013) and have therefore been excluded. The remaining 8068 structures

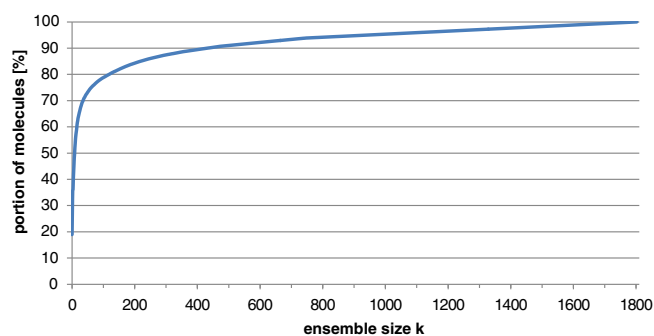
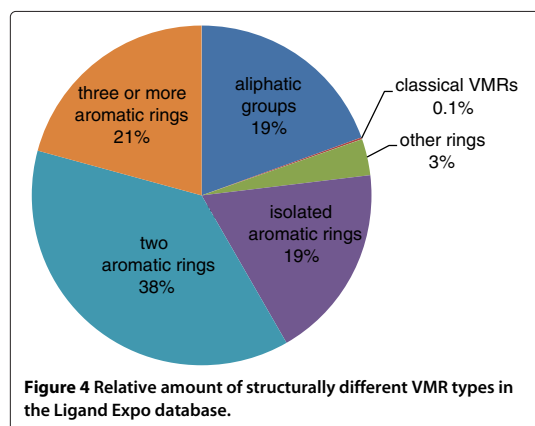


Figure 3 Dependency of the portion of Ligand Expo molecules whose protonational variability can be completely described by a set of the k most frequent VMRs on the ensemble size k .



were further processed by a clean-up step removing all existing hydrogen atoms, atom duplicates, and overlapping entities in order to reduce possible error sources which might bias the hydrogen prediction and validation experiments. This procedure comprises a series of atom entry filtering steps which were processed in the following order. First, all hydrogen atom entries were erased. Second, all residue entries were identified that overlap with the reference ligand. In this and all following cases, an overlap was defined as an atom distance of equal or less than 1 Å. Furthermore, an overlapping residue entry was defined to represent a part of the reference ligand if for each of its atom entries the closest atom of the reference ligand has a maximum distance of 1 Å and the same chemical element type. (This rather fuzzy matching criterion was chosen because some sc-PDB ligands are shifted or have a slightly different conformation compared to the original PDB structure). Otherwise the overlapping entry was removed. If an overlapping residue entry contains alternate locations we only kept that conformation which fits the reference ligand best. In case that the best conformation does not fulfill the matching criterion, the residue entry was only retained if the first alternate location has no overlap with the reference ligand. In the third step, all other atom entries were checked for alternate locations and only the first position per atom was kept. In a final step, all residue entries were dropped, which overlap with any preceding entry in the file or exhibit an internal atom overlap.

In 27 cases this cleanup procedure led to a partial or total removal of the reference ligand's heavy atoms, e.g. if the sc-PDB ligand, compared to the original PDB structure, exhibits a different conformation, deviating element, additional atoms, or an internal atom overlap. Therefore these structures were also removed.

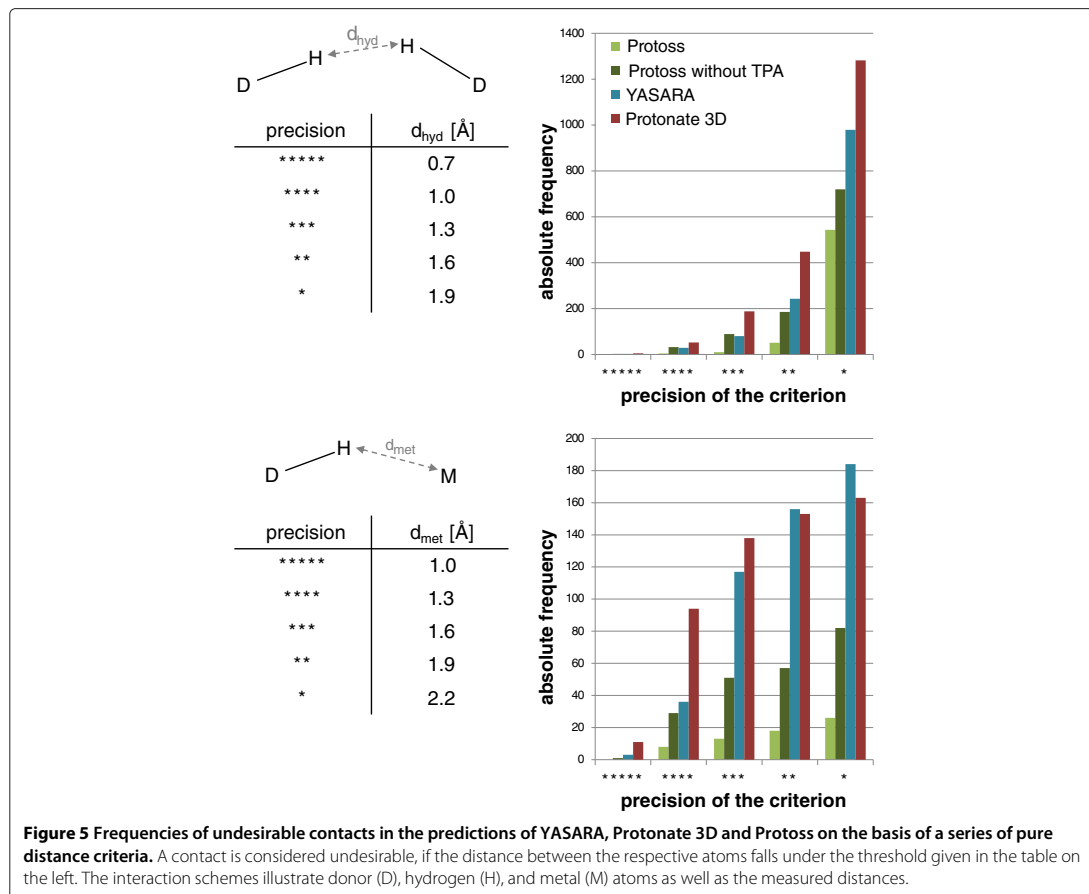
For the remaining set of 8041 files, all three tools were used to add new hydrogen atoms and to optimize the

hydrogen bonding networks. As the Yasara version used in this study shuts down during the prediction for one complex (3ptq), this structure was also excluded. Eventually, the results were scanned for undesirable contacts, which were defined as follows: All oxygen and nitrogen atoms of the ligand or the active site (6.5 Å around the ligand) which have at least one hydrogen bound were considered as hydrogen bond donors. Two hydrogen bond donors are defined to form an undesirable contact if the hydrogen atom distance is equal or less than a certain threshold. Likewise, an undesirable contact between a donor and a metal ion is determined on the basis of the hydrogen-metal distance (see Figure 5). For both cases, exactly one of the counterparts had to be part of the ligand. Beside this simple distance criteria, we also analyzed both types of contacts under consideration of additional measures, namely the heavy atom distance and the angles formed by both heavy atoms and one of the hydrogens (see Figure 6). We also defined different threshold sets to investigate the dependency of the error frequency on the precision of the interaction criterion. All used precision levels and their respective thresholds are listed in the tables in Figure 5 and Figure 6. Although an additional investigation of acceptor-acceptor contacts could provide further insights, we explicitly avoided this analysis, because acceptor orientations cannot be analyzed without interpreting the input data on the basis of geometric assumptions of an internal chemical model, which would compulsorily influence the evaluation. Overall, the possibly most conspicuous and expected finding is that the error frequency increases with decreasing precision of the interaction criterion. This effect can be observed for all prediction tools. The higher rate of undesirable contacts for the Protoss version without tautomer analysis throughout all precision levels clearly demonstrates the benefit of considering tautomerism and protonation states for the performance of hydrogen prediction.

Comparison to manual adjustment

Ultimately, a hydrogen prediction tool should be validated against experimental data. Unfortunately, there is only a very limited amount of experimental data that might be used for such an evaluation due the difficulties of determining hydrogen coordinates with X-ray crystallography.

As a result of the insufficient amount of experimental data, we intend to demonstrate the properness of our approach on the basis of the Astex diverse set [25] (Astex Set). This collection of 85 protein-ligand complexes, which was developed for the validation of docking performance, contains ligands which are manually adjusted with respect to their protonation and tautomeric states. Therefore, the Astex Set seems to be suitable for a verification of predicted ligand states. For each target structure in the dataset, the original file was retrieved



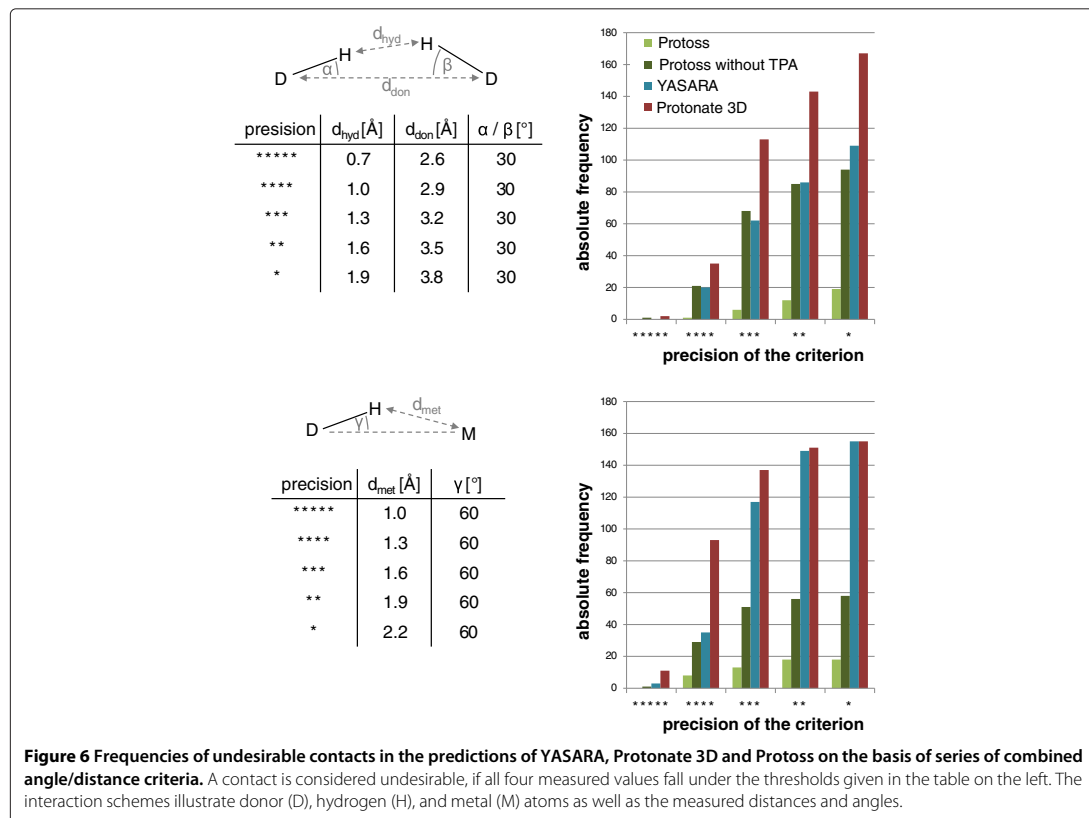
from the PDB, preprocessed as described in the previous section (removing existing hydrogen atoms, atom duplicates, and overlapping entities) and given to Protonate 3D, YASARA, and Protoss for generating new hydrogens as well as their coordinates. The results were then written to PDB files and compared to the ligand topology given in the Astex Set.

The topological ligand comparison was realized by a simple string comparison of Unique SMILES [26]. However, as the bond orders of the internal molecule representation that was used for the Unique SMILES generation are derived from PDB files, there is still a theoretical risk of misinterpreting the molecular topology. Therefore, all automatically detected deviations were additionally confirmed by visual comparison to the graphical molecular representations of the respective tools.

The deviating solutions are classified according to the deviation type, thus whether the solution constitutes a different tautomer, protonation state, or redox form.

Furthermore, the quality of the hydrogen bonding network with respect to undesirable contacts and missing a hydrogen bonds is analyzed. Since a different redox form constitutes a more serious problem, the latter aspect is only evaluated for deviating tautomers and protonation states. In contrast to erroneous redox forms, deviating protonation or tautomeric states are not necessarily incorrect. However, a worse hydrogen bonding network is at least a strong hint that the respective structure is inferior. A hydrogen bond was defined by a maximum heavy atom distance of 3.5 Å and a minimal donor-hydrogen-acceptor angle of 150°. Undesirable contacts were defined on the basis of precision level 2 (**) (see Figure 5).

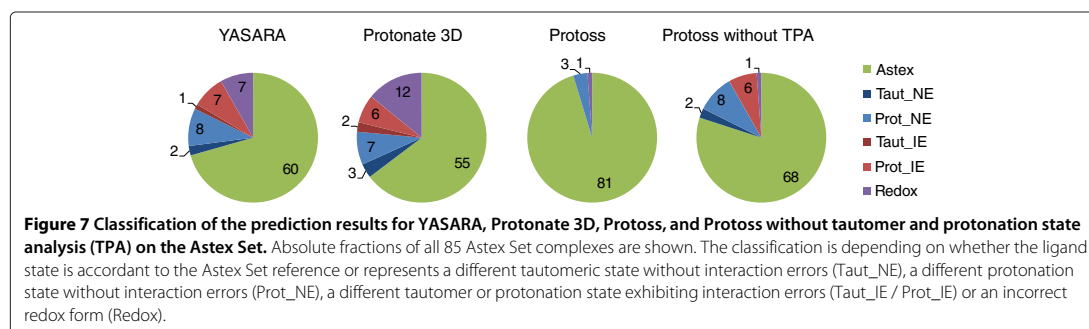
Figure 7 illustrates the amount of accordant and deviating ligand states as well as the five classes of the deviating solutions. For all of the three hydrogen prediction tools, the set of proposed solutions which are in accordance with the ligand states in the Astex Set (depicted in green)



constitutes the major portion. A closer look to the fractions of different tautomers and protonation states which form less interactions or even undesirable contacts (light and dark red), as well as incorrect redox forms (purple) demonstrate the importance of a comprehensive initialization of ligand molecules. A comparison to the Protoss version which does not execute an analysis of tautomers and protonations state (TPA) demonstrates the reduction

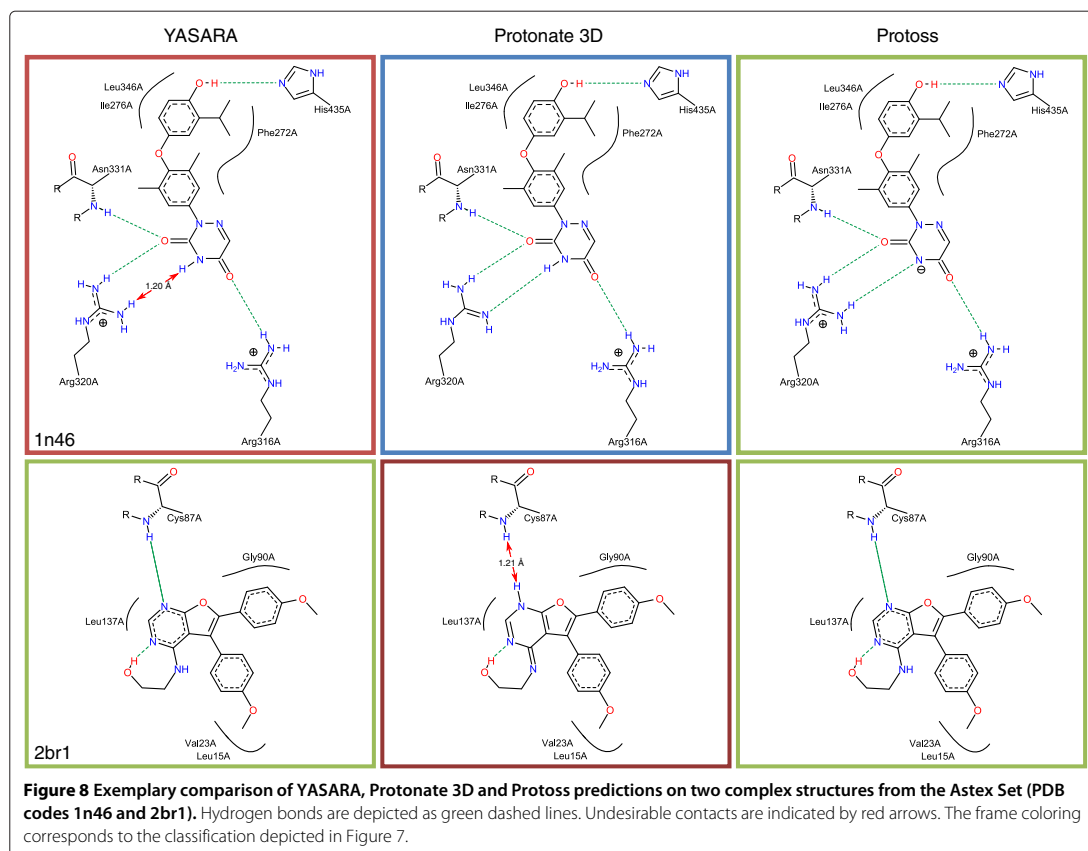
of critical cases and hence, also the ability of resolving erroneous prediction performance.

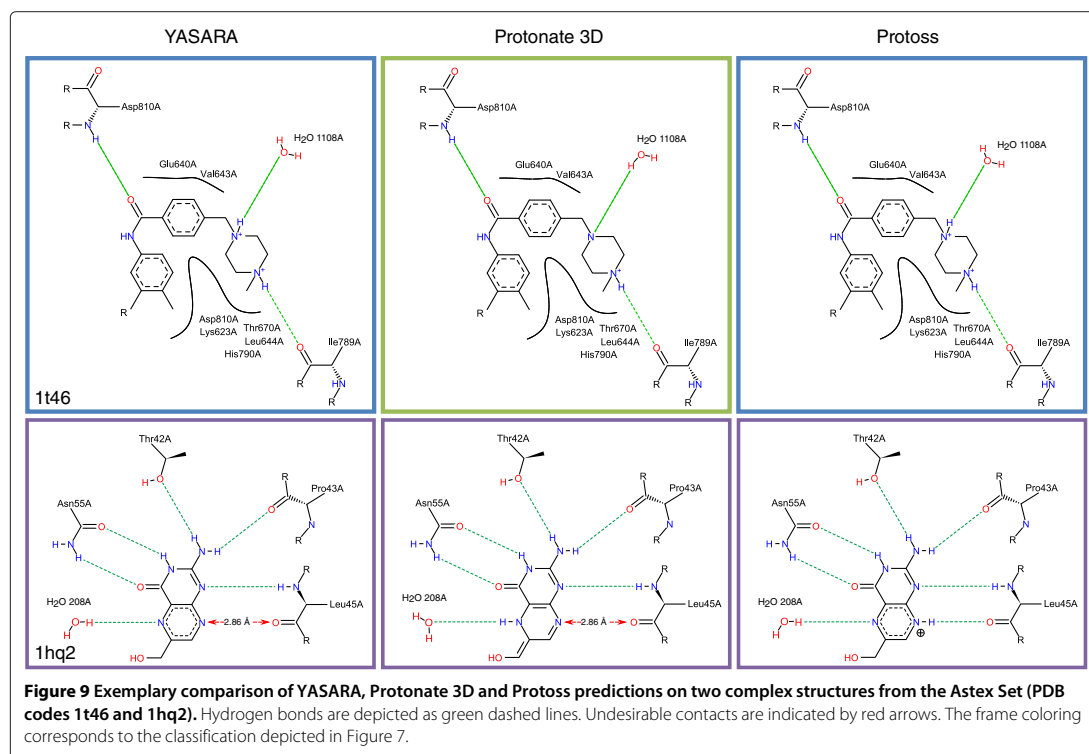
The classification is also illustrated by the following case studies taken from the Astex Set. Given the complex of the human thyroid receptor beta ligand-binding domain and its 6-azauracil-based ligand from PDB structure 1n46 [27], Protoss proposes a negatively charged state of the azauracil moiety which is able to form three



hydrogen bonds with the surrounding arginine residues (see Figure 8). This prediction is in accordance with the ligand state given by the Astex Set. In contrast to this, Protonate 3D chooses a neutral ligand state and deprotonates Arg320 instead. Although this leads to the same number of hydrogen bonds, considering the pK_a values of 6-azauracil ($pK_a = 6.9$ [28]) and the protonated arginine side chain ($pK_a = 12.5$ [29]) this solution seems to be less likely. YASARA neither deprotonates the azauracil moiety nor the guanidinium of Arg320 which leads to the loss of a hydrogen bond and simultaneously to the formation of a close donor-donor contact with a hydrogen distance of 1.20 Å. Figure 8 also depicts the solutions for serine/threonine-protein kinase Chk1 complexed with a furanopyrimidine inhibitor (2br1) [30]. While both Protoss and YASARA successfully reproduce the state of the reference ligand, which is stabilized by a hydrogen bond to the backbone of Cys87 and an internal interaction with a hydroxyl group, Protonate 3D selects a different tautomer. Thereby, the hydrogen bond to Cys87 is replaced by a contact of two donors with a hydrogen distance of 1.21 Å.

All in all there are only four cases where Protoss produces a ligand state that differs from the reference given by the Astex Set. However, we did not observe a missing hydrogen bond or an undesirable contact in any of these binding sites. For an inhibited thrombin complex (1oyt, not shown) [31] Protoss proposes a protonated nitrogen in contrast to a neutral state in the Astex Set. However, this does not change the quality of the hydrogen bonding network since this atom is not involved in a polar interaction. In case of an adenosine deaminase structure complexed with a non-nucleoside inhibitor (1uml, not shown) [32], Protoss protonates an imidazole ring of the ligand, which enables the formation of a hydrogen bond to Asp296. The same interaction can be found in the Astex Set structure, though here the hydrogen is located at Asp296 instead. In another example, shown in Figure 9, Protoss chooses a double protonated state of a piperazine ring (1t46) [33]. This can be explained by the fact that only conjugated ring systems are handled as a unit, while polar groups in others rings are treated separately. Here, only Protonate 3D identifies the more likely single protonated state. YASARA also





predicts the double charged piperazine ring. However, as one of the piperazine nitrogen only interacts with a water molecule, this deviation has no significant effect on the hydrogen bonding network.

The only critical solution produced by Protoss constitutes the complex of *E. coli* 6-Hydroxymethyl-7,8-dihydropterin pyrophosphokinase and its substrate. For this target, all three tools fail to produce the correct redox form of the ligand. This might be reasoned in the exceptionally short bond length of carbon C7 and nitrogen N8 with a distance in the PDB file of 1.35 Å (1hq2) [34]. Interestingly, there is another PDB structure of the same complex which contains the oxidized form of the ligand (3ip0) [34]. Here, the same bond has a length of 1.34 Å (see Figure 10). In this case, it is obviously a tough task to predict the correct redox form automatically only on the basis of heavy atom coordinates.

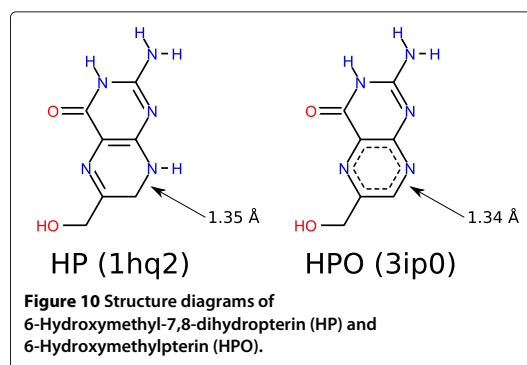
Computing time

On average, the hydrogen prediction by Protoss took 2.47 seconds for a complex from the Astex Set. The median of this prediction series is 0.93 seconds. This includes file IO, preprocessing, and hydrogen bonding network optimization for the whole protein-ligand complex with

all ligands, co-factors and water molecules. All runtime measurements were performed on a single core of an Intel Core i7-2600 with 3.4 GHz and 8 GB of memory.

Conclusion

There are several known cases in which a small change in the ligand molecule, resulting in a single additional hydrogen bond, makes a huge difference in binding affinity. Therefore, the correct assignment of the ligand's



tautomeric form, its protonation state and hydrogen orientations is a mandatory step in structure-based molecular design. Especially precise protein-ligand scoring functions, as a key component in docking and lead optimization procedures, rely on a correct protonation. Since validation procedures for docking and scoring are mostly based on carefully, hand-prepared test cases, the influence of wrong tautomerism and protonation is quickly overseen.

Several methods exist already addressing this important preprocessing step, however, most approaches lack a comprehensive model of ligand tautomerism. Here, we present a novel method for the placement of hydrogen coordinates in protein-ligand complexes under consideration of both tautomeric and protonation states. The method implements an optimization procedure designed to identify the best hydrogen bonding network based on a generic scoring function. Its main application is the automatic preparation of protein binding sites for structure-based virtual screening and large-scale statistical analysis of molecular interactions in biological systems. Our validation studies show that for this purpose our approach yields results which are in good agreement with manually adjusted ligand states. Numerous case studies demonstrate that the resulting molecular states are both comprehensible and chemically reasonable.

Availability

Protoss is available free of charge for academic use as a web service at <http://www.zbh.uni-hamburg.de/protoss>.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SB and SU contributed equally to this work. They developed the algorithmic concepts, implemented the software, tested it, and prepared the manuscript. BS contributed to the implementation and testing of Protoss. MR initiated the development and supervised the project. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to express their thanks to the development team of BioSolveIT for the long standing cooperation in software development, especially Martina Brümmer for her help in creating a reliable, stable and well-tested Protoss executable.

Author details

¹Center for Bioinformatics (ZBH), Universität Hamburg, Bundesstr. 43, 20146 Hamburg, Germany. ²Current address: BioSolveIT GmbH, An der Ziegelei 79, 53757 St. Augustin, Germany.

Received: 6 January 2014 Accepted: 17 March 2014

Published: 3 April 2014

References

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The protein data bank.** *Nucleic Acids Res* 2000, **28**(1):235–242.

