Structure-based synthesis planning: An integrative approach into an interactive, user-focused design cycle

Dissertation

with the aim of achieving the degree

Dr. rer. nat.

at the Faculty of Mathematics, Computer Science and Natural Sciences

submitted to the
Department of Informatics
of Universität Hamburg

submitted by
Kai Sommer

born in Bingen am Rhein

Hamburg, March 2019

Erstgutachter:	Prof. Dr. Matthias Rarey
Zweitgutachter:	Prof. Dr. Johannes Kirchmaier
Tag Der Disputation:	13. September 2019

Eidesstattliche Erklärung

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

Hamburg, den 25. März 2019

Kai Sommer

Danksagung

Ich danke Matthias Rarey für die Möglichkeit dieses interessante Thema bearbeiten zu dürfen. Ausserdem für die fachlich exzellente Betreuung, für ein stets offenes Ohr, die fachlich anspruchsvollen Diskussionen und jeden Rat, den er im Laufe der Jahre übrig hatte. Des Weiteren danke ich allen Mitarbeitern des ZBH, vor allem der Arbeitsgruppe AMD, für das angenehme Arbeitsklima. Ich danke auch meinen ehemaligen Kollegen für die entspannte Arbeitsatmosphäre, die zahlreichen fachlichen Diskussionen, aber auch die vielen lustigen Momente. Ich danke allen, die an der Entwicklung der NAOMI Bibliothek vor und während meiner Promotion beteiligt waren und ohne deren Beiträge diese Arbeit in dieser Form nicht möglich gewesen wäre.

Ich danke meinen Korrekturlesern Eva Nittinger, Emanuel Ehmki, Florian Flachsenberg und Florian Lauck. Darüber hinaus möchte ich mich bei Therese Inhester und Tobias Lang für die gemeinsamen Läufe um die Alster bedanken die oft zu neuen Ideen führten. Thomas Otto möchte ich für die stets fundierte Beantwortung jeglicher Fragen danken.

Ich möchte an dieser Stelle auch meinen Freunden und meiner Familie danken. Danke für die Unterstützung in all den Jahren. Besonders möchte ich meiner Freundin Theresa danken, die mich durch die gesamte Promotion begleitet hat. Danke für deine Unterstützung und Geduld, vor allem in der letzten Phase. Als letztes möchte ich noch meinem Sohn Karl danken, der mir durch seine unbeschwerte Art auch in schwierigen Phasen die nötige Motivation schenkte.

Abstract

In computer-aided drug design (CADD) computational programs have been used to assist researchers in finding novel drug candidates for decades. The ongoing gain in computational performance allows the development of new approaches with which the user can interact in semi-automated workflows. The design of easy to use graphical user interfaces (GUIs) combined with efficient algorithms is a field not considered thoroughly in CADD. Fully automated approaches, often resulting in so-called black boxes, are normally the first choice and mostly provide sufficient results. The obvious drawback of such methods is the lack to intervene or to support the algorithm with important knowledge to guide the optimization procedure to a desired optimum.

In this manuscript, the development of a computational tool called NAOMInext, which could improve the fragment-based drug discovery (FBDD) cycle, is introduced. In a semi-automated workflow NAOMInext facilitates synthetic accessible fragment growing within protein binding sites making use of as much user experience as possible. A novel approach for small molecule design is outlined which combines an automated optimization algorithm with user defined constraints through assistance of an interactive GUI. First, an algorithm for conformational search of small molecules within a protein binding site is described which was optimized with respect to speed and performance to facilitate interactive usage and accelerate screening campaigns. Implicit constraints allow the user to focus on crucial design aspects during drug discovery projects. The second part details an algorithm enabling synthetic accessible fragment growing in the context of FBDD. Here, synthetic reaction rules in a machine-readable format are used to conquer the chemical space and provide synthesis routes for each predicted small molecule.

The conformational space as well as the chemical space are vast and efficient heuristics are necessary to cope with this issue. The here described algorithms are specifically designed to improve the FBDD cycle, are validated on a large-scale data set, and efficiently combined into NAOMInext. Its applicability is exemplified in different case studies. In addition, NAOMInext offers medicinal chemists easy access to a powerful tool that provides new ideas for readily synthesizable molecules.

Zusammenfassung

Im Computer gestützten Wirkstoffdesign (CADD) werden seit Jahrzehnten Programme eingesetzt, um Wissenschaftler bei der Suche nach neuen Medikamenten zu unterstützen. Die kontinuierliche Steigerung der Rechenleistung ermöglicht die Entwicklung neuer computerbasierter Ansätze um den Benutzer in teilautomatisierten Arbeitsabläufen interaktiv einzubinden. Das Design von einfach zu bedienenden Benutzerschnittstellen in Kombination mit effizienten Algorithmen ist ein Bereich, der im CADD bisher nicht vollständig berücksichtigt wird. Vollautomatisierte Ansätze, oft auch "Black Box" genannt, sind in der Regel die erste Wahl und liefern meist ausreichend gute Ergebnisse. Der offensichtliche Nachteil solcher Methoden ist der Mangel an Intervention oder Unterstützung des Algorithmus mit wichtigem Wissen, um den Optimierungsprozess zu einem gewünschten Optimum zu führen.

In dieser Arbeit wird die Entwicklung eines Werkzeugs namens NAOMInext beschrieben. Dieses Werkzeug soll Wissenschaftler bei dem fragmentbasierten Wirkstoffdesign unterstützen. In einem halbautomatisierten Verfahren erleichtert NAOMInext die synthetisch zugängliche Fragment-Optimierung innerhalb von Proteinbindetaschen. Dabei wird so viel Benutzererfahrung wie möglich genutzt. In diesem neuartigen Ansatz für das Design kleiner Moleküle, wird ein automatisierter Optimierungsalgorithmus mit einer Schnittstelle für benutzerdefinierte Interaktionen versehen und in einem interaktiven Programm mit graphischer Benutzerschnittstelle bereitgestellt. Zunächst wird ein Suchalgorithmus für kleine Molekülkonformationen innerhalb einer Proteinbindetasche beschrieben, der in Bezug auf Geschwindigkeit und Leistung optimiert wurde. Dies soll die interaktive Nutzung erleichtern und Screening-Kampagnen beschleunigen. Implizite Randbedingungen ermöglichen es dem Anwender sich beim Wirkstoffdesign auf entscheidende Designaspekte zu konzentrieren. Der zweite Teil beschäftigt sich mit einem Algorithmus der synthetisch zugängliche Fragment-Optimierung ermöglicht. Hierbei werden synthetische Reaktionsregeln in maschinenlesbarer Form verwendet, um den chemischen Raum auf synthetisch zugängliche Moleküle zu Beschränken und Synthesewege für jedes vorhergesagte kleine Molekül bereitzustellen.

Sowohl der Konformationsraum als auch der chemische Raum sind riesig und effiziente Heuristiken sind notwendig, um dieses Problem zu bewältigen. Die hier beschriebenen Algorithmen wurden speziell zur Verbesserung des FBDD-

Zyklus entwickelt, in einem groß angelegten Datensatz validiert und effizient in NAOMInext kombiniert. Die Anwendbarkeit wird an verschiedenen Beispielen anschaulich belegt. Darüber hinaus bietet NAOMInext Medizinalchemikern einen einfachen Zugang zu einem leistungsstarken Werkzeug, das neue Ideen für einfach zu synthetisierende Moleküle liefert.

Contents

List of Figures

List of Tables

Abbreviations

β_2AR β_2-adrenergic receptor
2D two-dimensional
3D three-dimensional

AChE Acetylcholinesterase
ADMET absorption, distribution, metabolism, excretion, and toxicity
ALA alanine
API application programming interface
ASP aspartic acid
ATP adenosine 5′-triphosphate

BACE β-site amyloid precursor protein cleaving enzyme
BB building block
BCATm mitochondrial branched chain aminotransferase
BFS breadth-first search
BRICS breaking of retrosynthetically interesting chemical substructures

CA carbonic anhydrase
CADD computer-aided drug design
CAESA Computer Assisted Estimation of Synthetic Accessibility
CASF comparative assessment of scoring functions
CCG Chemical Computing Group
CDK2 cyclin-dependent kinase 2
CPU central processing unit
CSD Cambridge Structural Database
CV cross validation

Da Dalton
DB database
DFS depth-first search

EDIA Electron Density for Individual Atoms

FBDD fragment-based drug discovery
FBLG fragment-based lead generation
FN false negative

FPPS farnesyl pyrophosphate synthase

GLU glutamic acid
GUI graphical user interface

H2L hit to lead
HAC heavy atom count
HIV human immunodeficiency Virus 1
HSP90 heat shock protein 90
HTS high-throughput screening
HTT High-throughput technologies

LE ligand efficiency
LEU leucine

MC Monte Carlo
MCR multi component reaction
MCS maximum common subgraph
MOOP multi-objective optimization
MW molecular weight

NCE novel chemical entities
NMR nuclear magnetic resonance
NN nearest neighbor

PDB Protein Data Bank
PHE phenylalanine
PI(3)K phosphoinositide-3-OH kinase
PINGUI Python in silico de novo growing utilities
PK pharmacokinetic
PParγ peroxisome proliferator-activated receptor γ
PPI protein-protein interaction
PSA polar surface area
PTP1B protein tyrosine phosphatase 1B

QML Qt Meta-object Language

RAM random-access memory
RECAP retrosynthetic combinatorial analysis procedure
RMS root-mean-square
RMSD root-mean-square deviation
Ro3 rule of three
Ro5 rule of five

RR rigid rotor

SAR structure-activity relationship
SBDD structure-based drug design
SD structure-data
SIF simplified input file
SMARTS SMiles ARbitrary Target Specification
SMILES Simplified molecular-input line-entry system
SQL Structured Query Language

THR threonine

vdW van der Waals
VS valence state
VSEPR valence-shell electron-pair repulsion

1. Introduction

The primary focus in drug discovery programs is the discovery of new molecules for modulating biological function. In the early days of drug discovery, efforts were dependent on phenotypic approaches such as cells or whole organisms.[1] Emerging new technologies enabled new drug discovery approaches like structure-based drug design (SBDD) based on the development of X-ray crystallography.[2] SBDD approaches make use of the target (protein) structure to guide the drug design process, either experimentally or computationally based on 3D models. The ongoing technical progress and constant automatization lead to a multitude of new developments and the increased structural resolution of macromolecules.[3]–[5] In the last few decades computers gained more importance throughout all fields of current research. Recent developments in drug discovery would not have been possible without the usage of computers. Most of all, computational methods support the SBDD process.[6]–[8]

New technologies in related fields of research like biochemistry or computer science facilitate the development of new medicines and hence, enable new treatments for yet untreated diseases.[9] Most notably biochemistry had a large influence on drug discovery. The description and characterization of enzymes and receptors as drug targets during the first half of the 20th century was followed by the discovery of small inhibitors.[10]–[12] This led to a paradigm shift in the way drug discovery was used, from a phenotypic approach to a target-based approach. The more knowledge is accumulated, the more rises the understanding of how biological structure correlates with its function, thus, influencing the creation of novel chemical structures. Genome sciences, for example, make it possible to identify the genetic basis of diseases and pave the way for the further development of medical treatment options.[13]

After years of successful but protracted drug development, the pharmaceutical industry was looking for new opportunities to accelerate the drug design cycle. “Successful companies will be those that effectively integrate new technologies into their current drug discovery paradigms”.[14] The advent of automated high-throughput screening (HTS) alongside with genomic sciences, heralds a new era of drug discovery. In vitro or cell-based assays are exposed to a large number of compounds leading to an enormous amount of experimental data points. Compounds that induce a positive response within an individual assay, a so-called “hit”, should, in theory, precede to the development of more

potent leads.[13] Moreover, using this methodology should accelerate the drug design cycle. These expectations were not met as productivity of the pharmaceutical industry has not improved during the 1990s.[9], [14]–[18] One factor may be the correlation of high MW with poor solubility. Starting optimization with a high MW lead compound may result in molecules with even higher MW. Thus, leading to poor pharmacokinetic (PK) properties.[19] This does not imply that HTS is not of use in drug discovery contemplating the numerous success stories.[20] The field evolved and more specific applications, like target focused library screening, emerged. A full library screening may still be used as a last resort if focused strategies fail.[18] Although current drug discovery processes mainly rely on HTS efforts, this approach suffers from the limited coverage of drug-like chemical space.[21]

Another problem is the current trend in drug discovery projects focussing no longer on classical targets.[20], [22] Most compound libraries have been optimized for historical targets. This circumstance, however, reduces the chemical diversity in existing HTS libraries and thus, decreases the chances of finding novel leads especially for new targets, such as protein-protein interactions (PPIs).[17], [23] To solve this problem, the pharmaceutical industry implemented costly enhancement programs to fill compound libraries with high-quality structurally diverse chemotypes.[20] But all the efforts in compound library enhancement do not tackle the problem of the enormous size of the drug-like chemical space (estimated at $> 1 * 10^{30}$ compounds).[24] The need to cover more of the drug-like chemical space, especially for yet intractable or non classical targets, and enhancing the drug discovery process led to the development of new strategies in drug discovery, i.e. fragment screening with subsequent fragment-based drug discovery (FBDD).[25]

Back in 2009 Murray and Reese postulated "that screening 1,000 fragments (<16 heavy atoms per compound) might sample 'total chemical space' more effectively than screening 1,000,000 more typical, higher-MW HTS compounds (<36 heavy atoms per compound)".[9] This hypothesis was supported by earlier experiments of the industry performing fragment screens.[26] The initial hypothesis behind FBDD is that more complex molecules are less likely to bind to a given target. Hence, using smaller molecules (heavy atom count (HAC) <17) tend to increase the observed hit rate. Additionally, smaller molecules tend to form high potency interactions. This was theoretically analyzed by Hann *et al.* [27] in 2001 and experimentally supported by Teotico *et al.* in 2009.[28] In the past 20 years FBDD emerged as an alternative to classical HTS.[19], [29]–[31] The advantages of FBDD are evident in many stages of the drug design workflow. During hit discovery[27], [32], [33], lead identification[9], [27], and lead optimization[25], [34], [35] as well. After the identification of key interactions

important for binding, the optimization process can be focused on optimizing the molecule's physicochemical and pharmacological profile. Thus, leading to more focused inhibitors and a series of lead compounds based on the same core structure, i.e. a lead series.[36]

Fragment to lead optimization is still a challenging task and structural information is needed. However, it has a positive effect on the success rate.[17], [37] During hit to lead (H2L) optimization, SBDD may reduce the needed time (and cost) since a reduced number of compounds need to be synthesized.[6] Shuker *et al.* are among the pioneers in this field and were the first who reported a fragment-based approach for discovering high-affinity lead compounds using a structure-activity relationship (SAR) by nuclear magnetic resonance (NMR) approach.[25]They identified a couple of nanomolar inhibitors using a fragment linking strategy in order to combine fragments into lead compounds within only two months. However, others trying to use this approach described a significant potency loss due to suboptimal linking and thus, linker strain.[38], [39]

A more prominent method for H2L optimization is fragment growing. This method offers some advantages over fragment linking. First, the procedure is more straightforward than linking and, by design, does not lead to linker strain. Second, it can be used within *in silico* SBDD concepts, thus facilitating the lead optimization process. Of course, linking can also be performed using computational tools, but is by far more challenging and error prone due to the geometrical constraints of both fragments. Last but not least, fragment growing can be combined with synthetic reaction rules considering synthetic accessibility right from the beginning. Synthetic accessibility is a key issue in modern drug discovery projects and is known to be a bottleneck in the drug discovery process.[40] Incorporating synthetic accessibility into the design process, introduces more chemically relevant diversity into the compounds and, moreover, may accelerate the drug design cycle. Hence, a synthetic route is provided right from the beginning using well known reactions making time consuming retrosynthetic studies obsolete. Another benefit is the restricted chemical space through the usage of chemical reaction rules. Consequently, the to be sampled chemical space is limited to synthetically accessible compounds.

1.1. Project Overview and Main Contributions

In the present thesis, an integrative approach for structure-based synthetically accessible fragment growing, called NAOMInext, will be introduced. Medicinal chemists expert knowledge is important for successful fragment optimization. Thus, an efficient interactive workflow, incorporated into a graphical user in-

terface (GUI), is provided to empower users to incorporate their knowledge into the decision making process.[41] For that purpose, mainly three independent components have been developed. First, an easy to use GUI providing a high level of automation (automated structure pre-processing step) but at the same time empowering the user to guide the ligand optimization procedure. Herein, the right balance between automation and interactivity is mandatory. Second, an efficient and as accurate as possible sampling algorithm is needed to cover the conformational space.[42] Furthermore, this algorithm should be able to incorporate user constraints into its optimization process. And third, an intuitive and efficient fragment optimization procedure, i.e. fragment growing. Fragment growing is perfectly suitable for an interactive approach since after each optimization step subtle changes of the binding mode can be inspected using experimental data for verification. Additionally, synthetic accessibility is incorporated right from the beginning of the fragment growing process via the usage of machine readable synthetic reaction rules[43]. This reduces the chemical space to a desired and realistic optimum. "Ultimately, the aim is to offer support for hit and lead identification and widen the chemical horizon."[42]

The prime objective is the development of a well validated scientific workflow incorporated into an easy to use software tool to assist medicinal chemists in their day-to-day work. Therefore, a thorough evaluation is of great importance, to show the correct representation of chemical feasibility and gain trust by wet-lab experimentalists. Herein, wrong results may falsify the drug design cycle and lead to increased attrition rates during the drug discovery process. Therefore, a large-scale data set of related ligand pairs[44] is used to evaluate the conformational sampling performance of the developed algorithm. The integration of the synthetic reaction rules is validated using predefined input molecules and reaction results. About 25 % of the molecules in a database show tautomerism.[45] Hence, proper treatment of tautomerism is incorporated within the reaction process. The ability of the publicly available set of reaction rules to probe the bioactivity-relevant chemical space and generate a scaffold-diverse set of compounds has been shown by Hartenfeller *et al.* in a series of publications.[46]–[48]

1.2. Motivation

As the "low-hanging fruits" have been picked and targeted diseases become more complex, i.e. multi-target drug discovery is required[49], [50], computational methods are mandatory in supporting researchers in their day-to-day work. As soon as new technologies or experimental methods have been developed, software is designed to support researchers to obtain the best results out

of it. In recent years, many computational methods were developed, especially in the field of *de novo* drug design and alternative approaches in the context of *in silico* H2L optimization.[21], [37], [42], [51], [52] This trend seems to continue[53]–[60], which indicates that the optimization of initial screening hits remains a challenging task. [36], [37] However, most of the developed tools lack a statistically reasonable validation beyond their use exemplified on a small number of targets in prospective studies, i.e. a proof of concept.[37] For some tools, it gets even worse as they are validated retrospectively on a handful of targets.[37] Most of these tools have not been widely used and there seems to be no apparent acceptance in the community.[61], [62] It is either the lack of synthetic accessibility, which is mostly not covered by a wide range of computational methods, or due to their limited usability that hampers their usage. Another conceivable reason could be the absence of a large retrospective validation.[37], [51] An exception is the tool SPROUT[63]–[65], which is steadily extended[66], [67] (SynSPROUT[68]). Its importance is proved by numerous publications and success stories over the past twenty years.[69] Nevertheless, SPROUT is also not validated on a large scale data set[70] that would statistically proof its reliability on a wide range of different protein families.

In 2007 Vangrevelinghe and Rüdisser analyzed different computational approaches used for fragment optimization, namely *de novo* drug design, combinatorial docking, and interactive fragment optimization.[41] By summary publications from 1996 to 2006 they revealed that most of the successful fragment optimization projects were a combination of interactive optimization and structure-based design. Besides, Kumar and co-workers also recommend the combination of computational and experimental approaches.[19] FBDD is perfectly suitable for such an interactive approach. Fragment optimization is a key aspect of FBDD as experimentally discovered hits (fragments) have a low affinity and need to be further optimized into potent leads, i.e. fragment-based lead generation (FBLG).[71] Using *in silico* methods for fragment optimization enables efficient generation of potential lead series.[36], [72] The success of interactive optimization may be based on the interaction of experimental and computational methods in an iterative design cycle. Thus, the effect of each computationally predicted small modification on the compound's affinity is estimated quantitatively and unexpected binding mode changes are recognized immediately. Hence, facilitating an accelerated more rational drug design process.

Besides the inclusion of experimentalists into the computational design process, the graphical representation of results displays another important aspect. Providing scientific data in a graphical way is much more intuitive since the human brain processes images in parallel while writing is processed sequen-

tially.[73] The interactive inclusion of medicinal chemists in such a design cycle makes further demands on the software, especially considering the user interface. Scientific software does not only need to provide medicinal chemists with new ideas, the software itself needs to empower users to integrate their ideas easily into the software workflow. Thus, the software needs to be "well-thought-out, suitable for their needs, [and] able to generate useful, timely and valid results"[74]. To provide medicinal chemists with a valuable tool, NAOMInext has been developed. NAOMInext is an interactive software which provides an easy to use GUI. The software is based on the robust chemical model of the NAOMI framework, which is documented by numerous publications in the literature.[75]–[84] The newly implemented algorithms are validated on a large-scale data set to provide users with reliable results.

1.3. Outline

This thesis is structured as follows: Chapter 2 describes the background of FBDD and key methods of fragment optimization. Chapter 3 describes state-of-the-art chemistry driven computational lead optimization methods and tools. Since the methods for lead optimization are mainly based on methods from *de novo* drug design, this chapter also reviews parts of the *de novo* drug design field. Chapter 4 describes the used methods from the NAOMI framework. Hereafter, the made extensions and developed methods and algorithms in the course of this thesis are outlined. Chapter 5 describes the performed evaluation strategy and experiments. The results of this evaluation and any knowledge arising therefrom, are further analyzed and discussed in chapter 6. In chapter 7 the most important design decisions and implementations of NAOMInext are outlined. Finally, chapter 8 summarizes the described achievements and provides future prospects for NAOMInext.

2. Fragment-Based Drug Discovery

The work presented here is located in the field of fragment-based drug discovery (FBDD) or more precisely in the area of fragment-based lead generation (FBLG).The origins of the underlying technique go back to the 1980s, but experienced a rebirth in the mid 1990s.[25] At about the same time HTS reached its peak as productivity did not improve as expected while costs increased enormously.[15]–[18] Researchers started to question the success of High-throughput technologies (HTT) and searched for more rational alternatives to the often referred to as anti-intellectual and irrational HTS.[85] Considering the vast chemical space of possible drug-like molecules, which was estimated to be about 10^{63}[86], screening millions and millions of compounds will never be more than scratching the surface. A large HTS screen contains $10^5 - 10^6$ compounds, which samples only a fraction of the possible chemical space.[87] More actual estimations of the drug-like chemical space suggest it to be in the range of 10^{30} compounds.[24] Nonetheless, the number of leads obtained by HTS is still far from a real breakthrough.[85] Given the vastness of chemical space, finding good starting points is a key to the drug discovery process.

In this chapter, an introduction to FBDD is given alongside with the corresponding requirements. The basics of a fragment are given together with an introduction of the FBDD pipeline. Parts of this pipeline that are important for this thesis are further outlined.

2.1. Introduction

FBDD has emerged as a promising new approach to cover the chemical space more efficiently and suggest novel lead compounds, i.e. novel chemical entitiess (NCEs). Using small molecules, or just fragments of molecules as starting points and then optimize/elaborate the fragment into a lead like molecule with higher potency is the initial idea behind FBDD. The chance of finding a more promising starting point using fragments is based on the assumption: "the smaller the molecule, the fewer the possibilities."[88] In 2007 Fink *et al.* virtually explored the chemical universe for molecules with "up to 11 atoms of C, N, O, F".[89] The tangible number of 100 million molecules is significantly smaller than the possible drug-like chemical space of larger molecules. Even just a smaller

fraction, 13.9 million compounds, follow the rule of three (Ro3)[90], [91] for lead-likeness. Hence, screening just 10.000 fragments "captures substantially more chemical diversity space than a conventional high-throughput screen."[17] Another advantage of using fragments is the higher hit rate compared to traditional HTS. For decades the pharmaceutical industry has been seeking for new approaches in order to increase the rate of finding new starting points for lead discovery, thus, enhancing its productivity.[92] In 2001 Hann *et al.* showed theoretically, by using a simple model of ligand-receptor interactions, that the chance to observe a useful interaction drops significantly as the system complexity increases.[27] Thus, supporting the initial hypothesis of the key premise of FBDD. Luckily, several years later the Novartis group and Teotico *et al.* corroborate the hypothesis with experimental data where they showed significantly higher hit rates for fragment screens.[26], [28] One of the most important aspects is the extensibility of fragments. Due to a better binding efficiency, fragments with a molecular mass below 250 Da are easier to optimize and may improve the H2L design cycle.[41] FBDD is highly linked to SBDD but also used in solely ligand-based scenarios. However, structural information facilitates the fragment elaboration cycle by guiding the fragment optimization process to increase binding affinity as well as other properties.[93]

2.1.1. Definition of a Fragment

A fragment is simply a smaller part of a small molecule. Researchers at Astex[1] characterized fragments to be within the Ro3[90], [91] which is derived from the famous Lipinski rule of five (Ro5)[94].

1. Molecular weight $<$300 Da
2. Number of hydrogen bond donors ≤ 3
3. Number of hydrogen bond acceptors ≤ 3
4. cLogP (predicted) ≤ 3
5. Number of rotatable bonds ≤ 3
6. polar surface area (PSA) ≤ 60 Å^2

Fragments typically used in FBDD have molecular masses from 150 to 250/300 Da or a maximum heavy atom size of 18.[17], [30], [95] Depending on the context, a fragment may fulfill further restrictions considering other chemical or physicochemical properties like solubility.[96] In an attempt to increase the screening hit rate and the variety of chemotypes the Klebe group compiled a small fragment library not strictly following the use of the Ro3.[97]

[1] Astex Pharmaceuticals

During library construction they considered further chemical expansion of the library candidates using growing and merging techniques. Therefore, an increased number of functional groups, needed for chemical expansion, is required within the fragments. Strictly applying the Ro3, the authors would have missed seven out of eleven successfully crystallized fragments. It should be noted, that most of the introduced functional groups have hydrogen-bond donor and hydrogen-bond acceptor properties and perform interactions with the target. Hence, these groups are mostly not available for chemical extension using a fragment growing approach, unless the binding mode of the fragment changes during chemical elaboration.[44] In the context of fragment screening, fragments need to be more polar and soluble compared to larger "drug-like" molecules and are often thought to be better optimizable considering physicochemical properties.[72] This circumstance influences the entire drug design cycle, and thus differs substantially from the traditional drug design cycle.

2.2. The FBDD Pipeline

FBDD has different requirements compared to traditional drug discovery processes. An overview of the FBDD pipeline is shown in Figure 2.1. The library used for screening consists of small fragments which normally perform low nanomolar inhibition. Thus, specialized assays and techniques to detect such low affinities are needed (see biophysical toolkit in Figure 2.1 and Section 2.4). Usually, fragment screening is more about detection of binding instead of measuring affinity.[35] In a next step, the "fragment elaboration cycle", the fragment is elaborated into a more potent hit. This step is different to creating analogues of an HTS hit since structural modifications are mostly much more comprehensive.[35] In this long lasting way from a potential hit into a lead compound, different constraints, i.e. structure-based methods and/or synthetic accessibility consideration, may help medicinal chemists to make the right decisions to speed up the all over drug discovery process. Starting with the correct selection of fragments for the fragment library.

2.3. Fragment Libraries

Compiling reasonable fragment libraries is a crucial step for the success of a screening project. The aim of a screening library is to cover as much of the vast chemical space as possible and to provide potential starting points for FBLG.[98] Depending on the research group, a fragment library has to obey different requirements (see Section 2.1.1 for some examples). Hence, a

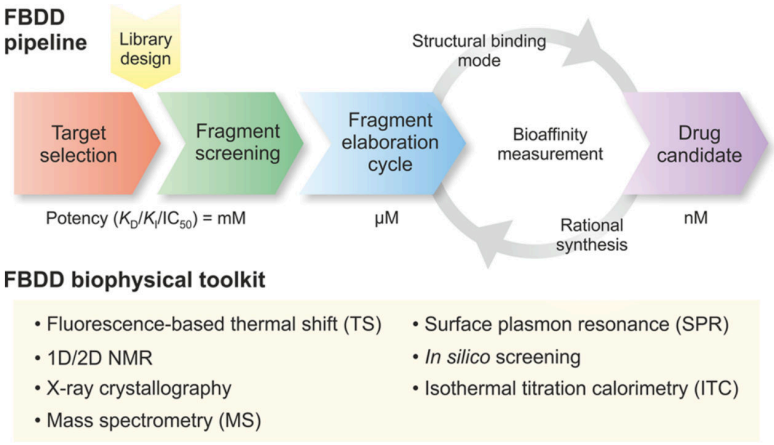


Figure (2.1) FBDD pipeline and available biophysical toolkit used for fragment screening. Reprinted with permission from Duncan E. Scott, Anthony G. Coyne, Sean A. Hudson, and Chris Abell. Fragment-based approaches in drug discovery and chemical biology. Biochemistry, 51(25): 4990–5003, 2012. Copyright 2018 American Chemical Society.

single optimal selection of fragments does not exist.[34] For example, Siegal *et al.* recommend a lower limit of 150 Da for fragments, and each fragment should possess a chemical "handle" that can be used for subsequent chemical elaboration.[99]

Depending on the used screening technique, high fragment concentrations are needed to detect the weak-affinity hits.[100] Hence, fragments need to possess high solubility that accompanies with additional functional polar groups. This may lead to fragments violating the proposed R03.[90] Schuffenhauer *et al.* described a method of a matching fragment library where each screening fragment is linked to a corresponding synthesis fragment.[26] Different other possible requirements on a fragment screening library have been reviewed by Mazanetz *et al.* [101] Furthermore, Keserű *et al.* published a collection of possible design principles for fragment library design.[93] These reviews exemplify that fragment library design is a field of ongoing research and is subject to various requirements. Specifically compiled fragment libraries are subsequently used in screening campaigns.

2.4. Fragment Screening

Experimental detection of low nanomolar inhibitors (weak binders) requires other techniques compared to traditional HTS assays which are designed to detect more or less efficient binders. As a result of the ongoing technological

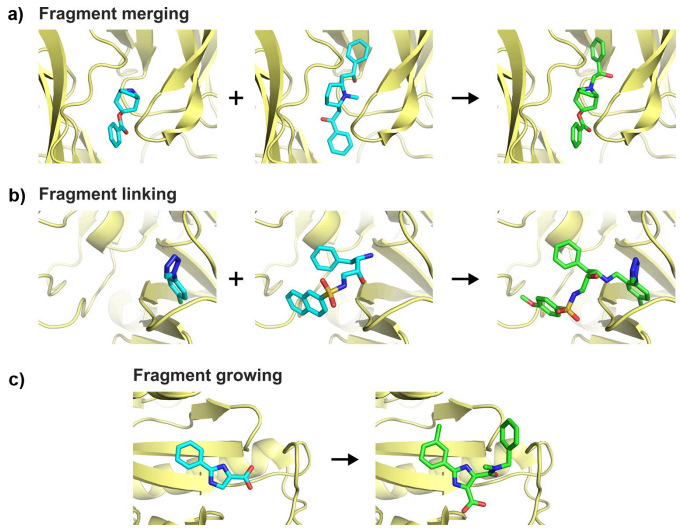


Figure (2.2) Different fragment optimization strategies. Reprinted with permission from Duncan E. Scott, Anthony G. Coyne, Sean A. Hudson, and Chris Abell. Fragment-based approaches in drug discovery and chemical biology. Biochemistry, 51(25): 4990–5003, 2012. Copyright 2018 American Chemical Society.

advancement different biophysical and biochemical techniques to detect weak binders are available today.[101] For example, high-throughput X-ray crystallography through soaking, NMR, and adapted biochemical assay methods to detect weak binding events.[102] Experimental fragment screening techniques have been extensively reviewed by Mazanetz *et al.* [101] A potential "hit" of these experimental methods is then used as input within the fragment elaboration cycle, i.e. target for fragment optimization.

2.5. Fragment Optimization

Fragment optimization is an important step within the FBDD pipeline with the aim of improving the affinity of a previously determined hit, e.g. from high throughput fragment screening or docking.[28], [103] Different optimization strategies exist, the first mentioned in 1996[25]: fragment linking, fragment growing, and fragment merging (see Figure 2.2). Additionally, a variant of fragment growing[104], i.e."SAR by Catalog", is also outlined. Either of these approaches have pros and cons and will be described below.

2.5.1. Fragment Linking

Fragment linking was the first approach for fragment optimization and was described by Shuker *et al.* [25] For an efficient joining procedure, two individually placed fragments in nearby sites are needed (see Figure 2.2b). Conceptually, this approach was already described in the early 1980s by the late William Jencks.[105] He proposed, that the Gibbs free energy (ΔG) changes of ligand-protein binding can be described as the sum of the "individual intrinsic binding energies" of the components of a molecule and a "connection Gibbs energy". Given that, linking two individually placed fragments should increase the overall binding affinity of the connected molecule. Howard *et al.* used this approach to link a small fragment, namely *p*-chlorophenlytetrazole with micro molar affinity on the trypsin-like serine protease thrombin, to another screening hit bound to an adjacent pocket (see Figure 2.2b). Usage of an amino methyl linker lead to an 50-fold improvement in affinity and a 1000-fold selectivity increase over other serine proteases like trypsin.[106] As promising as this approach may seem, success depends on choosing the right linker. Thus, if the chosen linker is suboptimal, in case of geometrical constraints, the induced linker strain leads to a significant potency loss.[38], [39] Therefore, a chemically more straightforward approach, namely fragment growing, emerged.

2.5.2. Fragment Growing

Growing describes the elaboration of an initial hit via extension at a defined vector. Fragment growing is the by far most often used H2L optimization technique[104] as judged by the number of published examples. This concept is also intuitive for most people since "growing is straightforward and closest to standard medicinal chemistry."[22] Growing has several advantages over other optimization procedures. First, since it is close to medicinal chemistry it can be combined with synthetic reaction rules to generate synthetic feasible compounds right from the beginning of the drug design cycle. Furthermore, most linking studies can also be achieved by growing the to-be-linked fragment. A successful example for fragment growing has been published by Potter *et al.* for a difficult oncology target (see Figure 2.2c).[107] The major problem was the identification of a promising hit as starting point. Using fragment screening, putative hits were identified and further elaborated to improve potency. Simple addition of different functional groups lead to a significant increase in potency.

2.5.3. Fragment Merging

The distinction between fragment growing and merging is mostly more theoretically than real, since the same molecule could be created via either of these approaches.[108] Given two fragments with given binding modes, chemists often think in enhancing the properties of a molecule by appending parts of the other molecule. This strategy may be exemplary if already known inhibitors bind to very close sub pockets and the ligands share a common part. Edink *et al.* successfully applied such a fragment merging approach during their designed fragment growing strategy.[109] Protein flexibility was a key issue in this optimization study. The aim was to induce a tyrosine-flip via fragment growing to target a sub pocket of an already known binder. To perform the growing step, the initial fragment was merged with the overlapping moiety of lobeline to address the desired sub pocket (see Figure 2.2a). In this study, as well as in many other fragment optimization studies, structural information is a key aspect for successful fragment optimization. With regard to putatively new forming interactions, the fragment elaboration may be guided to optimally match the target binding site. Hence, binding mode changes of the fragment and protein flexibility strongly influence the optimization result.

2.5.4. SAR by Catalog

Another fragment optimization method is "SAR by Catalog" and presumably one of the most common used techniques.[19] It can be simply described as a substructure or similarity search (using the initial hit) in available chemical vendor catalogs or in-house databases. Based on initial hits from a fragment screen, Jahnke and co-workers optimized an initial hit via similarity search using an in-house database and coarse pharmacophore criteria.[110] Further medicinal chemistry efforts lead to the development of a potent compound that target farnesyl pyrophosphate synthase (FPPS) with high potency. However, this approach is limited to commercially or in-house available molecules and thus, limited in the outcome of potential NCEs. Nevertheless, it allows for a quick survey of potential derivatives and has been successfully applied by others.[111] As a drawback of this method, synthetic accessibility can not be incorporated into the design cycle from the beginning. Thus, synthesizability of the predicted compounds needs to be estimated retrospectively.

2.6. Synthesizability

Synthesizability is an important issue in FBLG especially in the *de novo* design context. Since NCEs are generated *in silico*, subsequently they have to be synthesized and tested for affinity. Synthetic accessibility and feasibility are essential for efficient and successful follow up lead design. Both terms describe ease of synthesis and according to Baber and Feher[40] they can be described in the following way:

1. "Synthetic feasibility will be used to describe whether or not it is possible to synthesize a compound given a specified set of conditions."
2. "... synthetic accessibility is defined as the ease of synthesis under a specified set of conditions."

Given these definitions it is immediately clear how to distinguish both terms. Considering time as a condition for synthetic feasibility and if time would be an unlimited resource, nearly every compound may be synthetic ally feasible even for complex compounds such as natural products. Hence, including "conditions in the definition of synthetic feasibility effectively means that a synthetically feasible compound is one for which synthesis is practical rather than just theoretically possible."[40] On the contrary, synthetic accessibility is more of a measure of ease of synthesis. Baber and Feher mentioned a vivid example for the necessity of conditions for synthetic accessibility. Considering the scale-up process of syntheses in the pharmaceutical industry, it turns out that often simple syntheses in the laboratory are much more difficult or improper for large scale applications. Hence, synthetic accessibility largely depends on the underlying conditions.

3. State of the Art

This chapter describes state-of-the-art chemistry-driven computational H2L optimization methods, i.e. hit optimization methods in the field of fragment-based lead generation (FBLG), and reviews the multiple adopted approaches. The main focus is put on structure-based methods but most often the described technique can be used in both fields, ligand-based and structure-based design. The basis of fragment-based approaches is a good hit as starting point whereas *de novo* approaches design new molecules completely from scratch.[112] The aim of both methods is the generation of potential lead compounds that, in general, require further optimization. The basis (of both methods) are small BBs which are used to generate new drug-like molecules.[112], [113] Most of the optimization techniques used in fragment-based approaches, such as fragment growing or linking, are adopted from *de novo* drug design methods.[112] Due to this overlap in methodology, the terms "fragment-based" and "*de novo*" will be used interchangeably and further discussion of methodologies will also cover a part of *de novo* drug design techniques.