2. Forrest LR, Honig B: **An assessment of the accuracy of methods for predicting hydrogen positions in protein structures.** *Proteins: Struct, Funct, Bioinf* 2005, **61**(2):296–309.
3. Brünger AT, Karplus M: **Polar hydrogen positions in proteins: empirical energy placement and neutron diffraction comparison.** *Proteins: Struct, Funct, Bioinf* 1988, **4**(2):148–156.
4. Bass MB, Hopkins DF, Jaquysh WAN, Ornstein RL: **A method for determining the positions of polar hydrogens added to a protein structure that maximizes protein hydrogen bonding.** *Proteins: Struct, Funct, Bioinf* 1992, **12**(3):266–277.
5. Hooft RW, Sander C, Vriend G: **Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures.** *Proteins: Struct, Funct, Bioinf* 1996, **26**(4):363–376.
6. Word JM, Lovell SC, Richardson JS, Richardson DC: **Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation.** *J Mol Biol* 1999, **285**(4):1735–1747.
7. Li X, Jacobson MP, Zhu K, Zhao S, Friesner RA: **Assignment of polar states for protein amino acid residues using an interaction cluster decomposition algorithm and its application to high resolution protein structure modeling.** *Proteins: Struct, Funct, Bioinf* 2007, **66**(4):824–837.
8. Bayden AS, Fornabai M, Scarsdale JN, Kellogg GE: **Web application for studying the free energy of binding and protonation states of protein-ligand complexes based on hint.** *J Comput Aided Mol Des* 2009, **23**(9):621–632.
9. Labute P: **Protonate3d: assignment of ionization states and hydrogen coordinates to macromolecular structures.** *Proteins: Struct, Funct, Bioinf* 2009, **75**(1):187–205.
10. Molecular Operating Environment (MOE), 2013.08. Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7; 2013.
11. Daylight Theory Manual, version 4.9. Aliso Viejo: Daylight Chemical Information Systems, Inc.; 2008. [<http://www.daylight.com/dayhtml/doc/theory/index.html>] Accessed January 6, 2014.
12. Weininger D: **Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules.** *J Chem Inf Comput Sci* 1988, **28**(1):31–36.
13. Krieger E, Jr Dunbrack RL, Hooft RW, Krieger B: **Assignment of protonation states in proteins and ligands: combining pKa prediction with hydrogen bonding network optimization.** In *Computational Drug Discovery and Design*. New York: Springer; 2012:405–421.
14. Krieger E, Koraimann G, Vriend G: **Increasing the precision of comparative models with yasara nova—a self-parameterizing force field.** *Proteins: Struct, Funct, Bioinf* 2002, **47**(3):393–402.
15. Canutescu AA, Shelenkov AA, Dunbrack RL: **A graph-theory algorithm for rapid protein side-chain prediction.** *Protein Sci* 2003, **12**(9):2001–2014.
16. Urbaczek S, Kolodzik A, Fischer JR, Lippert T, Heuser S, Groth I, Schulz-Gasch T, Rarey M: **Naomi - on the almost trivial task of reading molecules from different file formats.** *J Chem Inf Model* 2011, **51**(12):3199–3207.
17. Lippert T, Rarey M: **Fast automated placement of polar hydrogen atoms in protein-ligand complexes.** *J Cheminf* 2009, **1**(1):13.
18. **PDB Format.** version 3.3. [<http://www.wwpdb.org/documentation/format33/v3.3.html>] (accessed Nov 6, 2012).
19. Urbaczek S, Kolodzik A, Heuser S, Groth I, Rarey M: **Reading pdb: perception of molecules from 3d atomic coordinates.** *J Chem Inf Model* 2013, **53**(1):76–87.
20. Gillespie RJ, Robinson EA: **Models of molecular geometry.** *Chem Soc Rev* 2005, **34**:396–407.
21. Urbaczek S, Kolodzik A, Rarey M: **The valence state combination model - a generic framework for handling tautomers and protonation states.** *J Chem Inf Model* 2014, **54**(3):756–766.
22. Rarey M, Kramer B, Lengauer T, Klebe G: **A fast flexible docking method using an incremental construction algorithm.** *J Mol Biol* 1996, **261**(3):470–489.
23. Feng Z, Chen L, Maddula H, Akcan O, Oughtred R, Berman HM, Westbrook J: **Ligand depot: a data warehouse for ligands bound to macromolecules.** *Bioinformatics* 2004, **20**(13):2153–2155.
24. Kellenberger E, Muller P, Schalon C, Bret G, Foata N, Rognan D: **sc-pdb: an annotated database of druggable binding sites from the protein data bank.** *J Chem Inf Model* 2006, **46**(2):717–727.

25. Hartshorn MJ, Verdonk ML, Chessari G, Brewerton SC, Mooij WT, Mortenson PN, Murray CW: **Diverse, high-quality test set for the validation of protein-ligand docking performance.** *J Med Chem* 2007, **50**(4):726–741.
26. Weininger D, Weininger A, Weininger JL: **Smiles. 2. algorithm for generation of unique smiles notation.** *J Chem Inf Comput Sci* 1989, **29**(2):97–101.
27. Dow RL, Schneider SR, Paight ES, Hank RF, Chiang P, Cornelius P, Lee E, Newsome WP, Swick AG, Spitzer J: **Discovery of a novel series of 6-azauracil-based thyroid hormone receptor ligands: potent, tr subtype-selective thymimetics.** *Bioorg Med Chem Lett* 2003, **13**(3):379–382.
28. Chang PK: **Synthesis of some 5-alkyl-6-azauracils 1.** *J Org Chem* 1958, **23**(12):1951–1953.
29. Schmidt CL, Kirk PL, Appleman W: **The apparent dissociation constants of arginine and of lysine and the apparent heats of ionization of certain amino acids.** *J Biol Chem* 1930, **88**(1):285–293.
30. Foloppe N, Fisher LM, Howes R, Kierstan P, Potter A, Robertson AG, Surgenor AE: **Structure-based design of novel chk1 inhibitors: insights into hydrogen bonding and protein-ligand affinity.** *J Med Chem* 2005, **48**(13):4332–4345.
31. Olsen JA, Banner DW, Seiler P, Obst Sander U, D'Arcy A, Stihle M, Müller K, Diederich F: **A fluorine scan of thrombin inhibitors to map the fluorophilicity/fluorophobicity of an enzyme active site: Evidence for c-f...c=O interactions.** *Angew Chem* 2003, **115**(22):2611–2615.
32. Terasaka T, Kinoshita T, Kuno M, Seki N, Tanaka K, Nakanishi I: **Structure-based design, synthesis, and structure-activity relationship studies of novel non-nucleoside adenosine deaminase inhibitors.** *J Med Chem* 2004, **47**(15):3730–3743.
33. Mol CD, Dougan DR, Schneider TR, Skene RJ, Kraus ML, Scheibe DN, Snell GP, Zou H, Sang B-C, Wilson KP: **Structural basis for the autoinhibition and sti-571 inhibition of c-kit tyrosine kinase.** *J Biol Chem* 2004, **279**(30):31655–31663.
34. Blaszczyk J, Li Y, Shi G, Yan H, Ji X: **Dynamic roles of arginine residues 82 and 92 of escherichia coli 6-hydroxymethyl-7, 8-dihydropterin pyrophosphokinase: crystallographic studies.** *Biochemistry* 2003, **42**(6):1573–1580.

doi:10.1186/1758-2946-6-12

Cite this article as: Bietz *et al.*: Protoss: a holistic approach to predict tautomers and protonation states in protein-ligand complexes. *Journal of Cheminformatics* 2014 **6**:12.

Publish with **ChemistryCentral** and every scientist can read your work free of charge

“Open access provides opportunities to our colleagues in other parts of the globe, by allowing anyone to view the content free of charge.”

W. Jeffery Hurst, The Hershey Company.

- available free of charge to the entire scientific community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
<http://www.chemistrycentral.com/manuscript/>



ChemistryCentral

An Integrated Approach to Knowledge-Driven Structure-Based Virtual Screening

[D7] A.M. Henzler, **S. Urbaczek**, M.Hilbig, and M. Rarey. An Integrated Approach to Knowledge-Driven Structure-Based Virtual Screening. *Journal of Computer-Aided Molecular Design*, 28(9):927-939, 2014.

<http://dx.doi.org/10.1007/s10822-014-9769-4>

Reproduced with permission from A.M. Henzler, S. Urbaczek, M.Hilbig, and M. Rarey. An Integrated Approach to Knowledge-Driven Structure-Based Virtual Screening. *Journal of Computer-Aided Molecular Design*, 28(9):927-939, 2014.
Copyright 2014 Springer Science + Business Media.

An integrated approach to knowledge-driven structure-based virtual screening

Angela M. Henzler · Sascha Urbaczek ·
Matthias Hilbig · Matthias Rarey

Received: 3 April 2014 / Accepted: 23 June 2014
© Springer International Publishing Switzerland 2014

Abstract In many practical applications of structure-based virtual screening (VS) ligands are already known. This circumstance requires that the obtained hits need to satisfy initial made expectations i.e., they have to fulfill a predefined binding pattern and/or lie within a predefined physico-chemical property range. Based on the RApid Index-based Screening Engine (RAISE) approach, we introduce cRAISE—a user-controllable structure-based VS method. It efficiently realizes pharmacophore-guided protein-ligand docking to assess the library content but thereby concentrates only on molecules that have a chance to fulfill the given binding pattern. In order to focus only on hits satisfying given molecular properties, library profiles can be utilized to simultaneously filter compounds. cRAISE was evaluated on a range of strict to rather relaxed hypotheses with respect to its capability to guide binding-mode predictions and VS runs. The results reveal insights into a guided VS process. If a pharmacophore model is chosen appropriately, a binding mode below 2 Å is successfully reproduced for 85 % of well-prepared structures, enrichment is increased up to median AUC of 73 %, and the selectivity of the screening process is significantly enhanced leading up to seven times accelerated runtimes. In general, cRAISE supports a versatile structure-based VS approach allowing to assess hypotheses about putative ligands on a large scale.

Keywords Structure-based virtual screening · Protein-ligand docking · Pharmacophore · Molecular properties · RAISE

Introduction

Virtual screening (VS) assists researchers in picking a few candidates from a vast amount of compounds giving a hint which chemical class of substances might be worth for optimization and further experimental testing. There exist various VS strategies [1]. Which strategy is deployed depends on the kind of information given in advance. Structure-based methods basically require a protein structure. Docking calculations predict the binding mode of a ligand that is assessed by scoring its protein-ligand interactions. In contrast to other VS approaches, the rather thorough assessment of compounds is at the expense of efficiency. Moreover, confronted with the well-known scoring problem, protein-ligand docking occasionally fails to predict the native binding mode [2] particularly, when protein flexibility is involved [3]. Pharmacophore-based strategies require a pharmacophore hypothesis given in advance. Meanwhile often applied in VS scenarios, the widespread feature-based models can be established from already known bioactive compounds, apoproteins, or protein-ligand complexes [4–8]. If a pharmacophore model compiles a few essential features representing commonly established protein-ligand interactions, the feature matching approach of pharmacophore-based VS is expedient to support fast compound selection. The scoring generally relies on geometric criteria assessing the alignment of the queried and matched features. Structure-based pharmacophore modeling offers the possibility to state excluded volume spheres and thereby to define a steric imprint of the

Electronic supplementary material The online version of this article (doi:10.1007/s10822-014-9769-4) contains supplementary material, which is available to authorized users.

A. M. Henzler · S. Urbaczek · M. Hilbig · M. Rarey (✉)
Center for Bioinformatics (ZBH), University of Hamburg,
Bundesstraße 43, 20146 Hamburg, Germany
e-mail: rarey@zbh.uni-hamburg.de

targeted protein. Since they geometrically limit the search space, VS gets more restrictive. However, in contrast to the atomic protein representation of classical structure-based methods, excluded volume spheres are generally porous and untyped i.e., they allow to roughly assess the shape but miss essential atom type information which is required to assess the electrostatic propensity of the screened compounds to bind to the target. If a pharmacophore hypothesis and a protein structure are both available, an integrated approach is motivated by observations made in several studies. It has been shown that combining docking with pharmacophore filtering improves binding mode predictions and the enrichment of actives [9–11]. Pharmacophore-based docking may therefore serve as an attractive alternative to substitute consecutive or parallel pharmacophore filtering and docking phases in screening projects. There already exist docking approaches that allow the propagation of pharmacophore hypotheses. Methods such as Gemdock [12], SP-Dock [13], and Gold [14] use the additional information to adapt their underlying scoring function. Additional terms examine the similarity of a posed ligand to the pharmacophore hypothesis giving rather similar poses a greater weight. As demonstrated by FlexX-Pharm [15], a pharmacophore hypothesis can also reduce the underlying search space. Incremental construction algorithms like FlexX [16] can discard partial solutions as soon as the given hypothesis cannot be fulfilled anymore. As a result, poses obeying the pharmacophore emerge and the guided approach can be applied in VS more efficiently.

Besides the observed synergetic effects with respect to prediction quality and efficiency, together with a pharmacophore-based docking engine, the highly interactive process of pharmacophore modeling can pave the way towards a user-directed VS process. With the development of cRAISE our main concern was to provide an externally controllable platform for structure-based VS. Herein we describe the methodology of cRAISE which is a completely redesigned adaptation of the TrixX approach [17, 18]. cRAISE is now based on the NAOMI framework [19], a robust chemical model which is designed to appropriately describe organic molecules relevant in the context of drug discovery. Nevertheless, cRAISE still captures the core idea of TrixX which postulates that a VS compares to a search that is only realized efficiently under the support of indexing-techniques. Essential search attributes, such as pharmacophore-like descriptors, are precalculated and stored in a way that allows to directly access relevant and omit irrelevant data during the search. The indexing requires costs and its benefit becomes apparent if multiple searches are performed. Moreover, an index requires that the prepared data remains unchanged throughout its complete lifetime. We assume that a typical large library, e. g.

collections of external vendor catalogs or in-house collections, hardly changes its content but is frequently queried with diverse target proteins—a screening scenario for that our approach is designed. Under this premise computational effort can be shifted to a preparative process that enables efficient, succeeding VS runs. The most probable conformations of compounds can be computed in advance, stored, and accessed later without traversing through the conformation space again. However, the aim of cRAISE to intervene in VS applications seems to be limited by the necessity of the TrixX approach to utilize a static compound library. Various screening projects may demand that the library content satisfies project-specific requirements, e. g. omit compounds that later will lead to experimental artifacts. Moreover, VS is often performed in iterations learning from and following-up on first round results. Once a screening result is obtained, analyzed, and the molecules retrieved show properties not corresponding with the expectations, it arises the need to adapt initially made hypotheses. Opposed to the former implementation cRAISE now offers a broad range of search possibilities in order to avoid a recalculation of the index with a restricted library in such situations. It enforces guided docking runs when a pharmacophore hypothesis is stated. The additional information tailors the search space as soon and as much as possible. Moreover, molecular library profiles about constitutional or topological ligand features can be stated and utilized to gain further external control over the VS process.

The results of a method requiring external knowledge strongly depend on the provided information. Nevertheless, in order to reveal insights how cRAISE can be controlled and how it reacts on the given information, we automatically derived pharmacophore models covering a range of strict to rather relaxed model definitions. Utilizing this data, we evaluated our method to demonstrate the directionality of the pharmacophore-driven approach i.e., its capability to suggest solutions that meet the externally made expectations, which was our major design goal. Within our study we could also observe the synergetic effect of pharmacophore-guided docking and thus confirm the results that have been already stated by others. The pharmacophore models were derived from standardized forms of the Astex Diverse [20] and the DUD [21] datasets that have been previously used to comparatively assess the most popular docking algorithms. A complete issue of this journal addresses the competition to which our results can be directly compared. [22–29] On the given datasets the predictions of those methods strongly depended on the data preparation and the utilized docking protocol, thus, mean AUC values ranging between 59 and 80 % were reported. Our guided redocking and enrichment studies on this data show that if a pharmacophore model is chosen appropriately, a binding mode

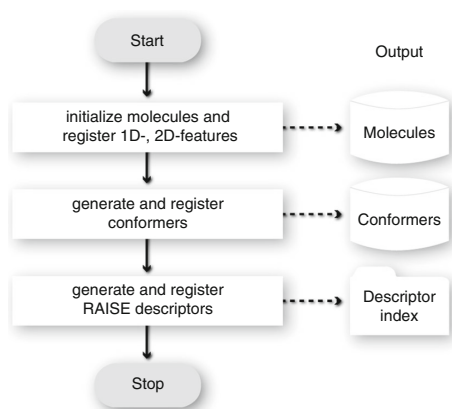


Fig. 1 Library preparation workflow: precalculation of constitutional and topological molecule features, conformations, and RAISE molecule descriptor indexing. The output is reused throughout succeeding VS runs

below 2 Å is successfully reproduced for 85 % of well-prepared structures. Compared to unguided predictions, we were able to increase enrichment with our hybrid screening approach resulting to a median AUC of 73 % with automatically derived pharmacophore models and there is still room to enhance the enrichment further with more sophisticated models. Benchmark studies on subsets of the ZINC database [30, 31] show that external knowledge in form of pharmacophore models and molecular profiles enhance the selectivity of the screening process leading up to seven times accelerated runtimes. All in all, our method provides a versatile tool to intervene in structure-based VS by means of pharmacophore hypotheses and library profiles. Thereby it allows to encounter the generally conflicting aims of structure-based VS that requires choosing a trade-off between accuracy and efficiency when utilizing large-scaled molecular libraries.

Methods

Overview

cRAISE is a two-tiered procedure. In the preparatory phase, molecular feature detection, conformational sampling [32], and descriptor generation for the given compound library is realized. The features and conformations are stored in a database, the descriptors in a bit-compressed index both remaining static throughout subsequent VS runs (see also Fig. 1).

As illustrated in Fig. 2, the screening phase derives combined spatial and physico-chemical RAISE descriptors from a given protein active site that are translated to

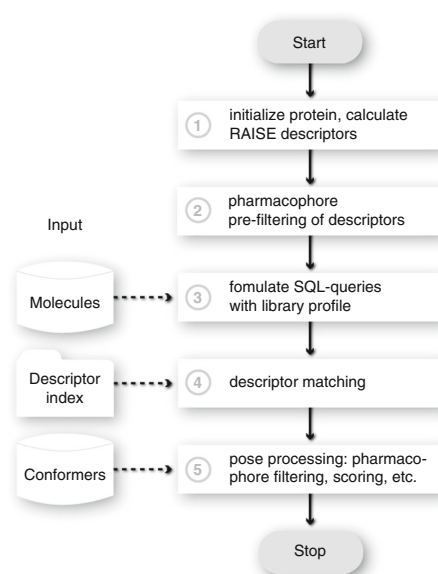


Fig. 2 Screening workflow: RAISE protein descriptor calculation, SQL-query generation, descriptor matching, pose generation, scoring. A pharmacophore hypothesis affects step 1, 2, and 5, a library profile step 3 and 4 of the workflow

SQL-like queries. Using the compressed bitmap index structure, the queries detect matching molecule descriptors. The conformers of the matches are fetched from the database, placed into the active site by descriptor superimposition, and scored keeping only the best pose of a compound for the final hit list of the screening run. An optional pharmacophore hypothesis is used to restrict the set of queries and furthermore, to early reject poses before they are actually scored. A library profile is directly encoded in the query such that violating molecules are never fetched.

The descriptor and concepts behind the indexing and matching phase have been described before [17]. Here we focus on the processing of pharmacophore hypotheses and library profiles to support externally guided VS.

cRAISE docking and virtual screening

cRAISE places ligands into a targeted protein active site by aligning complementary interaction sites. The determination of interaction sites is now based on the NAOMI model [19]: for hydrogen-bond donor or positively charged atoms, donor sites are placed at a distance of an idealized hydrogen bond (2.8 Å) away from the heavy atom center into the directions of their protons. Acceptor sites remain on the center of a hydrogen-bond acceptor or a negatively charged

heavy atom. Donor and acceptor sites possess attached directions indicating the orientation of protons or lone pairs and thus, possible interaction directions. Hydrophobic sites are undirected and reside on aliphatic and aromatic regions of small molecules. They are placed at carbons of acyclic aliphatic chains, at halogens, and on centers of aliphatic and aromatic rings. For adjacent carbon atoms, only a single site is created in the middle of the bond. In branched regions where a tertiary or quaternary central carbon is bound only to carbons, a single site is created in the center of the branch. The hydrophobic protein active site counterparts reside in volumes with a mostly hydrophobic environment. They are calculated by probing the active site volume with methyl-like representatives that are assessed by a classical Lennard–Jones (12, 6) potential. Surrounding hydrophilic atoms contribute to repulsion, but do not contribute to the attractive part of the potential. The top-scored representatives are selected and converted into hydrophobic interaction sites. Basically, cRAISE identifies an interaction in a protein–ligand complex if complementary interaction sites properly align in three-dimensional space i.e., if a donor site covers the site of an acceptor and if both sites possess roughly opposite interaction directions. Hydrophobic interaction sites do not have to fulfill the direction criterion.

Each triplet of interaction sites forms the corners of a triangle, the basis of the RAISE descriptor. A corner encodes the type (donor, acceptor, or hydrophobic) and obtains the associated interaction direction(s) of hydrophilic interaction sites. The descriptor additionally stores the lengths of the triangle sides. Some constitutional and geometric criteria ensure triangle angles being not too acute. Furthermore, each triplet must contain at least one hydrophilic corner. A special feature of the RAISE descriptor is that it encodes molecular shape relative to pharmacophoric features in a transformation invariant fashion. This is achieved with the lengths of 80 steric bulk rays that originate from the center of the triangle. The rays locally describe the van-der-Waals volume of a molecule or the interior volume of an active site. In order to decide whether a molecule fits into the active site, descriptor features can be simply compared. A descriptor match is recognized if complementary triangle corner types, opposite interaction directions, similar triangle side lengths, and an inclusion of all of the 80 ligand bulk rays in their respective active site descriptor counter-parts is detected. Then cRAISE accesses the molecule of origin designated by the descriptors compound/conformation ID from the molecule database. The coordinates of the triangle corners are used to calculate an affine transformation that superposes a pair of matching molecule and active site triangles. The transformation is applied to the molecule producing the actual pose. The basic idea of the RAISE screening

Table 1 Supported library profile features

Type	Features
Range ^a	Total charge, molecular weight, volume, topological polar surface area (TPSA), calculated octanol/water partition coefficient (logP), number of heavy atoms, hetero atoms, hydrogen-bond donors, hydrogen-bond acceptors, aromatic atoms, halogenic atoms, total number of bonds, rotatable bonds, maximum number of continuous rotatable bonds, number of ring systems, individual rings, aromatic rings, maximal ring size, maximal ring system size, number of stereo centers
Existence ^b	Chemical elements of the periodic table, any predefined molecular pattern (SMARTS), common functional groups (alcohol, ether, ketone, aldehyde, ester, amine, amide, amidine, guanidine, azide, nitrile, pyrrole, furan, thiophene, phenyl, pyridine)

^a Features registered and evaluated on a value range

^b Features registered and examined for existence

procedure is to avoid evaluating each molecule descriptor. This is achieved with an efficient bitmap indexing and compression system [33, 34]. Essentially, cRAISE performs rigid body docking. Molecular flexibility is introduced with conformers generated with an integrated conformer sampling method based on CONFECT [32]. CONFECT was reparameterized for the cRAISE docking methodology: for rather small and rigid compounds slight suboptimal conformers are generated in order to increase the chance of a shape fit. However, to keep a large-scale application tractable it is necessary to provide an upper bound for the number of generated conformers [35]. Thus, the conformation set is restricted to at most 250 conformers per compound. For rather large and flexible compounds, a k-medoid cluster algorithm using the TFD [36] as a distance measure, samples rotatable bonds rather granular and selects diverse representatives out of the conformation space.

Integration of library profiles

The molecular feature handling of cRAISE is based on functionalities of MONA [37], a tool for visualization and statistical analysis of molecular libraries. Constitutional and topological features of compounds are calculated during the library preparation step and stored in the molecule database. Basically, the registered features can be categorized according to the kind of supported query. Table 1 summarizes all supported molecular features.

In order to support a guided VS run, a library profile can be defined by an arbitrary combination of feature range and existence conditions. As soon as a profile is given, cRAISE determines the IDs of compounds that are in accordance with the conditions prior to descriptor matching. The IDs constrain the SQL-queries and enforce a fetching of

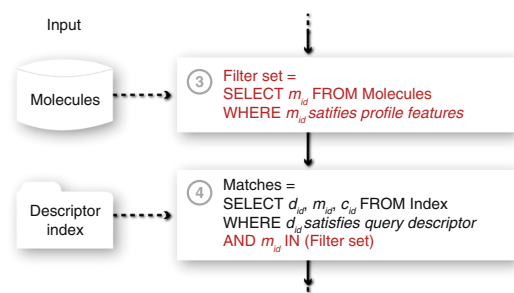


Fig. 3 For each protein descriptor, a library profile reformulates its SQL-query to select respective matches from the descriptor index (step 3 and 4 of the screening workflow). d_{id} , m_{id} , c_{id} denote descriptor, molecule, and conformer IDs, respectively. They enable proper compound/conformer selection for pose initialization

Table 2 Supported pharmacophore features and their interpretation during pose sampling

Type	Feature interpretation
Donor inclusion ^a	Place only an H-bond donor/cation center with proper proton direction here
Acceptor inclusion ^a	Place only an H-bond acceptor/anion center with proper lone pair direction here
Hydrophobic inclusion ^b	Place only a hydrophobic group here
Hydrophilic inclusion ^a	Place an H-bond donor/acceptor/cation/anion center with proper proton/lone pair direction here
Any inclusion ^b	Place any atom center here
Donor exclusion ^b	Do neither place H-bond donor nor cation atom centers here
Acceptor exclusion ^b	Do neither place H-bond acceptor nor anion atom centers here
Hydrophobic exclusion ^b	Do not place hydrophobic atom centers here
Hydrophilic exclusion ^b	Do neither place H-bond donor, acceptor, cation, nor anion atom centers here
Any exclusion ^b	Do not place atom centers here

^a Directed feature

^b Undirected feature

appropriate descriptor matches from the index. Figure 3 visualizes this process.

Integration of pharmacophore hypotheses

cRAISE supports the specification of pharmacophore-type inclusion and exclusion features directly influencing its pose sampling stage. An inclusion feature is a constraint defining a region in the protein active site where a ligand atom has to reside. Exclusion features define forbidden

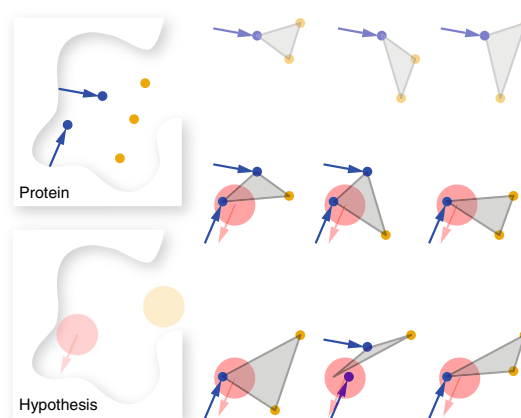


Fig. 4 A pharmacophore hypothesis is locally tested on query construction (step 2 of the screening workflow): an acceptor feature (red) restricts the calculation of query descriptors. Only active site descriptors with a complementary donor (blue) corner in the feature sphere and an opposite direction are built. Descriptors that would be built without the hypothesis are omitted (grayed out). Hydrophobic (yellow) features are not evaluated at this stage

regions of the active site. Each feature has either the type *donor*, *acceptor*, *hydrophobic*, *hydrophilic*, or *any* and is represented by a sphere defined by center and tolerance radius. Some features are directed to further constrain the location of a ligand atom with an appropriate proton or lone pair orientation. Table 2 summarizes all kinds of supported feature types and describes which constraints are enforced during the cRAISE pose sampling stage. A pharmacophore hypothesis can be defined by an arbitrary set of inclusion and exclusion features. Additionally, the number of essential inclusion features N_e states how many inclusion features have to be fulfilled simultaneously by a placed ligand.

Predefined pharmacophore features are used at two stages of the screening process: (1) prior to descriptor matching to reject pharmacophore violating query descriptors and (2) during the post-matching phase to reject pharmacophore-violating poses. Figure 4 demonstrates the effect of a pharmacophore hypothesis on query construction. Only triangles with at least one corner contained in a *donor*, *acceptor*, *hydrophilic*, or *any* inclusion feature are built. Since RAISE descriptors cover molecules only locally, false negative predictions could occur if one enforces more than one inclusion feature at this stage. Hydrophobic features do not restrict query descriptors but are evaluated in the post-matching phase. In consequence, hypotheses stating only hydrophobic features do not influence the query construction at all.

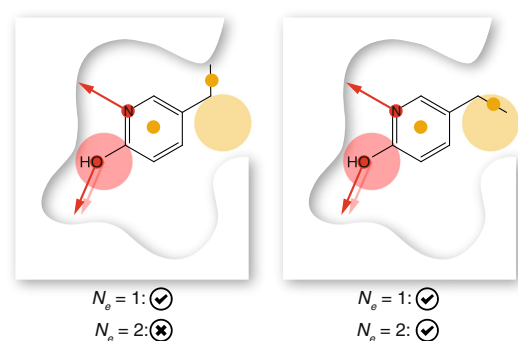


Fig. 5 A pharmacophore hypothesis is globally tested for poses (during step 5 of the screening workflow). *Left* the pose satisfies a single feature but violates the hypothesis if it requires two features. *Right* another pose satisfies the hypothesis

Figure 5 shows how a hypothesis is tested in the post-matching phase. All inclusion features are tested and if at least N_e features are fulfilled, the pose globally satisfies the hypothesis. This simple approach realizes the following task: If N_e equals the number of inclusion features defined by the user any of the generated poses need to obey the complete pharmacophore model, otherwise, for each pose all possible feature combinations of the given size are tested until either a single or none of them is fulfilled.

Hierarchical pose filtering and scoring scheme

A docking engine applied for large-scale VS produces a large amount of poses. The direct use of elaborate scoring functions too early without prior pose reduction hinders the throughput and eliminates the advantage in speed gained by the non-sequential screening paradigm of RAISE. The hierarchical pose-filtering scheme introduced here is intended to efficiently eliminate poses with sparse contacts, clashes, and pharmacophore violating poses as much and as soon as possible and to rapidly assess the quality of fit for succeeding poses. Initiating a VS run, cRAISE calculates information relevant for pose evaluation in advance. It determines an active site volume by computing the convex hull [38] from the active site atoms. A fine granular clash grid for hydrophilic and hydrophobic probes (0.25 Å voxel spacing) detects a clash for a grid point if the probe sphere contains any atom center. Moreover, a probe is assessed with its surrounding protein atoms and individual score contributions, namely possible protein counter interaction directions and Lennard–Jones-like potential values, are annotated at the grid. If a pharmacophore hypothesis states an exclusion feature, it is quasi seen as a protein atom sphere and grid points therein are flagged as clashes of the respective feature type.

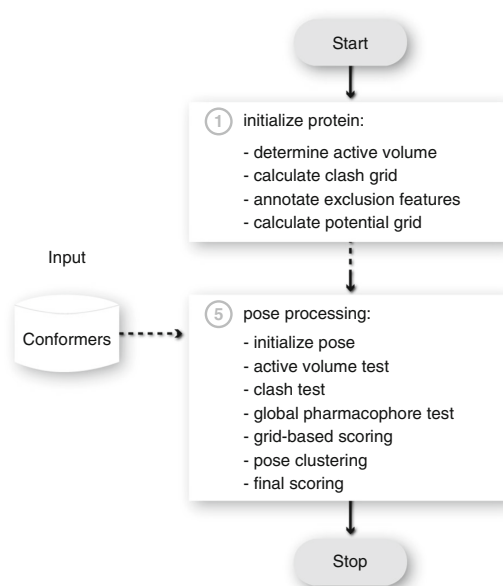


Fig. 6 Step 1 and 5 of the screening workflow: Relevant information is precalculated (*step 1*) and accessed later for pose evaluation (*step 5*)

The actual pose processing starts with two initial placement tests: to avoid sparse contacts with the protein, the majority of pose atoms has to reside in the precalculated convex hull after transformation. Ligand atoms clashing into the protein or intruding into exclusion volumes are rapidly identified using the clash grid. Then, each inclusion feature is tested until the number of fulfilled inclusion features N_e is achieved. The initial scoring stage accesses individual potentials for each atom and evaluates opposed interaction directions. These contributions compile the complete score for a ligand. It estimates the quality of fit by an empirical scoring function that accounts for hydrogen bonds, metal interactions, lipophilic contacts, and the loss of torsional entropy of the ligand. Essentially, the function is the Boehm scoring function [39], which was recalibrated on the Iridium Highly Trustworthy dataset [40] (v1.1) for which K_i/K_d values are published. Instead of piecewise linear penalty functions, Lennard–Jones-like potentials honor good and penalize close atom contacts (see Supplementary Material). The cosine assesses the angle deviations from the ideal geometry of opposed interaction directions. After ranking the poses by this score, similar poses closer than 0.5 Å RMSD to higher ranked ones are eliminated. Eventually, the pose-processing phase captures the scoring discrepancies that might occur due to grid mapping. It re-ranks the poses according to the cRAISE scoring function but this time evaluated on exact pose atom coordinates. The best pose for each compound

contributes to the final hit list of the screening run. Figure 6 summarizes the individual preparation and pose evaluation steps.

Results and discussion

Datasets

For evaluating the pharmacophore-guided binding mode predictions and screening performance, data sets provided by the organizers of the ACS docking symposium 2011 were used. Several docking tools and scoring functions have already been evaluated with these standardized sets [22–29]. In the following, we refer to the datasets as Astex_{ACS} and DUD_{ACS} respectively. The Astex_{ACS} set comprises crystal structures of 85 protein targets of the Astex Diverse Set [20] with rerefined protein heavy atom, hydrogen atom, and ligand coordinates. For monomeric structures only a single ligand is provided as reference for active site definition, while for multimeric structures all ligands are supplied. Taking protein atoms with a heavy atom distance of 6.5 Å from any ligand atom center into account, all in all 146 well-resolved active site definitions can be obtained from the dataset. The symposium organizers provided a non-crystallographic structure for each ligand as starting point for docking calculations in order to support an objective, comparable evaluation. We will report values that take all ($n = 146$) and only a single, namely the qualitatively best site ($n = 85$) into account. Since the quality of the multiple active site copies differs, the values will provide a hint at what the precision of our method is and moreover, will provide comparability to the other, already published methods that have been evaluated with this dataset and reported the values, as well. The DUD_{ACS} set covers active and decoy ligands for the 40 different targets of the DUD dataset [21]. The protein structures have been rerefined by the symposium organizers as well but opposed to the targets of the Astex_{ACS} set, the targets of the DUD_{ACS} set retained their key crystal waters. The Supplementary Material summarizes further corrections made to the datasets. Our large-scale studies were carried out with subsets of the ZINC database [30, 31]. From the ZINC clean leads subset [41] comprising 4,230,832 compounds at access time, we randomly selected one, two, and three millions of unique compounds. We further refer to these libraries as ZINC_{CL1M}, ZINC_{CL2M}, and ZINC_{CL3M} set, respectively.

Definition of pharmacophore hypotheses

Pharmacophore models introduce a strong bias in docking calculations. Although this is intended in practical

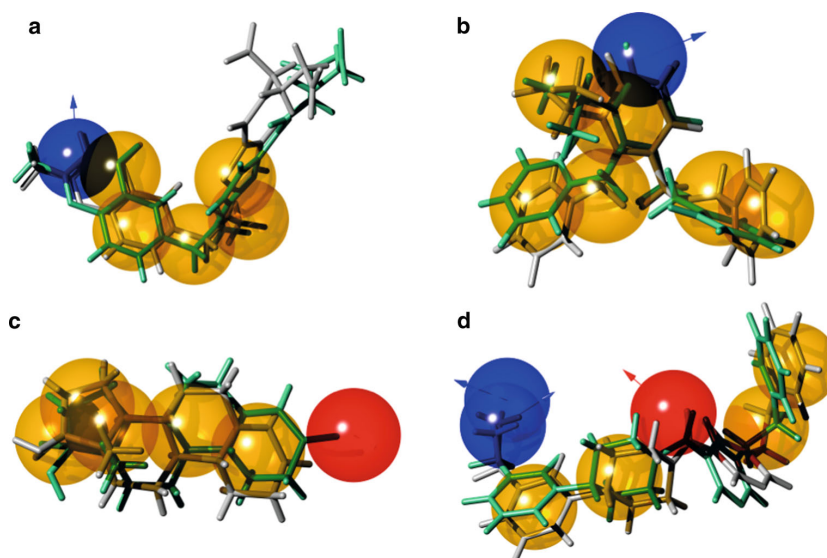
applications, it causes major deficiencies in validation studies with respect to objectivity and reproducibility. We therefore decided to derive them automatically from individual protein-ligand complexes. The method initially places inclusion features centered at sites complementary to protein interaction sites. Tolerance spheres are scaled to a 1.7 Å radius, a common default value for structure-based models. To identify features that propagate binding, then only those that are complementary to the given ligand with respect to interaction type and geometry are selected. In order to create a scenario close to practice, feature sphere centers are not positioned on ligand atoms. Instead, they are derived relatively to protein atom coordinates. We generated various versions of each model linearly decreasing the number of inclusion features N_e being essential, i.e. are expected to be fulfilled. A relaxation of N_e has two practicable applications: First, a structure-based model derived from protein atom coordinates can result in feature constellations that require tensed or even unrealistic ligand conformations for entire fulfillment. Second, usually not all features are obeyed by all binding ligands in a VS scenario. Relaxed models allow to implicitly account for both situations and to explore only subsets of feature combinations of size N_e during the pharmacophore matching phase.

Pharmacophore-guided binding mode predictions

With the Astex_{ACS} set and automatically derived models for these structures we performed pharmacophore-guided binding mode predictions. Figure 7 exemplarily depicts the guided top predictions found for four complexes of the Astex_{ACS} set (green). They were predicted on lower ranks if the placement procedure was not guided by any feature. cRAISE does not explicitly change the score of a pose based on pharmacophore information but discards poses that contradict the given features. As a result, guided predictions let pharmacophore fulfilling poses emerge on higher ranks if the unguided prediction does not already rank a fulfilling pose on top. These observations show how guided binding mode predictions implicitly exert leverage on pose ranking.

Sometimes the automatically derived, structure-based models require extremely tensed conformations for optimal fulfillment. We explicitly neglect to put features on reference ligand coordinates, a procedure that would capture such situations. Instead, we implicitly relieve some tension during pharmacophore matching by relaxing the number of demanded features N_e . If a few feature definitions geometrically contradict each other within the model, the relaxation enables a recovery of near native poses that would require unrealistic ligand coordinates for matching all features.

Fig. 7 Pharmacophore-guided binding mode predictions (*green*) are identified close to native binding modes (*gray*) on higher ranks. Donor inclusion (*blue*), acceptor inclusion (*red*), hydrophobic inclusion (*gold*). **a** 1hvy_3 at rank 1 (unguided 147), **b** 1jla_1 at rank 1 (unguided 5), **c** 1sqn_1 at rank 1 (unguided 5), **d** 2bm2_2 at rank 1 (unguided 143)



In order to quantify the effect of pharmacophore-driven binding mode prediction for the entire Astex_{ACS} set, we characterized a successful prediction as a reproduction of a pose with a root mean square deviation (RMSD) of less than 2.0 Å to the respective reference ligand. A partially predicted pose was characterized by an RMSD above 2.0 Å, but below 3.0 Å. The total success rate was defined as the percentage of successful reproductions on all complexes. Figure 8 plots the success rates of the top sampled cRAISE poses for linearly increasing the number of demanded inclusion features. A tendency to guide binding mode predictions is observed if the poses have to satisfy up to 80 % of the demanded inclusion features (blue bars). Partially predicted poses and docking failures are reduced (green bars). However, for some structures the models were already too strict to successfully recover a pose at all (red bars) and the trend to direct top predictions further by demanding more features being fulfilled is reversed. Those failures can be captured by appropriate N_e relaxation which is part of the pharmacophore elucidation process before a guided docking with cRAISE can be accomplished. In order to show what our method can potentially achieve if this task is realized properly, we selected a good model for each target i.e., N_e^g which minimizes the RMSD of the top prediction (N_e^g -bars). On the Astex_{ACS} set, the number of features was typically relaxed by 5–25 % to allow the poses to fulfill tensed models at least partially. Table 3 summarizes the guided success rates on the Astex_{ACS} set with appropriately relaxed models and compares it with unguided predictions. A paired *t*-test was used to assess the significance of the comparisons. Therefore, we compared

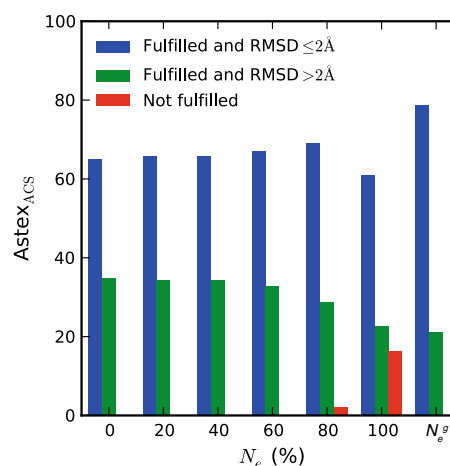


Fig. 8 Success rates of top predictions linearly increasing the number of demanded inclusion features N_e (blue bars). Less top predictions deviate from the native binding mode if they have to satisfy more pharmacophore features (green bars). Too strict model definitions lead to docking failures if the features cannot be satisfied by any pose (red bars). A relaxation of N_e recovers those failures (N_e^g -bars)

the paired RMSD differences within various rank cutoffs (for all protein-ligand complexes docked without any and with the use of the N_e^g -model). Under the assumption that the paired differences are independent and identically normally distributed, the probability p that the difference between the guided versus the unguided top predictions is random is far below 0.001. With increased rank cutoff value successively increases since the chance of finding an identical pose within those ranks by the unguided docking

Table 3 Pharmacophore-guided and unguided pose sampling success rates (%) of the Astex_{ACS} $n = 85$ ($n = 146$ in braces)

Rank	$\leq 1.0 \text{ \AA}$	$\leq 2.0 \text{ \AA}$	$\leq 3.0 \text{ \AA}$	p
Guided				
1	35 (32)	85 (80)	97 (95)	<0.001
5	41 (40)	91 (87)	97 (95)	0.002
20	48 (45)	93 (91)	99 (99)	0.151
32	51 (47)	93 (93)	100 (99)	0.227
All	52 (49)	95 (95)	100 (100)	0.562
Unguided				
1	29 (25)	71 (64)	84 (82)	–
5	38 (36)	86 (81)	94 (95)	–
20	46 (44)	87 (84)	97 (98)	–
32	47 (45)	91 (88)	99 (98)	–
All	55 (51)	97 (97)	99 (98)	–

Paired t -test p was determined for the complete dataset

runs increases as well. Moreover, the success rates reveal that the pharmacophore fulfilling poses with an RMSD $>2 \text{ \AA}$ (green bars) mostly correspond to partially docked ligands. They are the result of pharmacophore models guiding the prediction by features covering poses only locally (compare e. g. the model of Ihvy in Fig. 7a). As a result, these models allow the remaining flexible ligand portion to freely explore unconstrained regions of the binding site. In general, our observations suggest that pharmacophore-guided binding mode prediction directs pose sampling and appropriately influences pose ranking.

Pharmacophore-guided enrichment studies

To assess the enrichment performance under pharmacophore type constraints we automatically derived models from the initially given 40 protein-ligand complexes of the DUD_{ACS} set as described above. We performed pharmacophore-guided VS runs on the libraries consisting of the respective actives and decoys sets. Thereby, the enriched hits had to share at least a linearly increased amount of common features. The total area under (AUC) the receiver operating characteristic (ROC) curve served as a measure for the discriminative power of our method to separate actives from decoys. We additionally determined the true positive rate at a false positive rate of 1 and 2 % of the ROC showing the ability of our method to enrich actives early. The total enrichment performance was determined by averaging the AUC values of all 40 screening runs.

Since our models were derived from individual protein-ligand complexes i.e., determined from a single active, their features depict a superset of the real pharmacophore which is a particular feature combination therein. A

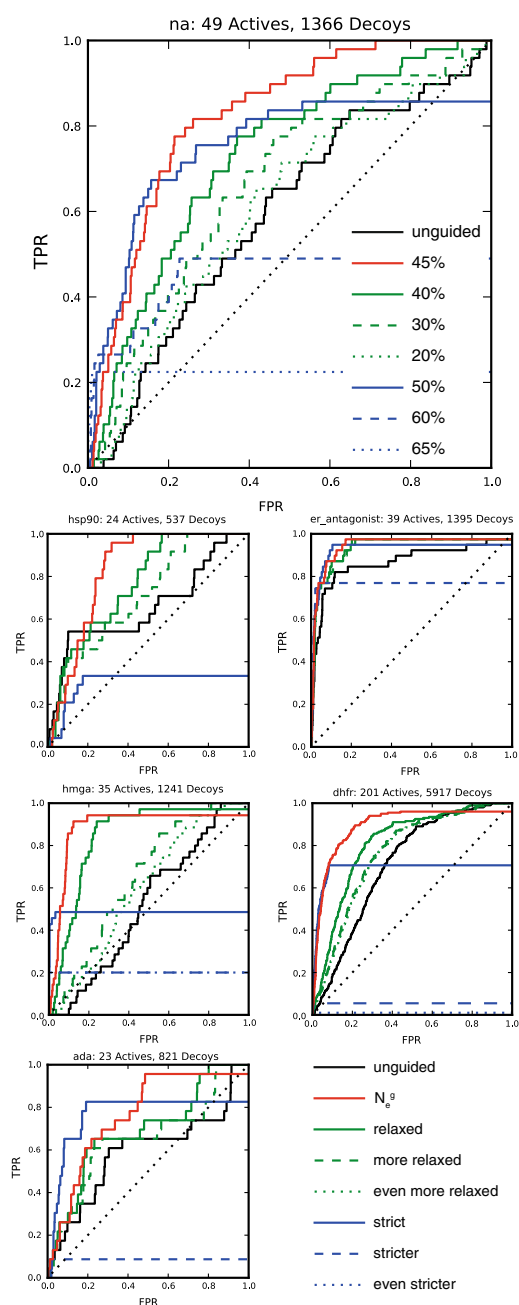


Fig. 9 Impact of strict and relaxed model definitions on enrichment behavior of the neuraminidase (na, $N_g^s = 45\%$), the human heat shock protein 90 kinase (hsp90, $N_g^s = 75\%$), the estrogen receptor antagonist (er_antagonist, $N_g^s = 65\%$), the hydroxymethylglutaryl-CoA reductase (hmga, $N_g^s = 65\%$), the dihydrofolate reductase (dhfr, $N_g^s = 55\%$), and the adenosine deaminase (ada, $N_g^s = 75\%$)

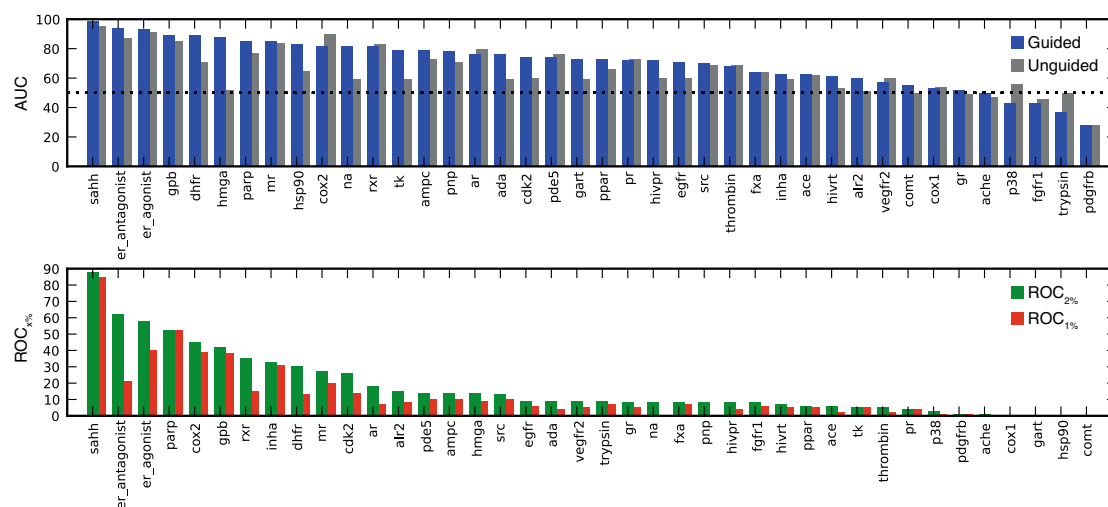


Fig. 10 ROC_{1%}, ROC_{2%}, and AUC for the N_e^g -models of the individual DUD_{ACS} sets

stringent model demanding all features being fulfilled identifies solely actives with that specific binding pattern. A relaxed model allows to introduce some tolerance and to explore various feature combinations during the pharmacophore matching phase. The degree of relaxations influences the enrichment of actives. This is demonstrated in Fig. 9 by example of six DUD_{ACS} targets. Demanding more features being fulfilled, the global enrichment with respect to the AUC metric is improved (green curves). In contrast, strict models enrich only subsets of actives—however, most often earlier (blue curves). All in all, these results show that cRAISE VS can be externally controlled by utilizing pharmacophore hypotheses.

It was not our goal to provide a tool for pharmacophore elucidation but for externally guided VS if well-prepared pharmacophore hypotheses have been already stated. In order to show what our screening method can potentially achieve in this case, in analogy to the above described guided docking experiments, we selected a good model for each target i.e., N_e^g , but this time with respect to maximize the global enrichment (Fig. 9, red curves). Basically, N_e^g represents an upper bound for the size of the actual pharmacophore. In our experiments this value ranged between 10 % and 95 % and there is still room to further improve the models, namely, if one determines the perfect N_e^g -combination that is able to recognize all actives. The following results represent realistic key figures what to expect from pharmacophore-guided VS. In Fig. 10 enrichment values for VS runs on the DUD_{ACS} sets guided by these models are plotted (blue bars). Most often guided VS significantly improves the early as well as the global

Table 4 Pharmacophore-guided and unguided enrichment performance on the DUD_{ACS} sets

	ROC _{1%}	ROC _{2%}	AUC
Guided			
Mean	0.123 (±0.055)	0.177 (±0.064)	0.704 (±0.052)
SD	0.173	0.201	0.164
Median	0.060	0.090	0.730
Min	0.000	0.000	0.280
Max	0.850	0.880	0.990
Unguided			
Mean	0.087 (±0.041)	0.142 (±0.059)	0.651 (±0.047)
SD	0.129	0.185	0.145
Median	0.050	0.090	0.610
Min	0.000	0.000	0.280
Max	0.520	0.700	0.950

Error ranges represent 95 % confidence limits

enrichment of actives with respect to unguided predictions (gray bars). In case of hmga the pharmacophore model turns screening towards a highly directive process resulting in an AUC that is maximally improved by 36 %, while unguided VS enriches actives only close to random. A few models decreased the AUC, most often, if our model generation was confronted with problematic complexes that lead to ambiguous feature definitions. Table 4 summarizes the performance on the complete dataset and shows the statistical information on the enrichment metrics including 95 % confidence limits on the mean metrics.

Table 5 Number of pharmacophore-guided and unguided query descriptors

	Guided	Unguided
sahh	5,678	10,677
gpb	15,433	37,579
hsp90	2,934	19,637
fxa	3,996	31,510
er_agonist	7,756	13,042
dhfr	7,491	36,943

Large-scale studies

We conclude with guided and unguided large-scale VS studies and runtime evaluations on subsets of the the ZINC comprising one, two, and three million compounds, respectively. For these lead-like libraries on average 239 conformations per compound and 108 descriptors per conformation were generated and stored in the ZINC_{CL1M}, ZINC_{CL2M}, and ZINC_{CL3M} indices. Since it affects the runtime of cRAISE, we chose representative targets of the DUD_{ACS} set reflecting lower and upper bounds with respect to the number of query descriptors. Pharmacophore models showing promising capabilities in the above enrichment studies were used to guide the following VS runs. Table 5 summarizes the target data employed in the experiments. Inclusion features of type donor, acceptor, or hydrophilic restrict the number of query descriptors of the targeted protein active site. Our models contain on average four of these features which reduce the ordinary queries (on average 30,000) by around three quarters. Basically, N_e does not affect the number of queries and increasing this number does not restrict the search space further. To verify how many inclusions are satisfied, the poses have to be actually built. However, increasing N_e reduces the kept poses forwarded to the scoring stage. Thus, a stricter pharmacophore model can save expensive scoring calculations.

All computations were performed in a parallel screening setting on a high performance cluster of 25 Intel Xeon CPU E5630 dual quad core nodes with 2.53 GHz. Each process consumed maximally 8 GByte of main memory. The ZINC_{CL1M}, ZINC_{CL2M}, and ZINC_{CL3M} indices were split into packages of 2,500 compounds (à 6.9 GByte) each and were distributed over the local hard drives of each cluster node in order to reduce the network load during a VS run. Our measured runtimes are given in form of wall clock times. The processing time t_c for a single compound is given by $t_c = t_{\text{total}}/N$, where N is the number of given compounds and t_{total} is the total VS time i.e. the sum of the wall clock times of all distributed jobs. t_c allows to

estimate the cRAISE screening time independent of the parallel setup and the size of the employed compound library. The processing time t_m for a single conformation is given by $t_m = t_{\text{total}}/M$, where M is the number of generated conformations. It allows to estimate the cRAISE screening time if externally provided conformers are processed. The parallel run time t_p is variable due to the current availability of compute nodes. Thus, it was estimated by the average of the VS time of individual jobs on basis of an optimal availability of 200 cores. Then, t_p reflects the best possible run time in a parallel setting of 25 freely available dual quad core nodes.