(Hit-to-)Lead optimization does not only imply improving ligand affinity but also improving desired PK properties. Synthesis of the predicted compounds is a prerequisite to perform further experimental testing of compound properties. In the last years, an increasing interest in synthetic accessibility of optimized compounds is discernible.[40], [46], [47], [56], [62], [114]–[119] This trend is supported by the increasing number of vendors of publicly available compound collections.[93], [120] Hence, using purchasable BBs as starting material for a fragment screening increases the likelihood that the optimized compound is synthetically accessible. Numerous *de novo* and fragment-based approaches addressing the difficult problem of H2L optimization exist and have been extensively reviewed.[21], [42], [51], [52], [121] Most of them do not incorporate synthetic accessibility in the drug design process. Due to the enormous amount of drug design methods, here, we focus on methods incorporating synthetic accessibility of the generated compounds. Nevertheless, also important and fundamental methods in the field (lacking synthetic accessibility) will not be overlooked. Ligand-based approaches will only be discussed briefly, as they are mostly based on the same methods as structure-based approaches, however, ignoring the structural context (e.g. TOPAS[122]). Fragment hit optimization strategies can be differentiated into two major categories: growing and link-

ing.[71] Linking is the term used when a potential lead structure originates from two fragment hits that are combined using a linker (see Section 2.5.1). A sequential elaboration of an initial hit to obtain a lead structure is called fragment growing (see Section 2.5.2). A third often mentioned fragment optimization method is fragment merging (see Section 2.5.3). However, merging can often be performed by fragment growing. Besides the algorithmic strategy, also the different validation approaches and the usability of the methods are discussed. Hence, most published workflows describe just a proof of concept or the usability is hampered due to missing automation or a clear user interface. In the following, fragment-based (and *de novo*) approaches and tools, are reviewed in four different sections:

1. Synthesis (chemical space coverage),
2. Search Strategy (conformational space coverage),
3. Validation Strategy, and
4. Interactive Interface

In addition to published tools and methods, there exist commercial approaches for fragment growing and linking as well as medicinal chemistry transformations. Most of the commercial software vendors do not disclose their algorithms and can not be discussed further here. One prominent commercially available software is MOE^{TM} by Chemical Computing Group (CCG)[1]. According to their website, fragment growing and linking can be performed within the protein binding site. Transformations are applied using reaction transform rules in the *.rxn file format. Custom transformations can be added to the reaction database. Since the described methods have not been published in any scientific journal they can not be discussed in the specific sections mentioned above.

3.1. Synthesis - Structure Generation

The first, and possibly most important step in *de novo* drug design is the generation of new molecules (NCEs). In this step, the only difference between *de novo* drug design and H2L optimization methods is the starting point used. In *de novo* drug design the tool itself places the initial fragment in the binding site. The fragment is then used for extension. H2L approaches use an already placed fragment (hit) instead. The fragment is then optimized into a lead like compound. For example, Chemical Genesis[123] and OpenGrowth[55] are tools offering both possibilities. The methodologies of both approaches overlap substantially and are therefore discussed together.[112]

[1] https://www.chemcomp.com

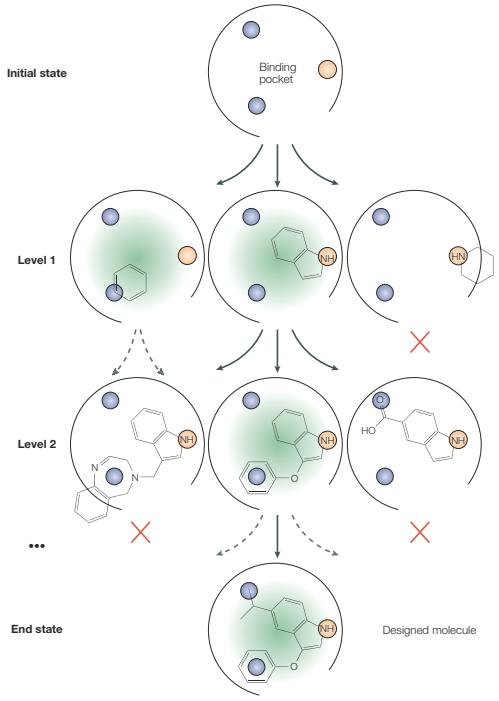


Figure (3.1) Search space exploration depicted as tree structure. Primary target constraints restrict the search space, either due to clashes or pharmacophore point (derived from the protein structure) violations. The picture shown here selects a single node for expansion (DFS), either by score or random. Reprinted with permission from Schneider, G.; Fechner, U. Computer-Based de Novo Design of Drug-like Molecules. Nat. Rev. Drug Discov. 2005, 4 (8), 649–663.

Combinatorial explosion in chemical search space is an issue which needs to be tackled by computational methods. As this problem is NP-HARD combinatorial search algorithms use HEURISTICS to solve this issue. See Table 3.1 for an example of different heuristics used to conquer the conformational space. Figure 3.1 exemplifies the search space of a *de novo* structure generation method as tree structure. The initial state (root node) is an empty binding pocket. In case of a H2L optimization method, the initial state would consist of an already placed small fragment (anchor) that performs key interactions with the binding site. Fragment evolution is exemplified as grow strategy. Each extension is depicted as node in the search tree. Green marked nodes represent valid extensions. Here, a depth-first search (DFS) is used to guide the growing process. Some tools use a combination of different search strategies to traverse the search tree much more efficiently, e.g. DREAM++[124], SPROUT[66], and FlexX[125]. Some combinations are listed in table 3.1. Here, we do not focus on *de novo* methods, but rather on H2L optimization methods which are mostly used in an interactive and iterative structure-based design cycle.[41] Hence, the different methods to

conquer the search space, either chemically or spatially, are discussed in more detail in the context of conformational space in Section 3.2. For an overview of different methods and a detailed analysis of the different search strategies to conquer the chemical space, the reader is referred to the review of Fechner and Schneider.[42] In the following section the different methods of structure generation are compared. According to Kutchukian and Shakhnovich[51] there exist six general strategies for compound generation:

1. R-group
2. Grow
3. Link
4. Evolve
5. Template
6. Synthesize

Synthetic accessibility assessment of a newly generated compound is an important issue in drug discovery projects. Some of the first drug design tools used an atom-based sequential buildup process, e.g. LEGEND[126], GenStar[127], CONCEPTS[128], and GrowMol[129]. GrowMol uses functional groups as well. Using atoms or small functional groups during structure generation ensures chemically diverse compounds but lacks synthetic accessibility. Hence, most of the designed molecules could not be synthesized which hampers subsequent experimental binding affinity measurements and PK property determination. To overcome this drawback the next generation of tools used a recombination of available fragments to generate NCEs. This procedure enhances the likelihood, that the new NCEs are synthetically accessible. Still, most of the molecules proposed by *in silico* methods are difficult to synthesize or require complex retrosynthetic analysis to provide a putative synthesis route. Hence, other methods for the design of synthetically accessible compounds are required. Difficult-to-synthesize molecules need more time and resources and may aggravate the lead optimization step.[116] Hence, it is worthwhile to incorporate synthetic accessibility of a compound as early as possible in the drug discovery process to save time, money, and resources. Different approaches exist to assess synthetic accessibility of molecules retrospectively but also early on in the development process. The terms synthetic accessibility and feasibility, have been outlined in Section 2.6.

Different methods based on retrosynthetic analysis and hard-coded reaction rules, encoded in a machine readable format, have been developed. For each of the mentioned compound generation strategies, synthetic accessibility methods can be integrated during the buildup process (implicitly or explicitly) or applied as post filter. The different compound generation methods, and possible extensions to increase the likelihood of synthesizability of the generated compounds,

are explained below. It should be mentioned that some tools use a combination of different strategies and may be part of more than one of the described categories. The R-group and grow strategy are discussed together since the methodical part overlaps significantly. Aged drug design methods are not covered in detail here and have already been reviewed in several publications.[41], [51], [113], [130]

3.1.1. R-group and Growing Approach

The perhaps simplest and most straightforward approach for compound generation is the connection of two fragments via single bond formation. This may be performed either through decoration of a core fragment (scaffold) with one or more R-groups or by replacing existing hydrogen atoms with a larger structure (BB, i.e. growing). The R-group approach can be used to enumerate large 'fragment spaces'[131] but also for lead optimization and *de novo* drug design.[53], [55], [129], [132]–[136] For lead optimization the scaffold is either created manually[53] by replacing unwanted groups and hydrogen atoms with R-groups or automatically[55], by using a randomly selected hydrogen atom as extension point. A drawback of this approach, if a fragment-based instead of an atom-based growing approach is performed, is the limitation of chemical diversity because only exo-cyclic bonds can be formed. Ring formation, introducing a lot more chemical diversity, is not possible. Thus, complex ring systems have to be provided as BBs. Compound generation is exemplified using the tool OpenGrowth[55] and works as follows: In an iterative procedure BBs are added to an initial fragment (anchor) within the active site of a target protein. The connection between the anchor and the new fragment is made by removing a hydrogen atom from each component and refilling the open valence of the connection atoms with a new single bond. The distance between both atoms is adapted according to the bond type of the connection atoms. Here, the crucial step is the selection of the proper fragment from the fragment library.

Synthetic accessibility is realized via the selection of the appropriate building blocks (fragments) similar to the retrosynthetic combinatorial analysis procedure (RECAP) approach.[137] OpenGrowth uses the FOG-algorithm[136] and selects fragments from a fragment library (using a Markov Chain approach) that have a high probability to form a bond to the anchor. To calculate the needed probabilities a library of known drugs (ChEMBL Drugstore[138]), called drug library, is used and cleaned to obtain a subset with desired properties. To compile a fragment library, the molecules of the drug library are fragmented by cutting all single bonds between rings and side chains to obtain rings and

ring systems that are added to the fragment library. Additionally, a predefined list of non-ring fragments is used.[55] The connection probability between two fragments is calculated by counting their occurrence in the drug library divided by the sum of all fragment counts in the training drug library. "Consequently, the produced molecules will statistically 'look like' molecules from the initial drug database".[55] However, the generated compounds show a similar synthetic accessibility to the ones used for training.[136]

3.1.2. Link Approach

In contrast to the above mentioned growing approach, other tools like LUDI[139], CAVEAT[140], BREED[141], CONCERTS[142], and LigBuilder[135] use fragment linking to generate new compounds. The fragment positions can be determined either experimentally[25] or computationally via docking. A subdivision of fragment linking is called 'scaffold hopping'. As this approach is a separate branch of research and reviewed elsewhere[143]–[151], it is not discussed further. Recently developed tools, specifically designed to perform linking, are not available. Nevertheless, fragment linking is still performed to generate new inhibitors.[152], [153] LigBuilder[135] uses a growing strategy to perform linking. Therefore, several preplaced seed structures (core fragments) are placed within the binding site of the target protein. The used fragments are extracted from a predefined BB library. In an incremental process a hydrogen atom is selected from the core as well as from the new fragment. The hydrogen bonds of either fragment are used to orient the new fragment to the core. Subsequently, the hydrogen atoms are removed and a new single bond is created to connect both fragments. This procedure is analog to the described growing approach in Section 3.1.1. Within LigBuilder, a link operation is performed if the growing process collides with another pre-placed core fragment in a reasonable way, i.e. when both fragments geometrically align properly and chemistry of atoms to be connected is compatible. In this case linking is performed via formation of a single bond.

A challenging approach of linking is applied in the study of De Fusco and co-workers.[153] Here, two inhibitors in different sub pockets of a common binding site should be linked to form a high-affinity inhibitor. Due to the problem of high flexibility of the needed linker (minimum length of nine atoms) an incremental growing approach was used instead.

LUDI[139], [154] uses a somewhat different approach which is similar to the strategy used by CAVEAT[140]. The used procedure is called 'fragment bridging' and uses small fragments (bridges) to connect two fragments within the binding site. Thus, the closest hydrogen atoms of both fragments are identified and

the respective bond to the adjacent heavy atom is used as alignment vector for the bridging fragment. The alignment is performed "by root-mean-square (RMS) superposition using the algorithm published by Kabsch[155]."[139] If the alignment is accepted (root-mean-square deviation (RMSD) below a given threshold) all fragments are merged into a single molecule.

Another fragment-based tool for *de novo* ligand design is GANDI[156]. Here, predocked fragments are also linked using a list of different fragments. However, the build up procedure uses a random approach, i.e. a genetic algorithm combined with a tabu search[157], [158].

Synthetic accessibility can also be incorporated during fragment linking using connection rules or synthetic reactions to join BBs. However, if other connections than simple single bond formation are used, the geometrical orientation of the to-be-linked fragment will most probably not be retained. A detailed description of this approach can be found in Section 3.1.5.

3.1.3. Evolve Approach - Evolutionary Algorithms

Another approach to generate new molecules in the context of *de novo* drug design or H2L optimization are evolutionary algorithms. Evolutionary algorithmic techniques may be divided into four major classes: genetic algorithm, genetic programming, evolutionary programming and evolutionary strategies.[159] All of theses classes share the basic concept of evolutionary algorithms described in more detail below. An evolutionary algorithm is an optimization method inspired by the basic concept of biological evolution, i.e. mutation, selection, reproduction, and recombination. To simulate the evolutionary process, at first, a population needs to be defined. The population may be any set of random candidates (individuals). In the *de novo* drug design context the individuals are simple chemical compounds. The initial selection of candidates from the population can be performed randomized or stochastic and is based on the used technique to simulate evolutionary pressure of selection.[159] Next, the selected candidates (parents) undergo the evolutionary process using "crossover" or "mutation" operators and a new variation (child) is generated. A scoring function, here often called fitness function, is used to assess the quality of the generated child over the current population. Iteratively using the described process over several generations, new descendants with different, mostly better, variations are obtained and added to the population (depending on the fitness function). This process will be continued until a user defined termination criterion is reached, e.g. number of generations. Tools based on evolutionary algorithms are, for example Chemical Genesis[123], LEA[160], LigBuilder[135],

[161], [162], SYNOPSIS[163], TOPAS[122], Molecule Evoluator[164], and AutoGrow[54], [165]. These tools mostly vary in the evolutionary technique, fitness evaluation, or selected operators.

One of the first *de novo* drug design tools incorporating evolutionary algorithms in a structure-based design concept was Chemical Genesis[123]. Chemical Genesis uses a random or available molecule as starting point and 'grows' the molecule within defined constraints using evolutionary operators. The algorithm works on substructures which are combined via single bonds. Crossover is performed by merging molecules which spatially share overlapping bonds (in terms of the position vector).[123] Defined constraints, for example scalar constraints like molecular weight (ligand based) and spatial constraints (receptor based) guide the 'growing' procedure of the new molecules.

Structural sampling may be performed within the evolutionary algorithm performing random translation and rotation mutations of the molecule (rotations also for single bonds). See Section 3.2.3 for more information. The tool GANDI[156] is based on an evolutionary algorithm using a genetic algorithm to join predocked fragments within a protein binding site. A set of 6882 fragments is available for docking using the program SEED[166], [167]. These fragments are also used as linkers. Heavy atom - hydrogen atom vectors of the fragments serve as connection points. Covalent bonds between fragments are generated using single bonds. In GANDI an individual is a single chromosome with multiple genes (docked fragments). A new child is generated using mutation and crossover operators using randomly chosen chromosome positions (genes). Hereby, crossover is performed using neighboring genes, all, or randomly selected ones. Linking fragments is performed using tabu search[157], [158]. All pairwise connections of all fragments of an individual are generated and a randomly selected pair is connected using a suitable linker. The described procedure is repeated until all fragments are connected or a termination criterion (maximum number of connections) is reached.

Synthetic accessibility is not considered in the above described workflow, since covalent bonds between fragments are formed by the replacement of hydrogen atoms. Similarly, AutoGrow[165] replaces hydrogen atoms with new fragments to generate 'mutant' ligands as well. In 2013 AutoGrow was updated[54] to incorporate synthetic accessibility into the drug design process. Thus, new fragments are added according to "click chemistry" rules (see Section 3.1.5).[114], [168] Of course other approaches like SMiles ARbitrary Target Specification (SMARTS)[169] encoded synthetic reaction rules, reaction vectors[61], or using BBs derived from RECAP rules[137] are also possible to facilitate synthetic accessibility.

3.1.4. Template Approach

The template approach is based on predefined small hydrocarbon skeletons called templates. Similar to atom-based tools like GenStar[127] and GrowMol[129], hydrocarbon structures are generated via bond formation and subsequent modification through the introduction of heteroatoms. Already in 1989 Lewis and Dean[170], [171] described a template based approach for the design of new structures within a protein binding site. Herein, spacer skeletons (planar ring systems) are used to address "the combinatorial problem of structure generation".[171] Later, a diamond skeleton is used as spacer to extend this approach into 3D space.[172] Based on these methods Todorov and Dean implemented an algorithm for *de novo* structure generation with the aim of controlling the diversity of molecular scaffold generation.[173]

SPROUT[63]–[65] uses a similar approach to the method described by Lewis and Dean[170], [171]. In the first phase hydrocarbon skeletons are generated fulfilling primary target constraints of the protein. Here, the combinatorial explosion is controlled using grouped molecular fragments which are represented by template structures and a combination of depth-first search (DFS) and breadth-first search (BFS) algorithms to efficiently sample the search space. Starting from a target site (hydrogen bond donor/acceptor group) of the protein an initial template structure is placed. Growing of the skeleton is performed via single bond formation. Cyclic templates are also able to join via fusion, bridging, and spiro joining to form more complex ring systems. Templates are further added until the remaining target sites are satisfied. In the second phase atom substitutions are performed in order to fulfill different constraints, e.g. hydrogen bond functionality, physical properties, and facilitate ease of synthesis. In most cases, the resulting molecules of such approaches are synthetically intractable. Which was, most probably, the cause why most of the programs were not often used.[61], [62]

Synthetic accessibility can not be incorporated explicitly into this approach, for example via synthetic reaction rules. Hence, Gillet and co-workers implemented Computer Assisted Estimation of Synthetic Accessibility (CAESA) to post rank structures based on synthetic accessibility.[66] Other approaches related to retrosynthetic analysis and synthesis planning are for example: WODCA[174], SYLVIA[175], [176] and the SAscore[116]. For a generated structure, CAESA uses a database of available compounds (publicly available or in-house) and automatically selects potential starting materials. Regions of the generated structure and selected starting materials are omitted during the synthetic accessibility estimate. A rule-based expert system (reaction knowledge base) "identifies and quantifies the molecular complexity that results

from the topology, the stereochemistry and the functional groups contained within the molecule."[66] Information about the starting material and the calculated complexity is then used to calculate ease of synthesis using causal networks[177]. A reasonable validation of the ranked molecules was neither performed retrospectively nor prospectively.

3.1.5. Synthesis Approach

Ease of synthesis has been an important requirement in *de novo* drug design since decades. The first tools incorporating reaction rules to enable peptide growing were GROW[132] and a special mode in LUDI[154]. In the following years others followed and incorporated more and more reaction rules into *de novo* drug design tools. For example, DREAM++[124] and SYNOPSIS[163]. In parallel with this development, another approach based on retrosynthetic analysis called RECAP[137] emerged. Based on chemical rules, molecules from a database of biologically active molecules are fragmented and suitable BBs for combinatorial library synthesis are obtained. Terminal atoms (interfaces) are annotated with the previous chemical environment in order to enable buildup of synthetically accessible compounds. Degen and co-workers improved the RECAP approach via "a new and more elaborate set of rules" called breaking of retrosynthetically interesting chemical substructures (BRICS).[178] The rule approach has been used in several tools like, Flux[179], [180], TOPAS[122], FlexNovo[181], and FSees[182]. Despite their extensive use and popularity, ease of synthesis is not ensured "since neither organic chemistry rules nor the availability of BBs were used during the process".[37]

A knowledge-based approach for the generation of synthetically feasible molecules is called 'Reaction Vectors' and described by Patel and co-workers.[61] Here, a database of available reactions in the MDL *.rxn file format is used to automatically derive 'Reaction Vectors' based on atom pair descriptors. The procedure automatically identifies atoms which need to be removed or added based on descriptor differences between the product and reactant site.

LeadOp+R[119] uses a similar strategy. First, for a given reaction the "reaction core" is identified and additional information (neighbor atoms) is gathered to extract a "reactant moiety" and "product moiety". Corresponding BBs (from a commercially available product library) are collected and stored in the LeadOp+R reaction database (DB). For a given molecule to react, the reaction DB is searched for a matching reactant and corresponding reactions. Appropriate BBs are gathered and the product is generated. The reaction is performed by joining each participating reactant and remaining parts of the molecule (excluding the reactant moiety) are reattached to the product molecule.[119]

A more elegant approach is the usage of organic chemistry rules according to "click chemistry" that allows explicit reaction handling right from the beginning.[114] *De novo* drug design tools incorporating such rules are for example, DOGS[47], AutoGrow[54], Virtual Chemist[56], DOTS[59], and Python in silico de novo growing utilities (PINGUI)[58].

Click Chemistry

"Click chemistry" rules are designed to follow nature's lead by linking fragments via single bond formation.[114] These rules describe reactions which should lead to substances that are easy to generate. A reaction should be:

- "modular, wide in scope, and have very high yields"
- "stereospecific"
- "simple reaction conditions"
- "readily available starting materials and reagents" and
- "simple product isolation".[114]

Further requirements have been defined and can be looked up in the publication of Kolb and co-workers.[114] The "click chemistry" philosophy has been widely used despite the disadvantage that there is no generic framework to use them in an *in silico* drug design approach. The framework AutoClickChem addresses this problem and facilitates *in silico* "click chemistry" reactions via a Python[183] script.[168] However, each "click chemistry" reaction needs to be implemented manually.[168]

In 2014 Massarotti used "click chemistry" to generate a database of triazoles called ZINClick to investigate the click-chemical space.[184] Just recently the ZINClick database was updated, thus expanding chemical space of 1,2,3-Triazoles. Additionally, a new description of the used implementation is provided.[185] Here, the *in silico* click reaction is implemented as SMIRKS rule [186] (SMARTS[169] notation) and performed using RDKit[187]. In conclusion, "click chemistry" and synthetic reaction rules are not mutual exclusive. In fact, the philosophy of "click chemistry" can be transcribed into a machine readable format.

Synthetic Reaction Rules

Synthetic reaction rules encoded in a machine readable format like Reaction SMARTS/SMIRKS[169], [186], Reaction-MQL[188], or CMLReact[189] provide easy access to synthetically accessible compound generation. Many recently published drug design approaches incorporate the Reaction SMARTS based organic reaction rule collection published by Hartenfeller and co-workers[43] or

use at least Reaction SMARTS/SMIRKS to perform *in silico* organic chemistry. Examples are DOGS[47], AutoGrow[54], VIRTUAL CHEMIST platform[56], PINGUI[58], DOTS[59], and ZINClick[185]. Furthermore, also new software modules to support synthetic reaction rules based on SMIRKS have been developed recently, e.g. Ambit-SMIRKS.[190] Synthetic reaction rules are mostly handcrafted rules describing real-world chemistry.[43] For more information about the functionality of synthetic reaction rules on the example of Reaction SMARTS the reader is referred to Section 4.3.2.

The first published *de novo* drug design tool incorporating several reaction expressions in a machine readable format was SYNOPSIS (70 different reaction types).[163] Molecule generation in SYNOPSIS is based on a functional group approach. For example, a NH_2 group may be oxidized to NO_2 if it is not part of a $N - NH_2$ group. Thus, an estimate of group reactivity is implemented through different rules. The reaction applied to a given molecule is selected by chance as well as the appropriate BB (if required). Thus, performing a broad sampling of chemical space. In later stages of the algorithm, if a given molecule should be optimized, a backtracking operator is used to generate analogues of a specific compound. Most of the predicted molecules could be synthesized without much effort, thus, proving the usefulness of the incorporated reaction rules.

Summary

Each of the described synthesis approaches have their Pro's and Con's. For example, synthetic reaction rules are hard to use during *rescaffolding*. Thus, an implicit consideration of synthetic feasibility using the RECAP approach is advantageous.[191] A disadvantage of this approach is the need to pre-process the generated fragment DB each time a new reaction is incorporated or new BBs are available. Furthermore, the algorithm may still generate compounds which may not be synthetically accessible, thus wasting (computer) time and resources. Using synthetic reaction rules has the advantage that each generated compound is served with a synthesis route. Although the generated compounds have a high likelihood of being synthetically accessible, it can not be assumed. Moreover, the generation of reaction rules is a time-consuming task and is mostly done by hand.[43]

3.2. Search Strategy

In SBDD different search strategies are used to conquer the conformational space of the newly designed molecules. As already mentioned earlier, struc-

ture generation (described in Section 3.1) and the used conformational search strategy are closely intertwined. For example, genetic algorithms (described in Section 3.1.3) can be extended to sample the conformational space whilst structure generation.[192]

3.2.1. Incremental Construction - Systematic Search

The initial idea of incremental growth algorithms was already described by Moon and Howe in the early 1990s and incorporated into a software package called GROW.[132] GROW uses precalculated conformations of amino acids to incrementally build up peptides from a user defined starting point. Compared to this, Leach and Kuntz use the incremental construction approach to explore the conformational space of a flexible ligand in greater detail.[193] Similarly to GROW, the incremental construction starts from a predefined anchor fragment which is placed using a variant of the DOCK algorithm.[194] Keeping the initial anchor fragment coordinates fixed, a systematic search algorithm (DFS) is used to search the conformational space. Using the concept of rigid rotor (RR) approximation, where bond lengths and angles are kept fixed[195], different dihedral angles for acyclic bonds and precalculated ring and ring system conformations, are used to cover the conformational space. Since the incremental construction approach can be described as a tree search problem[125], [132], [193], different search strategies can be applied (see Table 3.1 for a short overview).

Table (3.1) Overview of some published algorithms using incremental construction and corresponding search strategies.

Search Strategy	Algorithm or Author
Incremental construction (BFS) with greedy heuristic	GROW[132]
Backtracking (DFS)	Leach and Kuntz[193]
Incremental construction (BFS) with greedy heuristic	FlexX[125]
Hybrid algorithm combination of DFS and BFS algorithm	DREAM++ [124]
BFS in combination with DFS	TCG[196]
BFS (abort if a user-defined number is reached)	CONFECT[197]

Since the conformational space is vast and can not be enumerated exhaustively, different heuristics are used. For example, GROW and FlexX[125] (used in FlexNovo[178]) use a greedy heuristic where only the k best intermediate results are retained and used for further elaboration. Due to the incorporated heuristic, most of these algorithms can not be sorted into a certain category since mostly a combination of techniques is used. Most of the incremental construction algorithms are based on a BFS algorithm but use a specific heuristic to improve performance and trim the search tree as early as possible. For example, DREAM++ uses a combination of BFS and DFS to perform a hybrid conformational search.[124]

Incremental construction approaches benefit from the additional constraints present in SBDD. Each stage of the construction approach is stored as node in the search tree. Different strategies and/or binding site constraints, like clash or score contributions, can be used to guide the search space exploration and trim branches as early as possible. Thus, improving conformational search coverage and reducing computing time.

3.2.2. Metropolis Monte Carlo Algorithm

The Metropolis algorithm[198] is a stochastic method and used for different approaches in CADD, for example docking.[199], [200] Additionally, the GrowMol tool uses a Metropolis sampling criterion to decide if the newly grown atom will be retained.[129] On the other hand, SPROUT uses the Metropolis criterion to decide if a newly generated conformation, that is higher in energy, should be accepted.[63] Additionally, SPROUT uses the Metropolis Monte Carlo (MC) method to rapidly minimize high energy structures via simulated annealing.[201] Thus, a global energy minimum of the conformational space can be approached. Other tools using MC methods are: SMoG[202], FOG[136], and OpenGrowth[55].

Metropolis MC methods begin with an initial state of a conformation that is randomly altered. Each alteration (change of bond length or torsion angle) depends on a random number between -1 and 1.[198] Subsequently, the energy of the change is calculated. If the alteration leads to a state of lower energy ($\Delta E < 0$), the alteration is accepted. Otherwise ($\Delta E \geq 0$), the acceptance of the move depends on a probability which in turn depends on a user defined temperature value. If the probability is larger than a randomly chosen number between 0 and 1 the state is accepted else the algorithm returns to its old position. For more information on the Metropolis algorithm, the reader is referred to the work of Metropolis *et al.* [198]

3.2.3. Genetic Algorithm

Another stochastic approach is the use of a genetic algorithm to perform conformational sampling.[203] Different docking tools like GOLD[204], AutoDock[205], [206] used in AutoGrow[54], S4MPLE[192], and rDock[207] used in AutoCouple[57] use genetic algorithms to perform molecular docking, i.e. conformational sampling within a protein binding site. In a conformational sampling approach, Parent *et al.* encode a chromosome (used within the genetic algorithm) as a list of angles of the molecule.[203] Additionally, a weighting factor is assigned to each torsion angle to encode the impact of its rotation, because larger moving fragments have a higher influence. A start population is initialized with random torsion values. Descendants are generated using both, crossover and mutation operators. To perform crossover, the fittest individual selects a random partner and random chromosome locations. Furthermore, a tunable mutation rate controls the frequency of one-point mutations (random torsion angle change). To ensure optimization of the fittest members of a population, a tabu search[157], [158] is used. As usual for genetic algorithms, members of the final population are selected based on their fitness. Here, fitness is calculated as similarity to a reference individual using a geometric fingerprint-based similarity score.[208] If an individual is too similar to the reference individual, it is discarded due to its lower fitness. Convergence of the evolutionary approach is controlled by a user-defined similarity threshold. Docking tools like AutoDock[205], [206] and S4MPLE[54] extend the described approach by adding additional values for the ligand translation and orientation (rotation) within the Cartesian coordinate system.

3.2.4. Combinations in Structure-Based Designs

Different H2L optimization tools combine one of each of the described approaches to conquer the chemical space and the conformational space respectively. However, there are differences considering the type of combination. Some approaches combine individual tools in a script-based manner, e.g. using Python[183], whereas others integrate different approaches in a condensed workflow provided as user-friendly application.

User-friendly Application

An example for a tool which provides fragment optimization and *de novo* drug design is the tool SPROUT[63]. SPROUT uses a template based approach to generate new compounds from scratch (i.e. *de novo* drug design) directly within the protein binding site. The chemical search space is represented as search

graph. Each attached template is a node part of this graph. Thus, primary target constraints (clashes) can be used for early pruning of the search graph. The search for new compounds is directed by a cost score that is used to decide which node of the graph should be extended next: either using BFS or DFS. This heuristic is used to find the best solutions first. In a later version of SPROUT, synthetic accessibility of the generated compounds is incorporated using the CAESA program.[66] Hence, the generated compounds are ranked according to a synthetic accessibility estimate. The advantage of such a fused workflow is based on the increase in information in each step and the improved performance, since no conversion between different tools is necessary, which is mostly an error prone step.[75]

Workflows

Most of the implemented workflows for structure-based lead optimization are script-based and combine different tools to conquer the chemical and conformational space. Most of them use docking tools to transfer the structure generation process into 3D space. For example, the tool TOPAS uses the docking tool FlexX[125] to perform automated docking of *de novo* designed compounds.[122] The PINGUI toolbox uses two different docking tools (SEED[166], [167] and DOCK 3.6[209]) in a Python[183] workflow to perform fragment growing via merging of two individually docked fragments (BBs).[58] A robust collection of synthetic reaction rules[43] is used to conquer the chemical space and ensure synthesizability of the generated compounds to facilitate experimental validation hereafter. DOTS[59] also uses the reaction set from Hartenfeller *et al.* to generate a library of new compounds with desired physicochemical properties. Subsequent constraint docking using S4MPLE leads to a target focused library. The tool AutoCouple uses the tethered docking functionality of rDock[207] to efficiently dock the results of three implemented reaction schemes.[57] Tethering the core ensures compliance with the initial binding mode of the co-crystallized core fragment. AutoGrow implements several reactions following the rules of "Click Chemistry" and uses the docking tool AutoDock Vina to perform virtual screening.[54], [165] There exist also other "tools" which are based on different individual developments: LeadOp+R[119], DREAM++[124]

3.3. Validation Strategy

In this section, the different performed validation strategies of H2L (and *de novo*) algorithms are discussed. A fair comparison of each of these methods is not possible since most algorithms focus on different aspects of the *de novo*

drug design problem. For example, physicochemical property optimization[62], synthetic accessibility[43], bioisosteric replacement[47], energetic complementarity to a protein binding site[63], [129], a combination of several (but not all) optimization features (i.e. multiobjective optimization[62]), and different other constraints.[42] Moreover, an established validation strategy as it exists for docking[210], [211] is not available. Key aspect for a clean validation is a large-scale data set of diverse high-quality structures.[212] Several high-quality data sets in the context of docking or scoring function validation have been compiled.[212]–[215] Just recently Malhotra and Karanicolas[44], [216] published a large-scale data set of "related ligand pairs solved in complex with the same protein partner".[44], [216] In this data set, a smaller ligand (hit) and its putative successor (elaborated ligand) are analyzed for occurring binding mode changes during chemical elaboration. As traditional H2L optimization is based on the premise that the binding mode of the initial hit is preserved upon chemical elaboration, this data set is perfectly suitable as benchmark set for H2L optimization algorithms. Just last year Drwal and co-workers performed a large scale analysis of the PDB[217] to analyze fragment binding mode conservation. The data set (1832 drug-like ligands and 1079 fragments crystallized within 235 different proteins) is significantly larger than the data set published by Malhotra and Karanicolas.[218] As a drawback, the compiled data set is not available for download but can be queried via a web interface. Since the data set of Malhotra and Karanicolas is relatively new, none of the discussed H2L approaches was validated using a large-scale data set. Although some parts of a H2L approach may have been validated on a large-scale data set (e.g. the used docking tool), a thoroughly performed validation of the whole workflow in a real test scenario is missing. For example, the docking tool S4MPLE[192] from Hoffer and co-workers was validated on the Astex/CCDC 'clean' subset[213] and on the Astex Diverse Set[214]. In this validation the authors performed a re-docking experiment to assess success and compared the results to other common used docking tools like FlexX[125], GOLD[204], and Plants[219]. Moreover, S4MPLE is also able to successfully perform fragment docking[220], which supports its applicability in FBDD studies. However, a more realistic test scenario for H2L optimization approaches would be a cross-docking validation (see Section 5.2). Despite the above mentioned differences, the different performed validation approaches are summarized and discussed in the following section.

3.3.1. Proof of Concept

Most tools dealing with H2L optimization are validated using only a couple of examples (see Hoffer *et al.* [37] for a recent review of H2L tools incorporating

synthetic accessibility). As correctly stated by Hoffer *et al.*, this is more of a proof of concept.[59] In most cases the authors of a new algorithm/workflow have profound knowledge about the targets used for validation.[119], [124] Thus, the validation may be biased. Nevertheless, a prospective validation demonstrates the applicability of the algorithm to find a putative newly designed ligand (with similar properties/scaffold to a known ligand or experimentally determined binding affinity) for a specific target. A retrospective validation on the other hand, may use a known inhibitor and exerts the developed workflow to generate analogs of reference compounds or rebuild the reference compound. However, an experimental validation is not performed.[37]

Prospective Validation

The majority of the published H2L algorithms are validated prospectively on one or several (barely more than five) targets.[37] For example SYNOPSIS was used to design twenty eight new compounds where eighteen could be synthesized and tested. Ten out of eighteen tested compounds showed *in vitro* inhibitory activity at the human immunodeficiency Virus 1 (HIV) receptor. Other tools belonging to this class are for example, SPROUT[63], LeadOp+R[119], AutoCouple[57], and DOTS[59].

To validate the PINGUI toolbox Chevillard and co-workers computationally extended new ligands for the β_2-adrenergic receptor (β_2AR). Subsequently, the ligands are synthesized based on the proposed chemical reaction scheme and the binding affinity of the ligands is experimentally defined. Five fragments with experimentally determined binding affinity were initially docked into the binding site of the active (PDB id 4DLE[221]) and basal (PDB id 2RH1[222]) conformation of the receptor using DOCK 3.6.[209] Computational growing (PINGUI toolbox) was then started from each of the docked poses (hits) individually using BBs extracted from the *fragments now* data set of the ZINC database[223], [224]. To ensure synthesizability of the predictions and preservation of the reaction center (user constraint), reductive amination was used as exclusive organic reaction scheme. Visual inspection and manual selection of the best derivative products lead to eight auspicious ligands that were synthesized. 50 % of the predicted ligands showed an improved binding affinity compared to the initial core fragment. Furthermore, the binding mode of the initial core fragment was retained in all cases.

A special case is the tool SPROUT[63]. Although it was validated prospectively using four receptor-ligand complexes[70], beyond that several user publications (different protein targets) validate its usefulness for the community (list of

publications provided by Keymodule Ltd.[2]).

Retrospective Validation

A retrospective validation is performed by applying the developed workflow to a specific target and using available experimental data to evaluate success of the predictions. AutoGrow[54], [165], *E-Novo*[53], OpenGrowth[55], and LeadOp+R[119] are evaluated performing a retrospective validation. For example DREAM++[124], i.e. the program REACT++ and SEARCH++, was validated retrospectively by regenerating binding modes of HIV protease inhibitors. Therefore, the orientation of the anchor fragment was derived from the crystal structure of a reference compound (carbonyl group of inhibitor from PDB id 9HVP[225]), thus skipping the docking part (the program ORIENT++).[124] Conformational search and reactions of the attached fragments have been performed under defined user constraints (specific desired hydrogen bond interaction requested). Subsequently, the reference compounds with known activity have been docked to compare affinity and scores of these inhibitors. DREAM++ was able to find "most of the active compounds by the number of possible conformations."[124] RMSD values have not been specified.

Additionally, DREAM++ was used to find new types of inhibitors for HIV protease, thus testing the ability of the REACT++ module to design synthetically accessible ligands. Again, additional constraints have to be fulfilled during the validation.