The observed timings of our large-scale experiments are summarized in Table 6. We achieved an up to seven times accelerated run time with pharmacophore hypothesis guiding the screening process. Basically, the runtime varies from target to target. The cause is found in the selectivity of the query: If an index contains n descriptors, a single query descriptor has the potential to extract all of them. If a target possesses m query descriptors, in the worst case, a screening produces $n \times m$ descriptor matches. Even if this worst-case scenario never occurs, the run time depends on how many poses are actually processed in the post-matching phase i.e. on the amount of extracted index descriptors. Table 6 shows the observed selectivity values $\sigma = \text{\#matches}/\text{\#index descriptors}$ for the targets employed in the VS runs. A selectivity of 1 indicates that the whole index is extracted. These values correlate with the observed runtimes. Guiding a VS by a pharmacophore hypothesis generates queries that are more selective and explains the accelerated run time behavior of cRAISE.

Molecular profiles—an example

The definition of molecular profiles allows to further guide the VS process with respect to retrieve only hits satisfying user-defined molecular properties. For the sake of completeness we demonstrate here a simple screening scenario: The ZINC_{CL3M} library contains three millions of lead-like compounds. We defined a molecular profile with MONA restricting this library to molecules with a molecular weight of at most 300, maximally 5 rotatable bonds, and a logP of at most 3.5 (provided in the Supplementary Material). It was utilized to determine the runtime if the ZINC_{CL3M} library is simultaneously filtered during a VS run of er_agonist. The retrieved timings were as expected. The profile indirectly reduced the library by 75 to 707,770 % compounds and the timings ($t_c = 2.06$ s, $t_p = 9.25$ h) were reduced by approximately the same amount. This experiment verifies that library profiles can be used ad hoc during a VS run without the necessity to rebuild the static index for a restricted library.

Table 6 Timings on the ZINC_{CLIM/2M/3M} sets

	Guided					Unguided				
	t_m (s)	t_c (s)	t_p (h)	(1M/2M/3M)	σ	t_m (s)	t_c (s)	t_p (h)	(1M/2M/3M)	σ
sahh	0.01	1.20	1.66/3.32/5.00		0.07	0.01	2.75	3.65/7.13/10.40		0.13
gpb	0.01	2.80	3.73/7.73/11.70		0.35	0.04	9.73	12.58/26.22/38.78		0.87
hsp90	0.01	1.23	1.66/3.43/5.15		0.21	0.04	8.89	13.05/25.75/37.03		1.38
fxa	0.01	3.32	4.38/8.93/13.87		0.75	0.09	21.24	30.37/59.00/89.37		5.40
er_agonist	0.02	5.68	7.62/15.10/24.05		1.28	0.03	8.23	10.83/23.22/34.30		1.54
dhfr	0.02	4.49	5.98/12.15/18.70		1.06	0.10	24.81	34.12/63.32/103.37		6.10

Average per conformation t_m , per compound t_c , parallel runtime t_p , selectivity σ

Conclusion

We have described cRAISE, a VS tool that propagates additional knowledge to support pharmacophore-driven pose sampling and library profiling in structure-based VS. This is particularly useful if hypotheses about desired key interactions and/or physico-chemical features of the compounds are known beforehand. In such situations cRAISE allows to focus on predictions that are of major interest. Our pharmacophore-guided approach leads to an effective search space reduction and as a result, it reduces computational demand. Opposed to many other pharmacophore-guided docking approaches, it thereby provides a screening platform that allows the testing of hypotheses on a large scale.

Our results demonstrate that pharmacophores allow to externally direct the docking process. The implemented search space reduction does not lead to a loss of quality. To the contrary, if the models are prepared properly, they offer the chance to improve binding mode predictions. Poses that violate the given feature definitions are either not generated at all or rejected before they are actually scored. The procedure lets pharmacophore fulfilling poses emerge without the need to adapt the underlying scoring function. The presented enrichment studies reveal that early as well as global enrichment can be enhanced by this mean. Relaxed models offer the possibility to simultaneously evaluate different feature combinations. They enforce only some pharmacophoric commonality on the retrieved screening results. Nevertheless, by the use of strict model definitions it is possible to focus on compounds that reveal a specific binding pattern.

Confronted with millions of compounds, the once prepared cRAISE descriptor index basically enables fast information retrieval in succeeding VS runs. In order to benefit from our index-based VS technique the precalculated information needs to be permanently stored and the content needs to remain unchanged throughout its complete lifetime. The methods of cRAISE introduced here provide

a versatile interface to support flexible queries on this static compound library for different screening projects. Given pharmacophore definitions are utilized to guide cRAISE to extract only information of molecules with an improved chance to result in a pharmacophore-obeying pose. We showed that pharmacophore definitions can drastically accelerate the screening process. Moreover, cRAISE allows to state library profiles by constitutional and topological ligand conditions. The additional constraints restrict the index-based search further and filter out compounds without any loss of efficiency simultaneously during a VS run.

Our introduced hybrid method demonstrates how to gain external control over structure-based VS. Essentially, taking the best out of both worlds, it is a first step towards an integrated, synergetic VS platform combining structure- and ligand-based techniques. Relevant for any three-dimensional VS strategy is the consideration of tautomers and ionization states of query and library molecules. cRAISE provides the option to account for these degrees of freedom during pharmacophore guided and unguided VS. This extension accompanied with the respective results will be published separately. The cRAISE software is available for Linux operating systems (<http://www.zbh.uni-hamburg.de/raise>).

Acknowledgments We would like to thank Nadine Schneider for the fruitful discussions about protein-ligand interactions and scoring functions, moreover, Christin Schärfer for her commitments to conformer generation. This work was financially supported by the BMWI-ZIM Project KF2563701.

References

- Sottriffer C (2011) Virtual screening. WILEY-VCH Verlag GmbH & Co, KGaA, Weinheim
- Sottriffer C, Matter H (2011) The challenge of affinity prediction: scoring functions for structure-based virtual screening. In: Sottriffer C (ed) Virtual screening. Wiley-VCH Verlag GmbH & Co, KGaA, Weinheim, pp 177–221

3. Henzler AM, Rarey M (2010) *Mol Inform* 29:164–173
4. Sanders MP, McGuire R, Roumen L, de Esch IJ, de Vlieg J, Klomp JP, de Graaf C (2012) *Med Chem Commun* 3:28–38
5. Wallach I (2011) *Drug Dev Res* 72:17–25
6. Leach AR, Gillet VJ, Lewis RA, Taylor R (2010) *J Med Chem* 53:539–558
7. Yang SY (2010) *Drug Discov Today* 15:444–450
8. Dror O, Shulman-Peleg A, Nussinov R, Wolfson H (2004) *Curr Med Chem* 11:71–90
9. Tintori C, Corradi V, Magnani M, Manetti F, Botta M (2008) *J Chem Inf Model* 48:2166–2179
10. Muthas D, Sabnis YA, Lundborg M, Karl A (2008) *J Mol Graph Model* 26:1237–1251
11. Peach ML, Nicklaus MC (2009) *J Chem Inform* 1:6
12. Yang JM, Shen TW (2005) *Proteins* 59:205–220
13. Fradera X, Knegtel RM, Mestres J (2000) *Proteins* 40:623–636
14. Verdonk ML, Berdini V, Hartshorn MJ, Mooij WTM, Murray CW, Taylor RD, Watson P (2004) *J Chem Inf Comp Sci* 44:793–806
15. Hindle SA, Rarey M, Buning C, Lengauer T (2002) *J Comput Aided Mol Des* 16:129–149
16. Rarey M, Kramer B, Lengauer T, Klebe G (1996) *J Mol Biol* 261:470–489
17. Schlosser J, Rarey M (2009) *J Chem Inf Model* 49:800–809
18. Schellhammer I, Rarey M (2007) *J Comput Aided Mol Des* 21:223–238
19. Urbaczek S, Kolodzik A, Fischer JR, Lippert T, Heuser S, Groth I, Schulz-Gasch T, Rarey M (2011) *J Chem Inf Model* 51:3199–3207
20. Hartshorn MJ, Verdonk ML, Chessari G, Brewerton SC, Mooij WTM, Mortenson PN, Murray CW (2007) *J Med Chem* 50:726–741
21. Huang N, Shoichet BK, Irwin JJ (2006) *J Med Chem* 49:6789–6801
22. Neves MAC, Totrov M, Abagyan R (2012) *J Comput Aided Mol Des* 26:675–686
23. Spitzer R, Jain AN (2012) *J Comput Aided Mol Des* 26:687–699
24. Schneider N, Hindle S, Lange G, Klein R, Albrecht J, Briem H, Beyer K, Claußen H, Gastreich M, Lemmen C, Rarey M (2012) *J Comput Aided Mol Des* 26:701–723
25. Novikov FN, Stroylov VS, Zeifman AA, Stroganov OV, Kulkov V, Chilov GG (2012) *J Comput Aided Mol Des* 26:725–735
26. Liebeschuetz JW, Cole JC, Korb O (2012) *J Comput Aided Mol Des* 26:737–748
27. Brozell SR, Mukherjee S, Balias TE, Roe DR, Case DA, Rizzo RC (2012) *J Comput Aided Mol Des* 26:749–773
28. Corbeil CR, Williams CI, Labute P (2012) *J Comput Aided Mol Des* 26:775–786
29. Repasky MP, Murphy RB, Banks JL, Greenwood JR, Tubert-Brohman I, Bhat S, Friesner RA (2012) *J Comput Aided Mol Des* 26:787–799
30. Irwin JJ, Shoichet BK (2005) *J Chem Inf Model* 45:177–182
31. Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG (2012) *J Chem Inf Model* 52:1757–1768
32. Schärfer C, Schulz-Gasch T, Hert J, Heinzerling L, Schulz B, Inhester T, Stahl M, Rarey M (2013) *Chem Med Chem* 8:1690–1700
33. Wu K (2005) *J Phys: Conf Ser* 16:556–560
34. Wu K, Ahern S, Bethel EW, Chen J, Childs H, Cormier-michel E, Geddes C, Gu J, Hagen H, Hamann B, Koegler W, Lauret J, Meredith J, Messmer P, Otoo E, Perevozchikov V, Poskanzer A, Rübel O, Shoshani A, Sim E, Stockinger K, Weber G, Zhang Wming (2009) *J Phys Conf Ser* 180:1
35. Kirchmair J, Ristic S, Eder K, Markt P, Wolber G, Laggner C, Langer T (2007) *J Chem Inf Model* 47:2182–2196
36. Schulz-Gasch T, Schärfer C, Guba W, Rarey M (2012) *J Chem Inf Model* 52:1499–1512
37. Hilbig M, Urbaczek S, Groth I, Heuser S, Rarey M (2013) *J Chem Inform* 5:38
38. Barber CB, Dobkin DP, Huhdanpaa H (1996) *ACM Trans Math Softw* 22:469–483
39. Böhm HJ (1994) *J Comput Aided Mol Des* 8:243–256
40. Warren GL, Do TD, Kelley BP, Nicholls A, Warren SD (2012) *Drug Discov Today* 17:1270–1281
41. Zinc clean leads (2012) UCSF University of California, San Francisco. <http://zinc.docking.org/subsets/clean-leads>. Accessed 7 Dec 2012

Facing the Challenges of Structure-based Target Prediction by Inverse Virtual Screening

[D9] K. Schomburg, S. Bietz, H. Briem, A.M. Henzler, **S. Urbaczek**, and M. Rarey. Facing the Challenges of Structure-based Target Prediction by Inverse Virtual Screening. *Journal of Chemical Information and Modeling*, 54(6):1676-1686, 2014.

<http://pubs.acs.org/articlesonrequest/AOR-sGyGDQNmbZS4EtF4PCN>


Reproduced with permission from K. Schomburg, S. Bietz, H. Briem, A.M. Henzler, S. Urbaczek, and M. Rarey. Facing the Challenges of Structure-based Target Prediction by Inverse Virtual Screening. *Journal of Chemical Information and Modeling*, 54(6):1676-1686, 2014. Copyright 2014 American Chemical Society.

Facing the Challenges of Structure-Based Target Prediction by Inverse Virtual Screening

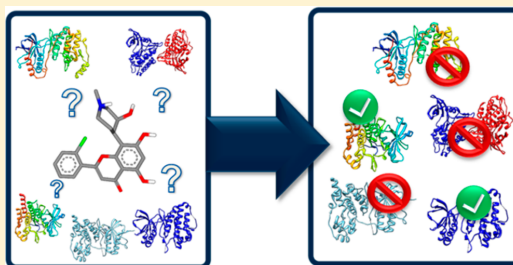
Karen T. Schomburg,[†] Stefan Bietz,[†] Hans Briem,[‡] Angela M. Henzler,[†] Sascha Urbaczek,[†] and Matthias Rarey^{*,†}

[†]Center for Bioinformatics, University of Hamburg, Bundesstrasse 43, 20146 Hamburg, Germany

[‡]Global Drug Discovery, Medicinal Chemistry, Bayer Pharma AG, 13353 Berlin, Germany

 Supporting Information

ABSTRACT: Computational target prediction for bioactive compounds is a promising field in assessing off-target effects. Structure-based methods not only predict off-targets, but, simultaneously, binding modes, which are essential for understanding the mode of action and rationally designing selective compounds. Here, we highlight the current open challenges of computational target prediction methods based on protein structures and show why inverse screening rather than sequential pairwise protein–ligand docking methods are needed. A new inverse screening method based on triangle descriptors is introduced: *i*RAISE (*i*nverse *R*apid *I*ndex-based *S*creening *E*ngine). A Scoring Cascade considering the reference ligand as well as the ligand and active site coverage is applied to overcome interprotein scoring noise of common protein–ligand scoring functions. Furthermore, a statistical evaluation of a score cutoff for each individual protein pocket is used. The ranking and binding mode prediction capabilities are evaluated on different datasets and compared to inverse docking and pharmacophore-based methods. On the Astex Diverse Set, *i*RAISE ranks more than 35% of the targets to the first position and predicts more than 80% of the binding modes with a root-mean-square deviation (RMSD) accuracy of <2.0 Å. With a median computing time of 5 s per protein, large amounts of protein structures can be screened rapidly. On a test set with 7915 protein structures and 117 query ligands, *i*RAISE predicts the first true positive in a ranked list among the top eight ranks (median), i.e., among 0.28% of the targets.



INTRODUCTION

Controlling the protein selectivity of a lead compound in drug discovery is crucial for avoiding adverse effects and, thus, lowering the high attrition rates of drugs during the past decade.^{1–3} For rational protein selectivity enhancement, the uttermost goal is the complete target profile for a compound on human targets. Furthermore, protein target predictions may reveal hidden opportunities in drug repurposing projects,^{4–6} support the difficult but promising design process of multitarget drugs,^{7–9} and reveal targets of drugs with so far unknown mechanisms-of-action (in the DrugBank,¹⁰ more than 80 drug entities are registered with unknown mechanisms-of-action). Other scientific fields, such as biotechnology, are profiting from target prediction methods, e.g., for the design of in vitro synthetic reaction pathways.¹¹

Strategies to predict targets for a small molecule are either computational or experimental.^{12,13} So far, the use of experimental activity assays for a broad range of targets still dominates in drug development processes. However, computational methods can complement or reduce—and even substitute—some costly and time-consuming experimental methods. In contrast to high-throughput screening of thousands of molecules for one target, no such time- and

cost-efficient experimental methods exist for screening thousands of proteins.

Depending on the available data, computational target prediction methods can be classified as ligand-based, network-based, side-effect-based, or protein-structure-based.

Ligand-based methods couple ligand similarity measurements with experimental data.^{14–18} Network-based methods exploit available data on ligand and target interactions for compiling networks and deduce thereof new predictions.^{19–22} Side-effect-based methods derive target predictions from phenotypic (adverse) effects of drugs.²³ Protein-structure-based methods use docking, pharmacophore searching, binding site comparison or protein–ligand interaction fingerprints to predict new targets.²⁴

Ligand-based, network-based, and side-effect-based methods show good results, if the molecules with available data are similar enough to those for which predictions should be made, following the paradigm of “*If something has been observed, knowledge can be deduced for similar things*”. However, these methods fail to predict effects that are outside the compound

Received: February 28, 2014

domain used to generate the respective model. Protein-structure-based methods are dependent on three-dimensional (3D) protein structures. Furthermore, pharmacophore searches, binding-site comparisons, and interaction fingerprints need at least one starting co-crystallized complex as input. Docking-based target prediction is the only method that is independent of such preliminary information, needing only the 3D protein structure and the active site location, e.g., identified by any co-crystallized ligand or a pocket identification algorithm. The amount of available 3D structures of proteins grows rapidly, promising increasing importance for this method in future.

In the following, we will focus on docking-based target prediction methods. These approaches have one further major advantage: simultaneously with predicting a target, the binding mode of a ligand to a protein is predicted. However, compared to classic protein–ligand docking, the reverse setup has different requirements. Four main challenges must be addressed in the development of structure-based target prediction methods:

- (1) Preprocessing and handling many protein structures: In classical screening, a single protein is used and the active site preparation is rather complex and time-consuming. For inverse screening, the method must be able to deal with at least 10^4 structures, calling for completely automated time-efficient processes.
- (2) Efficient and consistent handling of structural data: Protein structure data is storage-demanding defining a need for new approaches to handle large amounts of protein structures consistently and efficiently.
- (3) Ranking of targets: As has been stated and observed previously,²⁵ scoring functions that were developed for assessing protein–ligand complexes in classic docking are problematic when applied to intertarget ranking. Measures accounting for the diversity of protein pockets concerning shape and properties²⁶ must be included.
- (4) Significant evaluation methods: Prospective evaluation is expensive and not feasible for intermethod comparison. Therefore, reliable datasets for retrospective studies are needed. For inverse docking/screening, no standard evaluation datasets exist yet on which new methods can be evaluated and compared among each other, such as, e.g., the DUD²⁷ for classic docking and virtual screening. The main problem is the categorization of targets as true negatives for small molecules. Unfortunately, literature data rarely reveal negative results, i.e., if a molecule does not interact with a protein. In summary, a dataset is needed that contains a sufficient number of molecules and proteins with a reliable assignment of targets and nontargets.

So far, the available docking-based target prediction methods only barely account for the challenging requirements of the reverse scenario.

Invdock, which is the first published docking-based target prediction method, uses the DOCK docking algorithm.^{28,29} A threshold score is applied to avoid the high computing time of classic docking approaches for the reverse setup. Once any pose in a cavity is found with a score better than the threshold, the exhaustive search of the best pose is aborted. For evaluation, a redocking study on nine proteins,²⁸ as well as a screening of eight compounds against the TTD (therapeutic target database³⁰) of, at that time, 1040 structures of 38 proteins related to

side effects were conducted. Of the 43 experimentally documented protein–ligand interactions, Invdock finds 38.²⁹

TarFisDock, which is another inverse docking approach, was published in form of a web-service of inverse docking based on DOCK against the PDTD (Potential Drug Target Database³¹).³² For evaluation, Li and co-workers screened the PDTD of, at that time, 698 structures of 371 targets with two compounds. For vitamin E, 50% of reported targets were ranked among the first 10% of the targets and for 4H-tamoxifen among the first 5%.

Another inverse screening application utilizes DOCK for building a chemical-protein interactome for deriving relevant genes of adverse drug reactions, in particular for the identification of risk-alleles.^{33,34}

The sc-PDB is a subset of the protein–ligand complexes contained in the Protein Data Bank³⁵ relevant to drug design.^{36,37} Paul et al. inversely screened an early version of the database with 2148 binding sites of 1045 different proteins with five chemically diverse ligands with GOLD.³⁸ Of these, for four compounds, enrichment factors at 1% of the dataset between 26 and 102 are reported and, for one compound (AMP = adenosine monophosphate), poor performance was reported, leading the authors to recommend to use the inverse docking approach for selective ligands.

In an application study of Muller, the sc-PDB was also screened using GOLD with five combinatorial molecules sharing a 1,3,5-triazepan-2,6-dione scaffold.³⁹ Of five experimentally tested targets, one was confirmed as true target. Later, Kellenberger evaluated a combination of GOLD docking scores and an interaction fingerprint for the ranking of true targets for the same four ligands²⁵ on the sc-PDB consisting at that time of 4300 protein ligand binding sites of 1550 different proteins. For the four compounds, targets were predicted with AUCs between 0.7 and 0.95 for the GoldScore and AUCs between 0.45 and 0.9 for the interaction fingerprint scoring.

An evaluation of Glide in an inverse docking scenario on the Astex Diverse Set shows limitations in its intertarget ranking capability and coins the term “interprotein scoring noise”.⁴⁰ It was found that a correction of the standard Glide scoring function, considering protein properties, improved target predictions.

In the following, we introduce a new structure-based target prediction method, which is based on protein–ligand screening and applies measures to account for the requirements of the reverse setup. Special care is taken regarding the preparation and handling of protein structure data. In addition, the method is a true inverse screening approach, substantially reducing the computing time compared to the application of a classic docking method. Finally, to address the intertarget ranking issue, several special scoring measures are applied, taking into account the diversity of protein pockets.

METHODS

The overall screening process is divided into two parts: a preliminary registration procedure and the actual screening procedure (see Figure 1). The registration procedure enables fast screening by performing precalculations and data setup only once for a protein dataset. The screening procedure can then be performed recurrently on the prepared dataset. The basis of the screening technology is a descriptor-based bitmap search, called RAISE technology (where RAISE represents RApid Index-based Screening Engine). Since the RAISE

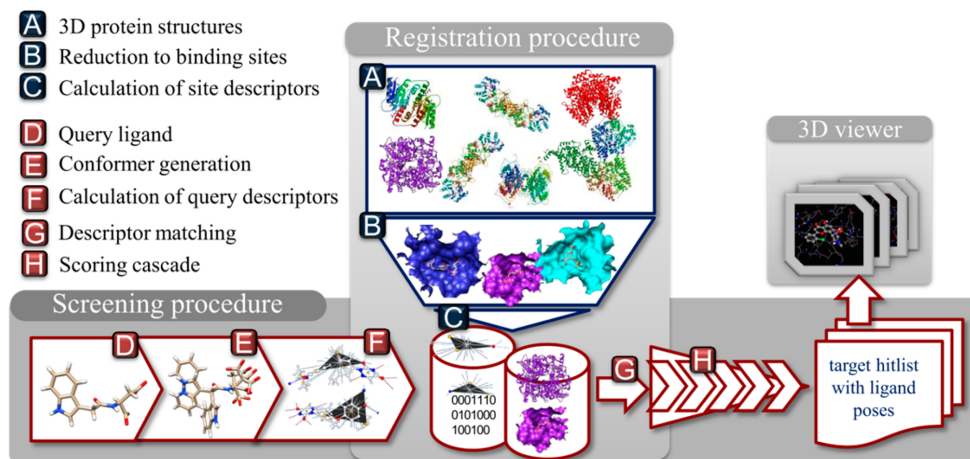


Figure 1. Workflow of the *iRAISE* inverse screening algorithm. Steps A–C of the registration procedure must be done only once for a dataset of protein structures. Steps D–H are part of the screening procedure.

technology, in this context, is applied in an inverse protein–ligand scenario, the tool is thus called *iRAISE* (*inverse-RAISE*).

Registration Procedure. Starting with a set of 3D protein structures (Figure 1A), first, the active sites are determined with a radius of 6.5 Å around a reference ligand (Figure 1B). The reference ligand can either be supplied by the user or identified in an automatic mode. In automatic mode, all ligands in a protein structure are used to build active sites except for co-factors, ions, crystallization agents, solution buffer agents, and ligands with covalently bound metals. The exclusion list is compiled by joining our own list with those from other publications^{37,41,42} and contains, in total, 1207 PDB HET codes (see the Supporting Information). Next, descriptors are calculated for all active sites (Figure 1C, see the section entitled “Triangle Descriptor”). Finally, the active sites and the protein structures are stored in a relational database (see the section entitled “Protein Structure Database”), enabling efficient and consistent data handling.

Screening Procedure. Initially, conformations for the query ligand are sampled with the CONFECT conformation generator⁴³ (Figure 1E). Triangle descriptors are calculated for each conformation (Figure 1F). Then, these descriptors are matched against the protein descriptors (Figure 1G). A descriptor match corresponds to one protein–ligand pose with at least three matching interactions and a rough shape fit. Each found pose is then scored by the Scoring Cascade (Figure 1H; see the section entitled “Scoring Cascade”). The result of the screening procedure is a ranked list of targets for the query ligand as well as poses of the ligand for all hit targets are the results of the screening procedure.

Triangle Descriptor. In order to obtain a rapid screening procedure, ligand–protein matching is abstracted by a descriptor representing pharmacophoric and steric features. With the descriptor, the time-consuming multiple sequential placing of each ligand into each active site is circumvented. In *iRAISE*, the same triangle descriptor that was published for the virtual screening tool TriX^{44,45} is used.^{44,45} In brief, the descriptor has the following properties. A triangle descriptor consists of interaction spots of type hydrogen bond acceptor, donor, or hydrophobic at the corners. Each hydrophilic interaction spot is

annotated with one or several interaction directions. The triangle side lengths encode the distance between the interaction spots. The shape of the ligand and, accordingly, the space of the active site around a triangle descriptor is encoded by 80 bulk rays originating from the center of the triangle limited by the surface of the ligand or the protein, respectively.

A novel feature of the triangle descriptor in *iRAISE* is the integration of flexibility of hydrophilic rotatable terminal groups (such as hydroxyl groups) of the active site and the query molecule. Rotatable groups are handled by interaction spots with multiple directions (for acceptors) or multiple possible interaction spots (for donors).

Unifying Query Descriptors. In molecular conformations where, e.g., only a terminal part of the ligand changes, many identical descriptors are generated. Therefore, a clustering procedure is applied reducing the total descriptor set to unique descriptors. Since the triangle descriptor contains only binned values, the clustering of identical descriptors is performed rapidly. The clustering reduces the number of descriptors significantly with which the index will then be queried: For an average of 200 query ligand conformations, the unique descriptor clustering reduces the amount to 35% of all descriptors.

Storing and Matching. The triangle descriptors of the active sites are stored in a FastBit compressed bitmap index^{44,46} In *iRAISE*, this index is extended by storing the coordinates of the triangle corners next to the triangle descriptors, enabling immediate superposition of descriptors. This modification is mandatory for inverse screening due to the time-consuming calculation of active site descriptors, in contrast to ligand descriptors in a standard virtual screening setup.

The triangle descriptors of the query ligand are matched complementary concerning interaction spots, interaction directions, shape, and triangle side lengths to those of the active sites in the bitmap index. Once a match is found, the transformation needed to superpose the triangles is applied to the respective conformation(s) of the query molecule, producing the pose(s) in the protein active site.

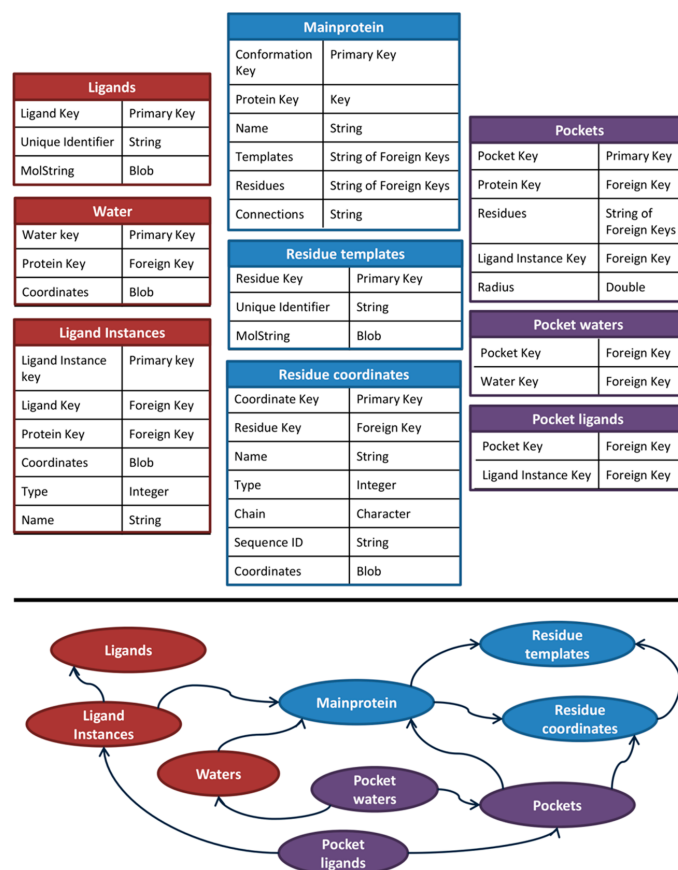


Figure 2. Database scheme of the protein structure database. Blue tables code the information on the protein. Red tables contain information on small molecules such as ligands, co-factors, metal ions, and water molecules. Purple tables code the information on the active site.

Protein Structure Database. For a storage- and time-efficient handling of the protein structure data as well as consistent representation of active sites calculated in the preparation procedure, a SQLite (www.sqlite.org) database is used. In Figure 2, the scheme of the database is sketched. The database consists of tables for protein data (blue tables); for data of ligands, water molecules, and metal ions (red tables); and for active site data (purple tables).

The protein data is represented in three tables: the *Mainprotein* table, containing general information, such as the protein name; and the *Residue templates* and *Residue coordinates* tables, containing amino acids. The *Mainprotein* table has two keys: a protein key and a conformation key, enabling the storage of protein ensembles. The *Residue templates* table contains each topological distinct amino acid of all proteins once. The USMILES⁴⁷ is used as unique identifier while the MolString contains the information on atoms, bonds, and valence states needed for reinitialization.^{48,49} While each amino acid is added only to the *Residue templates* table if a topologically identical one previously has not been registered there, its coordinates, name, type, chain, and sequence index are written to the *Residue coordinates* table. Each *Residue coordinates* entry is mapped with a key to a *Residue templates* entry. With

this setup of storing amino acids, the repeated information on the chemical composition of amino acids is stored only once in the database.

For storing ligands, the same concept of templates is used. The *Ligands* table contains a unique identifier in form of a USMILES and the MolString, coding the topology of a ligand. In this table, only a new entry is added, if the table yet does not contain the USMILES of the ligand. The coordinates of the ligands and distinct data such as the name and the corresponding protein key are stored in the *Ligand Instances* table. Water molecules are handled separately in the *Water* table containing the coordinates, a water key, and the protein key.

Active sites are stored in the *Pockets* table, which contains the key of the corresponding protein, the radius, the key of the ligand used as reference and the keys of the residues belonging to the active site. The keys of further ligands or metal ions contained in the active site are stored in the *Pocket ligands* table and keys of active site water molecules in the *Pocket water* table.

Storing the protein–ligand complexes in the database takes only 60% of the size needed to store the raw PDB files: For storing protein–ligand complexes of 100 random files from the PDB, the database size is 36MB (in comparison to 63MB for

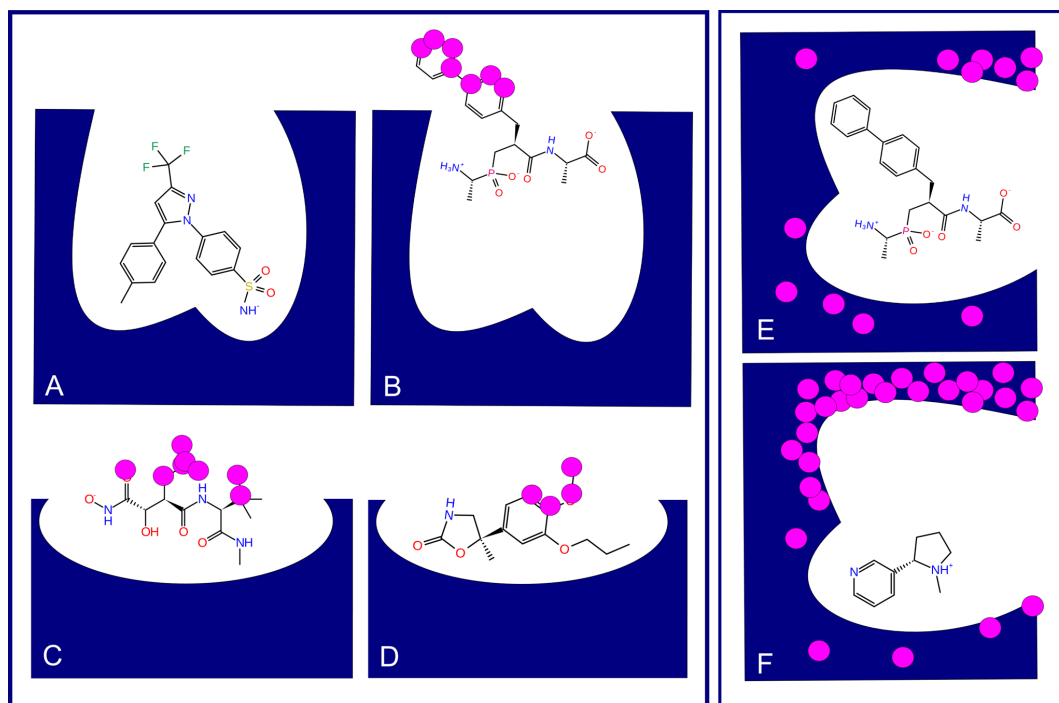


Figure 3. Schematic representation of the ligand coverage and pocket coverage in the Scoring Cascade. Panels (A–D) represent ligand coverage: (A) a reference ligand in a buried pocket with all atoms covered; (B) a ligand pose protruding into the solution with some noncovered atoms, highlighted by pink circles (this pose would be discarded due to insufficient ligand coverage); (C) a shallow pocket with a reference ligand with some noncovered atoms; (D) a docking pose in the shallow pocket with fewer noncovered atoms (thus, this ligand would not be discarded). Panels (E and F) represent pocket coverage: (E) a pocket with a reference ligand, which occupies most of the pocket and therefore has only a few noncovered pocket atoms; and (F) a small ligand in a large binding pocket with many noncovered atoms in the pocket, the score of which would be weighted down, because of insufficient pocket coverage.

the raw files), 333MB for 1000 random PDB files (in comparison to 558MB for the raw files), and 3.4GB for 10,000 random PDB protein files (in comparison to 5.7GB for the raw files). Reading a protein and the annotated active site from the database on average takes only 60 ms, independent from the database size as long as the database fits into main memory.

Scoring Cascade. A Scoring Cascade of five steps, accounting for active site diversity, is used to overcome interprotein scoring noise. The Scoring Cascade starts with all ligand poses obtained from the descriptor matches for one protein. It applies five steps to discard irrelevant poses and obtain a score comparable among proteins with diverse features:

1. **Clash Test.** The clash test discards clashing poses rapidly with a grid representation of the active site. This step already discards two-thirds of all poses from the descriptor matching.

2. **Interaction Score.** The second step is the scoring of each pose with a simple interaction score based on Lennard-Jones potentials for hydrophilic interactions, metal interactions, hydrophobic contacts, and hydrophobic–hydrophilic mismatches (see Supporting Information for Lennard-Jones parameters). Beforehand, for each pose, the best hydrogen bond network in the active site is calculated with Protoss.⁵⁰

Also, the reference ligands that were used to determine the active sites are scored with this simple interaction score.

3. **Reference Score Cutoff.** For each pose, its interaction score is compared to the interaction score of the reference ligand of that active site. If the score of the pose is less than 75% of the score of the reference ligand, then this pose is discarded. This step discards, on average, 50% of the target matches for a query ligand.

Taking the reference ligand into account in scoring renders the method dependent on reliable crystallized protein–ligand complexes. In addition, not each ligand of each co-crystallized complex binds with high binding energy. However, since this step is used only as a cutoff, the binding affinity variations are not problematic: A reference ligand with low binding energy is scored only lowly by the energy-based scoring function and, thus, fewer ligands are discarded by this cutoff step.

4. **Ligand Coverage Score.** This score measures how well a ligand is buried in a pocket and is used to discard poses that protrude with a large part into solvent. Consulting the reference ligand enables comparing pockets with different shapes, e.g., comparing scores of shallow to buried pockets. If the coverage of a ligand pose multiplied by a factor of 1.2 is less than the coverage of the reference ligand, or if <10% of all ligand atoms are covered in total, the pose is discarded.

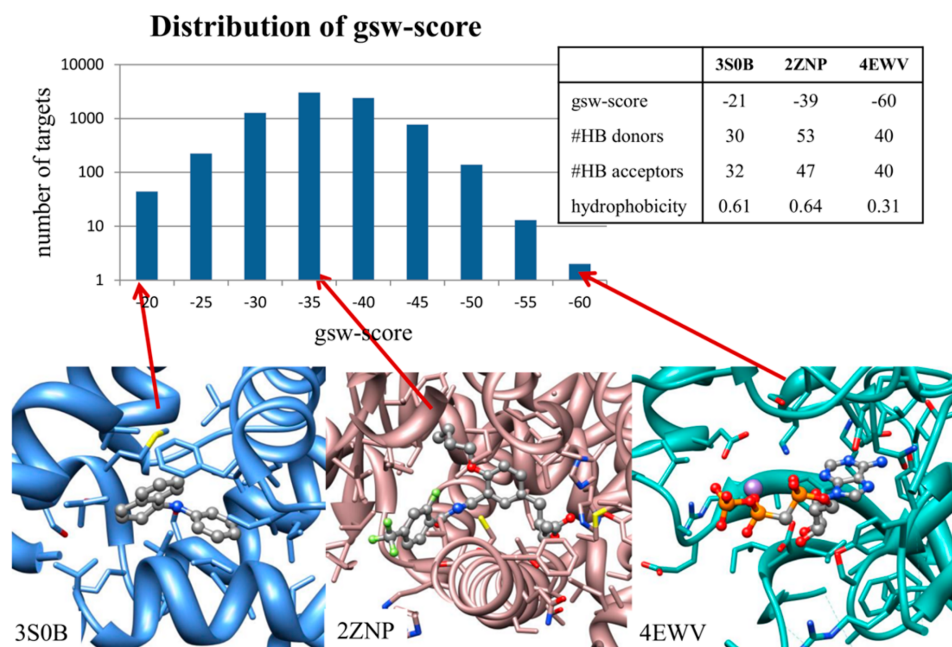


Figure 4. Distribution of target-specific gsw-scores. The complex 3S0B with a minor gsw-score of -21 is highly hydrophobic. The complex 2ZNP with an average gsw-score of -39 is a larger pocket containing many hydrogen bond partners. The complex 4EWV with a high gsw-score of -60 is hydrophilic and contains a metal ion. (#HB donors = number of hydrogen bond donors, #HB acceptors = number of hydrogen bond acceptors, hydrophobicity = number of hydrophobic amino acids of active site divided by total number of amino acids of active site.)

The *ligand coverage* is the average ligand atom coverage (where A is the set of heavy atoms of a ligand):

$$\text{ligand coverage} = \frac{1}{|A|} \sum_{a \in A} \text{coverage}(a)$$

with the coverage of an atom a given as

$$\text{coverage}(a) = \begin{cases} 1 & \text{if close receptor atoms} + \frac{1}{|N(a)|} \\ & \sum_{b \in N(a)} \text{coverage}(b) > 3 \\ 0 & \text{otherwise} \end{cases}$$

Close receptor atoms are all atoms of the active site, which are in a radius of 4.5 \AA of the ligand atom a . Furthermore, the average coverage of the covalently bound atoms $N(a)$ is added. As shown exemplarily in Figure 3, the ligand coverage is able to differentiate between binding scenarios to pockets with different shapes.

5. Pocket Coverage Score. The fifth and final step of the Scoring Cascade is the pocket coverage score, which addresses how well a ligand fills a pocket. Poses that produce insufficient pocket coverage in comparison to the pocket coverage produced by the reference ligand ($<80\%$ of the reference pocket coverage) are weighted down with a factor of 0.8 . The pocket coverage is calculated as follows:

$$\text{pocket coverage} = \frac{1}{|P|} \sum_{a \in P} \text{coverage}(a)$$

where P is the set of pocket atoms.

Thus, the *pocket coverage* is the number of covered active site protein atoms divided by the total number of active site protein atoms. The *coverage* of a receptor atom a is calculated using the following formula:

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

Since all pockets are determined with a cutoff distance of 6.5 \AA around the ligand, reference ligands have a pocket coverage of $\sim 5\% - 40\%$. In Figures 3E and 3F, a schematic representation of the pocket coverage demonstrates how query ligands are scored higher in pockets that they fill in a similar way as a reference ligand. This step is only used as a weight to the score instead of as a cutoff.

Gauss Cutoff Score. After the Scoring Cascade, the score of each pose is used to rank the proteins as targets for the query ligand. A final step is applied to further tune the ranking capability of iRAISE. A Gaussian score weight (gsw-score) is applied to the score of the Scoring Cascade (sc-score) to be able to decide if a score is statistically significant for a protein pocket. Taking the sc-score in relation to the gsw-score results in the final iRAISE score, which is statistically significant, if it is > 1 . The gsw-score is a characteristic score for each protein pocket. It is obtained using the following steps:

- (1) The complete protein pocket library of an iRAISE project is screened with the 84 chemically diverse ligands of the Astex Diverse Set.⁵¹

- (2) For each protein pocket, the sc-score is calculated for those of the 84 ligands, which could be placed into the pocket. The reference score cutoff-step (Step 3) of the Scoring Cascade is omitted to obtain a full spectrum of scores.
- (3) The gsw-score of each protein pocket is defined as the average score of all 84 ligand scores. Parameterization showed that using a score of the average plus twice the standard deviation as assessment of statistical significance was too strict.
- (4) All protein pockets with less than 20 of the 84 ligand scores are discarded, since the average score would be calculated from too few data points.

In Figure 4, the score distribution of the scores on the ~8000 targets of the sc-PDB dataset³⁷ screened with all 84 Astex ligands is shown. The scores are binned at units of 5. The gsw-score ranges between -21 and -60, and the median average score of all targets is -39. Between 0 and all of the 84 ligands can be docked to structures. The median is 67 ligands per target. Only 77 of the more than 8000 protein structures of the data set are scored with less than 20 ligands, leaving totally 7915 protein pockets. In Figure 4, a hydrophobic pocket with a minor gsw-score, a pocket with an average score and a hydrophilic pocket containing a metal ion are shown.

RESULTS

The evaluation of inverse screening tools still poses a huge challenge since no standard evaluation datasets or standard statistical metrics are established. Therefore, we focus on quantitative statistical evaluation, in comparison to other methods and evaluation of the gain of the individual steps of the Scoring Cascade.

In total, five evaluation experiments were performed:

- (1) Binding mode prediction evaluation: Redocking study with RMSD values
- (2) Evaluation of the ranking capacity of the Scoring Cascade
- (3) Comparison of the ranking capability of *iRAISE* with FlexX/Hyde and Glide
- (4) Comparison of ranking capabilities with a pharmacophore-based method on a large dataset
- (5) Computing-time analysis

Software and Data Sets. *FlexX/Hyde.* FlexX⁵² was applied as integrated in the LeadIT software suite (version 2.1, www.biosolveit.de) and for scoring the Hyde scoring function⁵³ was used. Default parameters were used. The docking was started with a conformation generated by Corina.⁵⁴

Glide. For Glide docking, the XGlide script (Version 3.3, provided by Schrodinger, Inc.) was used. XGlide automates the protein preparation and Glide grid generation step, based on the native X-ray ligand complex. Subsequently, XGlide performs Glide SP (Version 6.1) docking runs with default parameters. Starting conformations of the input ligands were generated by Corina.⁵⁴

Pharmacophore Search. For comparison of the ranking capability of *iRAISE* with a pharmacophore-based search strategy, we used the results published by Meslamani et al.⁵⁵

Astex Diverse Set. For experiments 1–3, the Astex Diverse Set⁵¹ was used. This dataset consists of 85 manually curated high quality diverse protein–ligand complexes. In the screening experiments, all 84 ligands were screened against all 85 protein structures (one ligand is present twice in the dataset). The objective of this experiment is to predict the co-crystallized

target for each of the 85 query ligands as a true target and to rank the true target to the first positions of the list of all targets. This experiment is suited for simple evaluation and for comparison to other methods, but some caution is necessary when interpreting the results: Redocking the ligands into the target with which they have been co-crystallized is an artificial use case and is only useful for proof-of-concept evaluation. For estimating practical applicability, experiments with structures not crystallized with the query ligand are necessary. Furthermore, for the Astex Diverse Set, it is not known, whether ligands bind to multiple of the 85 targets.

sc-PDB Diverse Set. For the fourth experiment, the sc-PDB Diverse Set⁵⁵ consisting of 157 diverse ligands and the sc-PDB protein structure data set was used. The sc-PDB is a subset of the Protein Data Bank filtered with quality and druggability criteria. Meslamani used the 2010 version of the sc-PDB for the pharmacophore searches. This version is no longer available; therefore, we used the 2012 version of the sc-PDB.⁵⁶ We downloaded the original PDB files from the PDB instead of using the preprocessed files contained in the sc-PDB. Of originally 8077 structures, we used 7992, since, of these, 51 were discarded due to several errors during initialization of the reference ligand or the protein, 25 due to a mismatch of the reference ligand provided in sc-PDB, and 9 due to obsolete PDB codes. Annotation of the 7992 structures with the gsw-score further reduces the number to 7915. The sc-PDB Diverse Set of ligands consists of 157 ligands, which are co-crystallized with targets of the sc-PDB. Of these, we took a subset of 117 ligands of which the co-crystallized PDB structure reported by Meslamani was also present in the sc-PDB 2012 version. The 7915 structures of the sc-PDB 2012 were clustered by UniProtPK ID, as provided with the sc-PDB 2012 version, resulting in 2879 different proteins. True positive structures for the 117 ligands were assigned by two protocols: First, proteins with the same UniProtPK ID as the co-crystallized protein are considered true positives only.⁵⁵ Second, also structures with the same EC number as the co-crystallized protein were considered as true positives.

1. Redocking Study. As an initial study, we evaluated the ability of *iRAISE* to predict binding modes comparing the poses generated by *iRAISE* with the crystal structures. *iRAISE* was started with a Corina-generated conformation of the Astex ligands and up to 200 conformations were sampled. In Figure 5, the bars show the sum of ligands that can be predicted with RMSDs lower than the value indicated on the abscissa. In the 30 best-scored poses of each ligand, a solution with RMSD

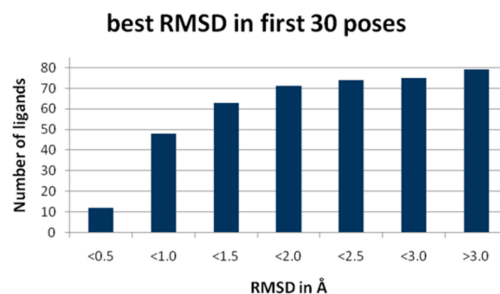


Figure 5. RMSDs for redocking each Astex ligand into its true target with *iRAISE*.

values of <2.0 Å were observed in $\sim 84\%$ of the cases. This value is slightly below the performance of optimized protein–ligand docking methods. One has to keep in mind, that *iRAISE* is a fully automated procedure enabling large throughput and therefore does not perform a post-optimization of poses. In such a scenario, the redocking performance is comparable to the state of the art.

2. Evaluation of the Ranking Capacity of the Scoring Cascade. In order to evaluate the effect of the Scoring Cascade, the rank of the true target for each Astex ligand was compared if only the simple interaction score, as a ranking measure, was used opposed to the full *sc-score*, based on the ligand poses obtained from descriptor matching (see Figure 6). This

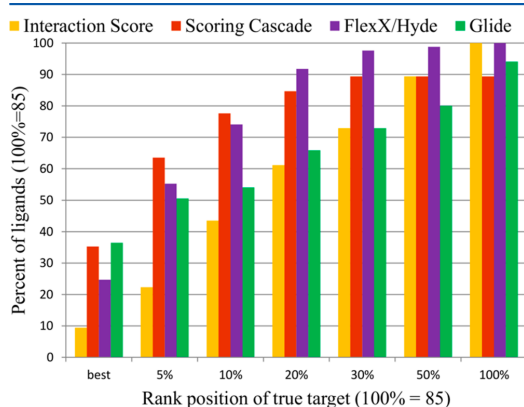


Figure 6. Ranking of true target for each of the 85 ligands of the Astex Diverse Set. The yellow bars show the ranks of the *iRAISE* poses scored only with the simple interaction score, the red bars show the ranking of the *iRAISE* poses scored with the full Scoring Cascade. The purple bars show the ranking if the FlexX docking with Hyde scoring is used and the green bars show the ranking of Glide. On the *y*-axis, the percentage of ligands is annotated, on the *x*-axis, the rank at which the true target was found. “Best” means the true target was found at rank 1 and the percentages show among which percent of the score-ordered list of targets the true target was ranked.

evaluation highlights two important points. First of all, the Scoring Cascade indeed ranks the true targets much better than the simple interaction score. With the Scoring Cascade, $\sim 35\%$ of all ligand queries result in a ranking of the true target at position one, and for more than two-thirds of all ligands the rank of the true target in the score-ordered list is among the best 5%, i.e. among the top four ranks. Second, since the Scoring Cascade tunes the selectivity, for some ligands the true target is not found in the target list. For these examples, the scoring is too strict, as can be seen in the diagram in Figure 6 in the fact that only 90% of the ligands get a rank for their true target.

3. Comparison of the Ranking Capability of *iRAISE* with FlexX/Hyde and Glide. As a third test on the Astex Diverse Set, we compared the ranking results to standard docking approaches to see how *iRAISE* performs in comparison (see Figure 6). The diagram shows that Glide and *iRAISE* predict almost the same number of targets at the first position ($\sim 35\%$). Ranking the true target in the first 5% of the target is accomplished by *iRAISE* for $\sim 63\%$ of the ligands, which is superior to standard docking. At 10%, *iRAISE*'s performance is

still marginally superior to the docking programs, while at higher percentages FlexX/Hyde shows better performance. This is due to the fact that *iRAISE*, as discussed in the previous section, does not generate a pose for each ligand in its true target, because of the selectivity enhancement of the Scoring Cascade. However, on large datasets, enrichment at the first percentages is important for the choice of targets to test experimentally. In summary, the diagram shows that taking the co-crystallized reference ligand into account, as in the Scoring Cascade, substantially improves inverse screening performance.

4. Comparison to a Pharmacophore-Based Method. Following the evaluation protocol by Meslamani,⁵⁵ we screened 117 ligands of the *sc-PDB* Diverse Set against the 7915 protein structures of the *sc-PDB* 2012 with *iRAISE*. The rank of the first true positive target in a score-ordered list of all targets is consulted as a measure of success. Therefore, Table 1 contains the median rank (median of the 117 ligands) of the first true positive target identified by the methods in absolute number, as well as in percentage of the dataset. In Table 1, the results of *iRAISE*, as well as the data extracted from the supporting information of Meslamani's publication, are shown. The medians of four different pharmacophore-based methods (rigid1, rigid2, flex1, flex2), and of Surfex⁵⁷ and Plants⁵⁸ docking were extracted for 117 ligands from Meslamani⁵⁵ (the names of the methods are adopted). Since the number of protein clusters (UniProtKB ID clusters) differs in the *sc-PDB* version used by Meslamani and the one screened by *iRAISE* (2556 vs 2879 different proteins), we calculated the percentage of the first rank on all clustered proteins. For screening with *iRAISE*, two protocols for ligand preparation were used: Initially, 200 conformations of the ligand were used without the co-crystallized ligand (called “*iRAISE* flex” in Table 1). Then, the co-crystallized ligand structure was used as input for *iRAISE* screening without generating conformations (called “*iRAISE* crystal” in Table 1). Clearly, the second experiment is artificial and much “easier” for the method, since the correct conformation already is used. We performed this experiment to be able to compare the results to the pharmacophore-based methods, since those deduce the pharmacophore from the co-crystallized complex, which, then again, is a true positive in the dataset.

For the experiment with the crystal ligand, *iRAISE* performs better than the pharmacophore methods, independent of the way how true positives are annotated, following Meslamani by UniProtKB ID or by the EC number. The first true positive is found at 0.07% of the database, while the best pharmacophore-based method ranks the first true positive at 0.16% of the proteins. For the *iRAISE* screening with conformations, the first true positive is ranked at 1.15% of the protein structures, following the assignment of true positives by UniProtKB ID, compared to 2.5% of Surfex and 4.4% of Plants. If the EC number is considered during assignment of true positives, *iRAISE* ranks the first true positives at the median at 0.28%. In contrast to the UniProtKB ID, the EC number is not organism-specific, but, nevertheless, does classify the same protein justifying the usage of EC numbers in this case. The comparison of both assignment methods shows that the targets ranked toward the beginning of the list, which are not true positives after the UniProtKB ID, are nevertheless frequently correct predictions. The analysis shows, in total, that the ranking of *iRAISE* of the first true positive is comparable to the pharmacophore-based method and clearly outperforms both docking-based methods Surfex1 and Plants1. The docking

Table 1. Median Ranking of First True Positive Identified by Pharmacophore-Based Methods (rigid1, rigid2, flex1, flex2), Two Docking Methods (Surflex1 and Plants1), Two Docking Plus Interaction Fingerprint-Based Methods (Surflex2 and Plants2), for iRAISE with Conformations (iRAISE flex), and for iRAISE with the Crystallized Ligand (iRAISE crystal)^a

method	median rank of first TP on 2556 proteins	median rank in percent of proteins	median rank first TP on 2879 proteins	median rank in percent of proteins	median rank first TP on 2879 proteins with EC-TP ^b	median rank in percent of proteins with EC-TP ^b
rigid1 (pharm)	4	0.16				
rigid2 (pharm)	4	0.16				
flex1 (pharm)	6	0.23				
flex2 (pharm)	4	0.16				
Surflex1	65	2.5				
Surflex2	11	0.43				
Plants1	113	4.4				
Plants2	29	1.13				
iRAISE flex			33	1.15	8	0.28
iRAISE crystal			2	0.07	2	0.07

^aAll results except those from iRAISE are extracted from the supporting information given in the Meslamani work.⁵⁵ Boldface font indicates numbers that are comparable and should be used for interpretation; other numbers are given for completeness. ^bEC-TP = annotation of true positives (TP) with EC numbers.

methods Surflex1 and Plants1 can only be compared to the iRAISE flex method, since the results of starting them with the co-crystallized ligand are not available. However, the results of Surflex2 and Plants2 can be compared to the iRAISE crystal results, since they take into account interaction fingerprints derived from co-crystallized complexes, which also helps to select the correct conformation. The ranks of the first true positives of all 117 ligands are listed in the Supporting Information.