The recently published tool LeadOp+R[119] was validated using two different targets to demonstrate its ability to optimize a compound and providing a potential synthesis route at the same time. A reference compound with high nM affinity was docked into the Tie-2 binding site (PDB id 2P4I[226]). The aminobenzoic fragment was manually selected as anchor fragment due to an important hydrogen bond. Application of the LeadOp+R[119] approach led to several newly designed compounds. Among these (rank 38 and higher out of 631), nine compounds were already published in the literature.[226] Three of these compounds have been investigated further because they have higher potency than the query compound and also a published synthesis route has been available.[226] Just like in the validation of DREAM++[124] no RMSD values of the designed compounds have been specified. The suggested reaction rules of LeadOp+R[119] matched the synthetic reaction steps in the literature. Hence, LeadOp+R[119] was able to optimize the anchor fragment providing a valid synthesis route.

[2]http://keymodule.co.uk/library/user-publications/sprout.html

3.3.2. Large-scale Analysis

As discussed in Section 3.3.1 a retrospective validation using just one or only a few examples has only little informative value about the applicability domain and performance of the tested approach. As described in the publication of DREAM++[124], the authors also set many conditions which need to be fulfilled in the validation experiment. Hence, an "unsupervised" large-scale analysis facilitates a more thorough validation. The data set of Malhotra and Karanicolas provides 297 ligand pairs incorporating 87 different proteins (i.e. unique UniProt IDs[227]). Thus, providing enough data to validate tools for a large applicability domain. Furthermore, an automated validation procedure hinders the usage of target specific constraints, thus allowing objective validation without bias.

Large-scale Validation of Docking Tools

Several *de novo* and H2L approaches use external docking tools. For example the toolbox PINGUI[58] depends on the fragment docking tool SEED[166], [167] and the molecular docking tool DOCK 3.6[209]. The DOTS workflow from Hoffer and co-workers incorporates the docking tool S4MPLE[192], [220] to perform conformational search of the grown ligands.[59] Both tools (PINGUI and DOTS) have been validated on a large-scale data set using a re-docking strategy. Although the findings of Shoichet and co-workers support the usage of docking tools for molecular docking of fragments[103], their usage in a H2L optimization scenario has not been validated using a large-scale data set.

In this work, the docking tool Glide[228]–[230] was evaluated in a re- and cross-docking study using the large-scale data set from Malhotra and Karanicolas[44], [216] (for further reading see Section 5.1.3). The analysis revealed a significant performance loss if cross-docking is used. Hence, to the best of my knowledge, a large-scale validation of docking tools (besides the work of Bursulaya *et al.* in 2003[210] and Li *et al.* 2018[215]) in a real H2L optimization scenario (i.e. cross-docking see Section 5.2 for a detailed description) has not been published.

Large-scale Validation of Fragment Optimization Tools

Common structure-based *in silico* H2L approaches use experimental data of a protein structure with a co-crystallized small fragment (hit) as starting point. This hit is then elaborated into a more lead like molecule. Either through growing or linking (see Section 2.5). Hence, the protein structure, co-crystallized with the initial hit, should be used to validate H2L optimization approaches. The data set of Malhotra and Karanicolas provides "related ligand pairs solved

in complex with the same protein partner".[44] A smaller ligand (hit) and a putative elaborated larger ligand. Their analysis revealed that in some cases chemical elaboration of the initial hit is leading to sterical clashes or new stronger interactions induce a different binding mode.[44] To the best of my knowledge no H2L optimization tool exists, whose workflow is validated on a large-scale data set. Please refer to chapter 5 for a detailed description of the validation strategy and Section 5.2 for the description of the cross-growing strategy.

Summary

The data set of Malhotra and Karanicolas is relatively new (published in 2017). Hence, older tools were not able to validate the H2L optimization approach on a previously published large-scale data set. Besides, the different validation strategies makes it difficult to compare one tool to another. Thus, a generic validation strategy (as available for docking) is of utmost importance. Chapter 5 contains a description of a validation strategy that could also be applied to other H2L optimization tools.

3.4. Interactive Interface

Nowadays computers provide great potential in all fields of current research. As an example, interactive workflow design using PipelinePilot[231] or KNIME[232], facilitates users without or little programming experience to compile workflows which automatically perform a specific task. For example, the structure-based lead optimization protocol *E-Novo* was prepared through Pipeline Pilot.[53] The *E-Novo* protocol facilitates "all-in-one" lead optimization which only requires a set of ligands, an anchor ligand with docked 3D coordinates and a target protein. In addition to academic and in-house solutions from pharmaceutical industry, there exist also commercial standalone tools from Schrödinger[3] or MOE[4] from CCG[5]. Here, the focus is placed on tools originating from academia. For example, SPROUT/SynSPROUT was developed in academia and can now be licensed from Keymodule Ltd.[6]. Commercial software will be discussed on the example of Schrödinger's Maestro Suite in the context of fragment growing, i.e. tethered cross-docking.

[3]https://www.schrodinger.com
[4]http://www.chemcomp.com/MOE-Molecular_Operating_Environment.htm
[5]https://www.chemcomp.com
[6]http://www.keymodule.co.uk

3.4.1. OpenGrowth

OpenGrowth is an open-source software implemented in C++ for the purpose of *de novo* drug design and H2L optimization. If an initial fragment is provided, growing is performed using this fragment as anchor, otherwise *de novo* ligands are generated using a user-defined BB library.

Installation

To use OpenGrowth, Open Babel[233] needs to be installed first to fulfill needed dependencies. Moreover, no pre-compiled binaries of OpenGrowth are provided (version 1.0.1), it has to be compiled by the user. Thus, only expert users are able to use the software.

User Interface

The provided GUI is used for automatic creation of files necessary to run OpenGrowth. No fragment growing is performed.[55] OpenGrowth is based on the FOG algorithm.[136] Thus, a training database is needed to calculate transition probabilities for the fragment extension process that increases the probability of synthetic accessibility. This training database can be generated using the provided GUI (see Figure 3.2). Generation of the training database is a crucial step because transition probabilities, based on the connection statistics of the fragments, are used to generate new compounds. For example, pyridine must be described three times in the final library (for each hydrogen position relative to the nitrogen in ortho, meta, and para position).[55] Hence, generation of a training database is a crucial and complicated step.

Usage of OpenGrowth is performed solely on the command line. Program control is made via a list of parameters, e.g. path to the input file, growth mode, etc., which is provided as input text file. Prior to the growing process the receptor needs to be pre-processed: completion of missing residues, water removal, and the addition of hydrogens. This can be done, for example, by the use of the Protein Preparation Wizard[234] of Maestro[235]. To define the center of the active site, the ligand needs to be extracted using Open Babel[233]. Next, a tool called "CenterOfMolecule.exe"(provided with OpenGrowth) should be called with the ligand file to define the center of the molecule. The center coordinates are used as parameter in the input file to define the active site.

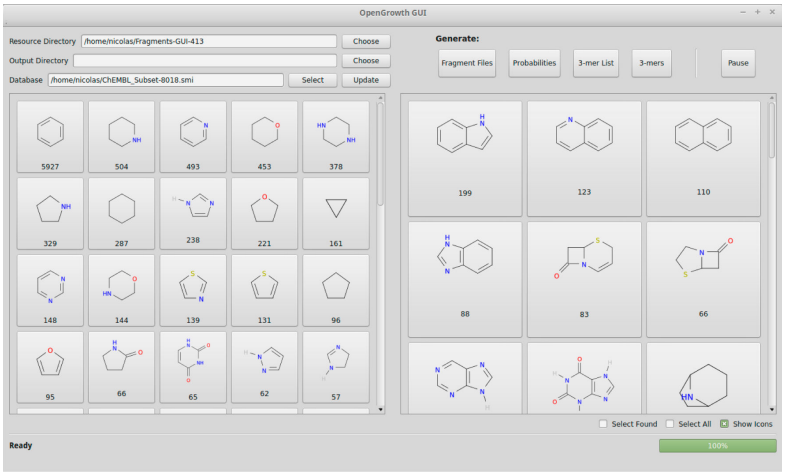


Figure (3.2) OpenGrowth graphical user interface for database generation. Non-ring fragments are provided but editable by the user. Ring fragments are extracted from a database of shredded drugs, e.g. from ChEMBL Drugstore[7][138].

3.4.2. Schrödinger Maestro Suite

Maestro[235] is a molecular modeling environment based on Python[183]. The user interface is modern and provides a wealth of functionality for a medicinal chemists day-to-day work. Furthermore, the provided Python[183] application programming interface (API) enables automatic workflows through Python scripts.

Installation

The Schrödinger Maestro Suite[235] is provided as a state of the art install package for windows and macOS and as tar archive for linux. Installation is straightforward and no further requirements are needed with the exception of a license server, which most likely needs to be installed by a computer administrator. This step is not examined further.

User Interface

As most commercial software suits, a lot of functionality is provided within one tool. Hence, it can sometimes be hard to find the desired button or workflow. Nevertheless, the user interface is well organized and a search function makes it easier to find the desired workflow (see Figure 3.3). To use Maestro in the context of fragment growing, tethered docking of a larger lead like ligand based on the input structure of a co-crystallized small fragment can be performed. A "Combine Fragments" GUI (CombiGlide) is available to elaborate pre-positioned

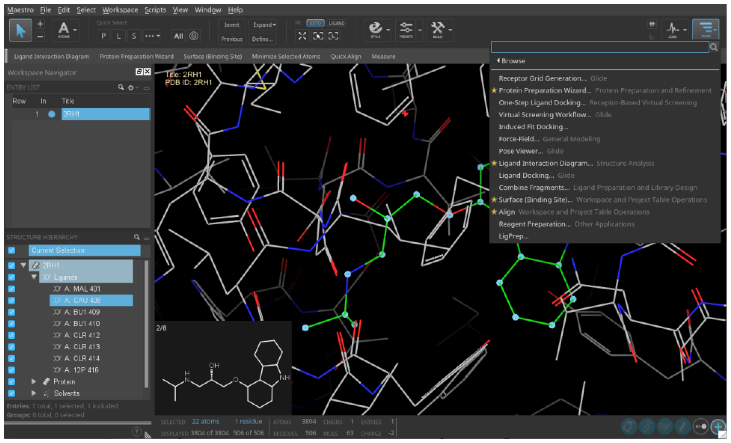


Figure (3.3) The user interface of the Maestro suite[235] is well organized and provides a search function to facilitate the usability.

fragments, either co-crystallized or derived from docking experiments. Moreover, Maestro provides several scripts that can be used to work with fragments. For example, tethered docking can be performed using the following steps:

1. Prepare complex (including the co-crystallized core fragment) using the Preparation Wizard[234]
2. Grid generation that defines a box to perform docking (extend box size to dock larger ligands)
3. Define atom mapping between ligand and core
4. Ligand docking using the defined grid and the smaller ligand as core

Using the described steps, tethered docking can be performed manually for an elaborated ligand based on a given core structure. A screening approach is hardly possible. For this purpose, the Python[183] API can be used to perform an automated core matching (e.g. MCS) and allow screening of larger libraries. Therefore, the ligand library to be screened must be compiled beforehand, if possible, including synthetic accessibility.

3.4.3. SynSPROUT

SPROUT[63] and its extension SynSPROUT[68] is a widely used *de novo* drug design tool, which is successfully used in the community in a number of publications.[69]

Installation

The installation of SynSPROUT needs to be performed on the command line. Afterwards, a shortcut, which opens a shell and starts SynSPROUT, is available on the desktop. SynSPROUT is available for Linux and macOS. No further requirements or additional software is needed.

User Interface

The user-interface of SynSPROUT is based on an ancient interface design and does not fit into the operating systems look-and-feel. The main purpose of the GUI is the connection of the different modules and pre-processing steps. Initially, a new job file needs to be generated and loaded. Then, the *CANGAROO* module is activated which indicates which step needs to be done next. After loading a complex in *.pdb file format, selecting a ligand defines the cavity which is used to constraint the position of newly generated structures. This needs to be saved as *.pdb file. Additionally, a receptor file needs to be defined which extracts the protein in a defined radius around the ligand. The result is saved as *.pdb file and is also used in the next module called *HIPPO* that becomes active if all necessary files have been defined. *HIPPO* defines important target sites (donor and/or acceptor groups) which may be used for the placement of the starting fragments. In this stage, the co-crystallized ligand is no more available in the 3D-view, which makes selecting the desired target sites tedious. Once target sites have been selected, they are used in the module *EleFANT* which docks initial fragments according to the selected target site(s). User-defined fragments can be loaded in PDB and MDL format. The docked orientations are visualized as a search graph at the bottom of the GUI (see Figure 3.4).

In Figure 3.4 the PDB id 5CVP[236] is loaded and processed as described above. The ligand, extracted with UNICON[237], is loaded in structure-data (SD) file format and docked using *EleFANT*. Unfortunately, during evaluation of the software, a crash occurred prior to docking (on macOS), which prevented to dock the standard fragment library (provided within SynSPROUT) and subsequent steps. In conclusion, SPROUT/SynSPROUT is one of the few drug design tools providing a GUI and whose methods are published. The search graph representation makes it easy to navigate through the chemical space. Interactively, individual branches can be deleted by the user to direct the search to a desired optimum.

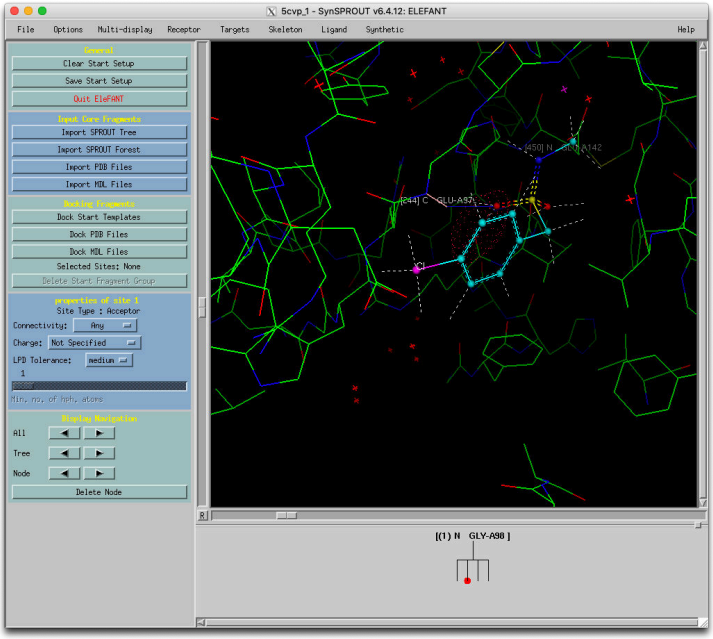


Figure (3.4) SynSPROUT user interface with docked start template (ligand from PDB id 5CVP[236] using the *EleFANT* module. The search graph is depicted at the bottom of the GUI displaying four start orientations at the first node level. Individual nodes can be deleted interactively.

4. Methods

This chapter describes both, the used and newly implement methods (and profound extensions made) during this thesis. First, previously developed concepts of the NAOMI framework used within NAOMInext, are shortly described. Second, newly developed methods besides the core methods - the constraint based sampling algorithm and the reaction workflow - are elucidated comprehensively. The sections describing existing functionality, which has been extended substantially during this thesis, contain the word 'Extended' in the subsection description to easily distinguish between existing and new/extended functionality. From section 4.2 down to the end of the chapter mainly own implementations are described except for subsection 4.2.6. In general, all used and newly implemented functionality deals with fragment growing, incremental construction within protein binding site, and synthesis reactions for *in silico* molecule design. Thus, a significant part covers the distinction between rigid and flexible molecule regions and their general representation (see Figure 4.1).

This distinction is needed in several areas such as: incremental construction, side chain torsion sampling, and *in silico* organic reaction chemistry.

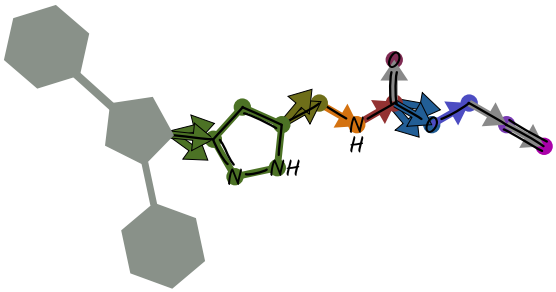


Figure (4.1) Depiction of a molecule divided into an anchor/unchanged part (schematic representation in grey) and a variable part (hand drawn labels). Multiple arrows in the same color are intended to illustrate possible variability in terms of different torsion angles, thus intimate the conformational space.

4.1. NAOMI Software Library

NAOMI is a well tested software library specifically designed to answer questions in the field of computer-aided drug design (CADD). Numerous publications validate its usefulness for accurate scientific research.[75]–[77], [79], [238], [239] Based on the published molecule model, new implementations benefit from a well validated basis. Besides own algorithmic implementations like ASCONA[240], the NAOMI library includes and makes use of several publicly available and well tested software packages and algorithms. For example, Eigen[241], nanoflann[242], InChI[243], and the Qt framework[244] to name just a few.

4.1.1. PDB Ligand Perception

The PDB[1][217] provides experimentally derived protein structures and is the largest source of publicly available structural data in this field. The file formats (*.pdb and *.mmCif[245]) provide atom entries with coordinates in the Cartesian space including element types and sometimes additional connection entries between atoms. Hydrogen atom positions are mostly not provided in X-ray crystallographic data. Besides, experimental conditions and additional meta information is given in the header section. The atom entries are grouped based on their amino acid or ligand membership. Atom type and bond order information is not included in the file format and must be derived from the given atom coordinates solely. The basic procedure used for molecules was described by Urbaczek *et al.* [77] A somewhat extended version is used for macromolecules (proteins) and was described by Bietz *et al.* [80] First, atom types and covalent bonds are assigned based on spatial criteria, i.e. distance and geometry. Second, an internal scoring scheme is used to derive the most probably correct valence state for each atom. In a final step, a heuristic is used to separate the identified macromolecules into protein chains and small isolated molecules, which do not meet the underlying requirement of being a protein chain. Next, hydrogen atom positions are determined using Protoss to optimize the all over hydrogen bond network.[80], [246]

The categorization of ligands and protein chains is based on an internal heuristic and may not coincide with the provided meta information in the PDB file. A shortcoming of this procedure is the identification of covalently-bound ligands, i.e. sugar molecules but also known inhibitors. This issue has been solved by Therese Inhester and was described in her PhD thesis.[247] Here, non-standard amino acids, connected to a protein chain, are identified and

[1] www.rcsb.org

grouped into connected components (covalent ligands) using a BFS algorithm. Covalent-bond formation is based on close atom contacts, thus, this procedure may be error prone since it can not differentiate between a real covalent-bond and shortcomings of the provided input data due to inadequate resolution of atom coordinates. Of course, errors in the provided data due to modeling failures are also possible.

4.1.2. Molecule Representation

The molecule library is the core of the NAOMI library. A molecule is represented as an undirected graph and comprises detailed bond and atom information. Consistent handling of molecules from different input file formats is ensured due to different layers of the implemented chemical model.[75], [81] Functions to check molecules for validity are provided by the NAOMI library, as well as initialization functions which calculate valence states, atom types, aromaticity, functional groups and so forth.

Chemical Model

The heart of NAOMI is an atom-centered chemical model.[75] This model ensures consistent handling of molecules from different input formats. Based on the valence state (VS) combination model, even handling of different tautomeric forms and protonation states is possible.[79] The valence state layer is of most importance considering modifications of molecules or when their validity has to be ensured. Each atom has a predefined set of valid valence states which are used to refuse incompatible valence state combinations of the molecule. Since molecule modifications, such as single bond cuts or more complex reactions, are possible, the modified molecule instance again needs to meet the requirements of the underlying model to ensure validity and consistency in subsequent steps and throughout the CADD workflow. Given that, invalid molecules are easily identified and discarded.

Atom Canonization

Atom canonization is used for several use cases, the most obvious one would be the canonical representation of a molecule like USMILES.[248] More important, the atom canonization procedure can be used to obtain an invariant representation of the input data for the used algorithms. Based on a variant of the Morgan extended sums algorithm[249], atoms are sorted in a canonical way. The implemented procedure in NAOMI differs only in the used CANON[248] algorithm, and has been published by Urbaczek *et al.* [79]

Superimposition of Atoms

A superimposition of atoms in the Cartesian space is often used in computational drug design, e.g. for evaluation purposes. The NAOMI library provides a function to perform the superimposition of molecules, as well as atom vectors, using the algorithm of Umeyama[250] provided via the Eigen library in version 3.4.[241] For symmetric molecules a simple canonical representation as described in section 4.1.2 is not sufficient. Hence, an isomorphism/automorphism analysis is performed to calculate the minimum RMSD (see formula 4.1) over all possible mappings of the molecular graph onto itself.[251] In formula 4.1, V and W are the two sets of atoms with the same length n, v_i is the currently considered atom from the set V matching to atom w_i from the set W. $\vec{v}_i$ and $\vec{w}_i$ are the position vectors of atom v_i and w_i, respectively. Based on the number of different automorphisms, the function 4.1 may be evaluated several times.

$$RMSD(V,W) = \sqrt{\frac{1}{n} * \sum_{i=1}^{n} ||\vec{v}_i - \vec{w}_i)||^2} \tag{4.1}$$

4.1.3. NAOMI Database Concept

NAOMI provides an interface to a Structured Query Language (SQL) based DB to store molecules including additional properties.[252] The DB uses an internal unique molecular string representation, the `MolString`, to identify different molecule-instances as the same molecule based on their respective topology. A description of the `MolString` can be found in the PhD thesis of Stefan Bietz.[251] Thus, different instances (conformations) of the same molecule can be stored and accessed efficiently. Furthermore, the canonical representation based on the atom canonization (see Section 4.1.2) allows to store different tautomeric instances as the same molecule.

4.1.4. Nearest Neighbor Search

The NAOMI library provides an interface to the nanoflann library.[242] Nanoflann indexes a set of points for nearest neighbor (NN)-matching via construction of a KD-tree and is optimized for two-dimensional (2D) or 3D point clouds. The interface allows for efficient parallel range queries in n-dimensional space (in this thesis solely 3D queries in the Cartesian space are used). Based on its template based design, the nanoflann library can be used with different data types such as atoms or interaction points. Range queries are performed using the euclidean (L2) metric.

4.1.5. Substructure Concept

NAOMI provides the concept of substructures, which is heavily used and extended substantially during this thesis. A substructure comprises atoms, bonds, and exo-bonds representing only a subset of a whole molecule. Exo-bonds describe bonds where only one atom is part of the substructure. Thus, describing an artificial barrier of the molecular sub-graph on the basis of a complete molecular graph. An example can be seen in Figure 4.2 where the substructure is marked in green and the exo-bond is depicted in purple.

Substructure Matching Method

The NAOMI software library supports the SMARTS[169] language to provide substructure searches and substructure mapping. A representation as line notation is able to sufficiently describe extensive molecular patterns. Recursive SMARTS enable detailed description of atom environments, for example describing specific neighboring groups. The assignment of labels facilitate the identification of specifically matched atoms to perform an atom-to-atom mapping between different substructures. The matching procedure first transforms the linear SMARTS representation into a topological graph structure. Atoms are transformed into nodes and bonds into edges. Available information like labels or bond types are mapped onto the corresponding nodes and edges, respectively. In the following thesis the terms graph, node, and edge may be used to describe SMARTS graphs and molecular graphs as well. This graph can then be used to perform a substructure search in a molecular graph using the VF2 algorithm.[238], [253] The functionality to interpret and apply SMARTS pattern to molecular graphs has already been published by others.[239], [254]

4.1.6. 3D - Coordinate Generation Procedure

Basic 3D-coordinate generation functionality for most organic compounds is also provided in NAOMI and is based on the master thesis of Therese Inhester. [255] The 3D-coordinate generation process is based on a tree structure, the so-called *Component Tree*.[197] Starting from a defined root node (part of the *Component Tree*) in the center of a molecule, new coordinates are recursively generated applying the RR approximation using fixed bond lengths and angles (valence-shell electron-pair repulsion (VSEPR)) of the NAOMI library.[255] A statistically relevant pose is achieved by the use of torsion angles derived from small-molecule crystallographic data.[78], [82] Possibly occurring clashes during the molecule buildup are solved using alternative torsion angles.

The Component Tree

To build up the `Component Tree` the molecular graph first needs to be converted into a tree structure (see Figure 4.7). Therefore, each rotatable acyclic bond represents an edge between rigid components (a so called node). A node contains a single atom or a complete ring (system). During the coordinate generation process, rings and ring systems, are treated as rigid components, and pre-calculated ring conformations (for rings with up to nine atoms) are applied.[197] Molecules with rings larger than nine atoms are discarded.

Node Torsion Data Concept

Transformations (ring conformations) and rotations (torsions), are annotated to each node of the `Component Tree`. For ease of use, the torsion angle data and ring conformation data is stored in a dynamic struct. The struct stores the derived peak angles including their score (relative frequency) and corresponding tolerance values deduced from statistically derived torsion angles.[78], [82] Additionally, a torsion flag comprising the values `PeakAnglesOnly, IncludeTolerance1, IncludeTolerance2, and AllTolerances` is stored to allow for node specific torsion angle extraction. Based on the enabled flag, all available torsion angles are generated on demand. For example, a struct containing two peak angle entries returns ten angles for the flag *AllTolerances* and two angles for the flag *PeakAnglesOnly*. Occurring duplicate angles are removed implicitly using a default threshold of 5.0°.

Incremental Construction Approach

For complete molecules, the coordinate generation starts from a root node (central node of the component tree). Default coordinates (1.0, 1.0, 1.0) in the Cartesian coordinate system are set if the root node is a single atom. If the root node contains a ring, pre-calculated coordinates, extracted from the internal ring template DB are used. Incrementally, new coordinates are assigned for each child node. Based on a recursively implemented DFS algorithm VSEPR geometries[256] and Cambridge Structural Database (CSD)[257] derived bond length for the correct geometry and distance between two atoms are used, respectively.[197], [255] If required, statistically derived torsion angles[78], [82] are assigned to each torsion bond to obtain statistically relevant 3D-coordinates. This process is performed until all heavy atoms are processed. Subsequently, valid hydrogen coordinates are generated based on each atoms geometry in 1 Å distance.

Clash Tester

During incremental construction of a molecule potential clashes are detected and solved based on the underlying *Component Tree* and torsion data. For each valid node pair, which is connected via at least three bonds, an internal clash value is calculated (see formula 4.4). An atom-atom clash (between non-hydrogen atoms) is detected if the interatomic distance is less than the sum of both van der Waals (vdW) radii multiplied by a constant softening factor k (0.7, which means 30 % vdW radii overlap allowed). The internal clash is summed up for each valid clash pair ij.[255]

4.1.7. Extended Coordinate Generator for Substructures

In the context of fragment growing, the anchor fragment already possesses valid 3D-coordinates, which also should not change dramatically during chemical elaboration. Thus, generating new ones would be a waste of time and resources and also a source of errors. Hence, a strategy to generate valid coordinates for just the attached part is implemented. The attachment type may be subdivided into the categories: 'Single bond attachment' and 'Ring fusion'. A special handling is needed if the connection point at the anchor fragment is a ring system. Even small changes in the ring system conformation can lead to significant changes of the growing vector orientation. Hence, these ring systems are incorporated into the coordinate generation and subsequent sampling process.

Covalent Bond Formation

Since the 3D coordinate generation procedure is based on a tree structure, implementing a substructure coordinate generation strategy is straightforward. Nevertheless, some hurdles have to be taken. First, the root node is modified and assigned to the target atom of the anchor fragment (unchanged part). Second, the substructure has to be extended to account for torsion bonds connected to the root node (see Figure 4.2 for an example). For the marked substructure in green, new 3D-coordinates are generated. The exo-bond (marked in purple) is used as connection bond between the part with valid 3D-coordinates (root atom in orange) and the substructure with invalid 3D-coordinates. To apply a valid torsion angle to this bond, the substructure is extended with additional atoms (marked in blue and orange). Finally, the recursive coordinate generation procedure proceeds as usual starting from the root atom (orange) in direction of the exo-bond marked in purple.

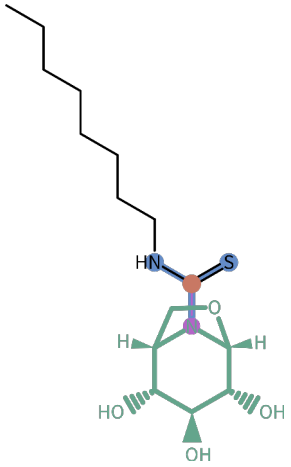


Figure (4.2) Example for a substructure extension to correctly perform root rotations. The substructure with invalid coordinates is marked in green. Needed extension atoms are marked in blue and orange (root atom). The bond and atom marked in purple connect the invalid substructure part and the root atom with valid 3D-coordinates.

Ring Fusion Attachment

NAOMInext provides the possibility to perform *in silico* chemical reactions. This also includes complex ring closure reactions. For a list of examples please refer to the Supporting Information of Hartenfeller and co-workers.[43] To generate valid 3D-coordinates for the attached reactant after a ring closure reaction that leads to a more complex ring system (e.g. *Pictet-Spengler* reaction), the already existing ring coordinates serve as template. Using a given ring as template for the 3D-coordinate generation of a ring system was not intended in NAOMI and therefore not part of the implementation. Hence, the interface of the ring conformation generator class was extended and needed modifications have been realized.

First, the available ring conformation from the template ring is transformed into an internal ring template DB structure. Second, all rings that are part of the ring system are sorted by order of highest connectivity. The provided ring is used as starting point for the 3D-coordinate generation procedure regardless of its connectivity. Finally, as usual during the ring assembly step, remaining ring templates are aligned iteratively to the initial ring template orientation, until the complete ring system is constructed. Rings, except spiro connected rings, are always connected at two atoms and share a common bond. Therefore, the second ring template 3D-coordinates are translated to the connecting atoms (common bond) of the initial ring. Based on the atoms geometry, the ring planes are aligned accordingly. The initial implementation of the `RingAssembler` class was not able to properly align the rings of a ring system using given template

coordinates. In case two identical templates should be aligned, e.g. benzene ring, an invalid rotation vector was derived from the template coordinates, which was then used during the ring alignment procedure. This issue was fixed using an orthogonal vector of the initial ring plane as rotation vector.

4.1.8. Fragment Combination Approach

Fragment combination is an important feature in this thesis. The NAOMI library just provides a function for simple single bond connections. For a given molecule with a defined target atom, another fragment can be attached via covalent single bond formation. This requires, that both atoms have at least one hydrogen or a linker atom each. Combinations are only possible without local atom geometry and bond count changes.

4.1.9. Extended Fragment Combination Approach

Organic synthesis is much more complex than only covalent single bond formation. Even if performed *in silico*, for example, to perform a grignard carbonyl reaction (see Supporting Information of Hartenfeller and co-workers [43]) *in silico*, the linear geometry of the carbon (part of the nitrile group) needs to be changed into a trigonal planar geometry. Moreover, bond types have to be adapted. Ring closure reactions are even more complex and need further treatment.

The input for a fragment combination is a number of fragments to be combined, and pre-calculated connection and modification rules. These rules may include connections between individual fragments, but also modifications of initial bond types. Connections can be created manually, as they are based on simple atom labels, but typically, they are derived from reaction rules (see Section 4.3.2). Initially, all fragment atoms and bonds are combined within one molecule object. Then, connections are processed consecutively and to be connected atoms are pre-processed to adapt charges or hydrogen count based on the new bond type. Afterwards, bonds between connection atoms are created and atom valence states are recalculated using preset bond types. In case of aromatic ring closures, additional adaptations are necessary. Atoms between two newly formed connections are gathered and checked for ring membership. If the ring count increased, bond types between those atoms is set to aromatic and valence states are adapted accordingly. Finally, the molecule is re-initialized assigning missing ring information, aromaticity, stereo descriptors, and so forth using existing NAOMI functionality.

4.1.10. Scoring Function

A ChemScore[258] based scoring function, implemented in the NAOMI library, is adapted and used in this thesis. See formula 4.2 for the original scoring function and used regression coefficients in table 4.1.

$$\begin{aligned} \Delta G_{binding} = \Delta G_0 + \Delta G_{hbond} \sum_{i,l} g_1(\Delta r) g_2(\Delta \alpha) + \Delta G_{metal} \sum_{aM} f(r_{aM}) \\ + \Delta G_{lipo} \sum_{IL} f(r_{IL}) + \Delta G_{rot} H_{rot} \end{aligned} \tag{4.2}$$

Table (4.1) Original ChemScore coefficients (in kJ/mol obtained by multiple linear regression.[258]

Coefficient	Value
Hydrogen bond coefficient	-3.34
Metal coefficient	-6.03
Lipophilic coefficient	-0.117
Rotatable bond coefficient	2.56
Intercept	-5.48

For a detailed explanation of the ChemScore function the reader is referred to the publication of Eldridge *et al.* [258].

The ChemScore empirical scoring function uses simple terms to estimate the free energy of binding. It is built from pairwise interaction contributions like a lipophilic term, metal-binding term, and a hydrogen bond term. Additionally, bond terms like a rotatable bond freezing term, and an internal strain term are used. The bond terms are only evaluated during the final scoring of compounds. All other (atom based) terms are used during the incremental construction process of the molecule. The scoring function was adapted to use NAOMI internal scoring contributions, i.e. the hydrogen bond interaction score and the general clash score. The NAOMI hydrogen bond interaction score function differs in terms of: distance, angles, and identified chem types of the interacting atoms. Moreover, additional scoring terms, needed for this thesis, have been implemented and are described below. These additional scoring terms are used to constrain the design of new compounds during the incremental construction process within the protein binding site. The individual scores are combined using the weighted-sum approach.[259] Therefore, the individual scores are

multiplied with trained regression coefficients. These coefficients were fitted based on the data set of the original ChemScore publication[258] (see table 4.2 for determined coefficients).

The Constraint Score function described in formula 4.3 is implemented to constrain the start pose sampling (see Section 4.2.2) and is also used for additional user defined constraints.

$$H_{constraintScore} = \frac{1}{2} * \sum_{i=1}^{n} k_i (x - x_0)_i^2 \tag{4.3}$$

The constraint term is based on a harmonic-oscillator model often used in force fields to model a bond in a molecule as a spring connecting two atoms. The potential energy $H_{constraintScore}$ is calculated according to formula 4.3 using the following variables: k is a proportionality constant, x is the interatomic distance between two atoms, and x_0 is the equilibrium distance (in case of a hydrogen bond constraint the optimal length of a hydrogen bond[83]). For ranking the initial start poses large deviations to the crystallized structure are penalized using the described constraint term (based on formula 4.3). Here, large deviations between the 3D-coordinates of the input pose and the sampled pose were penalized using an optimal distance of zero between corresponding atoms. The poses are then ranked according to their calculated score and increasing RMSD value compared to the initial (input) pose.

The Clash Score formula 4.4 is solely needed to evaluate the incremental construction algorithm, to check for internal clashes during the molecular build up process. The internal clash term is restricted to atoms, which are at least three covalent bonds apart. An atom-atom clash (between non-hydrogen atoms) is detected if the interatomic distance is less than the sum of both vdW radii multiplied by a constant softening factor k (0.7, which means 30 % vdW radii overlap allowed). The internal clash is summed up for each valid clash pair ij.[255]

$$H_{internalClash} = max(0, \sum_{i=1}^{m} \sum_{j=1}^{n} k * (vdW(x_i) + vdW(x_j)) - distance(x_i, x_j)) \tag{4.4}$$

The obtained regression coefficients for the implemented scoring function are as follows:

- Linear Regression, training R^2: 0.61
- Linear Regression, cross validation (CV) R^2: 0.46

Table (4.2) Trained coefficients for the ChemScore based scoring function implemented in the NAOMI software library. Training was performed on structures from the original ChemScore publication[258] using linear regression. Clash and constraint terms are not trained and a coefficient of 1.0 is used.

Coefficient	Value
Hydrogen bond coefficient	-2.89
Metal coefficient	-6.38
Lipophilic coefficient	-0.16
Rotatable bond coefficient	0.85
Intercept	-4.83

4.2. Conformational Sampling Algorithm

NAOMI provides a conformational sampling implementation, which is based on the procedure described by Schärfer *et al.* [197] but completely re-implemented using a template based software design concept and a DFS search algorithm instead of the initially described BFS algorithm. The implementation uses the RR approximation keeping bond length and atom angles fixed, solely modifying torsion angles and ring conformations from existing template libraries. The ring conformation library contains pre-calculated conformations for rings with up to nine heavy atoms. Molecules with larger rings are omitted. Using a predefined knowledge base of allowed torsion angles like the torsion library [78], [82] and pre-calculated ring conformations, systematically reduces the conformational space to cope with its complexity.[260]

The implemented recursive backtracking algorithm follows the incremental construction approach as described in section 4.1.6. Hence, for each torsion bond or flexible ring multiple potential spatial orientations are possible. Thus, the emerging sampling tree (describing the conformational sub space) leads to an ensemble of different conformations. Because of the combinatorial explosion persistent use of the DFS algorithm leads to results fast, but highly flexible molecules cannot be handled in appropriate time. Due to a maximum step limitation (recursive sampling steps), the algorithm may abort without enumerating all available conformations, thus, it may occur that the conformational space is not adequately sampled. Moreover, the development of a more sophisticated sampling algorithm is inevitable due to persistent constraints based on fragment growing within protein binding sites:

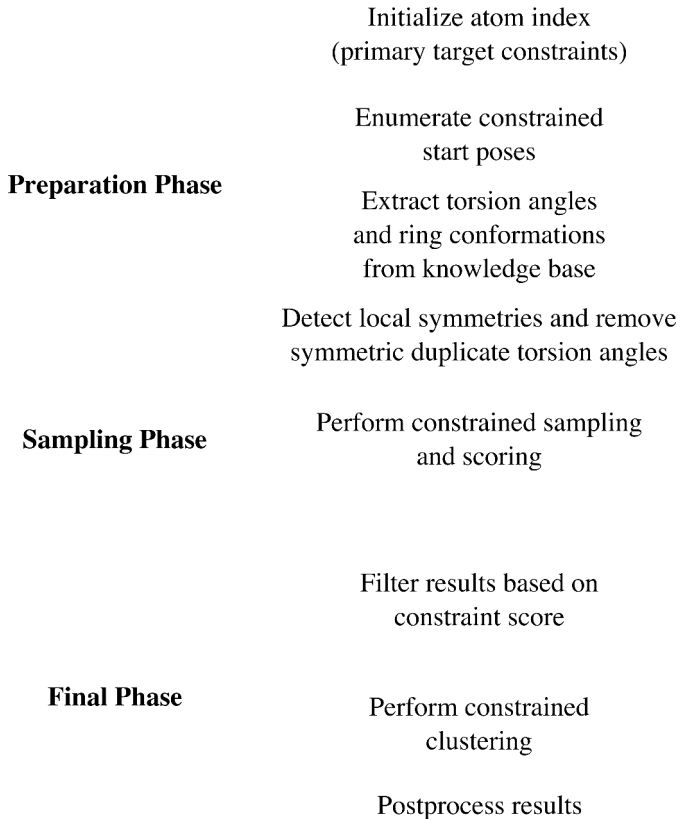


Figure (4.3) Different phases of the constrained sampling workflow. Comprehensive preprocessing is performed to reduce the needed sampling time, which is the time critical step.