5. Computing-Time Analysis. To evaluate the computing time of iRAISE, its two steps—the registration procedure and the screening procedure—are evaluated separately. The registration procedure, including the triangle descriptor generation and the protein database generation, takes, on average, ~7 s per target (all time measurements on a workstation with Intel Core i5/3570 CPU@3.4 GHz, 4 cores and 8GB RAM, single-threaded). The screening step requires, on average, 7 s per target (a median of 5 s per target) for a query ligand with an average conformation ensemble size of 200. However, the screening is highly dependent on the structure of the query ligand, ranging from, e.g., 1 s per target for a ligand with few triangle descriptors such as indirubin-3'-monoxime up to 38 s per target for a small hydrophilic ligand with many hits during the descriptor matching step such as pantoate (examples from the Astex Diverse Set). The iRAISE procedure is easily parallelizable with an automatic data partition during precalculation of data chunks of ~100 proteins. Therefore, with a small computing cluster of ~128 cores, a nonredundant PDB protein set with ~50 000 proteins can be screened within ~1–2 h.

CONCLUSION

With iRAISE, we introduce the first structure-based inverse screening method, which deviates from classic reverse docking approaches by applying several measures for facing the challenges of the reverse setup. An abstraction of protein–

ligand matching with a triangle descriptor breaks the sequential screening course and saves computing time. To handle huge amounts of protein structures efficiently and consistently, a protein structure database was introduced. By precalculating and storing descriptors and active sites only once for a set of protein structures, screening with a query ligand can be performed rapidly. The problem of interprotein scoring noise of common docking scoring functions is addressed by a five-step Scoring Cascade, which substantially increases selectivity of the target ranking. To assess the statistical significance of a score for a protein structure, we introduced a Gaussian-based weighting score. Weighting the iRAISE score with it, the ranking of proteins is further improved. The resulting score can be used as a cutoff to decide up to which ranks proteins should be tested experimentally. Such a dynamic approach is better suited than a fixed cutoff at, e.g., 10% of the ranking list, since experimentally testing many targets for a ligand is much more complex than screening the same amount of ligands for one target. Therefore, in target prediction, false positives have a worse effect than in ligand prediction. Adding selectivity in the true positive assignment led to missing some true target structures, because of the strict scoring scheme. Therefore, the balance of selectivity versus sensitivity is an area of improvement in iRAISE.

iRAISE has been evaluated thoroughly, concerning its binding mode prediction and ranking capabilities. On the Astex Diverse Set with 85 diverse high-quality protein–ligand complexes, it has been shown that the Scoring Cascade boosts the ranking of the true target at the first position from 9% to 35%. Furthermore, the ability of iRAISE to predict the correct binding mode was evaluated by root-mean-square deviations (RMSDs) on the Astex Diverse Set. Of the 85 complexes, 74 were redocked with a RMSD value of <2.0 Å. The comparison to classic docking methods shows that iRAISE outperforms these in ranking, because of its measures accounting for protein pocket diversity. Finally, we evaluated the performance of

iRAISE on a large data set of 7915 protein structures and 117 diverse ligands. The first true positive was ranked at 0.28% of the dataset, i.e., it is found among the first 8 ranks (median). In comparison to four pharmacophore-based protocols and two docking-based methods, iRAISE performs comparably and even better, if the same amount of preinformation is incorporated.

So far, iRAISE has only been evaluated on retrospective experiments. Prospective evaluation would be the next step to prove its usability. The iRAISE software is available for Linux operating systems (www.zbh.uni-hamburg.de/raise).

■ ASSOCIATED CONTENT

📄 Supporting Information

List of PDB HET codes excluded for pocket detection. Parameters of simple interaction scoring function. Ranks of first true positives of sc-PDB Diverse Set. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*Tel.: 004940428387350. E-mail: rarey@zbh.uni-hamburg.de.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding

This project is part of the Excellence Cluster in Excellence Initiative by the State of Hamburg "Fundamentals of Synthetic Biological Systems (SynBio)" (www.tu-harburg.de/synbio).

Notes

The authors have made the iRAISE software available at www.zbh.uni-hamburg.de/raise.

■ ACKNOWLEDGMENTS

The authors thank the BioSolveIT GmbH for the opportunity to use the HYDE scoring function and the LeadIT software suite, as well as the NAOMI framework underlying iRAISE. Furthermore, the authors thank Lara Kuhnke from Bayer Pharma AG for technical assistance with the XGlide script.

■ ABBREVIATIONS

TP, true positives; RMSD, root-mean-square deviation

■ REFERENCES

- (1) Khanna, I. Drug discovery in pharmaceutical industry: Productivity challenges and trends. *Drug Discovery Today* **2012**, *17*, 1088–102.
- (2) Azzaoui, K.; Hamon, J.; Faller, B.; Whitebread, S.; Jacoby, E.; Bender, A.; Jenkins, J. L.; Urban, L. Modeling promiscuity based on in vitro safety pharmacology profiling data. *ChemMedChem* **2007**, *2* (6), 874–880.
- (3) Huggins, D. J.; Sherman, W.; Tidor, B. Rational approaches to improving selectivity in drug design. *J. Med. Chem.* **2012**, *55* (4), 1424–1444.
- (4) Ashburn, T. T.; Thor, K. B. Drug repositioning: Identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discovery* **2004**, *3* (8), 673–683.
- (5) Liu, Z.; Fang, H.; Reagan, K.; Xu, X.; Mendrick, D. L.; Slikker, W., Jr; Tong, W. *In silico* drug repositioning—what we need to know. *Drug Discovery Today* **2013**, *18* (3), 110–115.
- (6) Ekins, S.; Williams, A. J.; Krasowski, M. D.; Freundlich, J. S. *In silico* repositioning of approved drugs for rare and neglected diseases. *Drug Discovery Today* **2011**, *16* (7), 298–310.
- (7) Roth, B. L.; Sheer, D. J.; Kroeze, W. K. Magic shotguns versus magic bullets: Selectively non-selective drugs for mood disorders and schizophrenia. *Nat. Rev. Drug Discovery* **2004**, *3* (4), 353–359.
- (8) Medina-Franco, J. L.; Giulianotti, M. A.; Welmaker, G. S.; Houghten, R. A. Shifting from the single to the multitarget paradigm in drug discovery. *Drug Discovery Today* **2013**, *18*, 495–501.
- (9) Bottegoni, G.; Favia, A. D.; Recanatini, M.; Cavalli, A. The role of fragment-based and computational methods in polypharmacology. *Drug Discovery Today* **2012**, *17* (1), 23–34.
- (10) Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. Drugbank: A comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **2006**, *34*, D668–D672 (www.drugbank.ca, accessed December 2013).
- (11) Nobeli, I.; Favia, A. D.; Thornton, J. M. Protein promiscuity and its implications for biotechnology. *Nat. Biotechnol.* **2009**, *27* (2), 157–167.
- (12) Ekins, S.; Mestres, J.; Testa, B. In silico pharmacology for drug discovery: Applications to targets and beyond. *Br. J. Pharmacol.* **2007**, *152*, 21–37.
- (13) Jenwithesuk, E.; Horst, J. A.; Rivas, K. L.; Van Voorhis, W. C.; Samudrala, R. Novel paradigms for drug discovery: Computational multitarget screening. *Trends Pharmacol. Sci.* **2008**, *29* (2), 62–71.
- (14) Nijima, S.; Yabuuchi, H.; Okuno, Y. Cross-target view to feature selection: Identification of molecular interaction features in ligand-target space. *J. Chem. Inf. Model.* **2010**, *51* (1), 15–24.
- (15) Keiser, M. J.; Roth, B. L.; Armbruster, B. L.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **2007**, *25* (2), 197–206.
- (16) AbdulHameed, M. D. M.; Chaudhury, S.; Singh, N.; Sun, H.; Wallqvist, A.; Tawa, G. J. Exploring polypharmacology using a ROCS-based target fishing approach. *J. Chem. Inf. Model.* **2012**, *52* (2), 492–505.
- (17) Kinnings, S. L.; Jackson, R. M. ReverseScreen3D: A structure-based ligand matching method to identify protein targets. *J. Chem. Inf. Model.* **2011**, *51* (3), 624–634.
- (18) Nettles, J. H.; Jenkins, A.; Bender, A.; Deng, Z.; Davies, J. W.; Glick, M. Bridging chemical and biological space: "target fishing" using 2D and 3D molecular descriptors. *J. Med. Chem.* **2006**, *49* (23), 6802–6810.
- (19) Hopkins, A. L. Network pharmacology: The next paradigm in drug discovery. *Nat. Chem. Biol.* **2008**, *4* (11), 682–690.
- (20) Gregori-Puigjane, E.; Mestres, J. A ligand-based approach to mining the chemogenomic space of drugs. *Comb. Chem. High. Throughput Screening* **2008**, *11* (8), 669–676.
- (21) Mestres, J.; Gregori-Puigjane, E.; Valverde, S.; Sole, R. V. The topology of drug-target interaction networks: Implicit dependence on drug properties and target families. *Mol. BioSyst.* **2009**, *5* (9), 1051–1057.
- (22) Nonell-Canals, A.; Mestres, J. In silico target profiling of one billion molecules. *Mol. Inf.* **2011**, *30* (5), 405–409.
- (23) Campillos, M.; Kuhn, M.; Gavin, A.-C.; Jensen, L. J.; Bork, P. Drug target identification using side-effect similarity. *Science* **2008**, *321*, 263–266.
- (24) Rognan, D. Structure-based approaches to target fishing and ligand profiling. *Mol. Inf.* **2010**, *29* (3), 176–187.
- (25) Kellenberger, E.; Foata, N.; Rognan, D. Ranking targets in structure-based virtual screening of three-dimensional protein libraries: Methods and problems. *J. Chem. Inf. Model.* **2008**, *48* (5), 1014–1025.
- (26) Gowthaman, R.; Deeds, E. J.; Karanickolas, J. Structural Properties or Non-Traditional Drug Targets Present New Challenges for Virtual Screening. *J. Chem. Inf. Model.* **2013**, *53* (8), 2073–2081.
- (27) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49* (23), 6789–6801.
- (28) Chen, Y. Z.; Zhi, D. G. Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule. *Proteins* **2001**, *43* (2), 21–226.
- (29) Chen, Y. Z.; Ung, C. Y. Prediction of potential toxicity and side effect protein targets of a small molecules by a ligand-protein inverse docking approach. *J. Mol. Graph. Model.* **2001**, *20* (3), 199–218.

- (30) Chen, X.; Ji, Z. L.; Chen, Y. Z. TTD: Therapeutic Target Database. *Nucleic Acids Res.* **2002**, *30* (1), 412–415.
- (31) Gao, Z.; Li, H.; Zhang, H.; Liu, X.; Kang, L.; Luo, X.; Zhu, W.; Chen, K.; Wang, X.; Jiang, H. PDTD: A web-accessible protein database for drug target identification. *BMC Bioinf.* **2008**, *9* (1), 104.
- (32) Li, H.; Gao, Z.; Kang, L.; Zhang, H.; Yang, K.; Yu, K.; Luo, X.; Zhu, W.; Chen, K.; Shen, J.; Wang, X.; Jiang, H. TarFisDock: A web server for identifying drug targets with docking approach. *Nucleic Acids Res.* **2006**, *34*, W219–W224.
- (33) Yang, L.; Chen, J.; He, L. Harvesting Candidate Genes Responsible for Serious Adverse Drug Reactions from a Chemical-Protein Interactome. *PLoS Comput. Biol.* **2009**, *5*, e1000441.
- (34) Yang, L.; Wang, K.; Chen, J.; Jegga, A. G.; Luo, H.; Shi, L.; Wan, C.; Guo, X.; Qin, S.; He, G.; Feng, G.; He, L. Exploring Off-Targets and Off-Systems for Adverse Drug Reactions via Chemical-Protein Interactome—Clozapine-Induced Agranulocytosis as a Case Study. *PLoS Comput. Biol.* **2011**, *7*, e1002016.
- (35) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, B. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
- (36) Kellenberger, E.; Muller, P.; Schalon, C.; Bret, G.; Foata, N.; Rognan, D. sc-PDB: An annotated database of druggable binding sites from the protein databank. *J. Chem. Inf. Model.* **2006**, *46* (2), 717–727.
- (37) Meslamani, J.; Rognan, D.; Kellenberger, E. sc-PDB: A database for identifying variations and multiplicity of 'druggable' binding sites in proteins. *Bioinformatics* **2011**, *27* (9), 1324–1326.
- (38) Paul, N.; Kellenberger, E.; Bret, G.; Müller, P.; Rognan, D. Recovering the true targets of specific ligands by virtual screening of the protein data bank. *Proteins* **2004**, *54* (4), 671–680.
- (39) Muller, P.; Lena, G.; Boilard, E.; Bezzine, S.; Lambeau, G.; Guichard, G.; Rognan, D. In Silico-Guided Target Identification of a Scaffold-Focused Library: 1,3,5-Triazepan-2,6-diones as Novel Phospholipase A2 Inhibitors. *J. Med. Chem.* **2006**, *49*, 6768–6778.
- (40) Wang, W.; Zhou, X.; He, W.; Fan, Y.; Chen, Y.; Chen, X. The interprotein scoring noises in glide docking scores. *Proteins* **2011**, *80* (1), 169–183.
- (41) Strömbergsson, H.; Kleywegt, G. J. A chemogenomics view on protein–ligand spaces. *BMC Bioinf.* **2009**, *10* (6), S13.
- (42) Boström, J.; Hogner, A.; Schmitt, S. Do structurally similar ligands bind in a similar fashion? *J. Med. Chem.* **2006**, *49* (23), 6716–6725.
- (43) Schärf, C.; Schulz-Gasch, T.; Hert, J.; Heinzerling, L.; Schulz, B.; Inhester, T.; Stahl, M.; Rarey, M. CONFECT: Conformations from an Expert Collection of Torsion Patterns. *ChemMedChem.* **2013**, *8* (10), 1690–1700.
- (44) Schlosser, J.; Rarey, M. Beyond the virtual screening paradigm: Structure-based searching for new lead compounds. *J. Chem. Inf. Model.* **2009**, *49*, 800–809.
- (45) Schellhammer, I.; Rarey, M. TriXX: Structure-based molecule indexing for large-scale virtual screening in sublinear time. *J. Comput. Aided Mol. Des.* **2007**, *21* (5), 223–238.
- (46) Wu, K. FastBit: An efficient indexing technology for accelerating data-intensive science. *J. Phys.: Conf. Ser.* **2005**, *16*, 556.
- (47) Weininger, D.; Weininger, A.; Weininger, J. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97–101.
- (48) Hilbig, M.; Urbaczek, S.; Groth, I.; Heuser, S.; Rarey, M. MONA—Interactive manipulation of molecule collections. *J. Cheminf.* **2013**, *5*, 38.
- (49) Urbaczek, S.; Kolodzik, A.; Fischer, J. R.; Lippert, T.; Heuser, S.; Groth, I.; Schulz-Gasch, T.; Rarey, M. NAOMI: On the Almost Trivial Task of Reading Molecules from Different File formats. *J. Chem. Inf. Model.* **2011**, *51* (1), 3199–3207.
- (50) Bietz, S.; Urbaczek, S.; Rarey, M. Protoss: A holistic approach to predict tautomers and protonation states in protein–ligand complexes. *J. Cheminf.* **2014**, DOI: 10.1186/1758-2946-6-12.
- (51) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein–ligand docking performance. *J. Med. Chem.* **2007**, *50* (4), 726–741.
- (52) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261* (3), 470–489.
- (53) Schneider, N.; Hindle, S.; Lange, G.; Klein, R.; Albrecht, J.; Briem, H.; Beyer, K.; Claußen, H.; Gastreich, M.; Lemmen, C.; Rarey, M. Substantial improvements in large-scale redocking and screening using the novel HYDE scoring function. *J. Comput. Aided Mol. Des.* **2012**, *26*, 701–723.
- (54) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-Ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000–1008.
- (55) Meslamani, J.; Li, J.; Sutter, J.; Stevens, A.; Bertrand, H.-O.; Rognan, D. Protein–ligand-based pharmacophores: Generation and utility assessment in computational ligand profiling. *J. Chem. Inf. Model.* **2012**, *52* (4), 943–955.
- (56) <http://cheminfo.u-strasbg.fr>, accessed January 2013.
- (57) Jain, A. N. Surflex-Dock 2.1: Robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *J. Comput. Aided Mol. Des.* **2007**, *21* (5), 281–306.
- (58) Korb, O.; Stutzle, T.; Exner, T. E. Empirical scoring functions for advanced protein–ligand docking with PLANTS. *J. Chem. Inf. Model.* **2009**, *49* (1), 84–96.

Fast protein binding site comparison via an index-based screening technology

[D10] M. v. Behren, A. Volkamer, A. M. Henzler, K. T. Schomburg, **S. Urbaczek**, and M. Rarey. Fast protein binding site comparison via an index-based screening technology. *Journal of Chemical Information and Modeling*, 53(2):411-422, 2013.


<http://pubs.acs.org/articlesonrequest/AOR-jYzSN5YNZ3RmrXcQznwG>

Reproduced with permission from M. v. Behren, A. Volkamer, A. M. Henzler, K. T. Schomburg, S. Urbaczek, and M. Rarey. Fast protein binding site comparison via an index-based screening technology. *Journal of Chemical Information and Modeling*, 53(2):411-422, 2013. Copyright 2013 American Chemical Society.

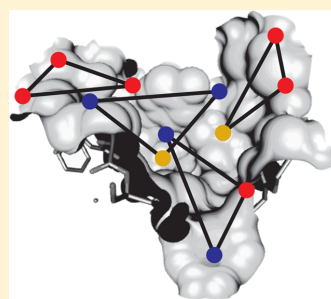
Fast Protein Binding Site Comparison via an Index-Based Screening Technology

Mathias M. von Behren, Andrea Volkamer, Angela M. Henzler, Karen T. Schomburg, Sascha Urbaczek, and Matthias Rarey*

Center for Bioinformatics, University of Hamburg, Bundesstrasse 43, 20146 Hamburg, Germany

 Supporting Information

ABSTRACT: We present TrixP, a new index-based method for fast protein binding site comparison and function prediction. TrixP determines binding site similarities based on the comparison of descriptors that encode pharmacophoric and spatial features. Therefore, it adopts the efficient core components of TrixX, a structure-based virtual screening technology for large compound libraries. TrixP expands this technology by new components in order to allow a screening of protein libraries. TrixP accounts for the inherent flexibility of proteins employing a partial shape matching routine. After the identification of structures with matching pharmacophoric features and geometric shape, TrixP superimposes the binding sites and, finally, assesses their similarity according to the fit of pharmacophoric properties. TrixP is able to find analogies between closely and distantly related binding sites. Recovery rates of 81.8% for similar binding site pairs, assisted by rejecting rates of 99.5% for dissimilar pairs on a test data set containing 1331 pairs, confirm this ability. TrixP exclusively identifies members of the same protein family on top ranking positions out of a library consisting of 9802 binding sites. Furthermore, 30 predicted kinase binding sites can almost perfectly be classified into their known subfamilies.



■ INTRODUCTION

Due to large scale structural genomics projects, the amount of available protein structures in databases expands at an exponential rate.¹ Experimental methods for feature annotation are not able to keep up with this data quantity due to time and cost limitations. Thus, computational methods for automatic analysis, e.g., annotation of protein function or druggability, are of high practical relevance for pharmaceutical and biotechnological industry. Exploiting structures with annotated function, knowledge can be transferred to unknown proteins. Therefore, elucidating similarities between proteins or binding sites can help in many drug discovery contexts, e.g., to address drug promiscuity and polypharmacology. Success stories exist for predicting cross-reactivity,² adverse effects,³ off-targets,⁴ and multidrug resistance.⁵ Furthermore, comparing active sites of enzymes can give hints about substrate specificity or potential mutation sites for enzyme optimization, and thus, assist in rational enzyme design for biotechnological research.

The number of computational methods for binding site comparisons is large^{6,7} and notably either based on sequence or structural similarities. For a long period of time, sequence-based homology transfer has been the gold standard for protein annotation. Knowledge is transferred based on the similarity agreement of multiple sequence alignments of the complete protein sequence (BLAST⁸) or sequence motifs (PROSITE,⁹ BLOCKS,¹⁰ PRINT¹¹). The fast progress in 3D protein structure elucidation, enhanced by the fact that structure was found to be more conserved than sequence,¹² recently promoted structure-based methods for protein comparison.

Classically, structure-based comparison methods rely on multiple structural alignments of complete structures (FAT-CAT,¹³ PAST,¹⁴ VAST,¹⁵ 3DCOMB¹⁶) or structural fragments (PROCAT¹⁷). Nevertheless, the amount of required overall sequence¹⁸ or structure identity to reliably transfer function information is debatable. Examples showed that nonhomologous proteins—in terms of overall sequence or structure—also share functions, which shifted the focus toward binding site analysis. Specific interaction partners and their arrangement are responsible for the recognition and binding of small molecules, hence, determining the protein's function. Thus, most approaches follow the assumption that similar ligands bind to similar cavities.¹⁹ As stated in several reviews,^{6,7} the three main components of methods for binding-site comparison are molecular recognition feature encoding, similarity searching, and scoring. In the first step, the complexity of the comparison problem is reduced by using simplified representations of the binding site encoded in structural features. Second, similarities are identified for these representations, mostly by structural alignment overlap or fingerprint comparisons. Third, a scoring function is applied to quantify the similarity between two sites.

The number of approaches trying to solve the comparison problem is manifold and can be rudimentarily divided into alignment-based and alignment-free methods. Alignment-based algorithms rely on superimposing two structures. Strategies used for structural alignments are mostly geometric matching

Received: October 1, 2012

Published: February 7, 2013

(ProSurfer,²⁰ SuMo,²¹ SiteBase²²), geometric hashing (SiteEngine²³), or clique detection (CSC,²⁴ Cavbase,^{25,26} eFsite,²⁷ eFseek,²⁸ IsoCleft,²⁹ ProBis³⁰). Cavbase, e.g., uses a grid representation of the binding site, in which cavity-flanking residues are mapped to pseudocenters, representing the chemical properties of the binding site. Cavities are compared using a clique detection algorithm identifying three-dimensional (3D) pseudocenter arrangements that are common for two cavities. Since the alignment of structures is computationally expensive, effort is undertaken to develop alignment-free methods. A common approach is to convert cavity properties into simple 1D fingerprints, facilitating high-throughput comparisons.^{31–38} Due to the speed of these methods, large screening scenarios with a few thousand up to a million binding site comparisons are feasible. One group of methods compares lists of sorted distances between, e.g., critical atoms (PocketMatch³¹), conserved atoms,³² and surface curvature.³³ Other methods analyze distances between centroids of fragment pairs³⁴ or property-encoded shape distributions (PESD).³⁵ Using pharmacophore-based fingerprints is another prominent approach (FLAP,³⁶ SiteAlign,³⁷ FuzCav³⁸). In SiteAlign,³⁷ a fingerprint overlap based on properties projected on an 80 triangle-discretized sphere is introduced. The mapping on the sphere comparison allows for easy alignments. A subsequent development, FuzCav³⁸ is completely alignment-free and performs comparisons based on a fingerprint of counts of pharmacophoric triplets. Moment-based methods use rotational invariant pocket representations by 3D mathematical functions that describe the protein surface space. Spherical-harmonics¹⁹ or 3D Zernike descriptors³⁹ are employed to represent the structure as a vector of coefficients of the function series. Repurposing methods from other fields like image processing⁴⁰ or word processing⁴¹ also proved useful. Merelli et al.,⁴⁰ e.g., used spin-images for surface matching. Pang et al.⁴¹ calculate a “visual words” descriptor and uncover similarities between binding sites based on a fast algorithm from the information retrieval area. Ito et al.⁴² achieved a good running time by representing the binding site as a bit string, combined with the application of ultra fast all pair similarity search methods.

Nevertheless, the speed of structural alignment free methods entails a lack of interpretability of the results. Besides the fingerprint-based similarity score, no information about the features responsible for the similarity is returned. Thus, the shortcomings of both—the slower character of alignment-based and the low interpretability of alignment-free methods—have recently been faced.^{43,44} BSAAlign,⁴³ e.g., finds the largest common subgraph based on clique detection and subgraph isomorphism with high-throughput. A sparse graph is built based on residues—together with geometric and physicochemical information—instead of point-based representations. An efficient algorithm has been invented to circumvent the computational expensive (NP-hard) problem of finding the maximum common subgraph.