- Anchor fragment constraint
- Primary target constraints (protein clashes)

4.2.1. Extended Conformational Sampling Algorithm

Within the protein binding site additional degrees of freedom, namely translational and rotational degrees of freedom, play an important role and further enhance the available conformational space. To handle difficulties with the anchor fragment constraint, together with primary target constraints, the initial pose is not treated strictly rigid and slight rotations of the anchor are allowed (see Section 4.2.2). This leads to an increase in the degrees of freedom in the sampling algorithm that enhances its complexity. Hence, this increase needs to be limited using a qualified set of symmetric free torsion angles (see Section 4.2.3) and, if required, a heuristic to further restrain the number of torsion angles to be probed (see Section 4.2.4). Figure 4.3 depicts the different phases

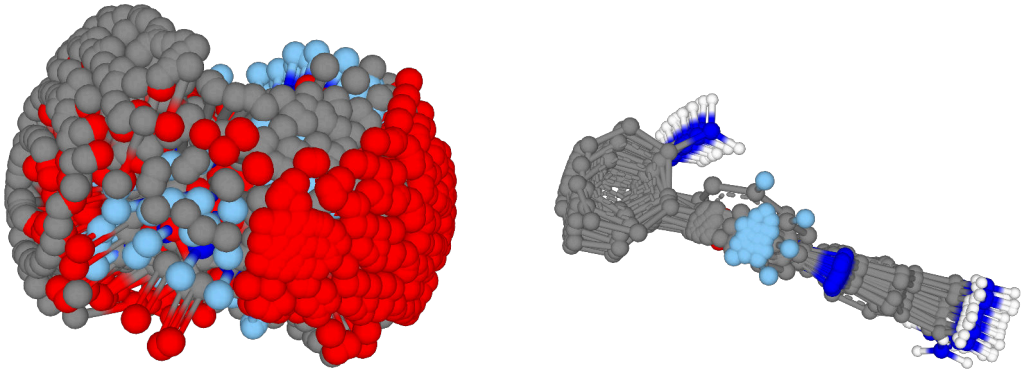
of the constrained sampling algorithm. The preparation phase is very important to reduce the needed runtime. First, removal of torsion angles leading to symmetric duplicates significantly reduces the number of to-be-generated poses. One duplicate torsion angle leads to a doubling of the possible number of poses (clashes are not considered). Second, the preparation of the atom index (primary target constraints) is used to trim branches of the search tree leading to unimportant poses for the given target.

The mentioned limitations and constraints significantly reduce the needed runtime (see Section 6.6) but do not necessarily lead to accurate results. Hence, a more sophisticated sampling strategy is needed. In consideration of the circumstances, sampling within the protein binding site, existent constraints facilitate the development of a dynamically adapting algorithm which probes the conformational space fast and exhaustively at once (see Section 4.2.5).

4.2.2. Start Pose Generation

Initial experiments revealed that treating the anchor fragment as a rigid constraint does not lead to the desired results (see Section 6.2.1). Hence, a small number of slightly modified poses is incorporated into the workflow. For fragments within the Ro3[90], [91] context - which means a molecular weight lower than 300 Da - rotations from $-180°$ to $180°$ on each axis of the coordinate system are applied using a $15°$ step size around its centroid. If the molecular weight is higher, the allowed rotations are reduced to the interval from $-5°$ to $5°$ (and a step size of $-5°$ is used) because we assume that larger rigid fragments perform more favorable interactions with the binding site and therefore will not change their binding mode significantly when a smaller fragment is attached. This assumption is supported by the findings of Malhotra and Karanicolas.[44] Their statistical analysis of related ligand pairs revealed, that compounds with fewer heavy atoms are more likely to change their binding mode. The rotations are performed using Tait-Bryan angles and lead to 27 start poses for a larger anchor fragment. The number of start poses for a smaller fragment would be significantly higher (7488) and are clustered to the best n scored poses (default: 50 and user customizable) to prevent excessive long runtimes. The anchor poses are scored using the scoring function described in section 4.1.10. Additionally, an additive harmonic potential (see formula 4.3) to calculate the RMSD to the crystal structure is used as a penalty term to privilege poses near their initial position. In Figure 4.4 examples for both cases are shown. a) shows start poses for a molecule with a MW lower than 300 Da. The blue sphere indicates a linker atom where the building block will be attached and additionally clarifies the large variation of the start poses. b) represents start poses for a ligand beyond

the Ro3. The linker atom placement indicates a much smaller variation of the start poses.



(a) Molecules within the bounds of the Ro3 (b) Molecules beyond the Ro3

Figure (4.4) Exemplary generated start poses for different types of molecules. a) molecule with MW below 300 Da and many distinct start poses. b) only slight rotations are allowed for molecules with a MW of more than 300 Da.

4.2.3. Pre-processing

The conformational sampling workflow uses pre-defined torsion angles and applies them in a defined order based on the used algorithm, for example DFS. Hence, modifications of the torsion angle data directly influences the outcome of the sampling algorithm.

Local Symmetry Detection Algorithm

A pre-processing step to remove torsion angles, which would otherwise lead to local symmetric duplicates, is performed prior to the sampling. At first, for each heavy atom of the molecule the symmetry class is determined in a canonical way using a variant of the Morgan extended sums algorithm (see Section 4.1.2).[249] Since torsion angles are applied in a DFS order based on the *Component Tree* the symmetry detection algorithm uses the same underlying data structures. Hence, starting from the root node of the *Component Tree* the symmetry classes of each child node are compared. In this step only nodes with symmetric VSEPR geometries — trigonal planar and tetrahedral — are considered. Because angles with a distance of 180° and 120° could lead to duplicate results in terms of symmetry (for identical child atoms or nodes). For example, if the angle 60° is used as torsion angle for an atom with a tetrahedral geometry the angle −60° would be omitted. In Figure 4.5 a simple example is shown where rotations

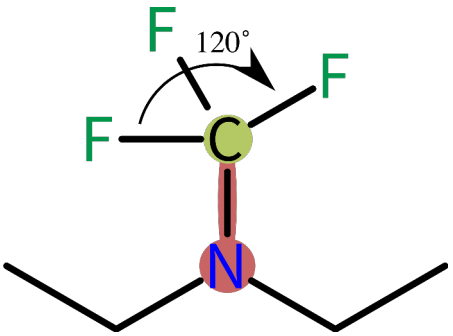


Figure (4.5) Exemplification of the symmetry detection algorithm. Considering the nitrogen atom as root and processing the C-N bond (marked in orange) in DFS order starting with the fluorinated carbon atom. Since all fluorine atoms (child nodes) are in the same symmetry class rotations of 120° and all multiples lead to symmetric duplicates. (Fluorine atoms are only used for simplification and may of course be replaced with larger groups having the same symmetry class, e.g. ethyl or phenyl group.)

of the C-N bond (marked in orange) in 120° steps would lead to symmetric duplicates using the nitrogen atom as root node.

However, using the fluorinated carbon atom as root node the symmetry detection cannot be applied because no rotation is performed for the trifluoromethyl group since torsion angles are applied in a DFS manner into direction of the child nodes. Hence, for the nitrogen node considering the ethyl moieties of the nitrogen atom as child nodes, no symmetry is detected. Since the nitrogen atom has a tetrahedral geometry the present lone pair is also considered as child node which would not match into the same symmetry class as the ethyl moieties. Thus, torsion angles which are multiples of 120° do not lead to symmetric duplicates.

Ring Symmetry Checker

A newly implemented algorithm that considers more complex symmetries of planar aromatic ring systems (see Figure 4.6) further reduces the number of duplicates during conformational sampling, thus, significantly reducing the needed runtime. Furthermore, reducing duplicate torsion angles leads to better results since more torsion angles are probed and a more diverse set of poses is obtained.

For a given ring and start atom, which is part of the ring and connected to the parent node (marked in red) of the *Component Tree*, the symmetry of the attached ring system is checked using a BFS algorithm. Initially, all ring atoms are traversed and atoms of the same level (distance to the start atom) are gathered. Then, the symmetry class for atoms of the same level are compared

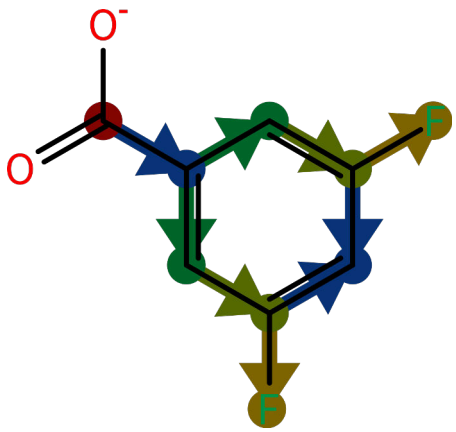


Figure (4.6) Exemplification of the ring symmetry checker algorithm. Starting from the root atom (marked in red) each level of the ring system is traversed using a BFS algorithm. Each level of the ring system is marked in equal colors as well as substituted exo-cyclic atoms of the same level.

as well as for substituted exo-cyclic atoms. If each level of the ring system is processed successfully, i.e. equality of the symmetry classes on each ring level, the ring system is marked as symmetric, asymmetric otherwise. The ring in Figure 4.6 has two *meta* substituents part of the same symmetry class. Thus, the complete ring is identified as symmetric and rotations around 180° are removed. *para* only substituted six-membered aromatic rings are also treated as planar.

4.2.4. Heuristic Approach

Finding an optimal solution in the conformational space is mostly impossible or impractical.[42] A heuristic is therefore used to speed up the search process of the solution space. The here described heuristic starts with a rough estimate of the molecules flexibility. This estimate is based on the number of possible conformations resulting from the enumeration of all pre-defined torsion angle values without considering potential clashes. Figure 4.7 shows annotated torsion data values for two rotatable bonds of a derived *Component Tree*. Using an iterative process, the possible number of conformations is altered through modifications of the pre-calculated and annotated torsion angle data for each node (see Section 4.1.6). For small inflexible molecules (possible number of conformations is lower than 500) the torsion data flag is altered to include tolerance values to enable a more subtle sampling. For highly flexible molecules the procedure is more complex. If the estimated number of conformations is larger than 10.000, we omit the first tolerance values and if still larger than allowed, we use the statistically most relevant peak angles only. In most cases this is sufficient to adequately reduce the required runtime. In cases where

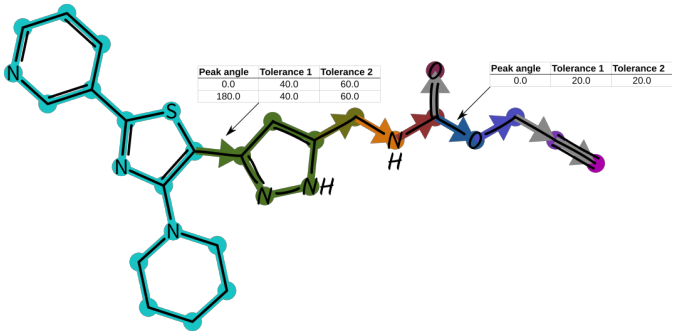


Figure (4.7) Example molecule with marked components of the *Component Tree* in different colors. The anchor fragment is marked in cyan. The processing order (DFS) of the attached fragment (depicted in hand-drawn font style) is marked with arrows starting from the root node (atom). Grey arrows mark non-rotatable bonds to terminal or linearly connected nodes where applying different torsion angles does not lead to a change in 3D-coordinates. Rotatable bonds are marked in color of the corresponding child node. Exemplary torsion angle data, extracted from the torsion library[78], [82], are shown.

this step is not sufficient and still more than 100.000 conformations could be generated (e.g. for the crystallized ligand in PDB file 1H22[261] at this stage $\sim$ 12 billion poses are possible), the peak data points are limited by score (relative frequency of occurrence). Each peak angle, whose score is lower than the half of the best scored angle, is discarded. Using this strategy, statistically least relevant torsion angles are removed and sampling molecules with many rotatable bonds in reasonable time is feasible.

The heuristic approach is not limited to adaptations of torsion data only. The start conditions, more precisely the number of available start poses, are also incorporated into the decision making step. In case the number of start poses multiplied by the possible number of conformations exceeds one million, the algorithm switches into a reduced sampling mode using at the very most the ten highest scored start poses. Since the torsion angle limitation is performed on a global level (all torsion data are treated equally) the sampling of larger flexible molecules is very coarse-grained and the conformational space may not be covered appropriately. Hence, the algorithm needs to adapt dynamically depending on the flexibility of the molecule and the condition of the binding pocket (large binding pockets lead to more start poses).

4.2.5. Dynamic Adaptation Procedure

As already mentioned above, processing the *Component Tree* and consistently sticking to the DFS order does not lead to sufficient results. Besides, the usage of primary target constraints enables the development of a more sophisticated

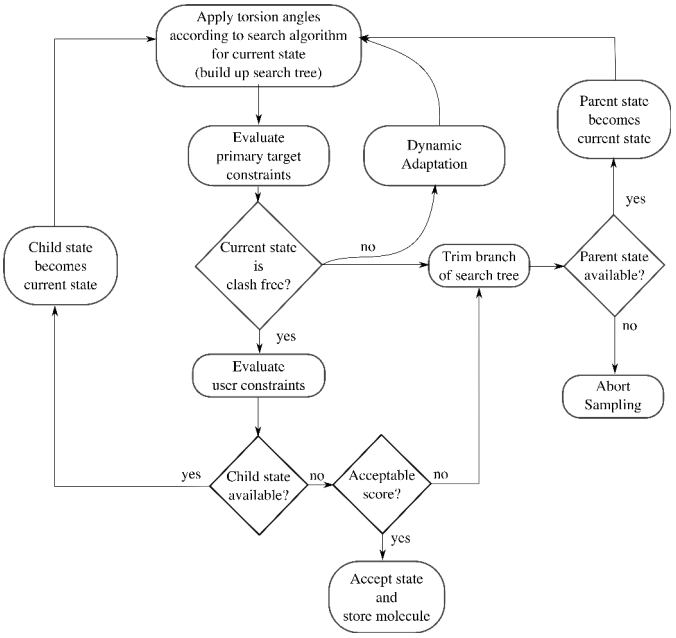


Figure (4.8) Workflow of the extended conformational sampling algorithm. The dynamic adaptation is implemented as feedback loop. The implemented heuristic is not represented in the workflow since it modifies the available torsion angles in the preparation phase (see Figure 4.3).

algorithmic procedure. Using a recursive backtracking algorithm (DFS) and starting from the root node, available torsion data values (defined during the pre-processing step see Section 4.2.4) are evaluated for subsequent child nodes following a BFS approach. For example, considering the molecule in Figure 4.7, initially the torsion peak angles 0.0° and 180.0° are evaluated for the pyrazole node (green). If both angles lead to clashes, the algorithm first probes tolerance 2 angles, which means peak angle $\pm\ 60^\circ$ leading to additional angles -60°, 60°, 120°, and -120°. Generated angles are normalized to be in the interval $(-180.0^\circ ... 180.0^\circ]$. If required, subsequently tolerance 1 angles are evaluated until at least two valid angles are available or all angles have been tested. Applicable torsion angles are then sorted by score and used in the next recursion step. If not a single angle fits, the branch of the search tree is trimmed at the current state (node of the *Component Tree*).

Figure 4.8 depicts a schematic overview of the sampling workflow. For each state available torsion angles are applied in a recursive procedure, which extends the search tree using a specified search algorithm, for example BFS. Evaluation of primary target constraints may trim the search tree for the current state. The dynamic adaptation procedure is implemented as feedback loop to influence the selection of the search algorithm (top of Figure 4.8). If the primary target

constraints (clashes) require are more subtle sampling of the current state, all available angles are probed using a BFS algorithm. Subsequently, the dynamic adaptation algorithm switches back to a DFS algorithm in order to keep the running times short.

The Early Bird Trims the Branch

Early branch trimming of the conformational search tree is important to improve the runtime of the algorithm. Therefore, clash and score values are calculated after applying torsion angles to each node. Clash values contain both, intra- and inter molecular clashes. Intra molecular clashes are calculated using the *Clash Tester* also used during 3D-coordinate generation (see Section 4.1.6). During the incremental construction an additional abort criteria, the convex hull, is used to prevent the algorithm from growing into solvent direction.

The Protein Clash Tester is a newly implemented inter molecular clash detector and implemented analogously to the already mentioned intra molecular `Clash Tester`. Based on internal data structures (the `Component Tree`) an efficient NN search is performed (see Section 4.1.4) using ligand atom coordinates as query to gather nearby protein atoms. For each atom in vicinity of the query atom the clash value is calculated according to formula 4.4. To account for valid close atom contacts due to strong polar interactions, for atoms generating valid polar interactions like donor-acceptor or metal-acceptor pairs, the allowed sum of vdW radii is reduced to 2.6 Å and 1.5 Å, respectively.[262] An interaction is valid if the calculated interaction score is above zero. [83], [251] All other contacts are treated as clashing atoms whereas a 15 % vdW overlap is allowed to mimic a soft-docking approach.[263], [264]

The Convex Hull Border is used to prevent the molecule to grow into the solvent during incremental construction (see 4.1.6). Since there might exist application scenarios where this behavior is desired, this parameter is user customizable. Besides, for shallow active sites on the protein surface (distance of the molecule's center of mass to convex hull surface is below 5 Å) the algorithm itself might deactivate convex hull usage to improve sampling results. For all other active sites the convex hull is used as sampling border to trim sampling branches as early as possible. Consequently, if a placed atom is outside the calculated convex hull, the sampling branch is trimmed immediately. The implemented convex hull is calculated via the QuickHull algorithm of Barber and co-workers based on protein atom spheres represented as icosahedron with 80 corners.[265], [266]

4.2.6. Sphere Exclusion Clustering

The sphere exclusion method is a clustering algorithm that selects compounds which most effectively cover the available property space.[267] The currently implemented version in NAOMI is optimized to reduce the needed pairwise comparisons between generated conformations.[268] Starting from a selected compound the RMSD distance to all other compounds is calculated. All compounds within a defined exclusion radius are removed. The remaining closest compound is used as a new cluster center. The described steps are repeated until no more compounds remain or a predefined maximum compound threshold is exceeded (number of cluster centers). If the threshold is exceeded, the exclusion radius is increased using a predefined step size. All clustering steps are repeated using the new threshold until no more compounds remain and the number of cluster representatives is below the given threshold. Otherwise, the exclusion radius is increased again in order to generate compound sets that do not exceed the set cluster size threshold.

Extended Constrained Sphere Exclusion Clustering is performed to optimize the clustering outcome for the specific use case of incremental construction within protein binding sites. As initial experiments have shown, that using the sphere exclusion clustering algorithm described above did not lead to the desired results. Hence, the algorithm needed to be restrained to efficiently work for this specific application. Normally, the clustering algorithm proceeds until the cluster limit is reached (maximum number of compounds), regardless of the used exclusion radius. Clustering compounds from ten thousands of conformations down to several hundreds is very common and the exclusion radius easily rises to more than 5 Å. Since the generated conformations are build up within a protein binding site and scored, the constrained version of the clustering uses the score of each conformation to ensure, that high scored conformations are privileged over low scored conformations. Thus, the restrained sphere exclusion clustering comprises an upper limit for the exclusion radius which is incorporated as follows. First, conformations are sorted according to their score. This guarantees, that the best scored conformation is always used as cluster representative. Second, the clustering proceeds as usual until the exclusion radius is exceeded. Consequentially, the cluster radius is increased stepwise until the predefined upper limit is reached. Then, the clustering is aborted and the intermediate cluster result is returned. All remaining conformations (not clustered ones) have a lower score and are putatively worse binders.

4.3. The in silico Reaction Flask

Synthetic accessibility is, and has been, an important factor in CADD.[269] In the following section, the implementation of a rule based procedure, by the example of SMIRKS, is described in detail. The described reaction algorithm is implemented in a separate library part of the NAOMI framework. The reaction procedure intensively uses the provided SMARTS library of NAOMI. SMIRKS are able to define simple transformations like nitrogen to oxygen or inverting stereo centers. These type of transformations, i.e. simple atom modifications, are not considered in this implementation. Here, only genuine organic chemistry reactions, which may be performed *in vitro*, are considered, i.e. reactions of the form $2 \rightarrow 1$ or $2 \rightarrow 2$. The following section is split in consecutive subsections alike the implementation of the algorithm is split in different independent work packages. First, the basic concept of SMIRKS is described.

4.3.1. The SMIRKS and Reaction SMARTS Concept

SMIRKS[186] is a publicly available reaction transform language used to describe chemical reactions in a computer readable format. The name is an acronym of the underlying technology Simplified molecular-input line-entry system (SMILES) and SMARTS. In the field of *de novo* drug design SMIRKS are used to manipulate existing molecules or to perform reactions to generate new molecules. Although SMIRKS is a hybrid of SMILES and SMARTS it is restricted by the supported SMARTS features. For example, the recursive feature in the described example (see 4.9) is not allowed in the SMIRKS language since the connectivity and bond order changes. The main SMIRKS rules are defined as follows:

- Pairwise atom mapping (equal number of labels on both sides of the reaction)
- Non-mapped atoms may be added or deleted (depending on the reaction side)
- No bond queries allowed

For a complete list of SMIRKS rules consult ref [186]. To overcome the limitations of SMIRKS, all available SMARTS features for reaction descriptions are supported in the here developed implementation. These extended SMIRKS are called Reaction SMARTS and have already been described and used by Hartenfeller and co-workers.[43], [169] In the following, the term SMIRKS is used to describe both variants. Nevertheless, both work in the same way and one has to discriminate between different types of atoms:

Tetrazole terminal

[CH0;$(C-[#6]):1]#[NH0:2]>>[c:1]1[n:2][NH]nn1

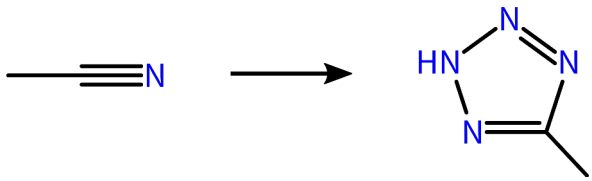


Figure (4.9) Example for illicit atomic expressions in the SMIRKS language where the connectivity of the described atom changes. However, this is a valid Reaction SMARTS.

1. Matched and labeled atoms
2. Matched but not labeled atoms (different treatment in reactant and product context) and
3. Unmatched and unlabeled atoms.

Figure 4.10 exemplifies the reaction process for a thiazole reaction from the publication of Hartenfeller and co-workers.[43] The first row shows the reactants with the different types of atoms. Atoms of the first type are marked with blue dotted circles and are part of the reaction center. For these atoms the connectivity and the bond order may change during the reaction. Atoms of the second type are marked with red dotted circles and are, since they are on the reactant site of the SMIRKS, removed during the pre-processing step. These atoms are either terminal heavy atoms or the complete branch is removed. If these type of atoms occur on the product site of the SMIRKS they are added to the product molecule and do not need to be terminal.

4.3.2. The in silico Chemical Reaction Implementation

In vitro preformed organic chemical reactions can be very complex and the *in silico* realization may be complex as well. A consistent molecular representation is important during this step. Therefore, a straightforward workflow with consecutive intermediate steps is implemented.

The SMARTS Matching Procedure

Already the first step in this workflow tends to be error prone. The SMARTS language is very flexible and different SMARTS pattern may describe one and the same substructure. Working with a discrete chemical representation of molecules may therefore lead to false negative (FN) matchings. Hence, the SMARTS matching procedure provided in NAOMI[239], [254] is applied to

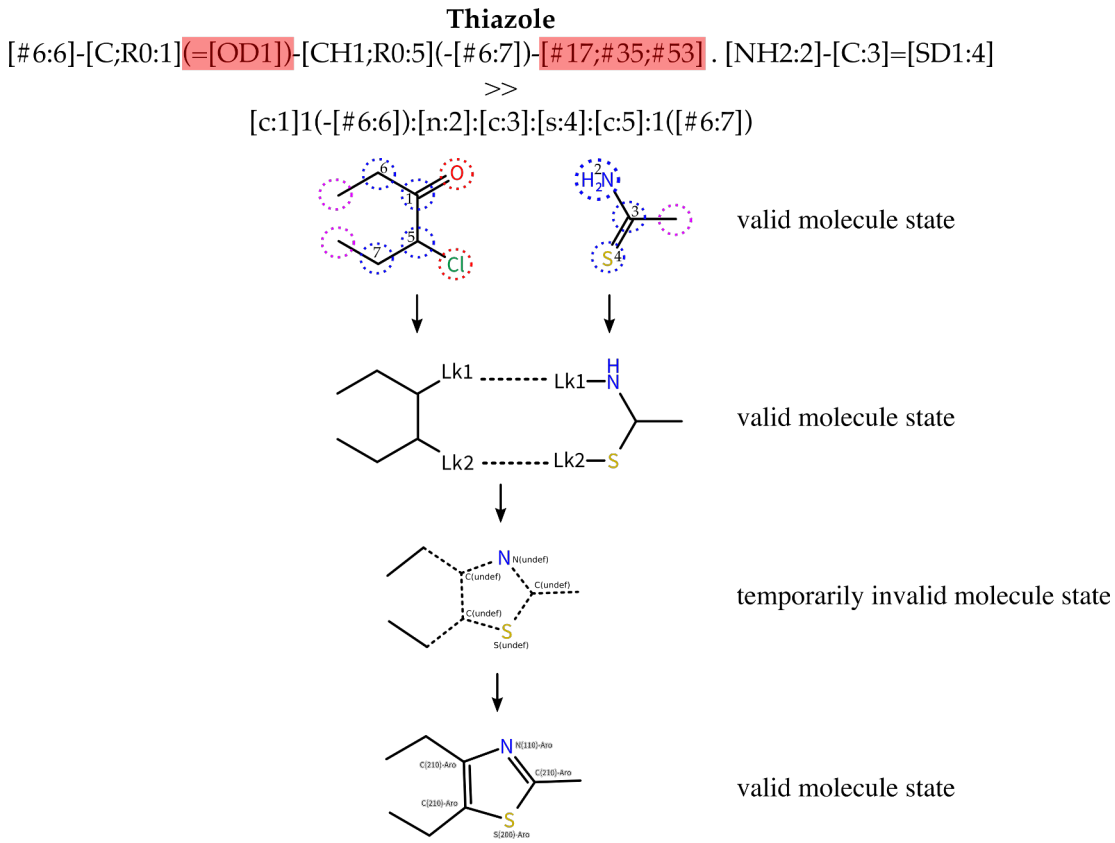


Figure (4.10) Thiazole SMIRKS slightly modified from Hartenfeller and co-workers to exemplify the implementation of the rule based chemical reactions.[43] In the first row the reactants are shown. Matched but unlabeled atoms are marked with a red dotted circle (corresponding SMARTS substring is also marked in red), matched and labeled atoms are marked with a blue dotted circle, and unmatched atoms are marked with a magenta dotted circle. The second row shows an intermediate state of the reactants after the pre-processing step. New covalent bonds are formed between linker atoms. Compatible linker atoms are depicted with dotted lines. The next stage of the reaction process (third row) shows the connected product as topological graph structure in a temporarily invalid state. Finally, NAOMI's molecule initialization procedure is applied to assign the correct valence states, ring membership and so forth.

Niementowski Quinazoline

[c:1](-[C;$(C-c1ccccc1):2](=[OD1:3])-[OH1]):[c:4](-[NH2:5])

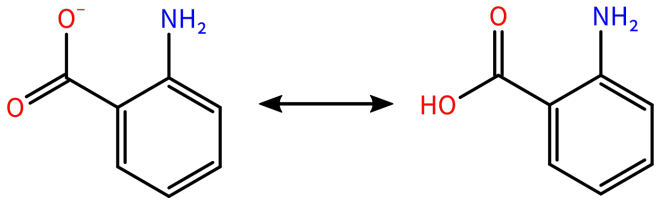


Figure (4.11) SMARTS tautomer and protonation state matching example extracted from a Niementowski quinazoline reaction from Hartenfeller and co-workers.[43] The given SMARTS pattern does only match the structure on the right (protonated form of the carboxyl group). Though it is known, that this group has delocalized electrons, the provided SMARTS pattern does not incorporate this information. A more generic SMARTS representation that matches both shown variants would be `[c:1](-[C;$(C-c1ccccc1):2](=[OD1:3])-[OD1]):[c:4](-[NH2:5])`.

different tautomer- and protonation states of the molecule (see Figure 4.11). Since the chemical molecule representation is based on a VS combination model (see Section 4.1.2), different VS combinations of a molecule can be enumerated efficiently. To be able to perform a SMARTS matching, the line notation of the SMARTS pattern is converted into a topological graph structure (see Section 4.1.5 and Figure 4.11 for an example). This graph structure is then used to extract all necessary information for subsequent pre-processing steps.

Derive Connections from Product SMARTS Graph

For each specific reaction covalent bonds (connections) to be formed are derived from the SMARTS graphs of the reactants and products either. Therefore, all edges of all available product graphs are iteratively processed. The nodes of each edge are checked for existing labels, since, based on the definition of SMIRKS (see Section 4.3.1), labeled and matched atoms describe atoms in the reaction center and the reactant side needs to provide the same number of labels. If the extracted labels originate from different reactants, the edge describes a to be formed covalent bond during the chemical reaction process. The corresponding atoms are extracted from the particular reactant, and bond type and available charges are extracted from the product graph as well.

The Surrogate Derivation (Provide Free Valences)

The next important step is the derivation of surrogates out of reactants and reaction information. This concept is also used within the PINGUI toolkit.[58]

Here, it is implemented differently and only used within intermediate steps. To ensure a consistent molecule representation, based on the underlying VS combination model, VSs and bond types may change based on the given SMARTS pattern. First, edges to-be-cut are determined based on the provided labels of the SMARTS pattern. Each edge from a labeled node to an unlabeled node is cut and unlabeled nodes are replaced using dummy nodes (linker). For example, the matched oxygen atom in Figure 4.10 is cut due to a missing label and replaced with a linker. The VS of the connected carbon atom is adapted accordingly. Linkers represent free valences for new covalent bonds and ensure a valid molecule representation during the pre-processing step. Subsequently, each node is checked for sufficient free valences to perform the needed number of connections. If there are more connections than free valences, different tautomers and protonation states are enumerated for the given molecule and the variant which provides a valid bond type (i.e. mostly a single bond to a hydrogen in case of a different tautomer) is then used to perform the reaction. The disposed hydrogen bond is also cut and replaced with a single bond to a linker (see thioacetamide reactant in Figure 4.10 row one to two). If still not enough free valences are available, as a fallback, bond types are modified (triple to double bond and double to single bond) and VS are adapted accordingly. For example, this fallback is used for the urea and thiourea reaction from Hartenfeller and co-workers.[43] The derived surrogate molecules are used within the subsequent step and are combined to form a new molecule.

The Surrogate/Fragment Combination

After the successful surrogate derivation, the surrogates/fragments are combined to build up a new molecule. Due to the build up of surrogates, all needed and complex modifications like valence state, bond type, and even geometry changes have been already performed. Combining fragments via a single bond is straightforward and an already implemented function in NAOMI is used. For more complex combinations, i.e. ring forming reactions, a completely new fragment combination procedure is implemented.

Ring Closures

Prior to the covalent bond formation, the connecting atoms are prepared. Based on annotated charges in the SMARTS pattern, the atom is either protonated or de-protonated and available stereo information is removed. At this point, the molecule is in a temporarily invalid state for the first time. Next, the covalent bonds are formed between the target atoms using the SMARTS derived bond

type. Subsequently, the valence states for the target atoms are invalidated and attached hydrogens are removed (see Figure 4.10 row 3). The correct VSs are then recalculated using the build-in molecule initialization function of the NAOMI library.[75] If the algorithm detects an aromatic ring closure, intermediary (yet unprocessed) atoms are gathered and incorporated into the recalculation procedure.

The Post-processing step

A successful fragment combination does not necessarily imply the correct reaction result. Therefore, the product SMARTS is used to verify the product. Hence, the applied SMARTS has to match the product molecule, otherwise the result is invalid and discarded. Finally, potential stereo isomers are generated if the reaction leads to a new stereo center or stereo bond.

Coordinate Invariance after Reaction

In this thesis, the described reactions are performed on molecules with a particular geometric orientation within a protein binding site. Hence, the initial fragments orientation is not allowed to change significantly during H2L optimization, i.e. fragment growing. For single bond forming reactions, the newly formed bond connects the anchor fragment with the attached fragment. Thus, the single bond also serves as border to discriminate fixed atom coordinates to retain and atom coordinates that should be re-sampled (exo-bond of a substructure see Section 4.1.5). For more complex reactions, like the thiazole reaction exemplified in Figure 4.10, the coordinate retention step is by far more complex. Atoms that are part of the reaction center may change their geometrical structure after a reaction is performed. For example, in Figure 4.10 the carbon atom that is connected to the chlorine changes its geometry from tetrahedral to trigonal planar. Hence, attached atoms need to be transformed to ensure validity of each atoms geometry. Every atom that is part of the anchor fragment, and whose geometry is kept during the reaction step is gathered. Moreover, aliphatic atoms becoming part of a ring are marked as changed and are not part of the invariant moiety of the molecule. Afterwards the largest connected component of the gathered atoms is determined. Finally a substructure with a single bond exit vector is build (see Section 4.1.5) for all atoms which are not part of the largest component (unchanged anchor part) and new 3D-coordinates are generated ensuring the invariance of the unchanged anchor atoms (see Section 4.1.7). About 50 % of the incorporated reaction rules from Hartenfeller and co-workers[43] are ring closure reactions. For those reactions, extending existing ring systems, a simple 3D-coordinate generation step via a single bond is not

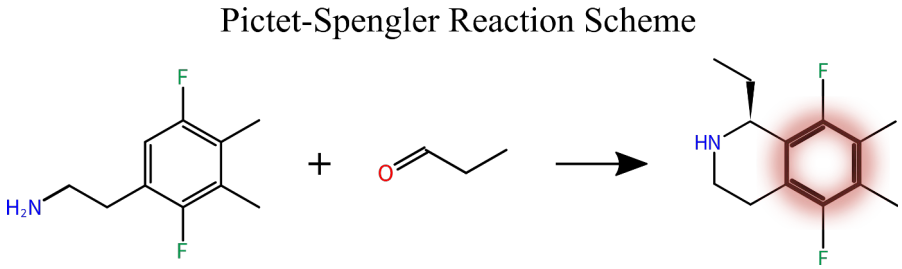


Figure (4.12) Ring template usage for 3D-coordinate generation exemplified on a Pictet-Spengler reaction from Hartenfeller and co-workers.[43] New 3D-coordinates for emerging ring systems are generated using the initial unchanged ring as template (marked in red) for the substructure coordinate generation procedure. Remaining attached atom coordinates at the ring template are kept (fluorine atoms and methyl groups).

possible since the derived substructure includes two exo-bonds that are part of the newly formed ring. In this case, the unchanged ring serves as template for the 3D-coordinate generation process (see Section 4.1.7 and Figure 4.12 for an example).

In Figure 4.12 the invariant part, i.e. the ring template, of the emerging molecule is marked in red. Attached unchanged atoms (the fluorine atoms and methyl groups) are not modified. During the 3D-coordinate generation procedure the new formed heterocyclic ring is aligned to the ring plane of the ring template.

Performed Adaptations and Extensions

The data set from Hartenfeller and co-workers provides many commonly used organic synthesis reactions.[43] Most of the described reactions can be performed using the described workflow above. However, to perform the described *3-nitrile-pyridine* reaction, an additional reactant needs to defined, since the implemented reaction library is not able to extract the complex reactant (*[?3?]N=C(-O)-C([?1?])-C#N*) from the product SMARTS (see appendix E for the performxed adaptation). The additional reactant is described as surrogate (see Section 4.3.2) and contains linkers to perform the needed connections. Hence, NAOMInext is able to perform such complex reactions with implicitly defined reactants. However, to definition is very complex and specific to NAOMInext, i.e. no standard SMIRKS.

4.4. Constraints

In *de novo* drug design constraints can be used to guide or restrict the development process for a specific use case. Most often, they are used to optimize the outcome of a method and or to improve the needed runtime. In the following the different types of constraints are described shortly and their integration into the H2L optimization workflow in NAOMInext is outlined. Two main classes of constraints can be differentiated: implicit constraints (no user interaction is needed) and user defined constraints.

4.4.1. Implicit Constraints

Implicit constraints are invariant and can not be influenced by the user. Nonetheless, they are helpful to facilitate usage and configuration of the algorithm and facilitate ease of use.

Primary Target Constraints

In a SBDD study, the primary target constraints are derived from the 3D-structure of the target protein.[42] The binding site is defined by the co-crystallized or docked anchor fragment. To define the binding site, by default, a radius of 20 Å around all ligand atom spheres is used to query the surrounding of the ligand for nearby protein atoms. The binding site atoms are then used within the protein clash tester for early branch trimming of the sampling tree (see Section 4.2.5). This improves the runtime of the conformational search and additionally, guides the sampling process into a target specific local optimum.

The Anchor Fragment Constraint

The anchor fragment is constrained by design since the developed algorithm uses the co-crystallized or docked fragment in its bound state. To consider for strain during the fragment growing process, the anchor fragment is used in slightly different start orientations (see Section 4.2.2) but its conserved binding mode is ensured.

4.4.2. Interactive User Defined Constraints

User defined constraints allow to influence the sampling workflow for a specific (target specific) use case, e.g. sub pocket targeting. Using modern scientific software, user defined constraints need to be provided in a simple intuitive way.[74] Here, a GUI facilitates users to provide their knowledge interactively and thus,

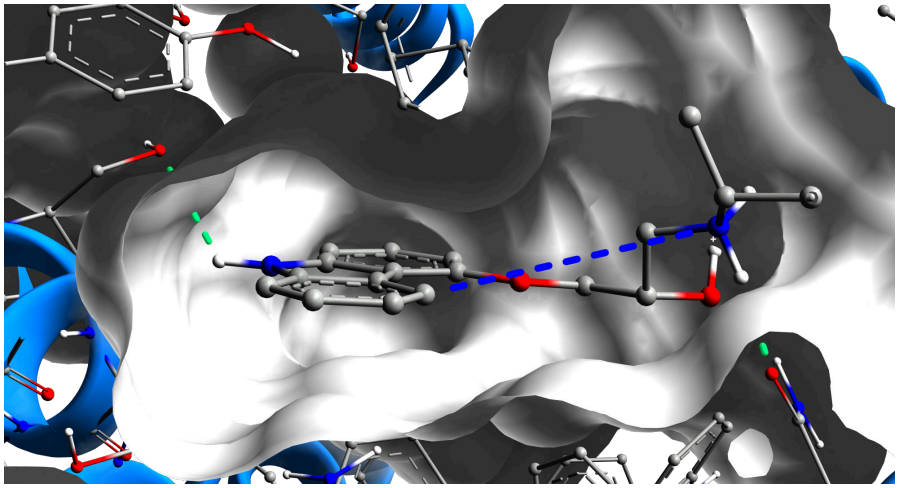


Figure (4.13) User defined constraints may be provided interactively via the GUI. Hydrogen bond constraints (inter molecular) are depicted in green and distance constraints in blue (intra molecular).

allows to influence the decision making step of the automated workflow. The constraint is evaluated as part of the scoring function (see Section 4.1.10) using the described formula 4.3. Compared to primary target constraints, user defined constraints are not discriminative, they are just used to higher rank poses which fulfill the desired constraint. Hence, poses with unexpected binding modes are not missed and can be inspected in the results.

Hydrogen Bond Constraints Part 1

Hydrogen bond constraints can be used if the anchor fragment should be further constrained at specific positions in the binding site. (see Figure 4.13 green lines). This type of constraint can be set by just click and hold the left mouse button on a residue atom (don/acc) and move the mouse to the specific ligand atom (don/acc) and release the left mouse button. This constraint can of course only be set for the already crystallized or docked anchor fragment.

Hydrogen Bond Constraints Part 2

A potentially more often used constraint is the second type of hydrogen bond constraint. This constraint can be placed by clicking on a residue atom with a don/acc functional group. During the incremental construction process the constraint is used as an additional positive contributing scoring term. Which means, if the constraint can be fulfilled, i.e. a hydrogen bond may be formed, the constraint score is calculated, otherwise 0.0 is returned. The optimal distance for

a hydrogen bond is derived from Nittinger and co-workers.[83] This constraint may be useful to higher rank results targeting a specific sub pocket or favor results which perform a desired interaction to increase specificity.

Distance Constraints

These constraints (see Figure 4.13 blue lines) are used to favor sampled poses of a desired user specific geometrical orientation. The set constraint is implemented using the user defined Euclidean distance as an optimum and are evaluated using formula 4.3. Hence, the relative atom position may change as long as the relative distance of the defined atoms stays the same.

The Reaction Rule Constraint

Strictly speaking, the incorporated reaction rules are also a certain type of constraint, since using predefined reaction rules constrains the available chemical space. The reaction rules are a mixture of implicit and user defined constraint. Only reactions which match the anchor fragment are performed. Additionally, the user may select only a subset of the reactions using the `Reaction View` of the GUI (see Figure B.4).

5. Evaluation Strategy and Experiments

This chapter comprises the used data set and performed experiments used to evaluate the formerly described methods. The evaluation chapter can be divided into two main parts. First, the coverage of the conformational space by using the developed sampling algorithm and second, the combination of the chemical reaction rule implementation and the sampling approach. For the latter, the focus is put on the implementation and testing of these rules since the coverage of bioactivity-relevant chemical space has already been proved by others.[46], [47] Additionally, a number of case studies are shown and discussed. The automatic evaluation workflow is implemented in Python[183].

5.1. Sampling Performance - Conquering the Conformational Space

A large-scale analysis of the conformational space coverage is performed using an already published data set of Malhotra and Karanicolas.[44] The sampling algorithm is evaluated in a self-docking and a cross-docking setting (see Figure 5.1 for a visual depiction of the workflow). A self-docking experiment is used to measure the sampling performance without any effects of the protein, like side chain flexibility and protein backbone flexibility due to possible induced binding. The cross-docking experiment is a more realistic use case since the initial fragment (hit) is used in its bound state. Moreover, this test incorporates the ability of the algorithm to consider protein flexibility within the evaluation procedure. Both experiments are also performed with the state-of-the-art docking tool Glide for comparison.[228], [270] The current measure of success of conformational sampling algorithms is the RMSD (see formula 4.1) to the reference crystal structure. As the reference ligand is crystallized within another structure of the same protein, a pairwise superimposition of the protein structures as well as the ligands is performed using SIENA[271] beforehand. All computations are carried out on a workstation with an Intel® Core™ i5 processor (i5-6500 CPU @ 3.20GHz) and 16 GB random-access memory (RAM).

The files are stored on the internal hard drive and no other disk or central processing unit (CPU) intensive jobs were running during computations.

5.1.1. Data Set Preparation

In 2017 Malhotra and Karanicolas examined in a large-scale study, when chemical elaboration of a small ligand induces a binding mode change.[44] The compiled data set contains 297 related ligand-pairs where the smaller ligand is a putative precursor of the larger lead-like ligand. Since fragment growing is a method used for chemical elaboration, this data set is perfectly suitable as benchmark for a fragment growing approach. The data set is published in the Supporting Information of Malhotra and Karanicolas [44], [216] as Microsoft Excel file of PDB id's and corresponding ligand identifiers including additional experimental data and information. For an example, please refer to Table A.1 (only relevant data is shown). The data set needs to be downloaded from the PDB server[2][217]. Thus, interpretation of the PDB files depends on the used cheminformatics toolkit (see Sectiontion 4.1.1 for further information on PDB file format interpretation). Using the PDB ligand perception method of NAOMI[77], fifteen pairs (at least one ligand of each pair) have been identified as covalently bound to the protein. This circumstance was not mentioned in the original publication from Malhotra and Karanicolas although the authors performed a visual inspection of all pairs using PyMol[273].[44] Despite the detected covalent bound ligands, several inaccuracies have been detected considering the provided ligand identifier. Table 5.1 lists the ligand pairs which have been changed. A complete list of identifiers, which have been used to extract the data from the PDB server, is provided in the appendix in section F. To perform an automatic evaluation of a fragment growing approach the provided data set needs to meet further requirements (discussed in the following sections). Thus, leading to a final subset of 271 ligand pairs for the evaluation procedure.

PDB Ligand Perception

At first, the data set was downloaded from the PDB[2][217] using the provided PDB ids and interpreted using the build in chemical model and PDB ligand perception functionality of the NAOMI library shortly described in section 4.1.1. This procedure identified several covalently bound ligands in the data set, which have not been mentioned in the original publication[44] (see Table A.1 for a list of invalid ligand pairs). Hence, these pairs are removed from the validation set. The tautomeric form of the co-crystallized ligand may be altered, since Protoss

[2]https://www.rcsb.org

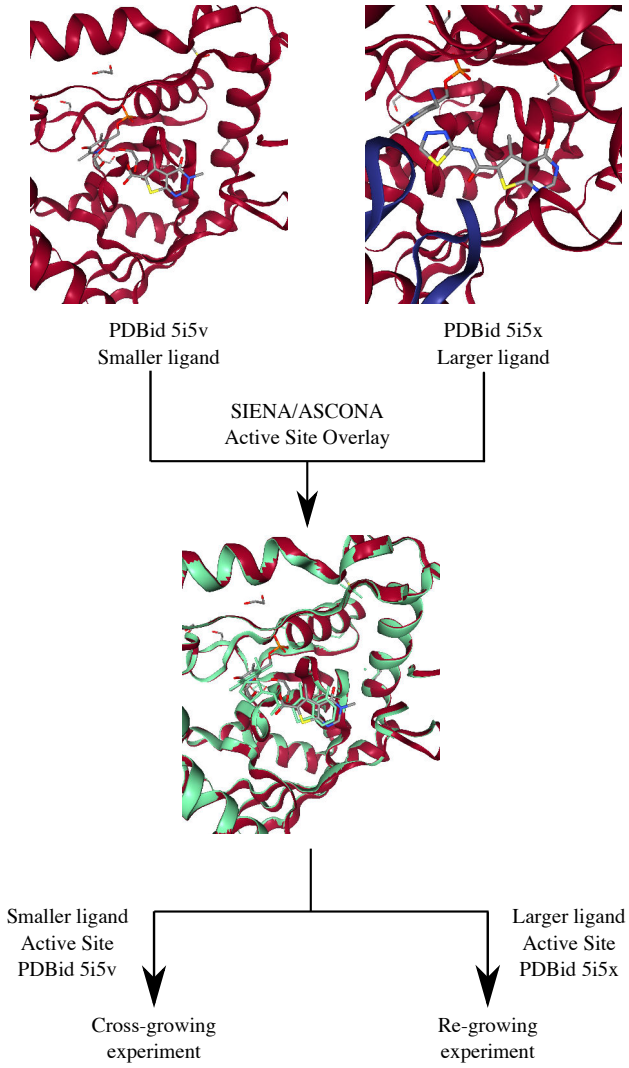


Figure (5.1) Exemplified data set preparation workflow for re-growing and cross-growing experiment. The PDB structures of the related ligand pair are superimposed using ASCONA[240] integrated in the tool SIENA.[271] Cross-growing/cross-docking is performed using the PDB structure of the smaller ligand. The MCS derived anchor fragment is superposed onto the smaller ligands coordinates. Re-growing/re-docking is performed using the PDB structure of the larger ligand. The MCS derived anchor fragment is used in its co-crystallized position. The pictures have been made using the ProteinsPlus[1][272] server.

Table (5.1) Changed ligand identifiers from the published data set of Malhotra and Karanicolas.[44], [216]

Changed Row	Changed Column	Performed Change
24	Smaller Ligand	WI3 → ZZ3
32	Smaller Ligand	NWA → CHH
151	Larger Ligand	4GF → 4GE
183	Smaller Ligand	JAK → 1RS
219	Smaller Ligand	DDM → DMJ

is used to calculate hydrogen atom positions and optimizes the hydrogen bond network.[80], [246] The corresponding ligand of the input structure is extracted using the provided ligand name from the published data set.[44], [216]

Binding Site Superimposition

In a next step, the binding sites of the ligand pairs are superimposed using ASCONA[240] integrated in the tool SIENA[271]. See figure 5.1 for the exemplified workflow. The superimposition of the binding sites is performed, based on the assumption that the anchor fragment and its counterpart in the larger ligand superimpose as well. Build up of the larger ligand, starting from the anchor fragment, should then match the larger ligand structure. Hence, the superimposition is necessary to be able to evaluate the sampling performance of both used methods (fragment growing and docking) using the RMSD as measure of success. Therefore, the larger ligand is extracted from the PDB file via the provided ligand identifier using UNICON[237] and stored in a single SD file. The extracted ligand file and the corresponding PDB file are then used as input for SIENA[271] and the protein file (PDB file) of the smaller ligand as template for the superimposition. Since SIENA identifies the binding site using the provided ligand SD file, this procedure failed for the ligand pair with PDB ids 1L2S[274] and 2HDQ[275] due to a failure during ligand extraction (see Table A.1 for more information). All other pairs have been processed successfully. SIENA is used with default parameters except for the shown parameters in Table 5.2.

Table (5.2) Used SIENA parameters for binding site superimposition of related ligand pairs.

Parameter name	Value
--identity	0.8
--filter_unwanted_ligands	false

5.1.2. Experimental Workflow - NAOMInext Fragment Growing

For the evaluation of a fragment growing approach at least one ligand is needed with a pre-defined anchor moiety and the remaining moiety to be re-attached to the anchor. Here, we use pairs of related ligands from a previously published data set[44], [216]. One smaller ligand, a putative precursor fragment, and a larger, more lead-like ligand (see Figure 5.3 for an example). The smaller ligand is used to identify and extract an anchor fragment (MCS between larger and smaller ligand). The extracted anchor fragment is used in its bound state and the difference of the larger ligand is extracted and used as BB during the fragment growing workflow. Initial 3D-coordinates of the BB are recalculated using UNICON.[237]

Maximum Common Subgraph Conundrum

The aim of the MCS calculation is to extract a substructure of the larger ligand to be used as anchor based on the smaller ligand. Some of the provided ligand pairs are based on real H2L optimization studies. Thus, the initially used core structure may be altered. Additionally, most ligand pairs in the data set do not share a large common substructure, hence, the MCS calculation is far from trivial (see Figure 5.2 in the middle). In this example a pyridine ring needs to be matched onto a phenyl ring. A simple MCS matching would not lead to the correct matching (see bottom in Figure 5.2) Therefore, several extensions to influence the atom matching are made and an additional spatial filter is used (see Paragraph 5.1.2).

The Extended Atom Matching is necessary to facilitate the MCS calculation for all the different types of related ligand pairs. For compatibility reasons to the NAOMI library the MCS algorithm is not modified. Hence, all modifications are performed on the basis of the vertex mapping only. The initial mapping is performed using the element type as identifier. To enable differentiation

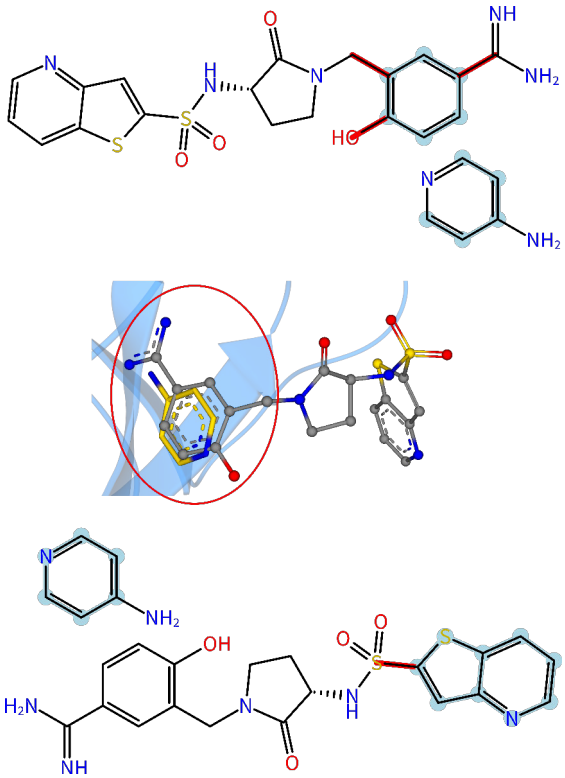
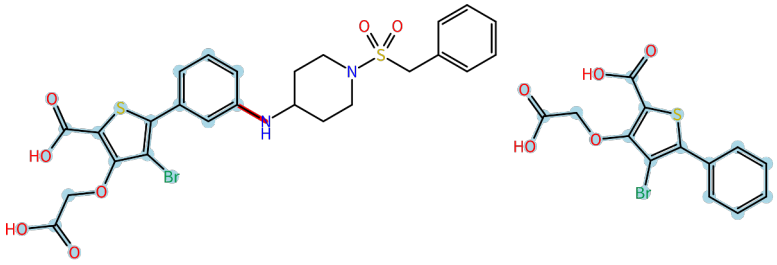


Figure (5.2) Ligand pair from PDB ids 1FoT[276] (larger ligand) and 3RXG[277] (smaller ligand). Top: Refined matching using the spatial filter (matching atoms are marked in light blue). Middle: 3D-view of superimposed ligands (ligand from PDB id 3RXG[277] superimposed in yellow stick model). Bottom: Ligand matching without spatial filter. The extracted anchor between both ligands is highlighted in light blue (anchor fragments are extended to match complete ring systems). The exit vector (connection bond) connecting the sampling fragment (BB) is marked in red.

between ring and chain atoms, a simple bit shift of the identifier is performed and 1 is added to the identifier if the atom is part of a ring. For example, the element identifier for carbon is 6. Bit shifted by 6 bits leads to the value: 384. Consequently, a carbon atom in a ring is assigned the identifier 385. The bit shift is needed to prevent collisions with existing element identifiers. For some ligand pairs, the core structure is altered through integration of a heteroatom (see Figure 5.2 in the middle). This is mostly done during H2L optimization for a number of reasons. Therefore, a ring topological MCS is provided, that ignores the element types of ring atoms to perform a topological ring mapping. A topological MCS is used as fallback if no initial mapping was found. Of course, this is a heuristic and may be invalid for some pairs. However, for some pairs there may not exist a single or optimal solution. Nevertheless, this procedure facilitates automatic processing of most of the provided pairs in the data set. Subsequently, a topological refinement step is performed to remove potential invalid matches. Since the MCS matching is only used for evaluation purposes, a runtime critical implementation is not necessary.

The Matching Extension for Ring Systems is performed to derive complete and valid substructures for the used anchor. As can be seen in Figure 5.2 at the top, only five out of six aromatic carbon atoms of the 4-Aminopyridine structure are matched to the benzene moiety of the larger ligand. This match is achieved using the extended atom matching functionality described above. Since this match does not cover a complete ring, a refinement procedure is performed to incorporate missing ring atoms. In figure 5.2 at the bottom, another example for ring system extension can be seen. Since a valid fragment growing evaluation example needs a single bond to correctly differentiate between anchor moiety and BB, incomplete matched ring systems are extended to incorporate the complete ring system. Hence, a valid substructure is derived, which can be used as anchor fragment in the subsequent described fragment growing evaluation workflow.

The Spatial Filter is implemented to avoid ambiguities during the matching procedure. This spatial filter is used to extract the most likely correct substructure from the ligand pair. This is based on the premise of FBDD, that fragments do not change their binding mode during chemical elaboration. The anchor fragment should, in theory, be a substructure of the larger molecule. Since the ligand pairs are superimposed, each atom sphere of the smaller molecule is used to query its surrounding performing a NN search (see Section 4.1.4). All neighboring atoms from the larger ligand, which overlap with an atom radii of the query ligand, are extracted (figure 5.2 within the red circle). The extracted



Ligand 527 from PDB id 2QBP Ligand 509 from PDB id 2H4K

Figure (5.3) Example ligand pair (PDB ids 2H4K[278] and 2QBP[279])from the compiled evaluation data set (see Section 5.1.1). The MCS between both ligands is highlighted in light blue and used as anchor fragment for the conformational sampling workflow. The exit vector (connection bond) to the sampling fragment (building block) is marked in red.

subset of atoms is subsequently used to perform the MCS mapping as described in section 5.1.2. In figure 5.2 an example match is shown that exemplifies the different outcomes between using and not using the spatial filter.

Building Block Extraction

The MCS derived moiety is used to extract a rigid anchor fragment from the larger ligand. The remaining part of the larger lead like ligand is extracted as BB. Therefore, the atom mapping from the MCS calculation is used to identify bonds where one atom is part of the matched anchor and the other is not. These bonds are called exo-bonds (see red marked bond in Figure 5.3). This exo-bond is cut from the larger ligand and the unmapped part of the molecule is stored as BB alongside with a mapping to the target atom of the anchor fragment for later re-attachment. The 3D-coordinates of the BB are recalculated using UNICON[237]. Based on the derived anchor fragment, multiple extracted BBs are possible.

Subgraph Mapping

Symmetric subgraphs are a problem during the substructure matching procedure because there are multiple possibilities for the subgraph placement and hence, may lead to an erroneous growing vector. For example, the ligand pair with PDB ids 1ROS[280] and 2WO8[281]. Due to a differently placed hydroxy group in the carbon chain that performs a backbone interaction, the whole structure is slightly shifted. Thus, the MCS (derived using the spatial filter) between both ligands contains a single aromatic six-membered ring (benzene). Benzene can be overlaid in multiple ways, due to symmetric substructure definition

(see Table A.1 for similar examples). Hence, pairs with ambiguous symmetrical substructures are excluded from the evaluation set.

Protein Preparation

Protein preparation is performed prior to the BB attachment and the subsequent sampling process. First, the binding pocket is extracted using a 20 Å radius around the anchor fragment. This large radius is used, because the anchor fragment is elaborated using BBs of unknown size. Hence, the ligand may get much larger. Finally, Protoss[80] is used to optimize the hydrogen bond network of the binding pocket including the anchor fragment.

Recursive Attachment Procedure

As a final step, the BBs are appended in a recursive procedure. First, they are sorted in increasing size, so that the largest BB is attached last. For each appended BB, the sampling method described in section 4.2 is applied. Due to possible emerging stereo centers, each attachment may result in two different intermediate molecules. Both intermediates are then used in the subsequent attachment step. Finally, the result poses for each molecule are rescored and ranked in increasing score (worst score last). The best 128 poses are used for evaluation purposes and an RMSD calculation against the reference pose (larger ligand) is performed.

5.1.3. Experimental Workflow - Glide Docking

An automated evaluation on a large-scale data is performed using Glide[228]–[230] via the Python[183] API of the Schroedinger MAESTRO Suite in version 2017.01[235]. Of course Glide can also be used with different constraints like tethering. Here, it is used without constraints and used as state of the art docking tool. The scripted docking workflow is implemented analogue to the workflow performed through the GUI and described in more detail below.

1. The protein including the crystallized ligand is optimized using the preparation wizard[234]. Here, water molecules are removed and the protein structure is minimized. See appendix C.1 `[prepwizard]` for the used python command. Hydrogen coordinates are calculated
2. The co-crystallized ligand (reference ligand) is removed from the processed input structure. See appendix C.1 `[extract_ligand_schrodinger]` for the used python command.

3. To prepare the docking workflow, the active site needs to be cleaned. Therefore, the volume overlap of each crystallized ligand in the input structure to the reference ligand (formerly extracted co-crystallized ligand) in the active site is calculated. If the volume overlap to the reference ligand is larger than 50 % the ligand is removed from the input structure. See appendix C.1 `[prepare_docking_schrodinger]` for the used python command.
4. To perform docking the active site needs to be represented via a grid. For this purpose the extracted reference ligand is used to define the dimensions for the `GridSite` needed to perform the Glide docking. The `GridSite` class is defined in `$SCHRODINGER/mm_share-v3.7/python /scripts/glide_gridgen_gui_dir` The grid center consists of a center, inner box, and an outer box. The grid center is set to the center of the reference ligand coordinates. The inner box has a fixed size of 10 Å around the grid center. The outer box size (default: 20 Å not user customizable) is added to the inner box size, thus leading to an outer box size of 30 Å. See appendix C.1 `[ref_lig_site_schrodinger]` for the used python command.
5. The grid file is generated using the `glide_sif.py` script. The Glide simplified input file (SIF) script aids in automating Glide workflows. The calculated grid dimensions from the former step are used to generate the grid file. See appendix C.1 `[glide_grid_input]` for the used python command.
6. Finally, the Glide input file is prepared, again using the `glide_sif.py` script providing all former generated files: grid file and reference ligand file. Further parametrization consists of: output type (SD file), to be used scoring function (SP), and the number of poses per ligand (32). See appendix C.1 `[glide_docking_input]` for the used python command.
7. Glide docking is simply performed using the generated input file from the former step. See appendix C.1 `[glide]` for the used python command.

5.2. Cross-Docking Evaluation

The cross-docking evaluation is performed analog to the described self-docking procedure including two differences:

1. The 3D-coordinates of the extracted (MCS derived substructure from larger ligand) anchor fragment are transformed onto the corresponding atom coordinates of the smaller ligand. Thus, simulating a real fragment growing experiment based on co-crystallized fragment coordinates.

2. The binding site of the smaller ligand is used to perform inter molecular clash calculations (see Section 4.2.5).

This evaluation is a more realistic use case since important input data from the smaller ligand (anchor fragment) is used as would be the case in a prospective study. The initial coordinate transformation of the atom coordinates is performed using a build in function of the NAOMI library.

5.3. Start Pose Evaluation

The influence of the used start poses on the outcome of NAOMInext is analyzed. Thus, experiments with different numbers of start poses are performed to estimate an upper limit of the individual sampling parts (start pose sampling and torsion driven sampling of the attached building block).

5.3.1. Sampling Performance

To evaluate the influence of the initial start pose sampling on the performance of the growing results, the above mentioned experiments 5.1 and 5.2 are performed without probing the conformational space (torsion driven sampling). Thus, performing the above described workflows but skipping the growing part described in section 5.1.2. Furthermore, only clashing start poses are removed without explicitly reducing the number of generated poses to a specific limit. Hence, an estimate of the upper limit of the possible performance of NAOMInext is gained. The performance of the pose sampling is then measured via RMSD calculation of the anchor fragment part only (substructure RMSD).

5.3.2. Scored Start Poses

To test the influence of the initial start pose sampling in combination with the scoring function and pose reduction, the above mentioned experiments 5.1 and 5.2 are performed without probing the conformational space. Thus, performing the above described workflow but skipping section 5.1.2. The performance of the pose sampling is then measured via RMSD calculation of the anchor fragment moiety only (substructure RMSD)

5.4. Statistical Analysis

The Mann-Whitney-U-test (as implemented in the SciPy package[282]) is used in IPython[283] to compute the significance of difference between distributions, e.g.

statistical difference between the RMSD result sets using different parameters. Depending on the tested hypothesis, either a one-sided or a two-sided test is performed. For example, to test if the result sets of the sampling output are significantly different if either one start pose is used or 50 start poses, a one-sided test was performed. Because it is expected that the obtained RMSD values decrease if the conformational space is probed more extensively.

5.5. Validation of Reaction Rule Implementation

The incorporated reaction rule implementation is evaluated in combination with the implemented sampling approach. To perform the validation, the used evaluation data needs to meet several requirements. First, a related ligand pair needs to be identified, where one ligand is a precursor, i.e. a substructure of the larger ligand, which can subsequently used as anchor fragment. Second, the anchor fragment needs to provide a reaction center for a supported reaction to perform an *in silico* chemical reaction. Third, the matching reaction product needs to be identical to the larger ligand. Last but not least, the binding mode of both ligands must be determined experimentally to be able to evaluate the sampling performance of NAOMInext based on a performed *in silico* chemical reaction.

The incorporated reaction rules from Hartenfeller and co-workers[43] serve as a basis for the challenge to generate synthetically accessible compounds. The ability of the reaction rules to cover the bioactivity-relevant chemical space has already been described earlier.[46] Here, the implementation of the *in silico* reactions is evaluated for its correctness. To evaluate the reaction mechanism implemented in NAOMInext, some case studies from the literature are performed in a retrospective study. Therefore, the MCS between the anchor fragment and the result molecule is extracted and used to perform an artificial fragment growing example. Thus, the difference between the MCS derived anchor fragment and the reference molecule is used as BB (see Figure 5.4 for an example). Additionally, the co-crystallized anchor fragment is used in its genuine form and a putative reaction, available as incorporated reaction rule, is applied to generate the related lead compound within the protein binding site.

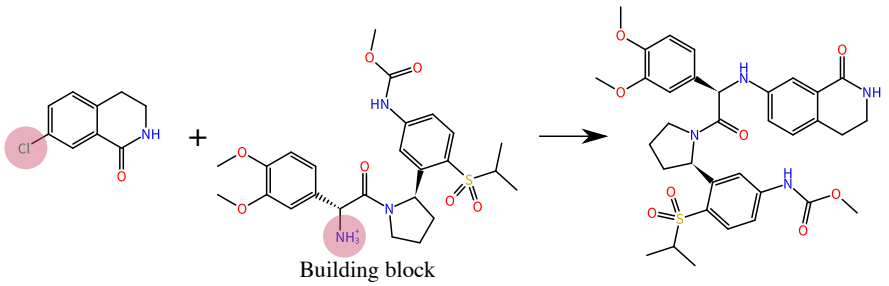


Figure (5.4) Validation case study to test the reaction implementation. The anchor fragment and the BB are used in two different forms. First, an artificial growing is performed using an MCS derived anchor fragment (without the chlorine atom marked with red sphere) and the extracted BB with an attached linker atom (linker added at the nitrogen atom marked in red). Second, the anchor fragment is used in its genuine form and the extracted BB as depicted. Synthetically accessible fragment growing is then performed using build in reaction rules (in this case a *Buchwald-Hartwig* reaction).

6. Results and Discussion

In this chapter the results of the performed experiments described in chapter 5 are shown and discussed. More specifically:

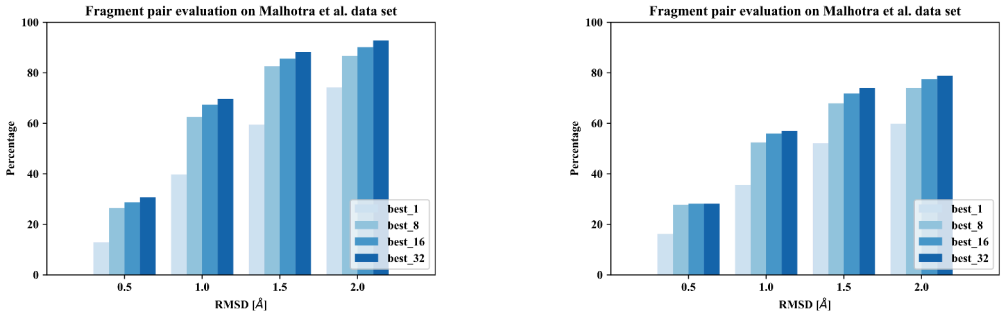
- Re-growing,
- Cross-growing,
- Start pose sampling,
- Runtime, and
- Organic synthesis reactions

are evaluated and discussed. At first, the sampling results are discussed in its entirety and compared to the baseline results generated with Glide. Furthermore, the torsion driven sampling and the start pose sampling are evaluated separately since both sampling algorithms are independent parts of NAOMInext's sampling strategy.

An important aspect in fragment growing, are binding mode changes of the initial hit during H2L optimization. Malhotra and Karanicolas analyzed related ligand pairs in terms of binding mode changes.[44] The compiled large scale data set contains additional information such as annotations about changes of the binding mode. Thus, this data set is used to analyze NAOMInext's sampling strategy and results in relation to binding mode changes. Several case studies are discussed in more detail and the sampling performance in terms of runtime is discussed depending on the flexibility of the used BB, i.e. number of rotatable bonds. Finally, the implementation of the reaction library is discussed based on hand crafted examples.

6.1. Sampling Performance - Re-Growing Experiment

In the first experiment (described in Section 5.1) the large-scale data set of Malhotra and Karanicolas is used to evaluate the sampling performance (conformational sampling) and compare it to a state of the art docking tool. Success is measured as a predicted pose with an RMSD to the crystal structure below 2 Å on the first 32 ranks.



(a) NAOMInext self-growing results on 271 related ligand pairs

(b) Glide self-docking results on 271 related ligand pairs

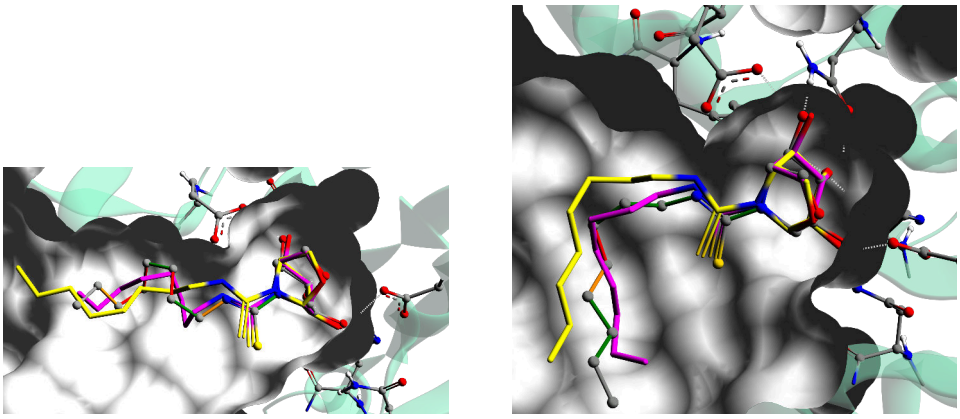
Figure (6.1) Self-docking results from NAOMInext and Glide on a subset of the data set from Malhotra and Karanicolas. Both, docking and fragment-growing were performed using the protein structure of the larger ligand. Each bar shows the achieved percentage of the test set (y-axis) for the best n ranked poses (n = 1, 8, 16, and 32) for a given RMSD threshold (x-axis).

In Figure 6.1 the results for both tools are shown individually. NAOMInext predicts over 90 % of the data set correctly and Glide about 80 %. The performance of both tools drops significantly (about 15 %) if just the highest ranked pose is used to measure success. This indicates, that both scoring functions can be improved in terms of pose ranking.

Docking using Maestro's[235] Glide[228]–[230] failed for four out of 271 ligand pairs from the data set. The targets with PDBids 4HV7[284], 2W8Y[285], 1YHS[286], and 1O5F[287] could not be processed using the Maestro[235] protein preparation wizard[234] due to failures during the protein minimization step. NAOMInext was able to process all ligand pairs from the data set. The superior performance of NAOMInext is expected, since a growing approach starts from an already optimally placed anchor fragment. Whereas docking approaches need to perform the initial placement by themselves (within a user-defined area (box)). Nevertheless, docking tools are often used in fragment growing studies to dock the formerly grown ligand. Subsequently, result poses are post-filtered using RMSD filters with respect to the anchor fragment. In the following, individual results will be further outlined and the results of both approaches are compared.

Beta-glucosidase Results for PDB ids 2VRJ[288]/2J77[289]

This pair of related ligands with PDB ids 2VRJ[288]/2J77[289] is complexed within Beta-glucosidase. The derived building block (BB) of this pair is an octane-carbothioamide moiety. The *n-octyl* chain of the reference ligand has



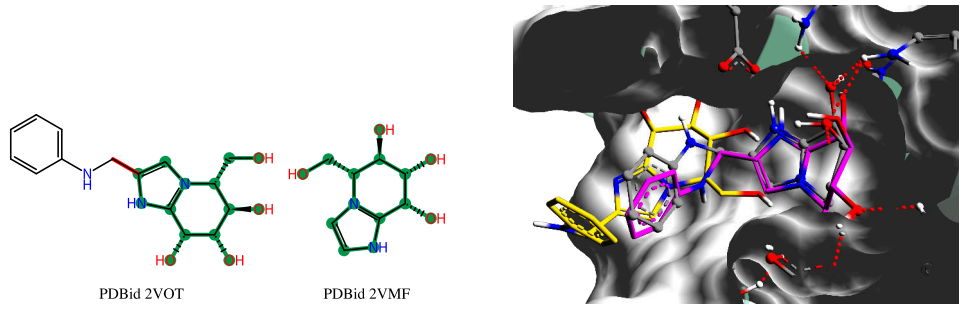
(a) Crystal structure of PDB id 2VRJ[288] with co-crystallized ligand in grey and superposed sampling results.

(b) Crystal structure of PDB id 2VRJ[288] with co-crystallized ligand (alternate location) in grey and superposed sampling results.

Figure (6.2) Sampling analysis including alternate location of PDB id 2VRJ[288]. The co-crystallized ligand is shown in grey including torsion bond quality (green: within first tolerance, orange: within second tolerance, red: outside statistically derived torsion angle tolerances).[78], [82] Poses of Glide and NAOMInext are depicted in yellow and magenta, respectively.

two different conformations, each with an occupancy of 0.5, which reflects the inherent flexibility of the chain.[290] In Figure 6.2 sampling results for both conformations are shown. Both, Glide and NAOMInext predict each pose correctly (below 2.0 Å RMSD to the reference crystal structure). The rotatable bonds of the reference ligand (grey) are color coded and reflect the statistical likelihood of the actual torsion angle, i.e. the torsion bond quality.[78], [82] In both conformations statistically unlikely torsion angles exist (marked in red). In Figure 6.2a Glide predicts a pose with an RMSD of 1.5 Å (yellow) whereas NAOMInext predicts a slightly more accurate pose (RMSD of 0.9 Å in magenta). Here, a smoother sampling approach which generates conformations on the fly, as incorporated in NAOMInext, seems to be superior to methods using a defined number of pre-calculated poses. Although NAOMInext does not generate poses with unlikely torsion angles (outside tolerance 2), the combination of the start pose sampling and using statistically less likely torsion angles (incorporating tolerance 2) is sufficient to achieve appropriate results, even for molecules with statistically less likely torsion angles.

Both approaches, Glide and NAOMInext, predict the alternate conformation of the reference ligand (see Figure 6.2b) accurately, although the *n-octyl* chain has a completely different conformation. This suggests that the available conformational space, within the protein binding site, is sampled efficiently and



(a) Ligand pair with PDB ids 2VMF/2VOT and MCS derived anchor fragment marked in green.

(b) Crystal structure of PDB id 2VOT with co-crystallized ligand in grey and superposed sampling results.

Figure (6.3) Sampling analysis for ligand pair with 2VMF/2VOT[291]. a) MCS derived anchor fragment of the related ligand pair is marked in green. b) Crystal structure of PDB id 2VOT[291] with best pose of Glide and NAOMInext in yellow and magenta, respectively.

diverse poses are obtained for both approaches.