In this work, we introduce TrixP, a new method for index-based binding site comparison which falls in the latter category of alignment-based but efficient algorithms. TrixP allows for fast structure-based screening of a query binding site against a library of precalculated sites. The method exploits the main advantages of TrixX,⁴⁵ a method for structure-based high-throughput screening. Pharmacophoric features present in the binding site are identified and a triangle descriptor—together with an 80-ray bulk spanned from the triangle center—is used

to represent physicochemical and spatial information of the binding site. The use of a bitmap index⁴⁶ and an efficient data partitioning scheme avoids the sequential evaluation of binding sites. Binding sites can either be identified by providing a reference ligand or automatically predicted by the built-in DoGSite⁴⁷ method. For a query protein, descriptors are calculated and binding sites with matching descriptors are returned from the bitmap index. The respective sites are superposed onto the query based on calculated clusters of matching descriptors. A scoring scheme, considering matching and mismatching pharmacophoric interaction sites, is introduced to rank the library binding sites by their similarity to the query. The method is evaluated on a set of 1331 pairs³⁸ and successfully retrieves 81.8% of the similar pairs while rejecting 99.5% of the dissimilar pairs. These results are in good agreement with the results achieved by FuzCav.³⁸ Furthermore, an index is built on 9802 structures from the sc-PDB⁴⁸ and screened against four different protein families. Querying the index with an estrogen receptor, e.g., delivers a ranked list of similar sites with 98.5% of the contained estrogen receptors among the top ranking positions. Next, the method is used to classify kinases into subfamilies, and achieves classifications similar to those of Cavbase and SCOP.⁴⁹ Furthermore, the quality and runtime of TrixP is compared to other recently published methods, on a data set containing eight protein pairs sharing only partial similarities which are hard-to-detect.³¹ TrixP finds similarities for seven of the eight pairs in a few seconds per comparison. Thus, the running time is comparable to BSAAlign, another alignment-based algorithm and faster than earlier alignment-based methods, which are in the order of minutes. Nevertheless, alignment-based methods are still slower than 1D fingerprint methods, which perform comparisons in the order of milliseconds. Finally, high-throughput screening studies are executed in parallel on the eight cores of an Intel(R) Xeon(R) E5630 @ 2.53 GHz machine with 32 GB RAM. Building an index takes 6.3 h for the sc-PDB data set with 9802 protein binding sites but has to be done only once. The estrogen receptor query on this library including query preprocessing, matching, and scoring takes 37.5 min.

The high recovery rate and the speed of the method show its importance in fields like protein function prediction, rational enzyme design, and polypharmacology.

DATA PREPARATION

Sc-PDB Data Sets for Method Evaluation. The sc-PDB⁴⁸ data set, released in 2011, containing 9877 entries, corresponding to 3034 different proteins and 5339 different ligands, has been downloaded and used throughout this work. Due to nonstandard PDB annotations or errors in the respective ligand mol2 files, 75 of the entries are discarded by NAOMI,⁵⁰ yielding a total of 9802 structures in our data set.

To determine a reliable score cutoff within TrixP, a similar and dissimilar pair sc-PDB subset is used. The pair sets have originally been setup by Weill and Rognan³⁸ to define a cutoff for FuzCav. Starting from the complete sc-PDB (Version 2008), Weill et al. clustered the entries according to their UniProt name. Subsequently, they randomly selected two entries (only cofactor-free ones) from each cluster based on the SiteAlign³⁷ distance value, yielding 769 pairs. The same number of dissimilar pairs has been selected from the clusters, with the requirement of having an EC number differing at the first level. Due to changes between the sc-PDB versions and some discards by NAOMI, only 683 similar and 648 dissimilar pairs

Table 1. Overview of All Protein Pairs of the Benchmark Data Set

first protein	protein family	second protein	protein family
1gjc	utpa ^a	1v2q	trypsin
1gjc		2ayw	trypsin
1gjc		1o3p	utpa ^a
1ecm	chorismate mutase	4csm	chorismate mutase
1m6z	cytochrome c4	1lga	peroxidase
1zid	enoyl-ACP reductase	2cig	dihydrofolate reductase
1v07	mini-hemoglobin	1hbi	hemoglobin
6cox	prostaglandin G/H synthase 2	1oq5	carbonic anhydrase 2

^aurokinase type plasminogen activator.

could be recovered from the 2008 version (used in SiteAlign) and are used within the TrixP study.

Since the entries of the sc-PDB are always annotated with a drug-like cocrystallized ligand, we used those ligands to determine the binding sites of the proteins within the sc-PDB experiments. Therefore, we selected every amino acid within a radius of 6.5 Å as part of the binding site. To evaluate the performance of TrixP on this data set, different protein families with a sufficient large amount of representatives are chosen. Iteratively, one randomly chosen representative of each of these families is used to query the complete data set. The chosen families are estrogen receptors (PDB codes: 1qkt, 1l2j, 2ewp), proteases (2q54), reverse transcriptases (1klm), and carbonic anhydrases (3bet). The sc-PDB contains 34 estrogen receptors α , 23 estrogen receptors β , and 5 estrogen related receptors γ . Furthermore, the sc-PDB contains five estrogen receptors, which are not further specified. According to the information present in the PDB entries of those proteins, two of them can be counted as estrogen receptors α (3l03 and 3hly). The remaining three estrogen receptors (2yat, 3os9, and 3osa) have to be labeled as “unknown form” during the result evaluation. The other three families consist of 174 proteases, 75 reverse transcriptases, and 105 carbonic anhydrases.

Kinase Data Set for Subfamily Based Classification. To show TrixP's sensitivity in protein family annotation, we perform a classification study on kinases, an enzyme class which is of special biochemical interest. In 2006,²⁶ Kuhn et al. collected a set of eukaryotic protein kinases to evaluate the performance of Cavbase. This data set consists of 30 binding sites of 28 kinases from five different kinase subfamilies. The challenge within this classification problem lies in the separation of closely related subfamilies, containing active and inactive conformations with significant differences in the local conformation of the ATP binding sites. In this experiment, binding sites are predicted using the DoGSite⁴⁷ method, likewise for holo- and apo-structures. To distinguish between the different families, an all vs. all comparison is performed. The resulting similarity matrix is used as input for a hierarchical clustering procedure.

Benchmark Data Set for Comparison Study. To directly compare TrixP with other recently published methods for binding site comparison, a data set originally introduced by Yeturu et al.³¹ and extended by Weill et al.³⁸ is used. The data set contains eight binding site pairs: three from the same SCOP family and five belonging to different SCOP families. The ligands present in 1v2q, 2ayw, and 1o3p are all bound to the same binding site, but in different orientations and interacting with varying residues. The four remaining pairs introduced by Yeturu et al. contain different folds, and therefore even belong to different SCOP families, but show similar binding sites as

reported in the literature. Furthermore, Weill et al. included an additional pair of two proteins showing a cross-reactivity, explained by the similarity of a small-sized subpocket within both binding sites. For this data set, we again used the respective ligands to determine the binding sites. The pdb codes of the eight pairs present in the data set can be seen in Table 1.

METHODS

TrixP is based substantially on the TrixX technology, a novel approach for structure-based virtual screening of large compound libraries. For a detailed description of the technology we refer to the original publications.^{45,51} Here, we briefly overview the basic concepts of TrixX and focus on the explanation of the adaptations, necessary to employ this technology for pocket similarity prediction. TrixP compares pockets, i.e., it screens a library of protein binding sites and identifies matching binding site descriptors. Therefore, TrixP follows the general TrixX proceeding of a first library indexing step followed by a screening step, in order to avoid repetitive calculations and to perform efficient virtual screening runs. Similarly in both methods, triangle descriptors for a given library are calculated and stored in a bitmap index, during the indexing phase. This bitmap index is created once and is reusable in subsequent virtual screening runs. In the screening phase, descriptors are derived for the binding site of a query protein and only matching descriptors and their associated structures are extracted from the index. Those hits are then transformed and scored according to their agreement of pharmacophoric features with the binding site of the query protein. The TrixP method mainly differs in four points from TrixX: First, instead of small molecules TrixP stores descriptors derived from protein binding sites in the index. Second, in order to account for the inherent flexibility of proteins, TrixP employs an adapted descriptor matching method. Third, instead of placing small molecules into the binding site, it aligns binding sites. And finally, TrixP uses a new scoring scheme that assesses the hits according to their similarity with the query protein. We introduce the new concepts in the following sections in more detail.

Recognition Feature Encoding. TrixP identifies the similarity between proteins by comparing pharmacophoric binding site features presumably responsible for the recognition of ligands. The starting point for the calculation of descriptors encoding these features is a binding site of a protein which can be determined using different strategies. The most straightforward way is the use of a reference ligand to identify surrounding residues or atoms. In this case, every amino acid within a distance threshold of 6.5 Å has been selected. Alternatively,

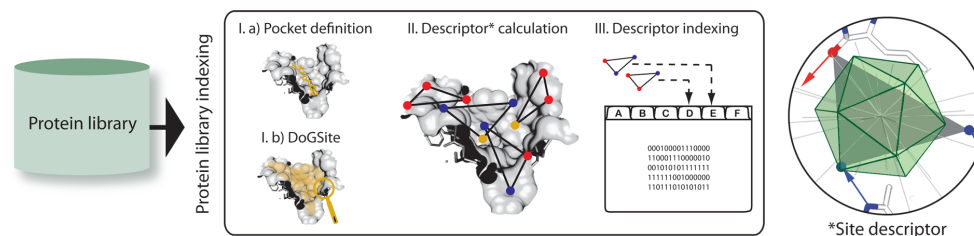


Figure 1. (left) Recognition feature encoding. For a given protein library, binding sites are detected and attributes of site descriptors are converted and stored in a bitmap index. (right) The Binding site descriptor encodes three types of pharmacophoric features in its triangle corners, three main interaction directions if the corners are of hydrophilic type, three triangle side lengths that describe the relative arrangement of the pharmacophores, and a set of 80 bulk rays through the 20 triangle faces of an icosahedron (four rays per triangle face) that locally describe the interior volume of a pocket.

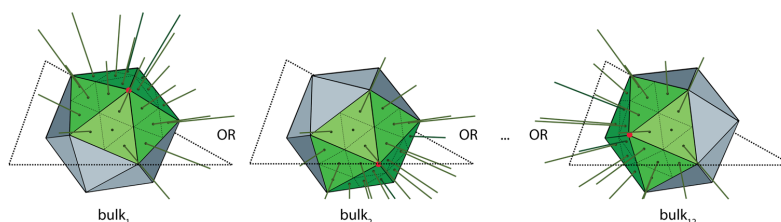


Figure 2. Partial bulk implementation: An icosahedron is orientated relative to a pharmacophoric triangle. For each vertex of the icosahedron, all rays going through a surrounding triangle are used to define a partial bulk. Combining them with a logical OR during descriptor matching indirectly introduces protein flexibility as only 25% of the shape of the compared binding sites have to match.

potential binding sites can be detected ligand-independently using the built-in DoGSite⁴⁷ method.

For each detected pocket, TrixX triangle descriptors are calculated based on present pharmacophoric features. The used triangle descriptors, hereby, resemble three-point pharmacophores, and the triangle corners have one of the three types: donor, acceptor, or hydrophobic. Hydrophilic features are generated from hydrogen-bond donor and acceptor atoms and possess potential main interaction directions. These directions indicate the locations of hydrogens or lone pairs, respectively. Hydrophobic features are derived from a grid placed in the binding site and are undirected. Grid points with a sufficient number of surrounding hydrophobic atoms represent hydrophobic regions and therefore become hydrophobic features. Since hydrophilic features are far more specific than hydrophobic features, triangle descriptors with only hydrophobic corners are not allowed. Nevertheless, hydrophobic features are of great importance since they increase the number of possible active site superpositions. Additionally, the descriptor is equipped with 80 steric bulk rays, aligned in an icosahedron, radiating from the center of the triangle. These rays represent the shape of the pocket relative to the triangle, since the length of each ray is the distance of the triangle center to the surface of the binding site. Due to the large number of possible triangle descriptors (on average 6090) per binding site, the derived data requires an efficient space management. For descriptors derived from the protein library, this is realized by binning the features of the descriptors and converting them into bitmaps. Thereby, the descriptors are separated and stored according to their descriptor attributes (types of corners, directions, lengths of triangle sides and of bulk rays) in the triangle descriptor index. Figure 1 summarizes the workflow of recognition feature encoding and depicts a descriptor of a binding site.

Similarity Searching. A comparison of query and library descriptors can reveal similarities between associated proteins. Therefore, descriptors are generated from the query protein and used to formulate logical expressions that directly access and extract only similar descriptors, and thus, similar pockets from the index. A query descriptor matches if the types of the corners, the directions, the lengths of the triangle sides, and the lengths of each of the 80 bulk rays are in accordance with a descriptor of the index. In order to allow a certain amount of structural flexibility during the matching procedure, tolerances are added to the lengths of the triangle sides and bulk rays. Then, the associated structures are identified as potential similar proteins. Due to data partitioning by type, the index structure avoids the evaluation of dissimilar descriptors and supports a rapid data querying. However, since even closely related binding sites exhibit differences in their overall shape, the bulk descriptor matching in its original form turned out to be a too rigorous matching criterion. Furthermore, the TrixX bulk descriptor ensures that the ligand completely fits into the binding site avoiding steric overlap. While steric overlap is forbidden in general, shape similarity considered in TrixP can also occur partially. Therefore, the shape-descriptor matching procedure is adapted to allow matches with only partial shape agreement. As depicted in Figure 2, this partial bulk implementation uses multiple subsets of rays as matching criteria for querying. A shape requirement of 25% is realized using only those rays going through triangles surrounding a particular icosahedron vertex. Since a vertex is enclosed by five triangles, only 20 out of 80 rays are selected as matching constraints at a time. Each vertex of the icosahedron defines a subset of rays leading to 12 possible sets of rays for a single query descriptor. These subsets are logically ORed during the evaluation of a query descriptor, i.e., it is sufficient if only 25%

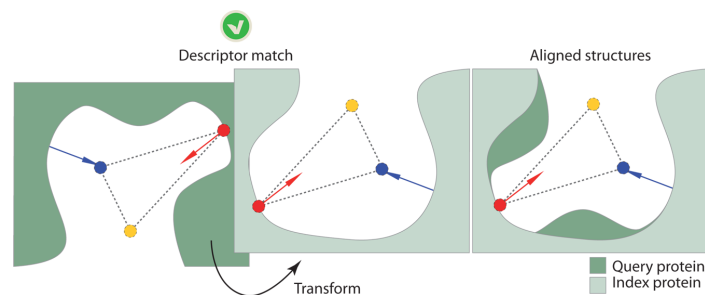


Figure 3. Structure alignment: Schematic superposition of binding sites based on a descriptor match. The colors of the triangle corners indicate their respective type: hydrophobic (yellow), hydrogen-bond donor (blue), or hydrogen-bond acceptor (red).

of the binding site shape of a library protein matches to indicate similarity with the query protein. Note that this matching 25% has to occur in one connected region of the active site.

Structure Alignment. The search procedure of TrixP results in a list of matching query and library descriptors. Each triangle descriptor match holds the information to superpose a pair of binding sites. The transformation of the query triangle onto the matching index triangle is calculated and applied to the coordinates of the query binding site. In order to reduce the number, as well as to improve the quality, of the transformations, matching descriptor pairs are clustered.⁵² The aim of the clustering is to identify groups of descriptor pairs whose transformation results in a very similar superimposition of the binding sites. A complete linkage clustering algorithm compares the descriptor matches on the basis of their transformation result. To evaluate the distance between any two descriptor matches, both query descriptors are transformed once with each of the two corresponding transformations. The RMSD between the resulting triangle descriptor corner coordinates of both transformations is used as distance measure for the clustering algorithm. In the end, only one combined transformation is calculated for each cluster by optimizing the simultaneous overlay of all included query pharmacophore points on their respective matching points in the index structure. This transformation is applied to superimpose the binding sites for subsequent scoring. Figure 3 illustrates the structure alignment resulting from a transformation based on a matching triangle descriptor.

Scoring. For each returned alignment of the query to a binding site of the library, named target in the following, a similarity score is calculated based on the compliance of pharmacophoric features. Therefore, let Q be the set of features q of a query protein and T be the set of target features t . Furthermore, let $A(Q, T)$ be the set of alignments of Q onto T gained by the clustering method. The similarity $S(Q, T)$ between Q and T maximizes the similarity scores $s_a(Q, T)$ of all structural alignments of $A(Q, T)$. $s_a(Q, T)$ is determined by scanning the environments of the query features for matching features in the target.

$$S(Q, T) = \max_{a \in A(Q, T)} \{s_a(Q, T)\} \quad (1)$$

$$s_a(Q, T) = \sum_{i=1}^{|Q|} (s_a(q_i, T_{\text{sphere}}(q_i))) \quad (2)$$

The function $\text{mtype}(q, t)$ discriminates between three different matching scenarios:

$$\text{mtype}(q, t) = \begin{cases} \text{dir:} & \text{if } q, t \text{ have the same directed type} \\ & \text{(directed match)} \\ \text{undir:} & \text{if } q, t \text{ have the same undirected type} \\ & \text{(undirected match)} \\ \text{mis:} & \text{if } q, t \text{ have different types (mismatch)} \end{cases} \quad (3)$$

Furthermore, we define $T_{\text{sphere}}(q_i)$ to be the set of target features with a maximum distance $d_{\text{max}} = 1.5 \text{ \AA}$ from q_i , i.e., $T_{\text{sphere}}(q_i) = \{t \in T \mid d(q_i, t) \leq d_{\text{max}}\}$. For each query feature q_i of an alignment a , $s_a(q_i, T_{\text{sphere}}(q_i))$ honors matching and penalizes mismatching features and, thus, reflects the similarity of feature q_i to its close environment:

$$s_a(q_i, T_{\text{sphere}}(q_i)) = \begin{cases} \max_{t_j \in T_{\text{sphere}}(q_i)} \{s_{\text{undir}}(q_i, t_j)\}: & \text{if only hydrophobic} \\ & \text{matches} \\ \frac{1}{n} \left(\sum_{j=1}^n (s_{\text{mtype}(q_i, t_j)}(q_i, t_j)) \right): & \text{otherwise} \end{cases} \quad (4)$$

Where n is the number of target features within $T_{\text{sphere}}(q_i)$. The individual similarity scores $s_{\text{dir}}(q_i, t_j)$ of directed hydrophilic matches, $s_{\text{undir}}(q_i, t_j)$ of hydrophobic matches, and $s_{\text{mis}}(q_i, t_j)$ of mismatches of the query feature q_i and the target feature(s) are defined as follows:

$$s_{\text{dir}}(q_i, t_j) = s_{\text{max}} \left[\left(1 - \frac{d(q_i, t_j)}{d_{\text{max}}} \right) + w \left(1 - \frac{\alpha(q_i, t_j)}{\alpha_{\text{max}}} \right) \right] \quad (5)$$

$$s_{\text{undir}}(q_i, t_j) = s_{\text{max}} \left(1 - \frac{d(q_i, t_j)}{d_{\text{max}}} \right) \quad (6)$$

$$s_{\text{mis}}(q_i, t_j) = p_{\text{max}} \left(1 - \frac{d(q_i, t_j)}{d_{\text{max}}} \right) \quad (7)$$

Figure 4 illustrates possible matching cases that might occur during the scanning and scoring of local query feature environments.

- (a) The similarity $s_{\text{dir}}(q_i, t_j)$ between directed hydrophilic features is determined linearly based on the distance $d(q_i, t_j)$ and the angle difference $\alpha(q_i, t_j)$ between the main interaction directions of the features. Therefore, we define the maximal score $s_{\text{max}} = 10$ and the angle weight parameter $w = 0.8$, resulting in an absolute score of 18 for

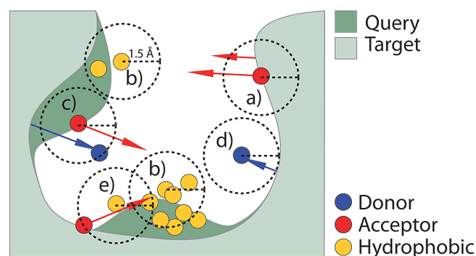


Figure 4. Schematic depiction of the scoring procedure: Around a query feature a sphere with 1.5 Å radius is placed and scanned for matching target features. (a) Match of hydrophilic features. (b) Match of hydrophobic features. (c) Mismatch of features. (d) No match within the 1.5 Å around a donor feature. (e) Mismatch between a hydrophobic and an acceptor feature, and a simultaneous match with a hydrophobic feature.

a perfect overlay of a query and a target feature. The hydrophilic score drops to 0 if $d_{\max} = 1.5 \text{ \AA}$ and $\alpha_{\max} = 65^\circ$ is reached.

- As hydrophobic features are undirected, the score $s_{\text{undir}}(q_B, T_{\text{sphere}}(q_i))$ only depends on the distance of the features. However, since hydrophobic features often appear in clusters of discrete features describing hydrophobic binding site volumes, a special case is introduced for the match of a hydrophobic feature with multiple other hydrophobic features. In this case, the score $s_{\text{undir}}(q_B, t_j)$ is maximized over all hydrophobic target features in $T_{\text{sphere}}(q_i)$.
- If the types of features differ but are located close to each other, this mismatch of query and target features is penalized by $p_{\max} = -2$.
- Generally, if no match or mismatch is identified in $T_{\text{sphere}}(q_i)$, there is no contribution to the score since there is no evidence for similarity or an explicit mismatch in such a case.
- Simultaneous matches or mismatches in the query feature sphere are seldom. However, in these rare cases their contributions are averaged in order to account for the heterogeneity of matched regions.

Finally, in order to grant comparability, the similarity score $s_a(Q, T)$ is normalized with respect to the query protein to reflect a value between 0 and 1. All chosen parameters within

the equations, e.g. maximal score, distance, and angle, as well as angle weight and maximal penalty, have been optimized on a small training set (see Supporting Information Material A) and proved to produce a reliable score for the overall similarity of two binding sites.

RESULTS AND DISCUSSION

In the following, different aspects of the presented binding site comparison tool are analyzed. First, TrixP is evaluated in terms of its ability to find similar sites while discarding dissimilar ones based on ligand-defined binding sites, with respect to studies from FuzCav.³⁸ Second, the methods capability to distinguish between subfamilies based on predicted binding sites is investigated and the results are compared to Cavbase.⁵³ Finally, several benchmark studies are executed comparing TrixP to other recently published efficient algorithms³⁸ and showing its potential as a high-throughput method.

Separating Similar from Dissimilar Protein Pairs.

In a first experiment, the pair data set introduced by Weill et al.³⁸ is used to determine a reliable cutoff value for the TrixP similarity score. Two indices are built containing all similar and dissimilar pairs, respectively. For each pair A, B, protein A is used to query the corresponding index.

First, the TrixP similarity score between each pair is investigated. The average score of all similar pairs is 0.46, while the average score of all dissimilar pairs is 0.17. A histogram of the achieved scores for similar and dissimilar pairs (see Figure 5) shows that a TrixP score of 0.3 is well suited to distinguish between similar and dissimilar binding sites. With this cutoff value, 81.8% of all similar pairs can be retrieved, while 99.5% of all dissimilar pairs are discarded.

Second, aside from the pairwise comparisons of proteins, the screening procedure allows to rank the respective partner relative to all other 1366 proteins in the index of the pair screening run. The first finding is that the method recovers the protein itself, which is also contained in the index, as top ranking hit (self-match) in all cases. This self-match is excluded from the following analysis. Figure 5 shows the distribution of the position at which the respective similar pair occurs within the result list. TrixP retrieves 69.4% of all similar pairs at the top ranking position. Furthermore, only 18.7% of the pairs are found on a rank below four. In total, 209 respective pairs do not occur at the first rank. In 54% of the cases, the best ranking hit as well as the query have an annotated EC number, which can be used to assess the quality of those matches. In the majority

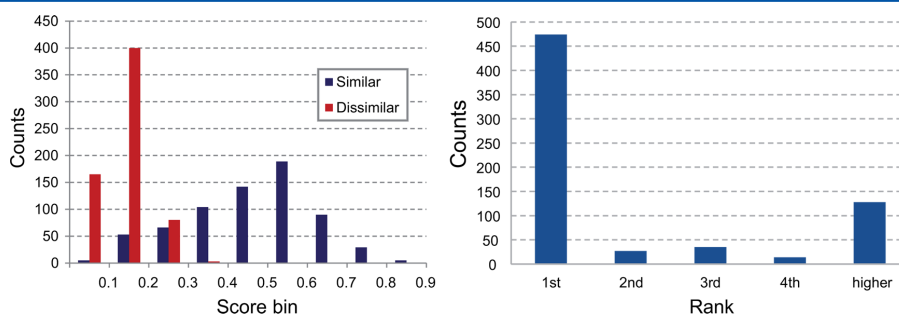


Figure 5. Results of the sc-PDB pair data set screening. (left) Distribution of the achieved scores for similar (blue) and dissimilar (red) pairs, displayed is the respective number of pairs within a certain score range. (right) Number of respective pairs found on a certain rank within the screening results sorted by similarity score.

of those cases, the top ranking match had at least the first three EC digits in common with the query. Only for as few as 0.7%, the method is unable to recover the respective pair at all. A further examination of those cases showed problems during the correct binding site determination, leading to disproportionately small binding sites.

In order to further demonstrate the discriminative power of TrixP, the scores of all similar (true positives) and dissimilar pairs (false positives) are sorted in descending order resulting in a ROC curve (see Supporting Information Material B). Generally, in the first 41.5% of the ranked data points true positives are exclusively found. An AUC of 0.96 is achieved for the performance of TrixP. Although the pair data set slightly differs (see above), these results are in good agreement with the data published by Weill et al.³⁸

sc-PDB Screening. Multiple screening runs are performed on an index containing all 9802 binding sites of the sc-PDB data set, measuring the potential of TrixP to select binding sites similar to a query site. Six proteins from four different protein families, i.e., carbonic anhydrase 2 (CA2), protease (PR), reverse transcriptase (RT), and estrogen receptor (ER), are chosen as examples to query the index. For all queries, in total only 1–9% of the binding sites in the data set are returned as matches with a TrixP score above 0.3 (Table 2). For each

Table 2. sc-PDB Screening Results Using Different Queries

protein family	PDB code	family hits	present in sc-PDB	members within top 50	general hits with score > 0.3
CA2	3bet	100	105	50	385
PR	2q54	147	174	50	151
RT	1klm	63	75	48	162
ER α	1qkt	36(66) ^a	36(67)	31(49)	843
ER β	112j	23(67)	23(67)	14(44)	467
ER γ	2ewp	4(63)	5(67)	4(45)	245

^aNumber in parenthesis represents the combined number of all estrogen receptors.

target, the number of family members present in the index is assigned beforehand and the recovery rate per target is analyzed. Between 84% and 100% of the contained family members can be recovered for the respective queries. Furthermore, the 50 top ranking positions are occupied by members of the same family in 88% up to 100% of the cases.