Beta-mannosidase Results for PDB ids 2VMF/2VOT

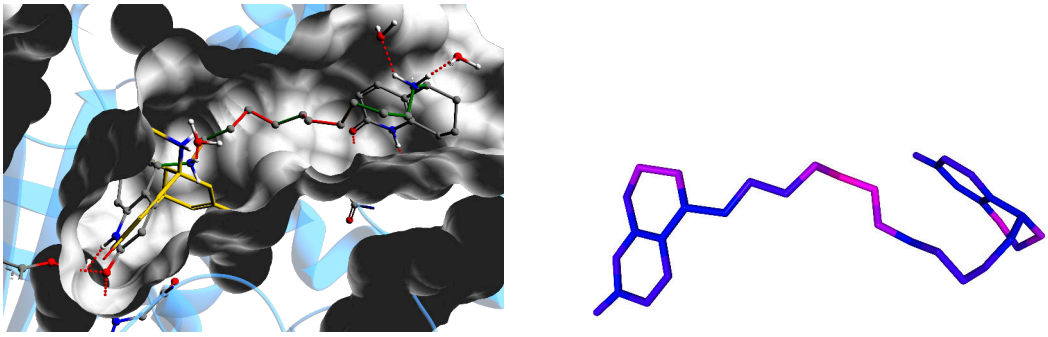
The following ligand pair is a good example for a genuine FBDD study, because both, the anchor fragment and the BB, obey the Ro3.[90] In the underlying study, the anchor fragment, a *Mannoimidazole*, was substituted with several BBs to generate new inhibitors.[291] The ligand from PDB id 2VOT[291] (see Figure 6.3a), is an example for a substituted mannoimidazole. The mannoimidazole fragment is highly strained within the β-mannosidase binding site performing several hydrogen bond interactions to surrounding residues. This tight binding may be the cause for the wrongly predicted pose of Glide since placing the anchor fragment in this small sub pocket may lead to close atom contacts, i.e. clashes. Thus, Glide predicts a completely different binding mode with an RMSD of 5.5 Å to the reference structure (see Figure 6.3b yellow pose). As already reported in the literature, finding new mannosidase inhibitors through HTS or *in silico* docking were unsuccessful.[291], [292] Thus, using a fragment-based H2L approach, e.g. fragment growing, seems to be an obvious alternative to generate a series of potential lead compounds ensuring the experimentally predicted binding mode of the anchor fragment. Performing fragment growing, as provided by NAOMInext, leads to a valid pose with an RMSD of 0.75 Å to the reference structure (see Figure 6.3b magenta pose).

6.1.1. Analysis of the Torsion Driven Sampling

In this section the re-growing results are analyzed considering the torsion driven sampling only. Therefore, the re-growing experiment is performed using the derived anchor fragment position only (without generating additional start orientations), i.e. the anchor fragment is tethered and no movement is allowed. In this scenario, NAOMInext was able to find at least one valid pose for all fragment pairs from the data set except for the pairs with PDBids 1P57[293]/1O5F[287], 2XDL[294]/4AWQ[295], 2XM2[296]/2WCA[297], 3FUH/3FUK[298], and 3VOZ[299]/3UO9[300]. For example, the pair with PDB ids 1P57[293]/1O5F[287] (hepsin) could not be processed due to close protein ligand atom contacts of the crystallized ligand (PDB identifier CR9), which was recognized as clash and no sampling was performed. Another pair, PDB ids 3FUH/3FUK[298] (Leukotriene A4 hydrolase), is crystallized within a very narrow, banana shaped binding site. The ligand from PDB id 3FUK[298] (PDB identifier 58Z) contains statistically less likely torsion angles, which differ significantly from the available angles in the used torsion library.[78], [82] Hence, the sampling algorithm was not able to enumerate a valid, non-clashing pose without altering the start pose orientation. It should be noted that the EDIA[301], [302] score (determined using the ProteinsPlus[1] server[272]) of the ligand is far from satisfying, hence, the ligand may be modeled wrongly into the binding site. This might be an explanation for the statistically unusual torsion angles. However, Glide is able to place the ligand within the binding site with an RMSD of 0.8 Å with respect to the reference structure.

Another interesting example is the pair 2XM2[296]/2WCA[297]. The to-be-sampled fragment contains an ester bond that, according to the torsion library statistics, should have an angle of 0°. However, the crystallized ligand has an angle of 180°. Such grossly distorted angles can only be compensated through introducing more flexibility to the anchor fragment, here, via incorporation of several start poses. Moreover, the shown case studies reveal the drawback of the torsion library approach. Even though the usage of a torsion library reduces the complexity of the conformational space to an amount which is negotiable, it is still a heuristic approach and not all possible solutions may be found. Using additional start poses during the fragment growing workflow, NAOMINext is able to process all of the 271 extracted ligand pairs form the data set. Subsequently, some interesting results are further outlined.

[1] https://proteins.plus



(a) PDB id 1H22 including anchor fragment from PDB id 1GPN (yellow)

(b) Ligand from PDB id 1H22 colored by EDIA score

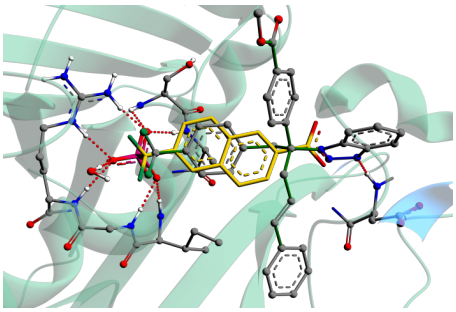
Figure (6.4) The ligand pair with PDB ids 1GPN[226]/1H22[261] crystallized within AChE is used to exemplify the torsion driven sampling performance of NAOMInext (using a tethered anchor fragment position). NAOMInext is able to grow the fragment including the long carbon chain in under 10 seconds considering binding site constraints. Several unlikely torsion angles are emphasized by EDIA calculations.

Acetylcholinesterase Results for PDB ids 1GPN/1H22

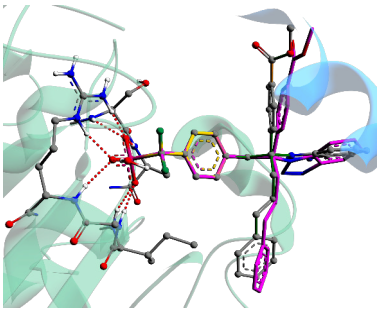
An interesting example is the ligand pair with PDB ids 1GPN[226]/1H22[261] complexed in Acetylcholinesterase (AChE) (see Figure 6.4). The difference to-be-attached, is a large fragment consisting of a long carbon chain with 13 rotatable bonds in a row and a flexible ring system. The carbon chain has statistically unusual torsion angles (see Figure 6.4a), which is hard to hit for an approach that is based on statistically derived torsion angles.[78], [82] The flexible ring system performs hydrogen bond interactions to the backbone of residue phenylalanine (PHE)288 and to a crystallized water molecule (see Figure 6.4a on the right). The best predicted pose (on rank 32 at most) consists of a relatively straight carbon chain comprising mostly, but not exclusively, statistically highly relevant torsion angles. The ring system is placed nearby its crystallized position but does not perform the earlier mentioned backbone interactions (see Figure 6.7a). The atoms of the flexible chain have a slightly worse EDIA score compared to the remaining atoms, indicating a highly flexible region (see Figure 6.4b). Hence, the most important molecule part to-be-placed, is the ring system which is achieved more or less reasonable.

Tyrosine Phosphatase 1B Results for PDB ids 1BZJ/2FJM

The ligand pair discussed here (PDB ids 1BZJ[303]/2FJM[304]) is crystallized within a protein tyrosine phosphatase 1B (PTP1B). The anchor fragment (a (difluoronaphthylmethyl) phosphonic acid derivative) performs important interactions to the catalytic site of the WPD loop (see Figure 6.5a).[303] The WDP



(a) PDB id 2FJM including anchor fragment from PDB id 1BZJ (yellow)



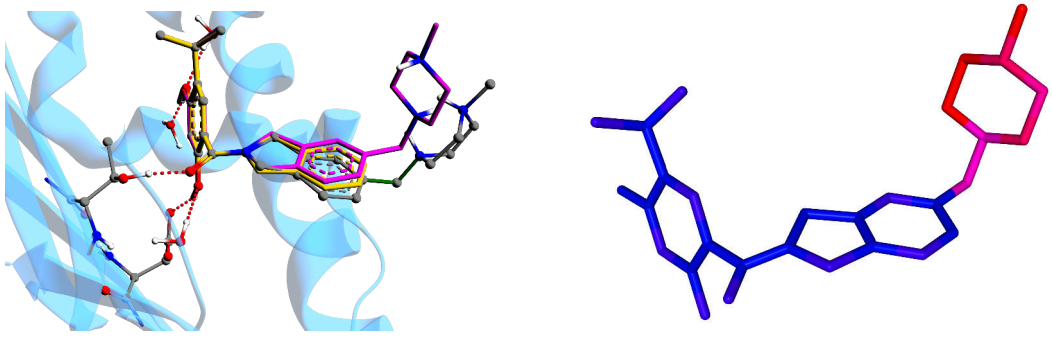
(b) PDB id 2FJM including MCS derived anchor fragment (yellow), reference pose (magenta), and NAOMInext result in grey

Figure (6.5) The ligand pair with PDB ids 1BZJ[303]/2FJM[304] illustrates efficient and accurate results for fragments incorporating statistically highly relevant torsion angles (green marked bonds in Figure 6.5a). Slight bond angle deviations from statistically relevant ones hamper an otherwise nearly optimal sampling result.

loop is a conserved loop and important for catalysis of PTP1Bs.[305] The to be attached fragment has nine rotatable bonds. The observed torsion angles of the larger reference ligand are all statistically highly relevant (see Figure 6.5a). Hence, the sampling result from NAOMInext has a low RMSD of just 0.9 Å to the reference crystal structure (see Figure 6.5b), although the grown fragment is very flexible (nine rotatable bonds). The observed RMSD deviations are mainly based on bond angle deviations of the fourfold substituted carbon atom. The 3D-coordinate generation procedure in NAOMI (which is used to generate initial 3D-coordinates for used BBs) uses standard bond angles based on the VSEPR theory.[256]

Heat shock protein 90 Results for PDB ids 2XAB/2XJX

Heat shock protein 90 (HSP90) is an interesting target in the treatment of cancer and of great interest in FBDD.[96], [294] The generated result pose has an RMSD of 1.5 Å to the crystal structure although the attached fragment has only two rotatable bonds with statistically relevant torsion angles. However, the *N-piperazine* group of the reference structure has an EDIA score of nearly zero since there is no electron density available (see Figure 6.6b). This may be caused either due to high flexibility or the absence of the *N-piperazine* group. Such data inaccuracies make exact statements about the sampling quality, which is based on RMSD calculations, impossible. Nevertheless, NAOMInext does generate a valid pose at 1.0 Å RMSD to the crystal structure on rank 34. Since



(a) PDB id 2XJX including anchor fragment from PDB id 2XAB (yellow) and best NAOMInext result pose (grey)

(b) Ligand from PDB id 2XJX colored by EDIA score

Figure (6.6) Ligand pair with PDB ids 2XJX/2XAB[96] performing strong interactions in the adenosine 5'-triphosphate (ATP)ase site of HSP90. The best predicted pose of NAOMInext differs in the position of the highly solvent exposed *N-piperazine* group which has little impact on enzyme activity.[96] Thus, different orientations are hard to detect for scoring functions.

the `N-piperazine` group is solvent exposed, the scoring function is not able to properly distinct poses with high RMSD differences. Another factor for the relatively high RMSD value is the differing ring system conformation of the 1,3,-dihydro-2H-isoindole moiety which is part of the anchor fragment. The attached fragment is connected to a non-aromatic ring system (at least not completely planar). Flexible ring systems may have a high influence on the relative orientation of attached fragments. Hence, if a fragment is attached at a flexible ring system, NAOMInext includes the ring system into the sampling process to increase the variability during the sampling. In this case, available ring system conformations differ from the co-crystallized form and lead to a higher RMSD.

Summary

Despite some pitfalls, the torsion driven sampling as implemented in NAOMInext is able to determine crystal structure poses within a short time frame even for BBs with many rotatable bonds (see ligand pairs with PDB ids 1BZJ[303]/2FJM[304] in Section 6.1.1). However, a prerequisite for good results is the quality, i.e. statistical significance, of the torsion bonds (see ligand pairs with PDB ids 1GPN[226]/1H22[261] in Section 6.1.1). Since this prerequisite is not always fulfilled, additional orientations of the anchor fragment, i.e. start poses, are used to improve the performance of NAOMInext's sampling strategy.

6.2. Influence of the Start Pose Sampling

An implicit constraint used in NAOMInext is the relative orientation of the anchor fragment within the protein's binding site. In this section the influence of additional orientations of the anchor fragment on the sampling performance is analyzed. First, the results discussed in Section 6.1.1 are taken up again and compared to obtained results using 50 start poses (default value). The analysis addresses two different questions:

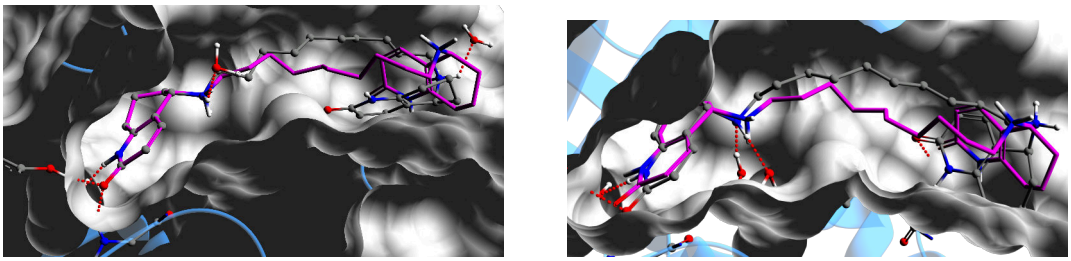
- Are multiple start poses necessary?
- Does introducing more variability (degrees of freedom) worsen the pose ranking?

6.2.1. Are Multiple Start Poses Necessary?

To address the above mentioned question, experiments with different number of start poses are performed. Therefore, the results from Section 6.1.1 are taken up again. Finally, results of the whole data set are discussed and extensively analyzed.

Acetylcholinesterase Results for PDB ids 1GPN/1H22

Using one start pose NAOMInext is able to generate a pose with an RMSD of 2.6 Å but fails to place the flexible ring system (*huperzine B* derivative) in its co-crystallized position. Hence, important hydrogen bond interactions to the backbone of residue PHE288 are missing (see Figure 6.7a).



(a) PDB id 1H22 including NAOMInext result in grey (1 start pose) and reference ligand (magenta)

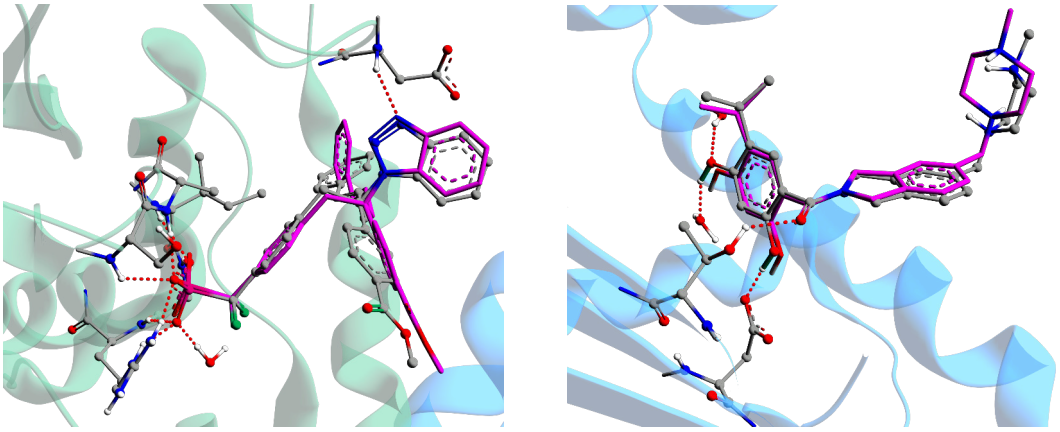
(b) PDB id 1H22 including NAOMInext result in grey (50 start poses) and reference ligand (magenta)

Figure (6.7) Sampling results of ligand pair with PDB ids 1GPN[226]/1H22[261] based on different number of start poses. Using multiple start poses improves the conformational sampling result. The placed flexible ring system performs important hydrogen bond interactions to the backbone of residue PHE [288].

Incorporating different start orientations of the anchor fragment into the sampling process, improves the performance and a RMSD of 1.6 Å to the crystal structure is obtained. But more importantly, the flexible ring system is placed in such a way, that it can perform hydrogen bond interactions to the backbone of residue PHE288 (as performed by the reference crystal structure). Thus, ensuring the binding mode of the crystallized reference ligand.

Tyrosine Phosphatase 1B Results for PDB ids 1BZJ/2FJM

In this example, introducing more start poses does not significantly improve the pose result, since using one start pose already lead to a valid pose with an RMSD of just 0.9 Å. Nevertheless, using 50 start poses leads to the best pose with an RMSD of 0.84 Å on rank seven. Although the RMSD does not change dramatically, the benzotriazole group performs an additional hydrogen bond interaction to the backbone nitrogen atom of residue aspartic acid (ASP)548 (figure 6.8a) (just as the reference pose (see Figure 6.5a)).



(a) PDB id 2FJM including NAOMInext result in grey (50 start poses) and reference ligand (magenta)

(b) PDB id 2XJX including NAOMInext result in grey (50 start poses) and reference ligand (magenta)

Figure (6.8) Sampling results for ligand pairs with PDB ids 1BZJ[303]/2FJM[304] and 2XAB/2XJX[96] using 50 start poses.

Heat shock protein 90 Results for PDB ids 2XAB/2XJX

Using one start pose, the best obtained pose (on rank 15) has an RMSD of 1.5 Å to the crystal structure. A slightly better pose is found on rank 34 with an RMSD of 1.0 Å. Performing the same experiment using 50 start poses, leads to an RMSD of 1.0 Å on rank three, thus improving the rank performance although

the used start pose that leads to the better RMSD value is only slightly different. As already mentioned above, slightly deviating ring system conformations have a significant influence on the obtained result pose. Here, a minor rotation of the start pose is enough to overcome the problem of the slightly different ring system conformation (compared to the reference crystal structure). This example shows that an initial pose minimization step may be beneficial and may lead to better overall results.

Overall analysis

The discussed case studies imply that using more start poses leads to better results. To test if both calculated RMSD sets (using either one start pose or 50) are statistically significantly distinct, the null-hypothesis is defined as: 'H_0 = *No RMSD change*'. This means, increasing the number of start poses has no significant influence on the results. Here, $\alpha = 0.05$ (5 %) is used as level of significance, which is a typically used threshold.[306] Performing a Mann-Whitney U-test leads to: $P = 0.35$. Thus, the null-hypothesis is not rejected, i.e. the data sets do not differ significantly ($0.35 \geq \alpha$). Consequently, increasing the number of start poses does not improve the results significantly.

For 124 of the 271 test cases the RMSD to the crystal structure improves, in case 50 start poses are used. For 115 test cases the RMSD values deteriorated (see Figure 6.9). For the remainder no RMSD change is detected. These numbers emphasize the results of the Mann-Whitney U-test. Considering RMSD changes of ± 0.5 Å as acceptable variation (see Figure 6.9 orange points), only seven cases got worse (see Figure 6.9 red points) but 63 cases improved (see Figure 6.9 green points). This shows that the use of additional start poses leads to an improvement of the overall outcome of NAOMInext. However, this improvement is statistically not significant. In the following, interesting results are analyzed and discussed.

Serine Protease Hepsin Results for PDB ids 1P57/1O5F

In this serine protease case study (PDB ids 1P57[293]/1O5F[287]), increasing the number of start poses leads to a valid result. The sampling algorithm failed if only one start pose is used. The co-crystallized ligand performs very short hydrogen bonds to the protein.[293] As a result, the ligand is strongly pulled to the protein surface and the implemented scoring function determines the initial growing step (fluorine attachment, see Figure 6.10) as clash. Using additional start poses, a slightly different start pose is sufficient to perform a valid growing step (see Figure 6.10 magenta overlay). To address such issues, an initial minimization procedure, slightly optimizing the molecule/anchor

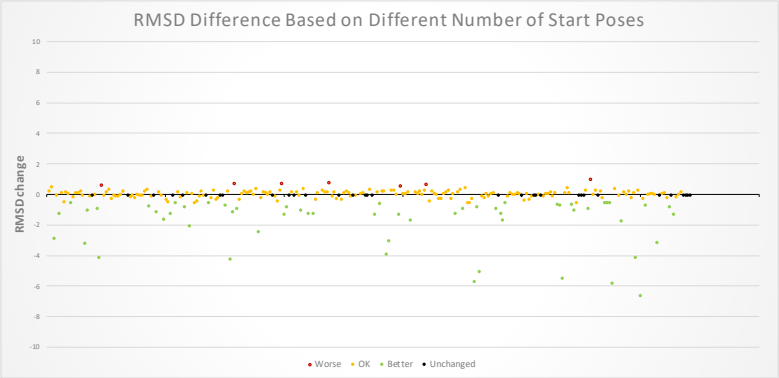


Figure (6.9) NAOMInext re-growing analysis between one and 50 (default) used start poses. The RMSD difference is color coded. If using more start poses leads to a more than 0.5 Å RMSD increase, the RMSD difference is plotted in red. If the RMSD decreases by more than 0.5 Å the difference is plotted in green. Changes between -0.5 Å and 0.5 Å are treated as acceptable variations and marked in orange. Unchanged RMSD values are marked in black.

fragment position according to the used scoring function, may prevent such numerical inaccuracies.

BACE-1 Results for PDB ids 3L5B/3L5D

Using only one start pose and performing a re-growing experiment (ligand pair with PDB ids 3L5B/3L5D[307]), the best NAOMInext result pose has an RMSD of 7.4 Å (figure 6.11a) compared to 1.6 Å if 50 start poses are used (figure 6.11b). Starting from the optimally placed anchor fragment NAOMInext grows into the S1' pocket of BACE-1 (figure 6.11a), which is a conventional binding site.[307] The highest ranked pose (using 50 start poses) has also a high RMSD, namely 9 Å. Again, the S1' sub pocket is targeted. Thus, high RMSD values originate, because the used scoring function (see Section 4.1.10) seems to prefer ligands targeting the S1' sub pocket due to better binding site complementarity. This result shows that depending on the binding site, a detailed inspection of result poses is inevitable for the success of subsequent optimization steps.

CAII Results for PDB ids 3SBI/3MYQ

In this example, a human CAII protein structure, the common core is a *benzenesulfonamide* group performing strong interactions with the zinc(II)-atom which is ligated by histidine residues.[308] Here, increasing the number of start poses also increases the RMSD value of the best found pose. Key interactions to the zinc(II)-atom, threonine (THR)199, and THR200 are kept. One possible issue may be the used scoring function. Since the NAOMI framework does not

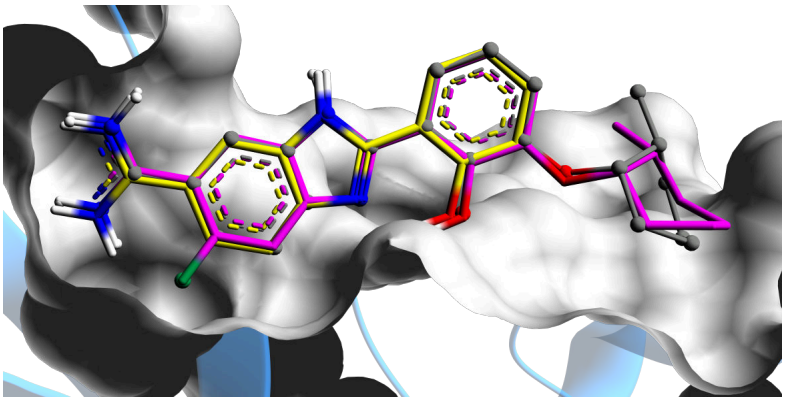
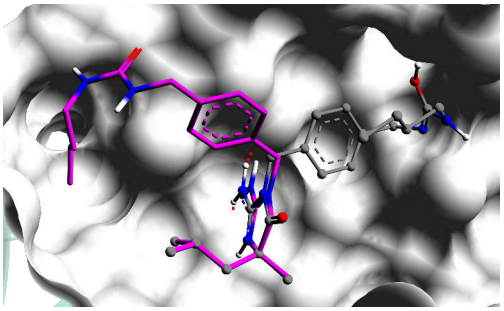
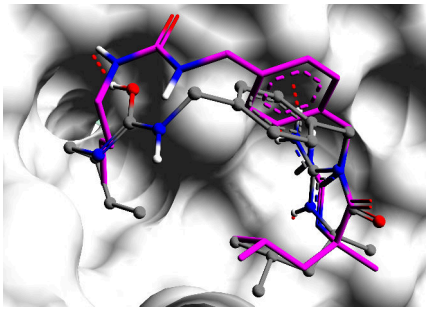


Figure (6.10) Start pose analysis for ligand pair 1P57[293]/1O5F[287]. The serine protease performs strong interactions to the crystallized ligand, thus, leading to immediate clash detection of the anchor pose during fragment growing of the fluorine atom (grey pose) and thus, abortion of the growing step. Using a slightly different start pose (magenta) allows growing the fluorine atom into the binding site without detection of clashes.

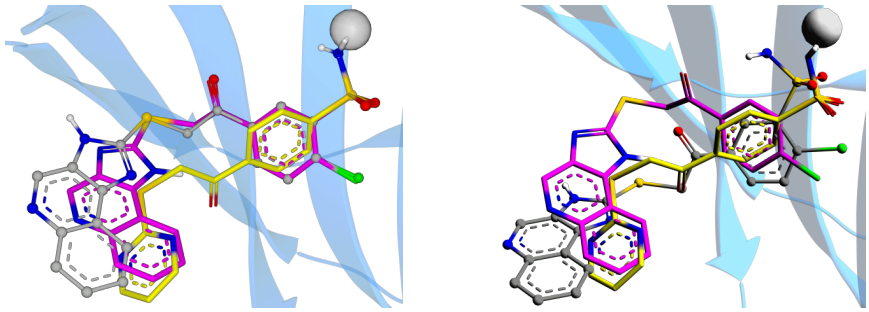


(a) NAOMInext result using one start pose (RMSD: 7.4 Å)



(b) NAOMInext result using 50 start poses (RMSD: 1.6 Å)

Figure (6.11) Start pose analysis for BACE-1 enzyme using the ligand pair with PDB ids 3L5B/3L5D[307]. Using 1 start pose the scoring function prefers targeting the S1′ sub pocket of the protein binding site. Using 50 start poses NAOMInext nearly finds an optimal solution performing key interactions within the S1 sub pocket.



(a) NAOMInext result using 1 start pose (RMSD: 1.3 Å)

(b) NAOMInext result using 50 start poses (RMSD: 2.3 Å)

Figure (6.12) Start pose analysis for CA II using ligand pair with PDB ids 3SBI[309]/3MYQ[310]. Anchor fragment (yellow), reference pose (magenta), and best NAOMInext result (grey) are shown. Increasing the number of start poses also increases the RMSD of the best obtained pose to the reference crystal structure.

determine the metal coordination geometry, interactions to metal atoms are calculated using a simple distance term. Including an angle term to the metal scoring may improve the pose ranking and most probably, will prefer the initial position of the anchor fragment.

6.2.2. Does Introducing more Variability (degrees of freedom) Worsen the Pose Ranking?

Incorporating more start poses into the sampling inevitably leads to more results, hence, places increased demands on the pose ranking ability of the scoring function. Figure 6.13 shows re-growing results for different number of used start poses. Using 50 start poses leads to the best overall RMSD to the crystal structure on rank 32 at most and is used as default in NAOMInext. To investigate the performance of the implemented sampling algorithm in dependence of the number of used start poses the experiment described in Section 5.3 is performed. As default, 50 start poses are used as described in Section 4.2. Here, additional numbers of start poses: 1, 10, 25, 75, and 100 are tested. The results are shown in Figure 6.13. It should be noted, if only one start pose is used, the extracted anchor fragment is based on coordinates of the reference crystal structure. Thus, the anchor part is already placed optimally (considering a re-growing experiment) and the results completely depend on the performance of the torsion driven sampling algorithm. For all other cases, the procedure described in Section 4.2.2 is applied. As can be seen in Figure 6.13, considering an RMSD of 2 Å as a threshold of success, using 50 start poses

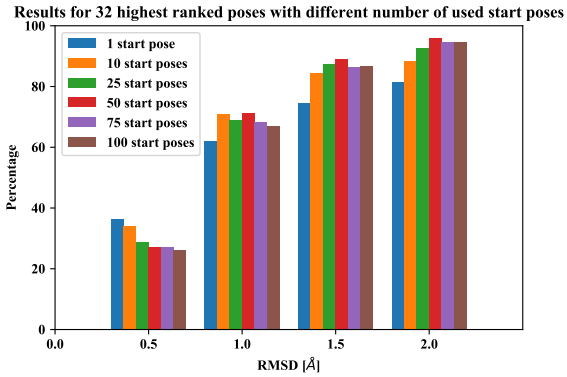


Figure (6.13) Re-growing results for the 32 best ranked poses using different number of start poses. The RMSD thresholds to the crystal structure are shown as bins on the x-axis. The ratio of the correctly predicted poses of the data set from Malhotra and Karanicolas[44] is shown on the y-axis.

leads to the best results.

Considering the results below 0.5 Å RMSD (see blue bar in Figure 6.13) using only one start pose seems to be superior compared to more start poses. An explanation could be the fact, that the crystallized pose (anchor fragment) is not part of the initial start poses and hence, the sampling algorithm only generates less good poses. This fact would point out the poor ranking ability of the scoring function. Regarding higher RMSD thresholds, this effect is compensated and even using a larger number of start poses leads to better results compared to one start pose. But this advantage has a peak at about 50 start poses. Further increasing the number of poses worsens results, which can be explained with the pose impaired ranking ability of the scoring function. Since the scoring function is not able to rank the crystal structure at the highest position, introducing more conformations further decreases the pose ranking performance.

This issue is compensated via clustering the result poses to maintain diversity. The clustering, described in Section 4.2.6), extracts diverse but still highly ranked poses. Omitting the clustering step deteriorates the results by 10 % (not shown). Nevertheless, considering the results on the whole data set, the use of multiple start poses is highly recommended. However, in specific cases it might be advantageous to just use the crystallized pose to improve the result ranking.

NAOMInext does not provide or use a pose minimization procedure. Hence, the generated poses are completely based on the available torsion angles of the underlying torsion library[78], [82] and the used start pose. For performance reasons, highly similar torsion angles are clustered to reduce the needed runtime. Furthermore, the derived angles from the torsion library, i.e. peak angle and tolerance values, describe only the area in which torsion angles are statistically

significant. Hence, a fine adjustment of individual angles is not intended in the workflow of NAOMInext. Therefore, using slightly varying start poses compensates for the downside of a stepwise torsion angle sampling.

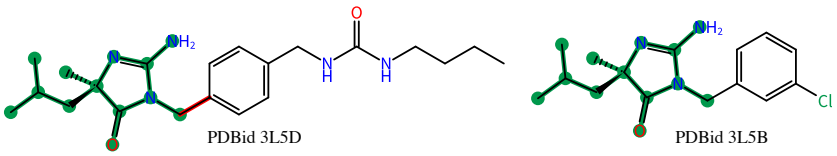
Summary

The sampling performance of the implemented sampling algorithm, mainly driven by applying torsion angles to rotatable bonds, is sufficient to reproduce the reference structure within its protein binding site for over 90 % of the data set (see Figure 6.13). The combination of the heuristic approach and the dynamic adaptation lead to results fast, but still with an acceptable performance. Incorporating different start positions of the anchor fragment is important for the performance of the algorithm. Compared to state-of-the-art docking, the described approach is superior considering the generated poses and using the described measure of success. The docking results may be improved using a constrained docking approach, e.g. tethered docking. Moreover, this requires profound knowledge of the used tool and usually has to be done manually. At least for Glide, this could not be performed (without a tremendous amount of scripting) in an automated fashion to perform a large scale analysis.

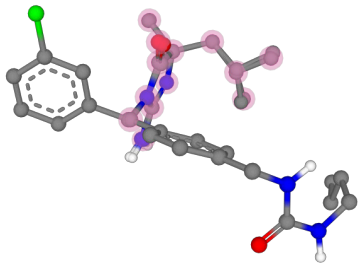
6.3. The Influence of the Spatial Filter

The initial step of the fragment growing experiment is a MCS calculation between the related ligand pair. The aim is a MCS derived anchor fragment, which is then used as starting point for the growing experiment (see Section 5.1.2). The implemented spatial filter (Section 5.1.2) facilitates a correct substructure determination (see figure 5.2 for an example). Figure 6.14 shows an example, where the derived anchor fragment omits a phenyl moiety which would be part of a correct MCS. The related ligand pair is shown in Figure 6.14a with the MCS derived anchor fragment marked in green. Because a spatially corrected MCS is calculated here, the correctly matching atoms of the phenyl moiety are omitted due to large distance error (see figure 6.14b).

In Figure 6.15 the EDIA score of both ligands is shown. The phenyl moiety of the ligand from PDBid 3L5B (Figure 6.15a) has a really low EDIA value indicating, either a wrong molecule, or a highly flexible part of the molecule. Considering the related ligand from PDBid 3L5D, depicted in Figure 6.15b, it's more likely that the phenyl moiety is highly flexible and may also point into the opposite direction. Reducing the MCS derived anchor fragment seems logically in this specific case. Moreover, this molecule does not obey the R03 for fragment like molecules as it has four rotatable bonds. Nevertheless, anchor fragments

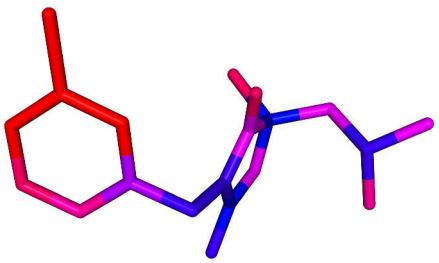


(a) Ligand pair from PDB ids 3L5B and 3L5D with marked substructure (green) and extension vector of the extracted BB in red.

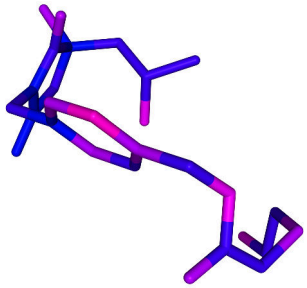


(b) Binding site superimposed related ligands from PDB ids 3L5B and 3L5D.

Figure (6.14) Binding site superimposed related ligands from PDB ids 3L5B and 3L5D[307]. a) The derived anchor fragment is highlighted in green b) 3D-view and corresponding anchor fragment atoms are highlighted using red spheres. A simple topological MCS between both structures would include the phenyl moiety. Here, a spatially corrected MCS is calculated which filters out strongly deviating atoms.

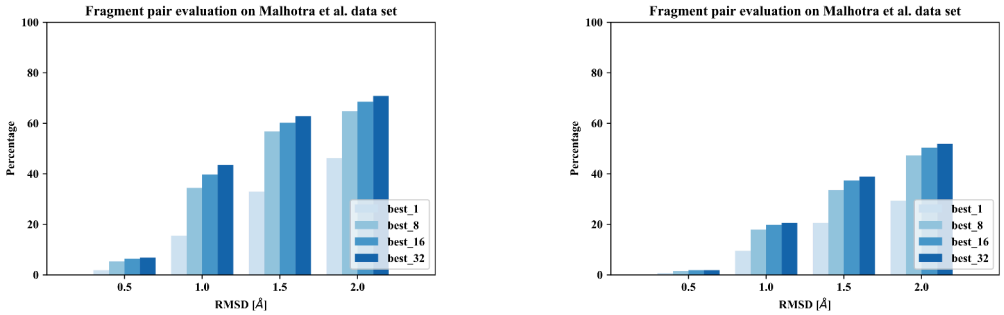


(a) Ligand BDO from PDBid 3L5B colored by EDIA score.



(b) Ligand BDV from PDBid 3L5D colored by EDIA score.

Figure (6.15) Related ligand pair from PDBids 3L5B and 3L5D colored by EDIA[302] score. The phenyl moiety from ligand BDO (PDBid 3L5B) has a really low EDIA score as there is barely electron density available. This indicates that the phenyl ring may be very flexible and a rigid positioning is kind of arbitrary. Pictures created with ProteinsPlus[2].[272]



(a) NAOMInext cross-growing results on 271 related ligand pairs

(b) Glide cross-docking results on 271 related ligand pairs

Figure (6.16) Cross-docking results a) NAOMInext and b) Glide on a subset of the data set from Malhotra and Karanicolas[44]. Each bar shows the achieved ratio of the test set (y-axis) for the best n ranked poses (n = 1, 8, 16, 32) for a given RMSD threshold (x-axis). Both, docking and fragment-growing experiments, are performed using the protein structure of the larger ligand.

incorporating flexible side chains should be avoided, because NAOMInext does not incorporate torsion flexibility of the anchor fragment during the optimization workflow. In such cases, tools performing docking of the complete structure may be advantageous. In other cases, i.e. using ligand pairs of non-conserved binding mode, the spatial MCS filter may prevent the algorithm from finding a valid anchor fragment. This does not falsify the results since NAOMInext is not able to predict binding mode changes anyway.

6.4. Cross-Growing/Cross-Docking Experiment

Besides the re-docking/re-growing experiment cross-growing and cross-docking experiments are performed. This experiment is a much more realistic scenario and incorporates putative changes of the protein conformation due to ligand elaboration. The test uses data from the smaller ligand of the related ligand pair, i.e. the protein structure and the binding mode. Therefore, the MCS derived anchor fragment is superimposed onto the smaller ligand coordinates to simulate a real fragment growing scenario. Furthermore, the soft-docking ability of both approaches is tested as well as their ability to work with experimental data from fragment screening experiments of FBDD projects.

Here, the performance compared to the re-docking results drops significantly for both tools (see Figure 6.16). The pose ranking seems comparable to the re-docking results with the exception of the larger drop between the best and best eight poses (on average 20 %). In this evaluation Glide and NAOMInext

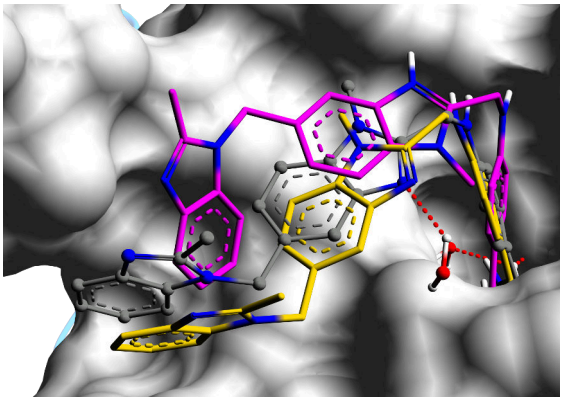


Figure (6.17) Cross-Docking results for trypsin and ligand pair with PDB ids 1G36[311]/3GY4[312]. The reference structure, the best Glide pose, and the best NAOMInext pose are shown in grey, yellow, and magenta, respectively.

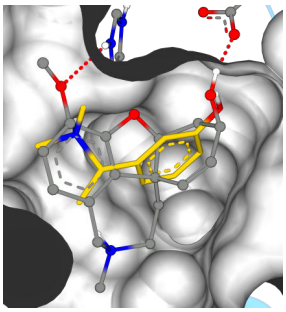
failed in 17 and 21 out of 271 test cases, respectively. A failure is, for example, a pair with a very different binding mode, hence, the growing could not be performed due to possible clashes with the binding site. For about 70 % of the test cases NAOMInext is able to predict a valid pose below 2.0 Å RMSD to the crystal structure on rank 32 at most. Glide correctly predicts about 50 % of the test cases using the same rank criteria. In the following, several case studies are discussed.

Trypsin Results for PDB ids 1G36/3GY4

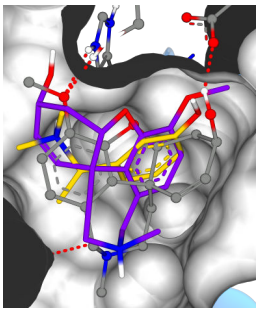
Open spaced binding sites, such as in the case of trypsin (see Figure 6.17), are more difficult for the sampling approach of NAOMInext since the conformational space is much larger and only less binding site constraints can be used to guide the sampling procedure into a local optimum. In the example shown in Figure 6.17, Glide obtains a valid pose with an RMSD of 1.5 Å (NAOMInext 2.8 Å) with reference to the crystal structure. NAOMInext misplaces the central benzimidazole group inverted by 180° performing a different hydrogen bond interaction to a water molecule (not shown).

Acetylcholinesterase Results for PDB ids 1GQS/1DX6

This ligand pair from the data set of Malhotra and Karanicolas is complexed within AChE and marked as binding mode changed.[44] Here, both ligands share a phenyl moiety as common core. Since NAOMInext uses the position of the MCS derived core structure, the phenyl moiety is misplaced and thus leads to a large RMSD with respect to the crystal structure of 4.3 Å. Nevertheless, the most important groups, performing hydrogen bond interactions, are placed in



(a) AChE structure from PDB id 1DX6



(b) AChE structure from PDB id 1DX6

Figure (6.18) Ligand pair with PDB ids 1GQS[313]/1DX6[314] annotated with binding mode change. a) PDB id 1DX6[314] and co-crystallized ligand in yellow. Larger ligand from PDB id 1GQS[313] superposed in grey. b) PDB id 1DX6[314] and co-crystallized ligand in yellow. Best glide pose in grey and NAOMinext pose in purple. The best Glide pose has an RMSD of 0.75 Å and the best pose of NAOMInext has an RMSD of 4.3 Å

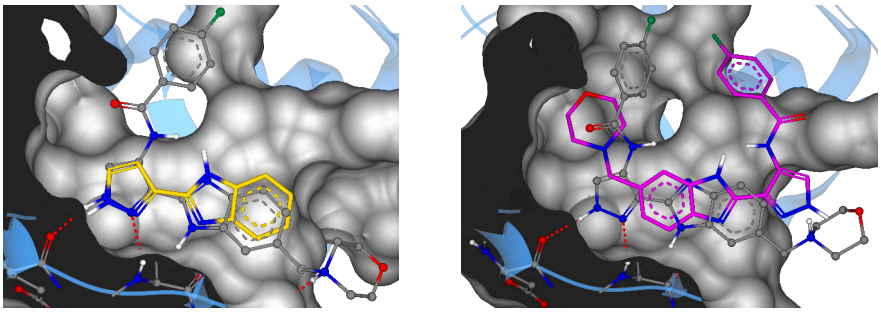
the right positions of the binding site (see Figure 6.18a). Glide does not use any information about the ligand binding position and perfectly docks the ligand into the binding site maintaining the binding mode of the reference structure with an RMSD of 0.75 Å.

Aurora Kinase A Results for PDB ids 2W1D/2W1C

This ligand pair is the result of a FBDD study for the development of an Aurora kinase A inhibitor starting from a pyrazole-benzimidazole fragment.[315] The pyrazole-benzimidazole fragment performs strong hydrogen bonds to the kinase hinge region (residues alanine (ALA)213 and glutamic acid (GLU)211 in PDB id 2W1D[315]) within the ATP-binding site of Aurora A kinase (see Figure 6.19).[315]

The elaborated ligand (PDB id 2W1C) contains an additional fluorinated benzamide group and a morpholine group (see grey structure in Figure 6.19a). The morpholine group (of the best NAOMInext pose) shows the largest RMSD deviation to the crystal structure. Since it is solvent exposed, several placements of the morpholine ring are possible. However, the morpholine ring in the reference structure performs a hydrogen bond interaction to the protein that is not targeted by the NAOMInext pose. Nevertheless, The introduced benzamide group is placed correctly within the ATP-binding site.

Glide predicts a completely different pose with an RMSD of 7.1 Å to the crystal structure. None of the key interactions to the hinge region is performed.



(a) Reference ligand from PDB id 2W1C in grey and fragment in yellow in crystal structure of PDB id 2W1D.

(b) Best poses of Glide and NAOMInext in magenta and grey, respectively.

Figure (6.19) Ligand Pair with PDB ids 2W1D/2W1C annotated with no binding mode change in crystal structure of Aurora A kinase (PDB id 2W1D[315]). a) Reference ligand from PDB id 2W1C[315] in grey and fragment in yellow. b) The best Glide pose (magenta) has an RMSD of 7.1 Å and the best pose of NAOMInext (grey) has an RMSD of 1.5 Å with respect to the reference structure (ligand from PDB id 2W1C[315]).

6.4.1. Start Pose Analysis

Analog to the re-growing evaluation, the influence of the number of start poses is also analyzed for the cross-growing scenario. Here, the usage of different start poses has a much bigger influence on the results as compared to the re-growing analysis (see Figure 6.20). Using the same α value as level of significance ($\alpha = 0.05$ (5 %)), the calculated $p - value$ of the performed Mann-Whitney U-test ($P = 3.2 \times 10^{-3}$) is two orders of magnitude lower as the calculated $p - value$ of the re-growing analysis ($P = 0.35$), i.e. both sets are significantly different. Hence, using more start poses is advantageous over one start pose.

Using 50 start poses improves the results for $\frac{2}{3}$ of the data set and worsens the results for $\frac{1}{3}$. Again, considering only RMSD changes above 0.5 Å, about 37 % of the results improve (see green dots in Figure 6.20) and only 4 % (ten ligand pairs) of the results deteriorate (see red dots in Figure 6.20). In the one start pose scenario NAOMInext failed to generate a result for 52 ligand pairs. After increasing the number of start poses, cases that cannot be handled dropped to 17 only. Thus, the incorporation of different start orientations of the anchor fragment significantly improves the outcome of NAOMInext and is an important component of the sampling algorithm. In the following, several case studies are discussed in more detail.

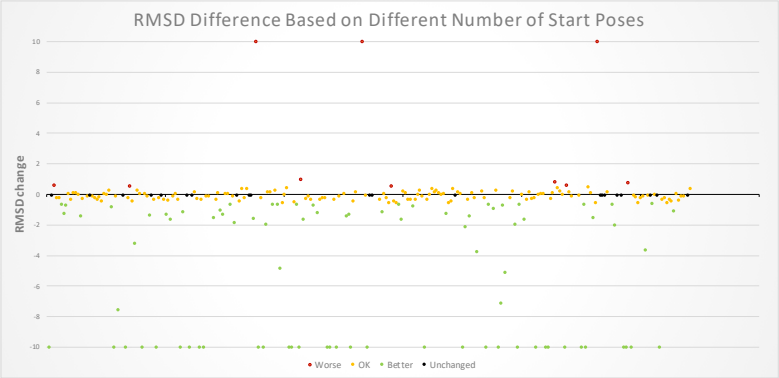


Figure (6.20) NAOMInext cross-growing analysis between one and 50 (default) used start poses. The RMSD difference is plotted color coded. If using more start poses leads to a more than 0.5 Å RMSD increase, the RMSD difference is plotted in red. If the RMSD decreases by more than 0.5 Å the difference is plotted in green. Changes between −0.5 Å and 0.5 Å are treated as acceptable variations and marked in orange. Unchanged RMSD values are marked in black. Cases where no difference could be calculated (no result in one case or the other) the difference value is set to 10 Å or rather −10 Å.

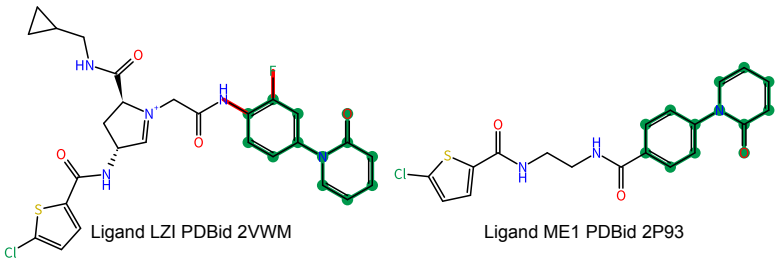
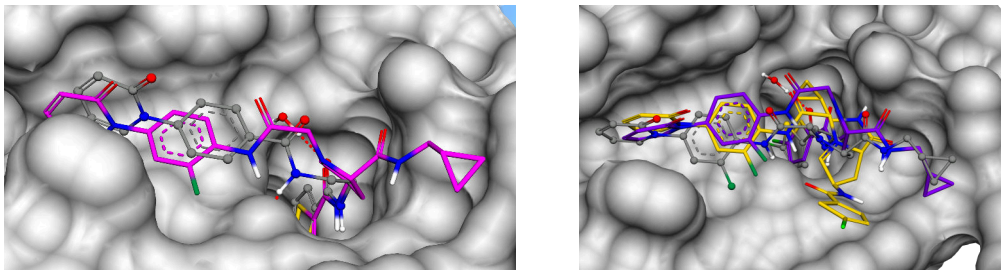


Figure (6.21) Ligand pair (PDB ids 2VWM/2P93) with marked MCS in green.

Factor XA Results for PDB ids 2VWM/2P93

The here discussed ligand pair is crystallized within the protein Factor XA, which is an important target for the treatment and prevention of thrombotic diseases.[316] Figure 6.21 shows both ligands side by side including the MCS derived anchor fragment marked in green. Besides the marked ring system, additionally both ligands share a *2-chloro-5-carboxamide-thiophene* group that is buried deeply inside the S1 pocket of the factor XA binding site (see Figure 6.22a).[316] Due to variation of the central chain (introduction of a *L-prolinamide* group), the *2-chloro-5-carboxamide-thiophene* group is not part of the derived anchor fragment, so this part must be sampled in the current evaluation. Figure 6.22b shows the obtained results using different number of start poses. Using only one start pose, NAOMInext is not able to correctly place the *2-chloro-5-carboxamide-thiophene* group within the S1 pocket (see Figure 6.22b yellow pose



(a) Overlay of related ligand pair from PDB ids 2VWM/2P93

(b) NAOMInext sampling results using different number of start poses.

Figure (6.22) Sampling results for related ligand pair with PDB ids 2VWM/2P93. a) smaller ligand in grey and larger reference ligand in magenta b) using one start pose (yellow) and 50 start poses (purple). Reference structure of larger ligand LZI (from PDB id 2VWM) is depicted in grey.

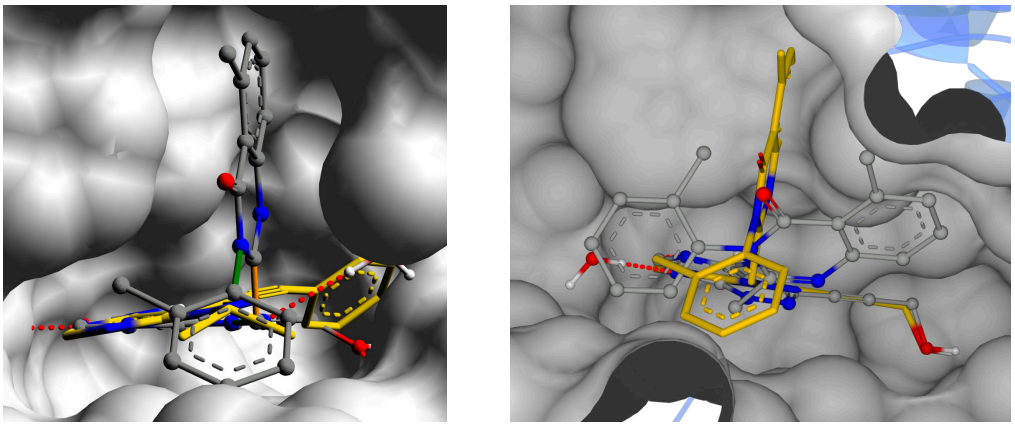
(RMSD: 6.3 Å)). Using a slightly rotated pose (see Figure 6.22b purple pose), the S1 pocket can be targeted correctly achieving an RMSD of 1.4 Å to the crystal structure. In this case study, the MCS derived anchor fragment has a relatively high initial RMSD of 1.4 Å (RMSD is calculated between the binding site superimposed related ligand pairs). Due to the fact that NAOMInext does not generate shifted start poses (see Section 4.2.2) a much lower RMSD is hardly possible. Using an initial pose minimization procedure may overcome this issue and may lead to better results.

Phosphoinositide-3-OH Kinase Results for PDB ids 2WXI/2WXN

Finding new structures to improve selectivity and potency for phosphoinositide-3-OH kinase (PI(3)K) inhibitors lead to the development of the ligand pair crystallized within PDB ids 2WXI and 2WXN.[317] Both ligands are categorized into different classes of PI(3)K inhibitors:

- flat inhibitors (PDB id 2WXN) and
- propeller-shaped p110δ-selective inhibitors.

The last class of inhibitors induce the formation of the specificity pocket (see Figure 6.23a). Hence, using a fragment growing approach (without considering protein flexibility) based on anchor fragments of the first class will not be able to generate results with the correct binding mode due to clashes with the binding site(see Figure 6.23b). Using NAOMInext with just one start pose does lead to a result with an RMSD of 1.6 Å to the crystal structure, but with clashes to the binding site. In cases with less available start poses, the algorithm tries to find



(a) Reference ligand from PDB id 2WXN with smaller ligand from PDB id 2WXI (yellow).

(b) PDB id 2WXN and NAOMInext results for one and 50 start poses depicted in yellow and grey, respectively.

Figure (6.23) Sampling results for related ligand pair with PDB ids 2WXI/2WXN a) larger ligand in grey and smaller ligand in yellow b) NAOMInext results using one start pose (yellow) and 50 start poses (grey). The yellow pose clashes with the protein binding site since protein flexibility is not considered.

a solution even though this would lead to slight clashes. However, using one start pose is for evaluation purposes only.

Using 50 start poses does lead to a valid and clash free pose, but with a larger RMSD of 2.6 Å to the reference structure. This example clarifies the drawback of growing approaches using a fixed protein representation. Incorporating side chain flexibility or using protein ensembles may resolve this specific issue and may lead to a valid pose with a small RMSD to the reference crystal structure. Thus, the user is referred to the command-line mode of NAOMInext providing the possibility to run the experiment on an ensemble of proteins.

6.4.2. Influence of the Number of Start Poses

As already discussed in Section 6.2.2, the number of used start poses has a significant influence on the outcome of NAOMInext. Here, this influence should be analyzed considering the cross-growing results. In Figure 6.24 the results for different numbers of start poses are shown. As in the re-growing case, using more start poses worsens the pose ranking for results below 0.5 Å RMSD. This effect is compensated if the RMSD threshold for success is increased to 1.0 Å. The higher the RMSD threshold for success is chosen, the more start poses can/should be used. In this thesis, poses below 2.0 Å RMSD to the crystal structure are considered as success. Thus, according to this analysis, 75 start

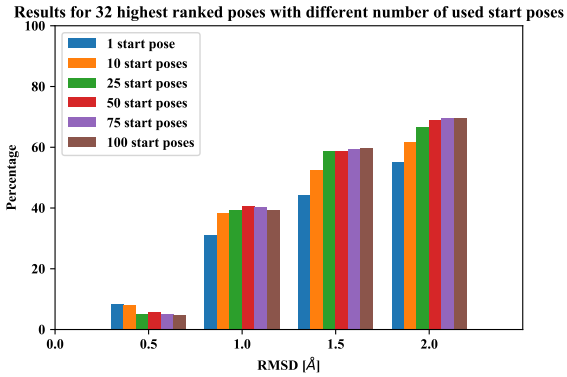


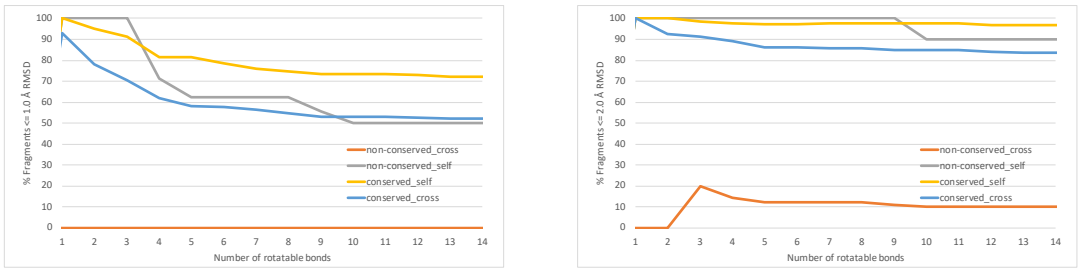
Figure (6.24) Cross-growing results for the 32 best ranked poses using different number of start poses. The RMSD thresholds to the crystal structure are shown as bins on the x-axis. The ratio of correctly predicted poses of the data set from Malhotra and Karanicolas[44] is shown on the y-axis.

poses should be used to gain the best results. However, as a default 50 start poses are used in NAOMInext since the results are not significantly worse, but the runtime is much lower. Depending on the desired level of accuracy, less start poses may be used. As a consequence some targets may not be able to be processed.

6.5. Binding Mode Analysis

According to Malhotra and Karanicolas binding mode changes during chemical elaboration of fragments are unfortunately relatively common.[44] 41 of the 297 investigated ligand pairs (14 %) have been found to change their binding mode. According to their statistical analysis, "Compounds that change binding mode upon elaboration typically have fewer (non-hydrogen) atoms than compounds that retain their binding mode".[44] Using the provided information from Malhotra and Karanicolas, subsets of conserved and non-conserved binding modes are compiled and analyzed separately. Figure 6.25 shows the performance of NAOMInext for fragment growing experiments (re- and cross-growing) using ligand pairs providing just one exit vector. Thus, the conserved binding mode subset incorporates 135 pairs which could be further analyzed and the non-conserved subset incorporates eleven pairs. As a consequence, the results of the non-conserved subset are not statistically significant and should be treated with caution.

In Figure 6.25a the results for both experiments, re- and cross-growing, are shown for the individually compiled subsets. The ratio of a successful growing



(a) Results with ≤ 1 Å RMSD to the crystal structure.

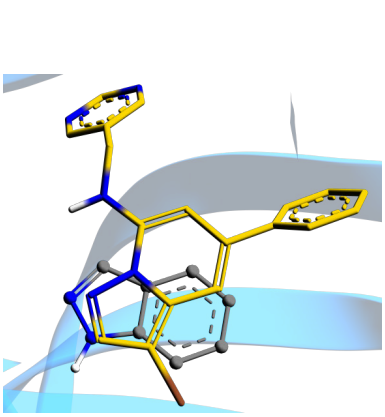
(b) Results with ≤ 2 Å RMSD to the crystal structure.

Figure (6.25) NAOMInext performance (ratio of a successful growing approach, minimum of 32 highest ranked poses) in relation to ligand flexibility (number of rotatable bonds) is plotted. The ratio of a successful growing approach for each individual set is plotted on the y-axis, the fragment flexibility on the x-axis. The conserved binding mode subset and the non-conserved subset contain 135 and eleven pairs, respectively.

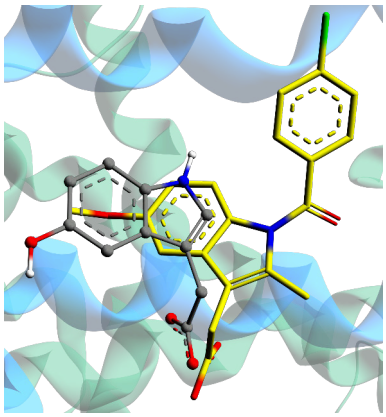
approach (RMSD ≤ 1 Å to the crystal structure, minimum of 32 highest ranked poses) is plotted with respect to the fragments flexibility, i.e. number of rotatable bonds. Considering the conserved subset (re-growing mode), NAOMInext is able to predict over 80 % of the set correctly for up to five rotatable bonds (yellow curve in Figure 6.25a). In comparison to this, the cross-growing results are on average 20 % worse. Considering a somehow less stringent threshold (see Figure 6.25b RMSD ≤ 2 Å to the crystal structure, minimum of 32 highest ranked poses), NAOMInext is able to predict a valid pose for over 80 % of the conserved subset (cross-growing mode). These results are achieved independently from the number of rotatable bonds. However, the number of examples of seven rotatable bonds and more is less than ten, i.e. statistically not meaningful (see Figure 6.29c). The same analysis is performed for the non-conserved subset. NAOMInext is not able to predict binding mode changes (figure 6.25 orange curve). This is not surprising since the binding mode of most ligand pairs differs significantly. Selected examples of related ligand pairs (extracted from the data set of Malhotra and Karanicolas[44]) with annotated binding mode changes are shown in Figure 6.26.

6.5.1. Cyclin-Dependent Kinase 2 Inhibitors and SBDD

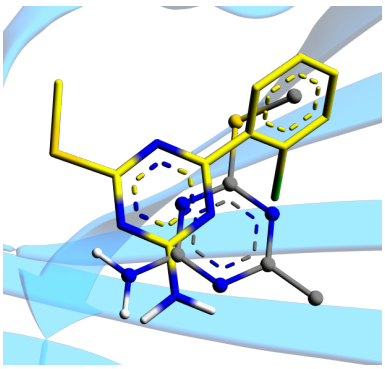
In Figure 6.26a, the indazole fragment, crystallized within cyclin-dependent kinase 2 (CDK2) binds to the ATP binding site performing key interactions to backbone residues at the hinge region (residues GLU81 and leucine (LEU)83).[319] The larger elaborated ligand occupies another position in the protein binding



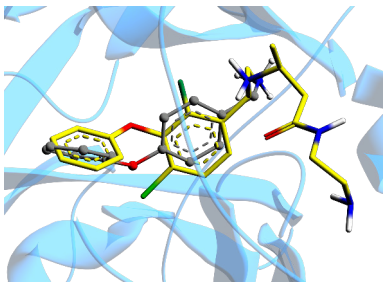
(a) PDB ids 2R3K[318] and 2VTA[319]. Pocket RMSD: 0.8 Å (all atoms); 0.16 Å (backbone)



(b) PDB ids 3ADS and 3ADT[320]. Pocket RMSD: 0.58 Å (all atoms); 0.16 Å (backbone)



(c) PDB ids 3B27[321] and 2WI3[322]. Pocket RMSD: 2.66 Å (all atoms); 2.51 Å (backbone)

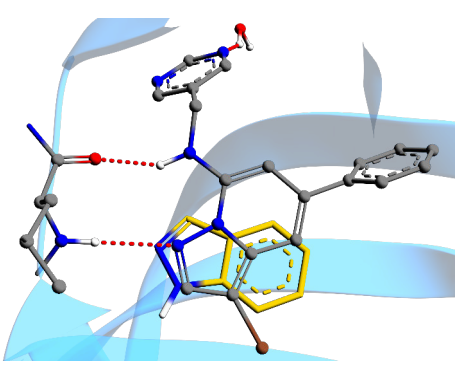


(d) PDB ids 4B74 and 4B6F[323]. Pocket RMSD: 0.87 Å (all atoms); 0.34 Å (backbone)

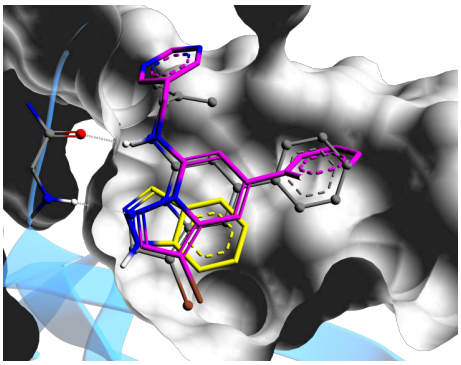
Figure (6.26) Related ligand-pair examples annotated with binding mode change extracted from the data set of Malhotra and Karanicolas.[44] The smaller fragment and the larger ligand are shown in grey and yellow, respectively. Pocket RMSD values are obtained after binding site alignment of the respective protein structures using the tool SIENA.[271]

site, however, still performing the former key interactions to the hinge region (see Figure 6.27a) Moreover, an additional interaction is formed that may be accountable for the binding mode change.

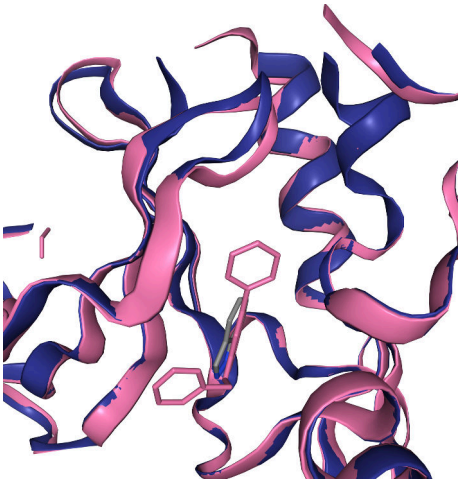
The binding pocket RMSD (all atoms) between both protein structures (of each ligand) is just 0.8 Å (see Figure 6.27c), i.e. there is most likely no significant



(a) Larger ligand interactions



(b) Generated start pose.



(c) Binding site superposition (top view).

Figure (6.27) Binding mode analysis of CDK2 using PDB ids 2R3K[318] and 2VTA[319]. a) Larger ligand from PDB id 2R3K[318] performing interactions to the hinge region. Fragment from PDB id 2VTA[319] overlayed in yellow. b) Generated start pose (grey) with larger ligand (magenta) and anchor fragment (yellow). Start pose generation is solely based on the anchor fragment structure, in this case, a topological MCS of the indazole structure c) PDB ids 2R3K[318] and 2VTA[319] superposed using SIENA[271] from ProteinsPlus[3] server.[272].

protein flexibility. As can be seen in Figure 6.27b, NAOMInext is able to generate a start pose matching the corresponding reference ligand position with an RMSD of 0.6 Å (substructure RMSD of the MCS derived anchor fragment only). However, sampling the attached fragment and subsequent scoring does not lead to a valid result with the correct binding mode. In this example, the binding site is very narrow and even a slightly suboptimal start pose leads to clashes during the torsion driven sampling process. One possibility to solve this issue would

be an initial start pose minimization step prior to the sampling to achieve a better start pose orientation.

The elaborated ligand retains key interactions of the initial binding fragment although in a different binding mode (see Figure 6.27a). This circumstance was also discovered during the fragment-based study performed by Wyatt and co-workers.[319]

In this specific case, the anchor fragment (core) is not only extended by additional substituents, moreover, the core is altered itself. Hence, the introduction of a different hydrogen bond donor (secondary amine) and removal from the donor group in the core moiety (change from 1H-indazole to pyrazolo[1,5-a]pyridin) lead to an expected change of the binding mode (see Figure 6.27a). However, the deprotonated nitrogen atom of the pyrazole still performs the same interaction to the backbone nitrogen of residue LEU83 of the hinge region.

6.5.2. PPARγ

The next ligand pair is crystallized within peroxisome proliferator-activated receptor γ (PParγ) and pictured in Figure 6.26b. The ligand pair is subject to a significant binding mode change during chemical elaboration, which can not be approached by any generated start pose since a translation of the anchor fragment would be necessary (NAOMInext only performs rotations around the anchor centroid see Section 4.2.2). The binding site of PParγ is very open spaced and derivatives (ligands from PDB ids 3ADV, 3ADW, and 3ADU[320]) of the ligand, complexed in PDB id 3ADT[320], nonspecifically bind over the complete binding site (data not shown). Hence, this protein target may not be the best choice for FBDD and traditional docking or HTS may be beneficial.

6.5.3. Does the Binding Mode Change Correlate with the Pocket RMSD?

On average, related ligand pairs with annotated binding mode change seem to have a higher pocket RMSD (see Figure 6.28). Statistically, there is no difference between both sets (p value = 1.0, determined according to Section 5.4). The only statistically meaningful difference found between both sets is the number of heavy atoms of the smaller ligand (according to Malhotra and Karanicolas).[44] Typically, compounds with fewer heavy atoms change their binding mode upon elaboration. Anyway, there is no threshold to be used to filter out fragments that most probably would change their binding mode. To be on the safe side, there is no other way as to determine the binding mode of potential new lead like ligands experimentally.

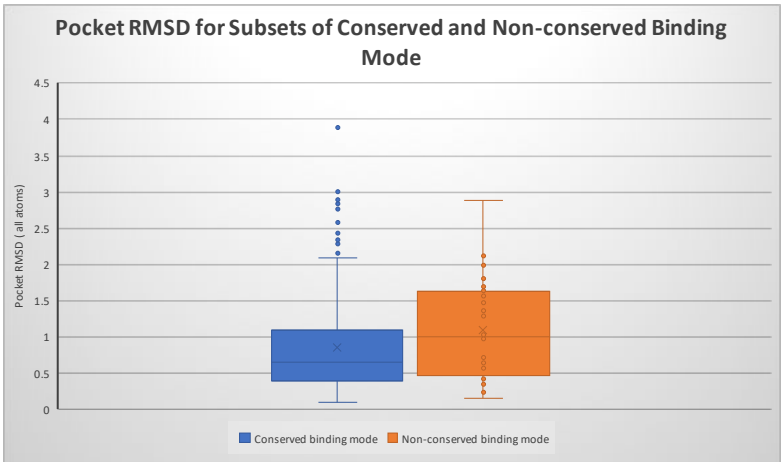
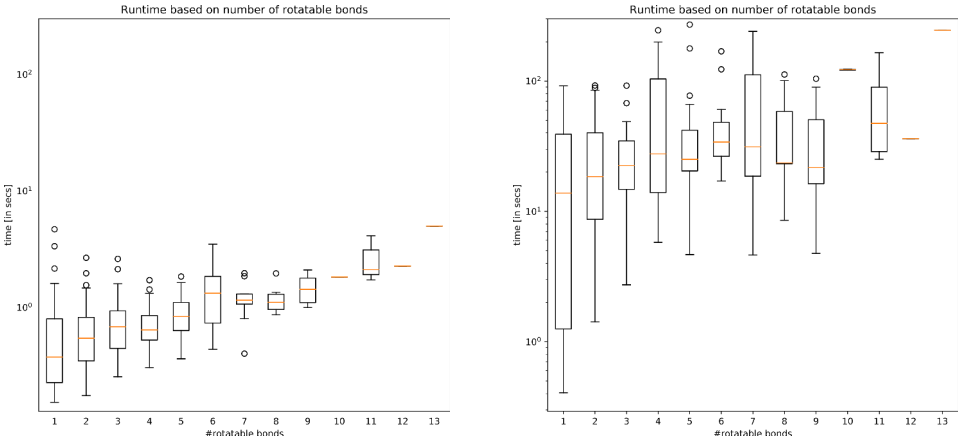


Figure (6.28) Pocket RMSD statistics for conserved and non-conserved subsets. Pocket RMSD values of the non-conserved binding mode set seem to be higher on average. However, the difference between both sets is not statistically significant (p value = 1.0)

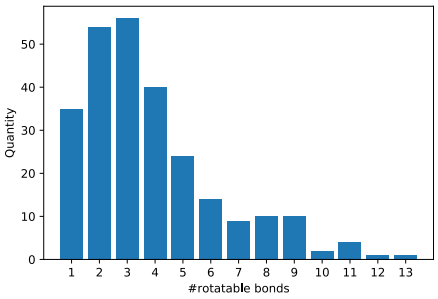
6.6. Runtime

To provide an interactive user experience during the usage of NAOMInext, great value was placed on runtime, i.e. runtime of the conformational sampling approach. Since the implemented sampling procedure is based on two independent parts, namely the torsion driven sampling and the start pose sampling, runtime results are analyzed independently. In Figure 6.29a the results for the torsion driven sampling (one start pose only) are shown, binned by rotatable bond count. For up to five rotatable bonds the needed runtime is below 1 s, which will cover most of the attached BBs if they obey the Ro3. See Figure 6.30 for an example of a rotatable bond analysis of two different vendor BB catalogues downloaded from ZINC.[120], [223], [224] For BBs with more than five rotatable bonds the needed runtime slightly increases, but is still below 10 s for up to 13 rotatable bonds. It should be mentioned that the runtime results for fragments with more than six rotatable bonds should be treated with caution, since the low amount of data does not allow for a statistically sound statement (see Figure 6.29c). Figure 6.29b shows the needed runtime for a usual NAOMInext experiment using at most 50 start poses (default upper bound), which are needed to account for strain during fragment growing. As expected, compared to the usage of only one start pose, the runtime increases to a median runtime of about 25 s for fragments with up to nine rotatable bonds. The only sub-linear increase in runtime is achieved by using a heuristic approach which adapts to existent states and constraints (see Section 4.2.4). Hence, the implemented approach adapts to the available number of start poses by limiting the



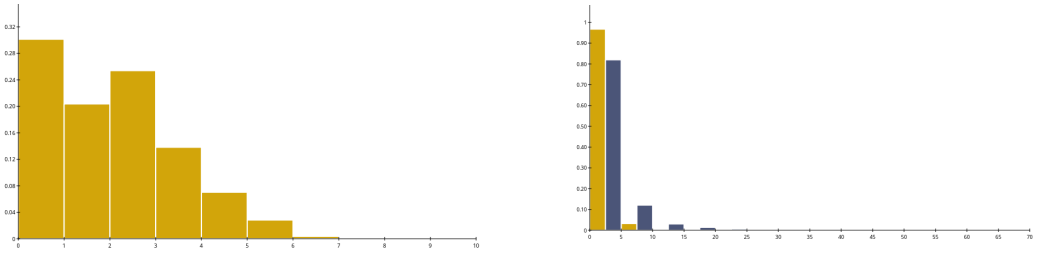
(a) Runtime per rotatable bond for one start pose

(b) Runtime per rotatable bond for upper bound of 50 start poses



(c) Quantity of fragments grouped by rotatable bond count

Figure (6.29) Sampling runtime analysis of NAOMInext. a) Runtime for one start pose to analyze the needed runtime of the sampling approach. b) Runtime for upper bound of 50 start poses to analyze the ability of the heuristic approach to ensure short run times. c) Quantity of fragments grouped by rotatable bond count.



(a) BBs from ChemBridge data set binned by rotatable bond count

(b) BBs from ChemBridge and Sigma Aldrich data set binned by rotatable bond count

Figure (6.30) Rotatable bond analysis of vendor BB catalogues. a) BBs of ChemBridge data set binned by rotatable bond count, one bond per bin. b) BBs of ChemBridge (mustard yellow) and Sigma Aldrich (blue) data set combined and normalized.[120] Over 90 % of the BBs have five rotatable bonds at most. The Sigma Aldrich BB catalogue is by far more complex and comprises larger BBs (over 80 % of the BBs below six rotatable bonds).

number of eligible torsion angles used for the torsion driven sampling of the attached BB. Thus, optimizing the tradeoff between runtime and accuracy.

6.7. Reactions

The implemented reaction workflow is validated using a unit test framework based on Qt[244] and the described approach in Section 5.5. The coverage of the bioactivity-relevant chemical space has already been described earlier.[46] Besides the unit testing procedure, the reaction workflow is tested in combination with the subsequent conformational sampling approach. Since the required test data need to meet several requirements, an automated large scale analysis is not possible. Thus, a few hand curated examples are tested and discussed.

6.7.1. Reaction Unit Tests

All incorporated reactions are tested within a unit testing framework of Qt[244]. Unit tests are used to test individual modules of a software framework or software tool. The publicly available reaction set from Hartenfeller *et al.* incorporates educt smiles (reactants) for each reaction (see Supporting Information of Hartenfeller *et al.* [43] and appendix E). Thus, each reaction is tested using the provided reactants. As a result:

- reactant matching,
- reaction execution, and

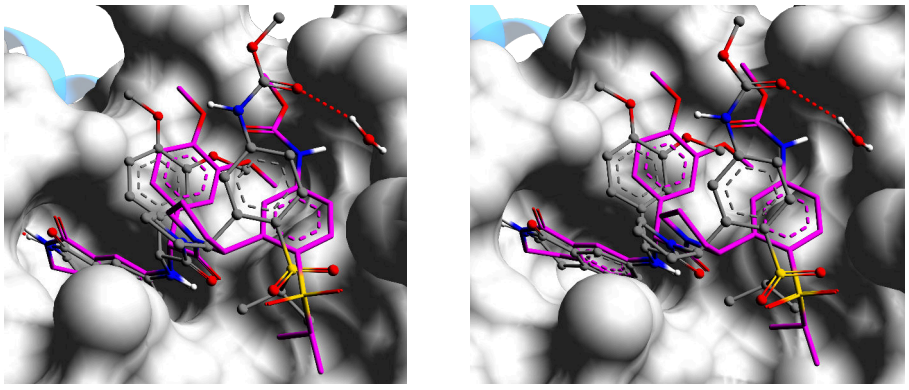
- result validation

is performed within the test. Invalid reactions and wrong reaction execution is detected at an early step, because the product SMARTS of a given SMIRKS is not only used to derive to-be-formed covalent bonds and required atom state adaptations, but also to match the resulting product to ensure correctness of the reaction. All described reactions from Hartenfeller and co-workers can be performed. However, small adaptations have been performed (see Section 4.3.2 and appendix E).