Similar to an experiment performed in the evaluation of SiteAlign,³⁷ the ER α , ER β , and ER γ queries are further investigated. The query with an ER α receptor (1qkt) retrieved in total 98.5% of the ERs present in the library, more precisely all ER α , all ER β , all ER γ structures, and two out of the three nonspecified estrogen receptor (1qkn) achieved a score of 0.20. In contrast to the query, which had been crystallized in complex with the estrogen estradiol, the antagonist raloxifene is bound to 1qkn. These two ligands differ significantly, especially concerning their size. Since the bound ligands have been used to determine the binding sites of the proteins, the significant differences of the bound ligands might be the reason for the low similarity score in this case. Using an ER β structure (112j) as query retrieved all 23 ER β structures, and additionally all 67 other present estrogen receptors. Finally, for an ER γ (2ewp) structure as query, four of five ER γ , 34 of 36 ER α , all ER β structures, and two out of three nonspecified estrogen receptors are recovered. The results for ER α are further analyzed, with respect of high-scoring family and nonfamily members. Two out of the four missed ERs during this screening run, ER γ 2gpp and nonspecified ER 3os9, still achieved a score higher than 0.29. Even if the threshold of 0.3 had not been exceeded in these two cases, both receptors still show a relatively high similarity to the query. Note, that similarity is always rated with respect to the query protein. In the case of the two missed ER α proteins, their ligands differ from the ligand bound to 2ewp and might cause the low similarity. Figure 6 shows the top ranking binding sites up to rank 150. The 16 top ranking positions are exclusively occupied by ER α s, which are still dominant on ranks up to 40.

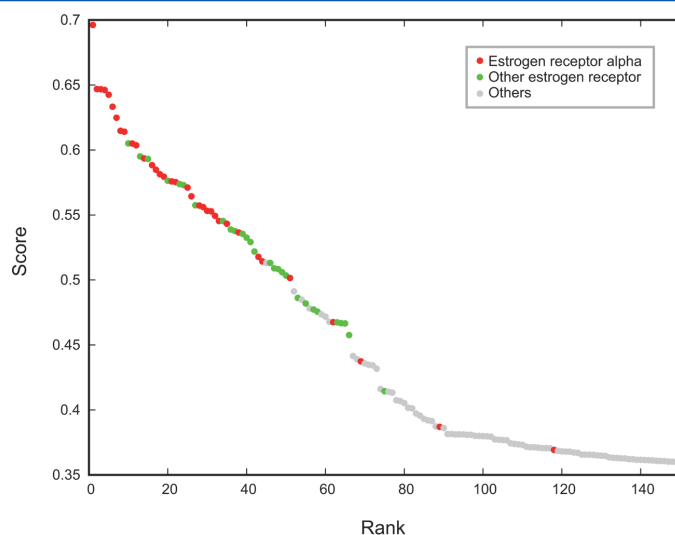


Figure 6. Results for the query with an ER α (PDB code: 1qkt) ranked by TrixP score. ER α and other estrogen receptors are colored in red and green, respectively. Nonestrogen receptor family matches are colored in gray.

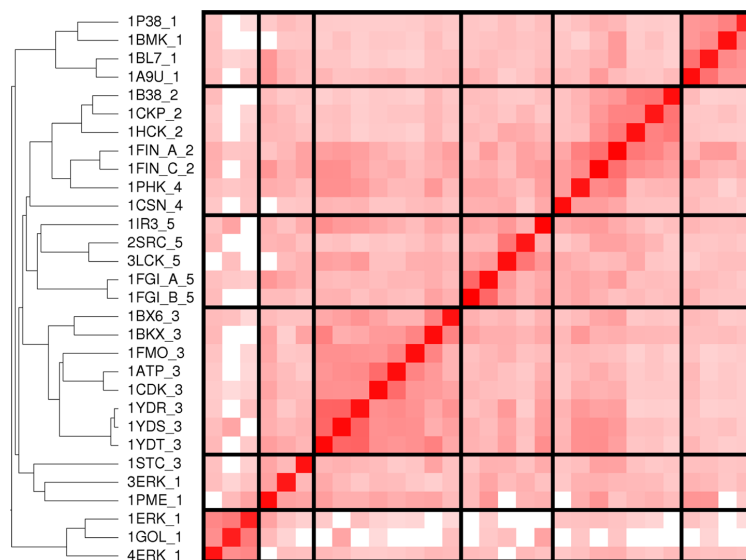


Figure 7. Agglomerative clustering of 30 kinase pockets by TrixP similarity score. SCOP annotation of the structures is indicated as a number: MAP kinases (1), CDK2 (2), PKA (3), Ser/Thr kinases (4), and Tyr kinases (5).

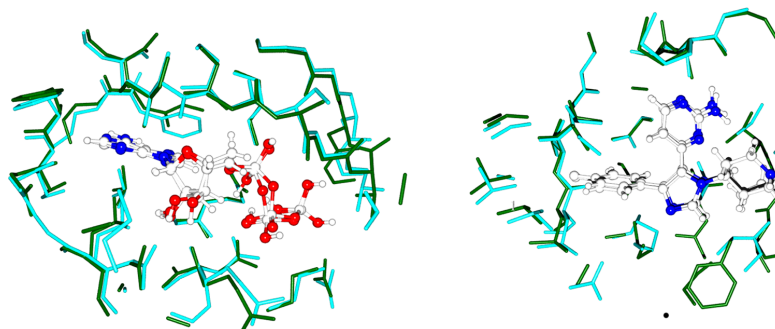


Figure 8. Superimposition of the two cyclin-dependent kinases 2 (CDK2) 1b38 (cyan) and 1hck (dark green) on the left and of the two mitogen-activated protein (MAP) kinases 1bmk (cyan) and 1bl7 (dark green) on the right.

ER β s and ER γ s capture most of the following ranks up to 70. Furthermore, the eight non-ER proteins found within these ranks belong to other human nuclear receptors also binding steroid hormones like progesterone (four), aldosterone (one), glucocorticoid (one), and mineralocorticoid (two).

Kinase Subfamily Detection. In a third experiment, the kinase data set introduced by Kuhn et al.⁵³ is used to evaluate the ability of TrixP to distinguish between closely related binding sites. Binding sites are predicted with the built-in DoGSite method,⁴⁷ and the resulting 30 kinase pockets are stored in the index. The index is queried with all pockets and a hierarchical clustering is performed on the resulting similarity matrix. As previously described in the scoring section, the score of TrixP is calculated with respect to the query binding site and, therefore, not symmetric by design. In order to account for the fuzziness in the definition of the pocket boundary, the maximum of the two respective scores for comparing A vs B and B vs A is used. Applying an agglomerative clustering results

in six clusters (Figure 7). SCOP annotations are indicated by number: MAP kinases (1), CDK2 (2), PKA (3), Ser/Thr kinases (4), and Tyr kinases (5). Clearly, the clustering based on TrixP similarity is in good agreement with the SCOP classification. All Tyr kinases (5) and PKA (3) structures aggregate within one cluster, respectively. The Tyr kinase cluster hereby contains two active (1ir3 and 3lck) as well as three inactive structures (1fgi_a, 1fgi_b, and 2src), which were nevertheless correctly classified as members of the same family. The only exception, hereby, is the PKA structure 1stc, which ended up in a cluster mostly occupied by MAP kinase structures. The binding of a large rigid inhibitor (STU) to 1stc may have introduced a change in its binding site conformation, causing the misclassification. Similar to the findings by Cavbase⁵³ on this data set, TrixP is able to distinguish between the different activation states of CDK2s. The CDK2 (2) main cluster contains two subclusters: one is occupied by inactive CDK2s (1b38, 1ckp, 1hck), and the other one contains active

CDK2s (1fin chain A and C). Furthermore, the two Ser/Thr kinases (4) are assigned to the CDK2 cluster. Although the ligand is not considered within this experiment, all structures (except 1ckp) of this cluster contain a bound ATP, and thus, the method detects the common interaction points within similar distances present in these structures. The MAP kinases (1) span over three clusters. One cluster exclusively contains all structures from the p38 α subfamily (1bmk, 1p38, 1bl7, and 1a9u). A second MAP kinase cluster holds only structures from the Erk2 subfamily (1gol, 1erk, and 4erk). The third cluster contains the remaining two Erk2 structures (1pme and 3erk), paired with the only miss-annotated PKA kinase 1stc. The correct classification of active as well as inactive structures within certain families, like the Tyr kinases and CDK2 kinases, proved the flexibility of TrixP regarding local changes within overall similar binding sites. Figure 8 shows the superimpositions of the CDK2 structures 1b38 and 1hck and of the MAP kinases 1bmk and 1bl7, as examples.

Qualitative and Quantitative Comparison to Other Methods. To compare the performance of TrixP with other recent methods, a small set of eight difficult targets is investigated. Yeturu et al.³¹ and Weill et al.³⁸ evaluated the performance of multiple recent methods on this data set, concerning their ability to identify similarities as well as their run time requirements. As shown in Table 3, TrixP was able to

Table 3. Comparison of TrixP to an Extraction of Other Recently Published Binding Site Comparison Tools^a

PDB1–lig1	PDB2–lig2	efficient alignments		fingerprints	
		TrixP ^b	BSAlign ^c	PocketMatch ^d	FuzCav ^e
pairs of proteins belonging to the same SCOP family					
1gjc–130	1v2q–ANH	0.18	31.77	50.17	0.19
	2ayw–ONO	0.27	31.51	52.29	0.18
	1o3p–655	0.65	42.26	88.01	0.18
pairs of proteins belonging to different SCOP families					
1ecm–TSA	4csm–TSA	0.16	×	55.56	0.18
1m6z–HEC	1lga–HEM	0.24	×	63.85	×
1zid–ZID	2cig–IDG	0.19	×	56.01	×
1v07–HEM	1hbi–HEM	0.43	×	61.42	0.18
6cox–SS8	1oq5–CEL	×	×	×	0.16
speed order		s	s	ms	ms

^aThe full list can be found in the publication of Weill et al.³⁸ ^bTrixP similarity score. ^cBSAlign alignment score.⁴³ ^dPocketMatch PMScore.³¹ ^eFuzCav similarity score.³⁸

assign a similarity score to seven out of those eight difficult pairs. Regarding the three pairs of proteins belonging to the same SCOP families, TrixP like most other methods detects similarities between the sites and exhibits a similar score trend as BSAlign⁴³ and PocketMatch,³¹ by assigning a higher score to the pair of urokinase type plasminogen activators (1gjc and 1o3p). For the five pairs of proteins belonging to different SCOP families, TrixP and PocketMatch are the only methods able to derive a score for four out of the five present pairs, while FuzCav³⁸ is the only method assigning a score to the new pair of a prostaglandin G/H synthase 2 (6cox–1oq5). Furthermore, the TrixP and PocketMatch comparably assign higher scores to the pairs of a cytochrome c4 with a peroxidase (1m6z–1lga) and of a mini-hemoglobin with a hemoglobin (1v07–1hbi). Using the determined threshold of 0.3 for the TrixP similarity score would only yield two cases of possible cross-reactions or

related function among the eight pairs present in this data set. Nevertheless, the results of TrixP show a certain degree of similarity for five of the six remaining pairs and therefore confirm the possibility of the observed cross-reactions. Furthermore, TrixP is able to reproduce the same score trends among the different pairs as BSAlign. Figure 9 shows the superimposition of mini-hemoglobin 1v07 and hemoglobin 1hbi as calculated by TrixP. The calculated score of 0.43 indicates high similarity between the two binding sites even if they belong to different SCOP families and therefore have different folds. The figure shows an almost perfect superimposition of the two heme groups with an RMSD of 0.93 and reasonable alignments of some residues common in both sites. In terms of run-time requirements, the pairwise comparison of the eight protein pairs, using TrixP, takes on average 19 s, including index querying and scoring. Thus, TrixP performs in the same speed order (seconds) as BSAlign, another method designed for efficient alignment-based comparison. Furthermore, both methods are faster than general alignment-based methods executed on this data set (ProFunc,⁵⁴ SitesBase,⁵⁵ SuMo,²¹ SiteEngine⁵⁶), as can be seen in the extended table within the publication of Weill et al. But clearly, the performance of alignment based-methods is still slower than the millisecond range of efficient fingerprint-based methods, such as PocketMatch and FuzCav, which on the other hand often produce results with a lack of interpretability.

Pocket-Based High-Throughput Screening. The TrixP high-throughput screening process can be parallelized by splitting up the data into subindices, simultaneously screening each on one CPU core. As a test scenario, the index in this study is split into eight equal parts, and TrixP is run on the eight cores of an Intel(R) Xeon(R) E5630 @ 2.53 GHz with 32 GB RAM.

The most time-consuming task within TrixP is the initial creation of indices. Calculating descriptors for one binding site and writing them into the bitmap index takes on average 14.25 s. Thus, building the sc-PDB index containing 9802 structures, when equally split onto eight cores for parallel screening, takes 6.3 h, but has to be done only once. The time for screening an index with a protein query depends on two components: First, the number of structures in the index and second, the size of the query's binding site. The first part within TrixP is the matching phase. The average time needed to evaluate the sc-PDB index is 1.76 s per query descriptor. The efficiency within TrixP arises from the usage of the index technology. First, the sequential screening scheme is overcome by efficient horizontal data partitioning based on the descriptor's triangle corner types. Second, the number of binding sites to be scored is greatly reduced to the number of matches returned by the initial index query. The second part of TrixP captures postprocessing—from reinitialization to superposition and scoring. Hence, postprocessing can be done on average in 1.18 s per matching binding site returned by the index query. For ER α (1qkt), 5179 descriptors are calculated, a number close to the average number of 6090 descriptors per binding site for the sc-PDB data set. Using this ER α to query the sc-PDB index, the parallel screening finishes after 37.5 min. Note that the scoring time needed to screen each of the subindices could be further reduced by splitting the data either more reasonable or onto more cores.

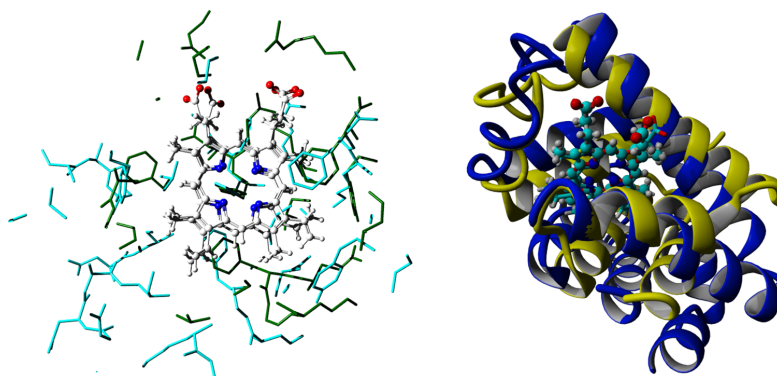


Figure 9. (left) Superimposition of the binding sites of mini-hemoglobin 1v07 (cyan) and the hemoglobin 1hbi (dark green). The bound hemoglobin of each structure is shown in ball-and-stick modus, color-coded in white (except oxygens (red) and nitrogens (blue)). (right) Three-dimensional depiction of the superposed secondary structures of 1v07 (yellow) and 1hbi (blue). The bound hemoglobins are color-coded in cyan (except oxygens (red), nitrogens (blue), and hydrogens (white)).

CONCLUSION

Due to the growing amount of available protein structures, computer methods are required to efficiently tackle the annotation problem. In this study, we introduced TrixP, a new method for fast binding site comparison and function prediction based on structural alignments of the binding sites. The main invention is hereby the representation of binding sites by chemical and structural triangle descriptors, stored in a bitmap index technology, allowing for time-efficient screening.

In multiple experiments, the ability of TrixP to efficiently produce reliable results, comparable or partially superior to other state of the art methods, is shown. Screening two data sets containing known similar and dissimilar binding sites, a reliable cutoff value for the TrixP similarity score is determined. With this cutoff value, 81.8% of all similar pairs can be recovered with TrixP, while rejecting 99.5% of all dissimilar pairs. Furthermore, 69.4% of all similar pairs have been ranked at position one of 1331 screened binding sites. Large scale screening experiments using four different protein families as a query against the sc-PDB index containing 9802 structures are performed. TrixP is capable of identifying similar binding sites to the respective query, to assign an appropriate score to them, and thus, rank related above unrelated binding sites. For each tested protein family, TrixP recovers at least 84% of all family members present in the library. Another experiment on a small data set containing representatives of five kinase subfamilies proved TrixP's ability to distinguish between closely related binding sites.

Besides the quality assessment of TrixP, the efficiency of the method is investigated on a prereleased comparison study on eight binding site pairs. The experiments showed that TrixP is able to perform pairwise comparisons in a few seconds while recovering similarities between so-classified difficult binding sites. Parallel screening, using eight cores, allows TrixP to build the index for the whole sc-PDB database within 6.3 h and afterward to screen it within only 37.5 min.

The application scenarios show the assistance of binding site comparison tools like TrixP to solve important and challenging tasks of today's biochemical research. Nevertheless, as some studies indicate, geometric rearrangements of some amino acid side chains result in different similarity scores. As demonstrated

with the kinase data set, TrixP already is able to take into account a certain amount of protein flexibility by its representation of rotatable hydrophilic interactions as well as by using tolerance values for the matching of the lengths of triangle sides and bulk rays. However, there is still room for further improvement. Especially, large changes of the structure like different possible folds could not be handled by the recent version of TrixP. Another improvement of TrixP might be to also value the shape similarity of two binding sites during the scoring procedure.

ASSOCIATED CONTENT

Supporting Information

Data set used for parameter training (A) and the ROC curve for the similar and dissimilar pair data set (B). This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: rarey@zbh.uni-hamburg.de.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The first two authors, Mathias von Behren and Andrea Volkamer, contributed equally to this work. We thank Christin Schärfer who originally developed the method for partial matching of bulk rays during her diploma thesis in 2008. Furthermore, we thank Didier Rognan for providing us with the FuzCav pair data and Daniel Kuhn for help with the Cavbase kinase cluster experiment. Components of the presented work emerged from the COMPASITES project from the Biokatalyse2021 cluster and were funded by the BMBF under grant 0315292A.

REFERENCES

- (1) Sleator, R.; Walsh, P. An overview of in silico protein function prediction. *Arch. Microbiol.* **2010**, *192*, 151–155.
- (2) Stauch, B.; Hofmann, H.; Perkovic, M.; Weisel, M.; Kopietz, F.; Cichutek, K.; Munk, C.; Schneider, G. Model structure of APOBEC3C reveals a binding pocket modulating ribonucleic acid interaction

required for encapsidation. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 12079–12084.

(3) Xie, L.; Wang, J.; Bourne, P. In silico elucidation of the molecular mechanism defining the adverse effect of selective estrogen receptor modulators. *PLoS Comput. Biol.* **2007**, *3*, e217.

(4) Xie, L.; Bourne, P. Structure-based systems biology for analyzing off-target binding. *Curr. Opin. Struct. Biol.* **2011**, *21*, 189–199.

(5) Kinnings, S.; Liu, N.; Buchmeier, N.; Tonge, P.; Xie, L.; Bourne, P. Drug discovery using chemical systems biology: repositioning the safe medicine Comtan to treat multi-drug and extensively drug resistant tuberculosis. *PLoS Comput. Biol.* **2009**, *5*, e1000423.

(6) Nisius, B.; Sha, F.; Gohlke, H. Structure-based computational analysis of protein binding sites for function and druggability prediction. *J. Biotechnol.* **2012**, *159*, 123–134.

(7) Kellenberger, E.; Schalon, C.; Rognan, D. How to measure the similarity between protein ligand-binding sites? *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 209–220.

(8) Altschul, S.; Madden, T.; Schaffer, A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402.

(9) Henikoff, J.; Greene, E.; Pietrokovski, S.; Henikoff, S. Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res.* **2000**, *28*, 228–230.

(10) Attwood, T. The PRINTS database: a resource for identification of protein families. *Briefings Bioinf.* **2002**, *3*, 252–263.

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

(12) Illergard, K.; Ardell, D.; Elofsson, A. Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins* **2009**, *77*, 499–508.

(13) Ye, Y.; Godzik, A. Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics* **2003**, *19* (Suppl 2), ii246–255.

(14) Taubig, H.; Buchner, A.; Griebisch, J. PAST: fast structure-based searching in the PDB. *Nucleic Acids Res.* **2006**, *34*, W20–3.

(15) Gibrat, J.; Madej, T.; Bryant, S. Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* **1996**, *6*, 377–385.

(16) Wang, S.; Peng, J.; Xu, J. Alignment of distantly related protein structures: algorithm, bound and implications to homology modeling. *Bioinformatics* **2011**, *27*, 2537–2545.

(17) Wallace, A.; Laskowski, R.; Thornton, J. Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Sci.: Publication Protein Soc.* **1996**, *5*, 1001–1013.

(18) Rost, B. Enzyme function less conserved than anticipated. *J. Mol. Biol.* **2002**, *318*, 595–608.

(19) Morris, R.; Najmanovich, R.; Kahraman, A.; Thornton, J. Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. *Bioinformatics* **2005**, *21*, 2347–2355.

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

(21) Jambon, M.; Imbert, A.; Deleage, G.; Geourjon, C. A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins* **2003**, *52*, 137–145.

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

(23) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. Recognition of functional sites in protein structures. *J. Mol. Biol.* **2004**, *339*, 607–633.

(24) Milik, M.; Szalma, S.; Olszewski, K. Common Structural Cliques: a tool for protein structure and function analysis. *Protein Eng.* **2003**, *16*, 543–552.

(25) Schmitt, S.; Kuhn, D.; Klebe, G. A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.* **2002**, *323*, 387–406.

(26) Kuhn, D.; Weskamp, N.; Schmitt, S.; Huellermeier, E.; Klebe, G. From the similarity analysis of protein cavities to the functional classification of protein families using cavbase. *J. Mol. Biol.* **2006**, *359*, 1023–1044.

(27) Kinoshita, K.; Furui, J.; Nakamura, H. Identification of protein functions from a molecular surface database, eF-site. *J. Struct. Funct. Genomics* **2002**, *2*, 9–22.

(28) Kinoshita, K.; Murakami, Y.; Nakamura, H. eF-seek: prediction of the functional sites of proteins by searching for similar electrostatic potential and molecular surface shape. *Nucleic Acids Res.* **2007**, *35*, W398–402.

(29) Najmanovich, R.; Kurbatova, N.; Thornton, J. Detection of 3D atomic similarities and their use in the discrimination of small molecule protein-binding sites. *Bioinformatics* **2008**, *24*, i105–11.

(30) Konc, J.; Janezic, D. ProBiS-2012: web server and web services for detection of structurally similar binding sites in proteins. *Nucleic Acids Res.* **2012**, *40*, W214–221.

(31) Yeturu, K.; Chandra, N. PocketMatch: a new algorithm to compare binding sites in protein structures. *BMC Bioinf.* **2008**, *9*, S43.

(32) Binkowski, T.; Joachimiak, A. Protein functional surfaces: global shape matching and local spatial alignments of ligand binding sites. *BMC Struct. Biol.* **2008**, *8*, 45.

(33) Yin, S.; Proctor, E.; Lugovskoy, A.; Dokholyan, N. Fast screening of protein surfaces using geometric invariant fingerprints. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 16622–16626.

(34) Xiong, B.; Wu, J.; Burk, D.; Xue, M.; Jiang, H.; Shen, J. BSSF: a fingerprint based ultrafast binding site similarity search and function analysis server. *BMC Bioinf.* **2010**, *11*, 47.

(35) Das, S.; Kokardekar, A.; Breneman, C. Rapid comparison of protein binding site surfaces with property encoded shape distributions. *J. Chem. Inf. Model.* **2009**, *49*, 2863–2872.

(36) Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for Ligands and Proteins (FLAP): theory and application. *J. Chem. Inf. Model.* **2007**, *47*, 279–294.

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

(38) Weill, N.; Rognan, D. Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites. *J. Chem. Inf. Model.* **2010**, *50*, 123–135.

(39) Sael, L.; Chitale, M.; Kihara, D. Structure- and sequence-based function prediction for non-homologous proteins. *J. Struct. Funct. Genomics* **2012**, (epub).

(40) Merelli, I.; Cozzi, P.; D'Agostino, D.; Clematis, A.; Milanese, L. Image-based surface matching algorithm oriented to structural biology. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2011**, *8*, 1004–1016.

(41) Pang, B.; Zhao, N.; Korkin, D.; Shyu, C.-R. Fast protein binding site comparisons using visual words representation. *Bioinformatics* **2012**, *28*, 1345–1352.

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

(43) Aung, Z.; Tong, J. BSAIalign: a rapid graph-based algorithm for detecting ligand-binding sites in protein structures. *International Conference on Genome Informatics*, Gold Coast, Australia, Dec 1-3, 2008; pp 65–76.

(44) Desaphy, J.; Azdimousa, K.; Kellenberger, E.; Rognan, D. Comparison and druggability prediction of protein-ligand binding sites from pharmacophore-annotated cavity shapes. *J. Chem. Inf. Model.* **2012**, *52*, 2287–2299.

(45) Schellhammer, I.; Rarey, M. TriX: structure-based molecule indexing for large-scale virtual screening in sublinear time. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 223–238.

(46) Wu, K. FastBit: an efficient indexing technology for accelerating data-intensive science. *J. Phys.: Conf. Ser.* **2005**, *16*, S56–S60.

- (47) Volkamer, A.; Griewel, A.; Grombacher, T.; Rarey, M. Analyzing the topology of active sites: on the prediction of pockets and subpockets. *J. Chem. Inf. Model.* **2010**, *50*, 2041–2052.
- (48) Meslamani, J.; Rognan, D.; Kellenberger, E. sc-PDB: a database for identifying variations and multiplicity of 'druggable' binding sites in proteins. *Bioinformatics* **2011**, *27*, 1324–1326.
- (49) Hubbard, T.; Murzin, A.; Brenner, S.; Chothia, C. SCOP: a structural classification of proteins database. *Nucleic Acids Res.* **1997**, *25*, 236–239.
- (50) Urbaczek, S.; Kolodzik, A.; Fischer, J.; Lippert, T.; Heuser, S.; Groth, I.; Schulz-Gasch, T.; Rarey, M. NAOMI: On the Almost Trivial Task of Reading Molecules from Different File formats. *J. Chem. Inf. Model.* **2011**, *51*, 3199–3207.
- (51) Schlosser, J.; Rarey, M. Beyond the Virtual Screening Paradigm: Structure-Based Searching for New Lead Compounds. *J. Chem. Inf. Model.* **2009**, *49*, 800–809.
- (52) Rarey, M.; Wefing, S.; Lengauer, T. Placement of medium-sized molecular fragments into active sites of proteins. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 41–54.
- (53) Kuhn, D.; Weskamp, N.; Schmitt, S.; Hullermeier, E.; Klebe, G. From the similarity analysis of protein cavities to the functional classification of protein families using cavbase. *J. Mol. Biol.* **2006**, *359*, 1023–1044.
- (54) Laskowski, R.; Watson, J.; Thornton, J. ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res.* **2005**, *33*, W89–93.
- (55) Gold, N.; Jackson, R. Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships. *J. Mol. Biol.* **2006**, *355*, 1112–1124.
- (56) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. Recognition of functional sites in protein structures. *J. Mol. Biol.* **2004**, *339*, 607–633.

Declaration

Eidesstattliche Versicherung

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

Declaration on oath

I hereby declare, on oath, that I have written the present dissertation by my own and have not used other than the acknowledged resources and aids.

Hamburg, den

Sascha Urbaczek