6.7.2. Reaction Results for Factor VIIa using PDB ids 4X8T/4X8V

Cheney and co-workers obtained this ligand pair using a fragment-based screening approach to achieve a new inhibitor for Factor VIIa.[324] One of the derived screening hits, a lactam derivative, was used as starting point and elaborated into a new potent inhibitor. In this case, the larger ligand can be obtained by applying a *Buchwald-Hartwig* reaction to the initial hit. Thus, performing an *in silico* chemical reaction within the protein binding site. The appropriate BB is generated based on the reference structure to test the reaction implementation of NAOMInext. See Figure 5.4 in Section 5.5 for an example of the BB extraction. This case study is of course artificially created. The original ligand from PDB id 4X8V[324] was synthesized differently performing several reaction steps including a multi component reaction (MCR) (see Supporting Information from Cheney *et al.*).[324] However, the case study is able to exemplify the reaction workflow in combination with subsequent conformational sampling incorporated into NAOMInext.

Figure 6.31 shows the obtained results for both approaches performing the experiment described in Section 5.5. The result of Figure 6.31a is obtained from the artificial growing approach and the result of Figure 6.31b using a built in reaction rule from the published Reaction SMARTS set of Hartenfeller and co-workers.[43] Both results are nearly identical except for the pyrrolidine ring conformation (see Figure 6.31a and 6.31b in the middle which has a slightly different conformation. This difference occurs because NAOMInext does not perform a canonization procedure for substructures (due to implementation issues), thus, leading to the generation of a different ring conformation which depends on the atom order of the given input. Nevertheless, the difference is not dramatic since the pyrrolidine ring has no further substituents and the same binding mode is obtained. In case the pyrrolidine ring had further substituents, the effect could be more dramatic but should be absorbed due to the smooth sampling procedure.



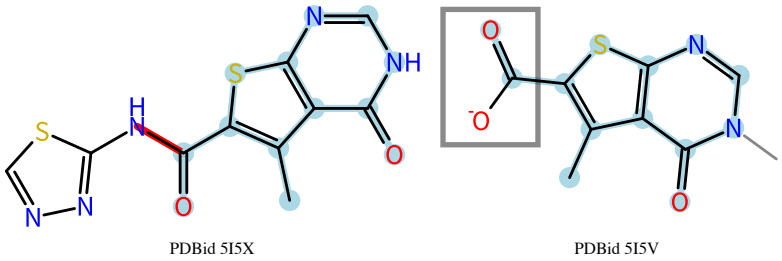
(a) Artifical growing result using extracted lactam derivative as anchor.

(b) Sampling result using a *Buchwald-Hartwig* reaction and ligand from PDBid 4X8V as anchor.

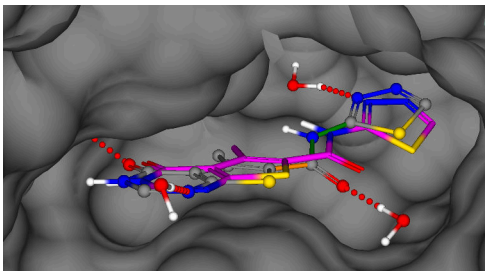
Figure (6.31) Comparison of sampling results using artificial building block and genuine *in silico Buchwald-Hartwig* reaction mechanism within protein structure of factor VIIa (PDB id 4X8T[324]). a) Best artificial growing pose in grey achieved an RMSD of 1.65 Å with respect to the reference structure. b) Best pose in grey achieved an RMSD of 1.8 Å with respect to the reference structure. The pose seems to be nearly identical except for the conformation of the pyrrolidine ring.

6.7.3. Reaction Results for Mitochondrial Branched Chain Aminotransferase using PDB ids 5I5V/5I5X

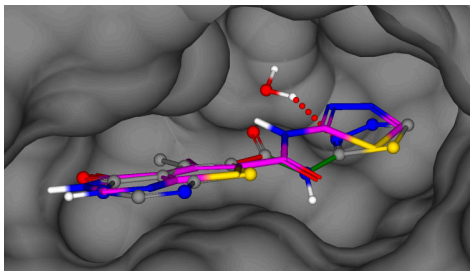
Performed fragment screening and subsequent amide reaction lead to this ligand pair crystallized within mitochondrial branched chain aminotransferase (BCATm).[325] To be able to perform RMSD calculations of the result poses, the substituted methyl at ligand 68A (PDBid 5I5V) has been removed from the ring system (for RMSD calculation only, see Figure 6.32a grey bond). The reaction center is depicted in Figure 6.32a (grey box). Though the carboxyl group is represented with explicit double and single bonds, it has a delocalized character. Hence, different explicit representations (protonation states and tautomers) are possible and either of these bonds may have a double bond or single bond character.[79] The implemented SMARTS matching procedure (see Section 4.3.2) makes use of this circumstance and performs a correct reaction (in this case a *Schotten-Baumann amide* reaction). Since the carboxyl group is the reaction center, the subsequent substructure determination (needed for the conformational sampling procedure) includes the bond between the carboxyl group and the ring system within the torsion driven sampling procedure. Both, the artificial growing procedure and the reaction implementation lead to identical results (see Figure 6.32b). The artificial growing procedure strongly depends on the given user input. In case a different tautomeric form is used, a different result



(a) Ligand pair with PDB ids 5I5V/5I5X and derived MCS marked in light blue.



(b) Sampling result using a *Schotten-Baumann amide* reaction and modified ligand from PDBid 5I5V as anchor.



(c) Sampling result using different artificial tautomer.

Figure (6.32) Comparison of sampling results using artificial building block and genuine *in silico Schotten-Baumann amide* reaction mechanism within protein structure of mitochondrial branched chain aminotransferase (PDB id 5I5V[325]). a) Ligand pair with PDBids 5I5V/5I5X[325] and derived MCS marked in light blue. The extension vector (exo-bond) is marked in red. The reaction center is depicted in a grey box. The removed methyl group is marked as grey bond (smaller ligand with PDB id 5I5V[325]). b) PDB id 5I5V[325] and best artificial growing and reaction pose (both in grey) achieved an identical RMSD to the reference structure (0.59 Å). c) Best artificial growing pose using wrong extension vector in grey achieved 1.28 Å RMSD to the reference structure.

pose is obtained (see Figure 6.32c). In this setting, the sampling algorithm takes the user input for granted and does not extend the to-be-sampled substructure. In conclusion, if using the artificial (i.e. manual) growing procedure, the user is responsible for the correct tautomer usage. Thus, using the implemented reaction procedure is highly recommended, since tautomer and protonation state mapping issues are implicitly solved by NAOMInext.

7. NAOMInext

In recent years, more and more methods have been published providing medicinal chemists with the needed set of tools to develop new drug candidates. Especially in the field of FBDD new tools, which serve as idea generators, have been developed recently [55], [57], [58]. This fact shows, that there still exists the need for new developments since not all user requirements are solved yet.

As part of this thesis a software application called NAOMInext is developed to support medicinal chemists, during H2L optimization in the field of FBDD. NAOMInext, especially its conformational sampling engine, is validated using a large-scale data set. Incorporated synthetic reaction rules facilitates subsequent chemical synthesis of the predicted compounds. The modern and simplistic interactive user interface and the performed large scale validation distinguishes NAOMInext from other currently available programs from academia. Moreover, NAOMInext combines several aspects of the FBDD design cycle within one condensed workflow. This eliminates the tedious manual application of various tools.

This chapter describes the requirements that must be met by an interactive program in the context of H2L optimization within FBDD. Interactive usage and efficiency are important aspects of the tool, since users should be able to incorporate their knowledge into the workflow by means of an easy-to-use user interface. Moreover, during implementation hardware limitations of current desktop computers have been incorporated into the software design. The requirements and the implementation of the different features is described in the following. Moreover, a short overview of the software architecture and the underlying implemented software libraries is given.

7.1. Requirements

This section describes the different requirements that need to be solved by NAOMINext to provide users with a useful tool in the field of FBDD.

Interactivity An important aspect of NAOMInext is interactivity. Users should be able to incorporate their knowledge (additional constraints) into the design process of new compounds. Hence, an easy-to-use interface is a requirement that must be met.

Hardware limitations NAOMInext should be used by medicinal chemists to generate new lead series including potential synthesis pathways. Generating huge compound libraries may exceed the main memory of current desktop computers. This limitation should be considered during the software design.

Parallelism Since interactivity is an important prerequisite of NAOMInext, parallelism is inevitable during the design of computationally demanding calculations. Thus, different threads are used to asynchronously perform CPU-intensive calculations like conformational sampling.

Ease of use Complicated user interfaces or installation obstacles prevent users from using useful software. Non-expert users should not be overwhelmed from a wide variety of possible settings and tool functionality. Hence, easy installation and a clean simplistic interface should be considered during the software design.

7.2. Software Architecture

NAOMInext is based on the NAOMI framework, which is implemented using *C++*. Different functionality is divided in individual (independent) libraries. Some libraries may depend on functionality of other libraries (see Figure 7.3).

Mainly two different libraries have been developed during this thesis: the *ConstraintSampling* and the *Reactions* library (see Figure 7.3). Many other functional implementations have been integrated into already existing libraries, e.g. *Coordinates3d* and *TorsionLib*. The *ConstraintSampling* library depends on many basic NAOMI libraries like the *Molecule* library (implicit dependency). However, the main functionality is derived from the *Coordinates3d* library, which received a significant number of enhancements during this thesis. The *ConstraintSampling* provides the needed functionality to perform conformational sampling within protein binding sites. Needed pre-processing steps are implicitly performed by underlying library functionality.

The *Reactions* library is a new implementation as well. The most important dependencies are the *SMARTS Matching* library and the *Molecule* library. The *SMARTS Matching* library is used to interpret the synthetic reaction rules (SMIRKS) and to perform the substructure matching, which is needed to obtain a pairwise atom mapping between reactant and reaction products. The *Molecule* library provides basic functionality to alter molecular graphs (single bond formation). However, several extensions have been incorporated into the *Molecule* library to perform ring closure reactions and more sophisticated modifications as well. The *Reactions* library provides a clean interface and performs all necessary steps to perform a reaction between two molecules or transformations for a

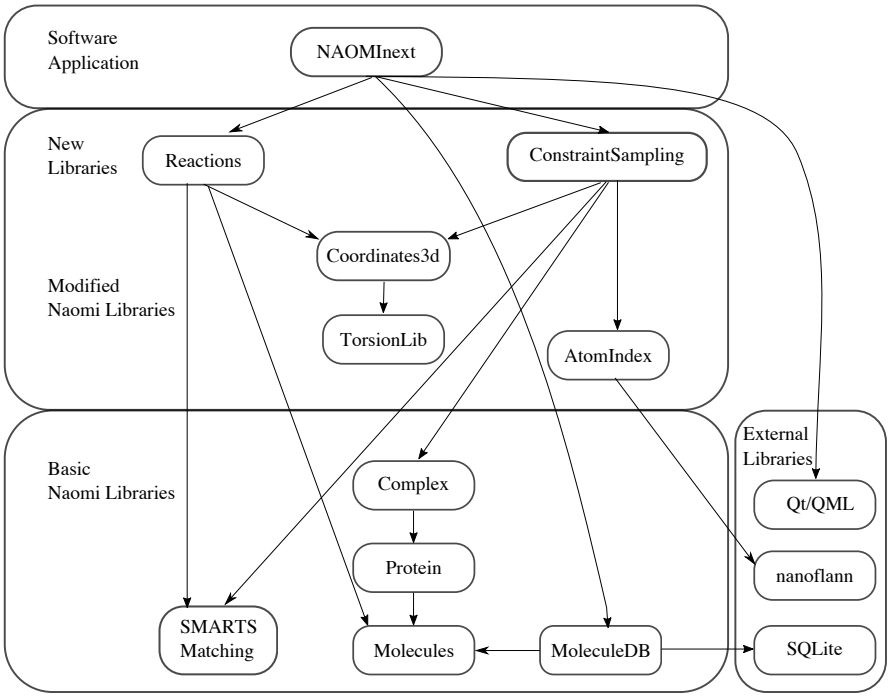


Figure (7.1) Software architecture dependencies of NAOMInext. For simplification, only relevant libraries and dependencies are shown.

single molecule. The interface function only needs two molecules and a SMIRKS or a reaction name to perform already incorporated reactions from Hartenfeller and co-workers.[43] SMIRKS parsing, reaction execution, and reaction result validation is completely performed using library internal functionality.

7.3. Implicitly solved Requirements

During the development of NAOMInext, great emphasis was put on correctness, efficiency, and ease of use. The basis of NAOMInext, the NAOMI framework, is a validated cheminformatics toolkit based on a consistent chemical model.[75] The main cause to develop a new tool for FBDD is based on NAOMInext's intended use. Because it should be used in an interactive and iterative way, a clean and simplistic user interface is of the utmost importance. Moreover, the combination of synthetically feasible fragment growing and virtual HTS in an integrated approach allows for a more efficient search space coverage. NAOMInext does most of the time consuming pre-processing without user interaction based on the following published algorithms:

- reading and parsing input files in different file formats[75], [237]

- correction of invalid input molecules or removal[75]
- buildup of macro molecules based on 3D-coordinate files and small molecule (ligand) extraction[77]
- calculating hydrogen bond networks and determination of tautomer and protonation states[79], [80], [246]
- calculating missing 3D-coordinates for loaded building blocks[197], [255]
- implicit duplicate detection based on a unique molecule description[75], [252]

Using NAOMInext spares a lot of pre-processing time and tedious manual application of different tools and file format conversion. The usage is further simplified through implicit constraints, e.g. tethering the anchor fragment during conformational space exploration and primary target constraints.

7.4. Graphical User Interface

The GUI is a key element during the work of this thesis. The main interface is implemented using QtQuick within the Qt Meta-object Language (QML) provided by the Qt-Framework and allows for efficient user interface design and platform interoperability.[244] Visualization of proteins and ligands is performed using the 3D visualization library developed by BioSolveIT[1]. The design is guided by:

- ease of use
- interactivity and
- fluidity

7.4.1. Ease of use

The first challenge prior to using and testing the software is the installation process. Thus, NAOMInext is equipped with an installer, provided by Qt[244] that allows the user to install the software as any other software on the system. Furthermore, NAOMInext is provided as a compressed archive which only needs to be unpacked: no installation is required. System dependencies are reduced to a limit, thus, the software is compatible to most operating systems and versions. NAOMInext is supported on Windows, Linux and macOS. The basis of NAOMInext is NAOMI, a well tested molecular framework.[75]–[77], [79], [238], [239] All standard chemical file formats are supported, hence, no additional conversion tools are required.

[1] www.biosolveit.com

The interface is designed in a clean and simplistic way. The GUI is divided in three main parts. The `Tool Bar`, the `Ligand View`, and the `3D-View`. An additional logging section (`Output log` in Figure 7.2) is provided to support the user with important information. The `Tool Bar` provides load and clean up functionalities as well as several buttons to manipulate the appearance of the loaded data in the `3D-View` (see Figure 7.2). Further settings or filters may be set in the appropriate views (see NAOMInext user guide in Section B).

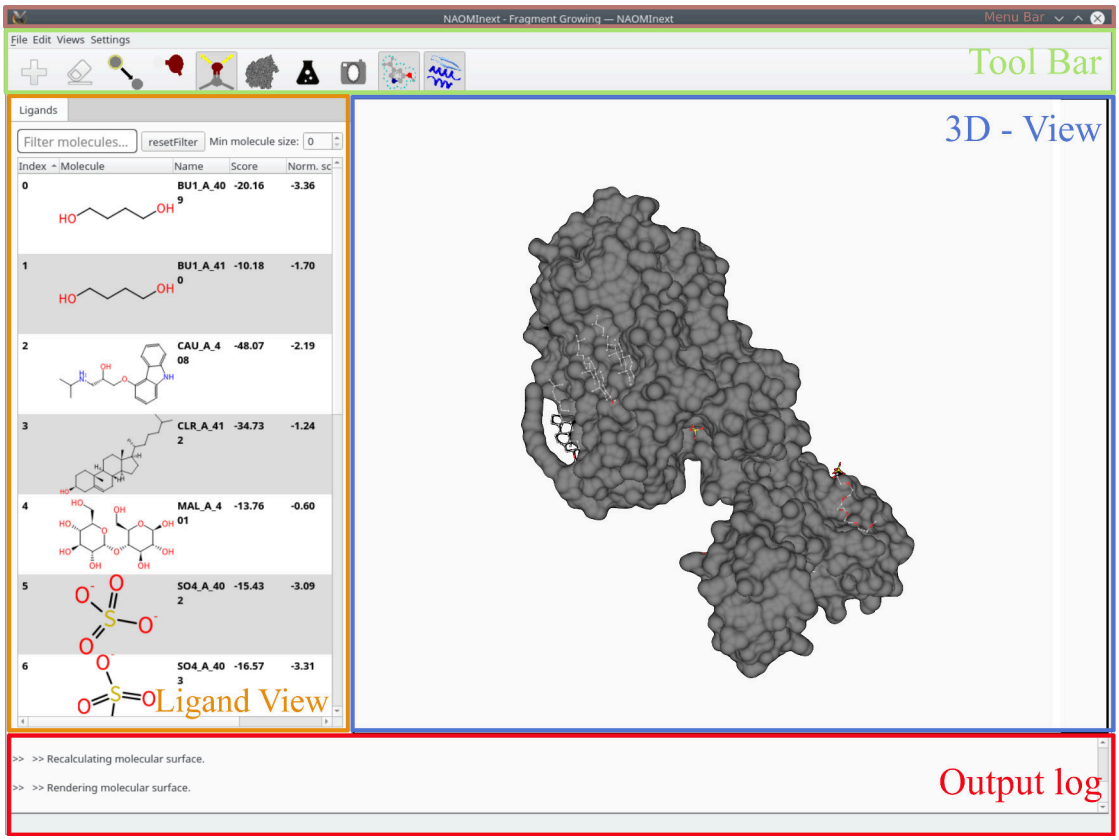


Figure (7.2) NAOMInext main view after loading a protein (PDB id: 2RH1[222]). The main window is divided into different independent parts: 3D-view, tool bar, ligand view, and output log.

3D-View

The `3D-View` visualizes the loaded protein including small molecules, fragments, metals, and water molecules. Moreover, it provides interactive manipulation of the protein orientation and allows for visual inspection of important protein-ligand interactions. Besides the inspection of the surrounding of the ligand, it

also facilitates the identification of useful constraints.

Fragment Growing

Fragment growing can be performed using traditional docking tools (e.g. AutoDock[326], FlexX[125], Gold[204], Glide[228], [270]) combined with additional constraints to maintain the binding mode of the anchor fragment. This requires profound knowledge of the used tool and is the most time consuming step. Hence, the most important implicit constraint in NAOMInext is the tethering of the initial crystallized anchor fragment. Thus, no additional preprocessing steps are necessary. Interactively, growing vectors can be defined and performed with just one click. Moreover, synthetically accessible fragment growing can be performed just as easy. NAOMInext already incorporates published robust organic reaction rules from Hartenfeller *et al.* [43] These rules can be used for fragment growing and available reaction centers can be investigated interactively (see Figure B.6). User provided synthetic reaction rules can be provided and are checked using an internal reaction validation procedure.

Providing User-Defined Constraints

User-defined constraints may be defined to either limit the conformational space, e.g. interactively define desired interactions to the protein, or to limit the chemical space. Distance and interaction constraints can be set by clicking on the appropriate atoms in the `3D-View`. To influence the chemical space new reaction rules may be added or only a subset of the available reaction rules can be used. All these constraints may be defined without much effort. Either through the GUI or by editing a text file.

7.5. Memory Management

Implicit target focused library design may lead, depending on the used target, reaction set, and BBs, to a large amount of results. These results may not be stored solely in the main memory (RAM) of the desktop computer. Thus, an SQL based DB to store the result molecules including additional poses is used (see Section 4.1.3). The DB file is stored on the hard drive. Since access to the hard drive is significantly slower, a cache is used to keep up the performance of the tool. Qt provides a cache class *QCache* which provides all needed functionality.[244] Each ligand which is stored in the DB is added to the cache which is initialized with a fixed size. The used size is just an estimate to allow caching of about 1000 molecules including additional poses. If the used size is

exceeded, the least used element is automatically removed from the cache but still available in the DB and available on user request. Thus, an efficient caching procedure is provided without user intervention.

7.5.1. Databases

The NAOMI MoleculeDB and PropertyDB[252] are used to store the input molecules, results, and BBs as well. Due to possible memory issues during molecule enumeration, the results are stored inside a SQL database file. For example, BB files are mostly not that large, but are stored in a SQL DB file as well. This has the advantage, that if a BB file without 3D-coordinates is loaded, 3D-coordinates are generated on the fly and stored in the database file alongside the molecule information. Hence, the 3D-coordinate generation step needs to be performed only once, and the BB database can be saved and loaded instantaneously on subsequent runs. Moreover, the BB library may be based on different input files and the used DB ensures a duplicate free insertion process (see Section 4.1.3).

7.6. Parallelism

NAOMInext is designed to perform synthetically feasible fragment growing within a protein binding site. Thus, facilitating implicit target focused library design. Since the intended use of the software is thought to be interactively, each automated growing cycle may be investigated by the user. Hence, each subsequent step, i.e. design decision, may be user defined. To ensure interactive usage of the GUI most of the processes (at least the most resource intensive ones) are performed in different asynchronous threads to keep the main thread (GUI) responsive.

Figure 7.3 shows the most important threads in a sequence diagram. The central component is the user interacting with the main (GUI) thread. The *ComplexLoader* thread is used to load protein complexes but also additional input data, e.g. input molecules, building blocks, and fragment DBs. The input file location is send from the main thread to the loader thread. Parsing the file and construction of the input molecules, i.e. conversion into internal data structures is performed asynchronously.

If the user initiates the fragment growing procedure, either by provided reaction rules or by simple single bond formation, a *Worker Thread* is created. This thread uses the input molecule and the loaded building block DB to perform the requested extensions. Database access is performed sequentially and single threaded. Readily assembled molecules are then send to the *Sampling*

Thread Manager where each molecule is prepared for conformational sampling before it is send to an individual *Sampling Thread*. The number of parallel threads is derived from the number of available cores (default) and handled by the *QThreadPool* of Qt.[244]. Thus, each molecule is sampled individually in its own thread or waits within the *QThreadPool* for the next free slot. If the conformational sampling is finished, the thread is reused for further samplings or destroyed. The *Sampling Thread Manager* collects the results until a user defined threshold is reached. Thus, preventing overload of the *Worker Thread* caused by continuous notifications of finished sampling results. The *Worker Thread* redirects the input to the main thread in defined intervals for storage and visualization.

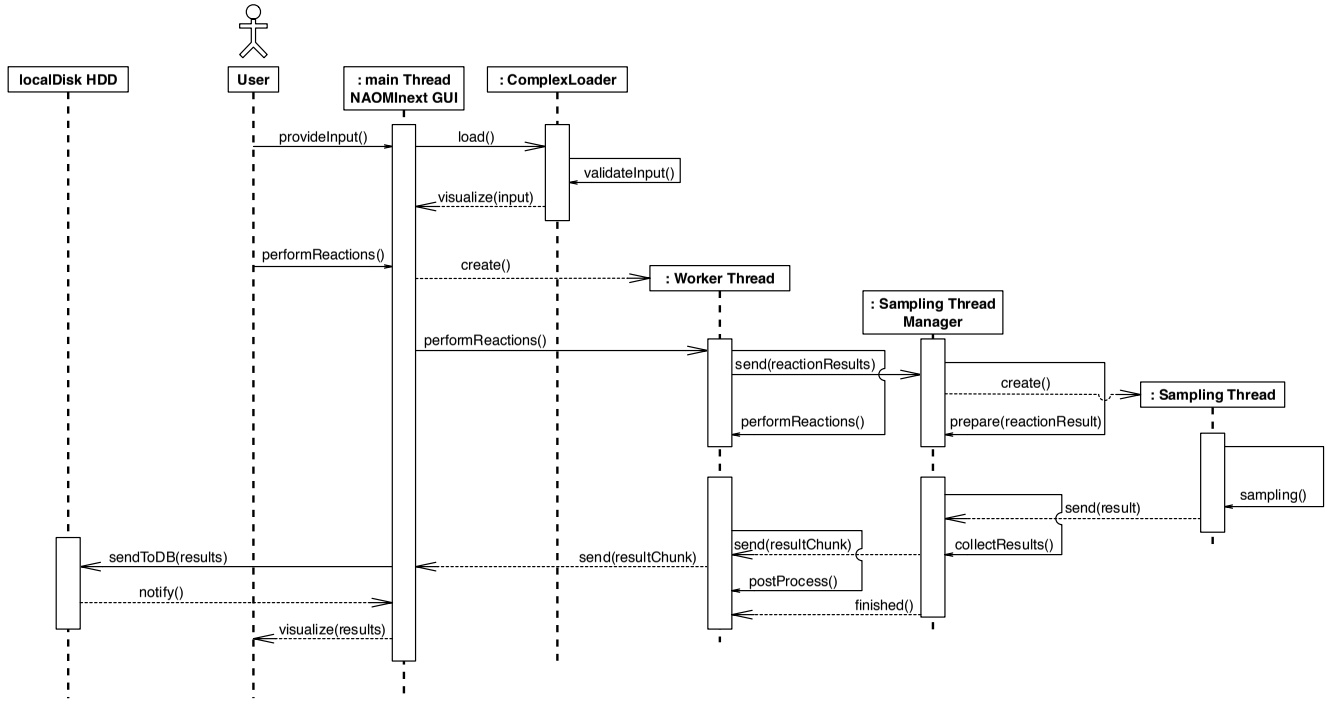


Figure (7.3) NAOMInext Sequence Diagram to exemplify the used threading to keep the GUI responsive. The user only interacts with the main (GUI) thread. Computationally demanding calculations are performed asynchronously in different threads: the *ComplexLoader* and the *Worker Thread*. The *Worker Thread* starts an additional thread, the *Sampling Thread Manager*, which performs the conformational sampling of newly generated compounds in individual threads. The main purpose of the *Sampling Thread Manager* is to manage all the started conformational sampling threads and to collect the individual results. Result packages are send to the *Worker Thread* in defined intervals for post processing, redirection to the main thread, and subsequent storage. Storage on a secondary memory (local HDD) limits main memory consumption.

8. Conclusion

The implicit combination of structure-based synthetic feasible fragment growing, thus, generating a target-based compound library, which is screened on-the-fly, enables medicinal chemists to focus on the most promising candidates. Consequently, the hit rate is increased and at the same time costs reduced.[8]

The intuitive usage of scientific software and thus, providing medicinal chemists with computational support at an early stage of their projects may be a key factor in generating synthetically accessible lead compounds in a shorter period of time. The discussed results, for example the use case in Section 6.4 (depicted in Figure 6.19), show that NAOMInext is able to reproduce the binding mode of a larger related/elaborated ligand based on a small co-crystallized fragment. Based on this findings, the generation of a lead series, starting from the same fragment, is also in the scope of NAOMInext. Hence, *in situ* local chemical space coverage using pre-defined reaction rules is now possible without much effort. The example from Section 6.4 is based on a genuine FBDD study providing additional structural information of potential lead compounds[315], which NAOMInext is able to predict as well (results not shown). Hence, using NAOMInext in a H2L optimization study may provide fast and reliable results to get a first impression of potential leads for a given target. And as a special treat, all proposed lead compounds are provided with synthesis routes based on a set of robust organic reaction rules[43] that are commonly used in pharmaceutical R&D laboratories.[327]

Although there exist other structure-based tools supporting the reaction set from Hartenfeller *et al.*, e.g. PINGUI[58] and DOTS[37], based on its simple usage and thoroughly performed validation, NAOMInext is a valuable contribution to the field and may support medicinal chemists in their day-to-day work.

8.1. Achievements

The number of tools targeting the problem of H2L optimization is vast, but the number of new implementations increased anyway. Thus, there seems to be the necessity for new tools to support researchers in FBDD. Most of the available tools or workflows are designed for a specific purpose and are mostly used of

only a handful of people. A key issue is the mostly very cumbersome usage of the academic tools and time consuming pre-processing steps. Limitations of existing tools in the field of fragment growing are:

- lack of automation
- lack of profound validation
- lack of synthetic accessibility
- lack of a graphical user interface

NAOMInext covers all of the mentioned points in a condensed workflow, which is validated on a large-scale data set. A new constrained based conformational sampling algorithm, specifically designed to perform synthetically accessible fragment growing within protein bindings sites, has been developed to take the specific requirements into account.

The newly implemented reaction framework is able to process all synthetic reaction rules published by Hartenfeller and co-workers.[43] Thus, also complex ring forming reactions (about 50 % of the provided reactions from Hartenfeller and co-workers) can be performed, enabling the generation of diverse NCEs.

An intuitive user interface is provided and relieves the user during important pre-processing steps (see Chapter 7). Furthermore, the user interface allows to influence the underlying workflows using provided user constraints, i.e. user information to guide the sampling process. NAOMInext works with current file formats as input as well as output. Thus, intermediate results can be used in other cheminformatics toolkits and facilitate interchangeability of the produced results.

8.1.1. Usability

The main contribution to the field is the combination of a conformational sampling algorithm and synthetically accessible fragment growing within a user-friendly and easy-to-use graphical tool. The developed GUI allows users to perform fragment growing with just a few clicks. Publicly available data, e.g protein structures, ligands, and vendor catalogues (building blocks) can be used in any common cheminformatics file format. The condensed workflow combines several steps of the fragment elaboration cycle, i.e. medicinal chemistry, focused library design, virtual screening, and compound ranking. The implemented reaction framework, which supports organic synthesis reaction rules in a machine readable format (SMIRKS), allows user provided extensions of the reaction library. Thus, users are not dependent on steady updates and may incorporate their own (in house) reaction rules into NAOMInext.

The "correct" definition of a SMIRKS is not trivial and the protein environment influences the molecular state of a ligand (e.g. tautomerism). NAOMInext

helps to avoid SMARTS matching errors due to enumeration of different tautomeric forms and protonation states of the provided ligand. Hence, users do not have to incorporate different tautomeric forms or protonation states into their SMIRKS rules, for example `[OH, O-]`. Thus, the definition of SMIRKS is kept as easy as possible.

8.1.2. Validation

NAOMInext is validated on a large-scale data set compiled for binding mode analysis of chemically elaborated ligands and their respective putative precursor.[44] A subset of 271 ligand pairs, incorporating 87 different proteins, is used to validate NAOMInext. The results show that NAOMInext is able to predict correct binding poses and has a wide applicability domain.

8.2. Limitations

Despite the made achievements described in the previous section, there are still limitations which are discussed in more detail below.

8.2.1. Start Poses

The analysis of the obtained results in the re-growing Section 6.2 especially in the cross-growing Section 6.4.1 revealed that the start pose sampling can be further improved. The extension to also allow small shifts, rather than just rotations, would in some cases lead to improvements in pose prediction. The resulting problem is the combinatorial explosion which needs to be handled. One possibility would be the usage of a local grid around the anchor fragment. Rotated poses could then be placed on each grid point, thus, reducing the number of generated poses, i.e. reducing the combinatorial complexity. This procedure is also used by docking methods to perform the placement of the initial pose, e.g. Glide.[228], [328]

8.2.2. Scoring Function Validation

The scoring function used within NAOMInext is based on the ChemScore[258] scoring function (see Section 4.1.10). Trained regression coefficients are used to account for differences that may occur due to different implementations and to assess the "scoring power"[215]. However, a large-scale assessment of the scoring function is not performed. Thus, the performance considering compound ranking is not evaluated.

8.2.3. Multi-step Reactions

Since the here described development is based on a fragment growing approach that screens a database of fragments, multi step reactions (in one growing step) are not possible. Of course, covering multi step reactions may increase the diversity of the generated compounds.

8.2.4. Multi Component Reactions

MCRs are currently not supported due to the intended workflow of NAOMInext. Fragment growing, as implemented during this thesis, uses one start fragment and "grows" this fragment into a larger lead-like molecule screening a DB of potential BBs. Hence, one BB is attached at a time. MCRs need several fragments to generate the larger ligand within one reaction step (see ref [329] for an example). To support MCRs in NAOMInext, these reactions need to be encoded as two to one reaction using the remaining components implicitly (see Supporting Information of Hartenfeller *et al.* [43]).

8.2.5. Ring Opening Reactions

FBDD projects often start with a fragment screening to identify a potential hit.[330] Fragment libraries used for this purpose usually contain small rigid compounds with ring systems of size one or two. Thus, including ring opening reactions would increase the chemical diversity and allow the generation of more diverse lead series. Nevertheless, ring opening reactions are not supported in NAOMInext since they are not often used.[43]

8.2.6. Protein Flexibility

Depending on the target, protein flexibility is a big issue during fragment growing.[109] Sometimes single flexible side chains prevent the growing step to place the fragment within the correct sub pocket. In some cases, the ligand itself opens a ligand-inducible sub pocket of the binding site.[109] In other cases, flexible backbone loops move and enlarge the binding site. Hence, the growing procedure may fail due to side chain flexibility of the protein, which is not considered in NAOMInext.

8.2.7. Binding Mode Changes

A significant issue are binding mode changes of the used anchor fragment during H2L optimization, i.e. fragment elaboration through fragment growing

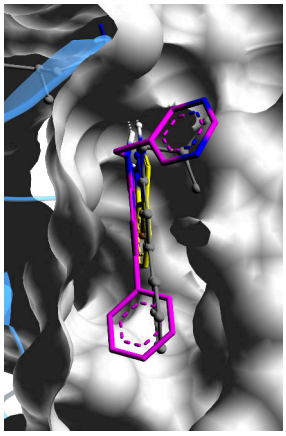


Figure (8.1) Top view of the kinase binding domain of PDB id 2VTA[319]. Based on the anchor fragment (yellow) NAOMInext generates a start pose with an RMSD of 0.6 Å (grey) considering the indazole (anchor fragment) moiety only. The reference structure (ligand from PDB id 2R3K[318]) is overlaid in magenta.

(see examples shown in Figure 6.26). In most of the cases, the initial start pose sampling is able to find a good start pose. For example, the anchor fragment 1H-indazole crystallized within CDK2 (see Figure 6.27b) has an RMSD of 0.6 Å to the crystal structure (considering the indazole moiety only). However, the conformational sampling algorithm is not able to find a valid result. This may be due to clashes because of the relatively narrow binding site (see Figure 8.1) or a just suboptimal placement of the anchor fragment as well. There are also other examples of binding mode change (see Figure 6.26) where NAOMInext is not able to predict an acceptable start pose and no growing result can be obtained.

8.2.8. Geometrical Inaccuracies or Deviations from the Standard

Geometrical inaccuracies or deviations from the underlying model (see the ring system in Figure 6.6a for an example) are a major issue for methods using the RR approximation. Slight geometrical deviations may accumulate in large deviations from the optimum. Thus, a post optimization procedure, allowing bond length and bond angle variations, should improve results in highly constraint binding sites and incorporate for usual geometrical deviations from the standard. Needless to say, torsion angle variations should be incorporated as well.

8.2.9. Additional Features and Interactivity

NAOMInext does not calculate other properties of the generated molecules as potency (via a ChemScore[258], [328] based scoring function see Section 4.1.10) and an estimate of ligand efficiency (LE). Of course, during optimization other properties like lipophilicity, polarity, charge, stability, etc. need to be considered.[23] An important extension of NAOMInext would be a visually indication of each molecules property and a possibility to provide property filters. This would improve the usability (clarity of results) of the software and provide medicinal chemists with the needed information to make the right decisions.

To do so, Hilbig *et al.* designed a cheminformatics platform called Mona for interactive compound library processing.[252], [331] Mona is based on the NAOMI framework and is able to process the outcome of NAOMInext. Since the data export of NAOMInext is performed in the SD file format, the synthetic route, score, performed reaction, and reactants of each resulting molecule are stored within the file format as additional information. As Mona[252], [331] is able to display those additional data users are able to filter the compound collection based on different criteria. For example, used reaction, binding site score, substructure filter, Lipinski's Ro5[94], and of course individual in house filters. Based on those filtered sets the compounds may be clustered.

8.3. Outlook

With reference to the aforementioned limitations, several improvements/extension are possible to further increase the applicability of NAOMInext.

8.3.1. Binding Mode Changes

Changes of the binding mode during chemical elaboration of an anchor fragment, may be addressed using a minimization procedure of the initial start position of the anchor fragment. Using a sampled start pose, a local optimization procedure may place the anchor fragment slightly more optimal considering the used scoring function and binding site conditions. Hence, the growing step may be guided in a slightly different, more target specific direction and minor binding mode changes may be solved.

Another attempt at a solution could be the incorporation of additional translational degrees of freedom to the initial start pose sampling (see Section 4.2.2). However, this would lead to an enormous increase in runtime as this would add three more degrees of freedom (translation in x, y, and z direction) and

blow up the number of to-be-tested start poses. A more expedient way would be the optimization of the initial spatial orientation of a small number of chosen start poses. However, the premise of FBDD is an invariant binding mode of the co-crystallized fragment.

8.3.2. Optimization of Start/Interim Solutions

Optimization of the start pose (anchor fragment) or interim solutions (with respect to the scoring function) is not performed within NAOMInext. In case of slightly overlapping atoms with the receptor or if a new interaction is found due to chemical extension, an optimization of the spatial orientation of the fragment (anchor fragment or interim solution) may improve the outcome of NAOMInext. Moreover, continuous small adjustments of interim solutions may supersede a significant number of currently used start poses. Thus, possibly further improving the runtime and performance of the sampling algorithm.

8.3.3. Scoring Function Validation

A scoring benchmark would be helpful to identify domains, i.e. protein targets, which could not be adequately scored using the implemented scoring function. For this purpose, the comparative assessment of scoring functions (CASF) benchmark[215], [332] can be used to estimate the "scoring power" of the implemented scoring function on a large-scale data set. The CASF benchmark is specifically designed to evaluate scoring functions independent from docking or growing approaches. Despite the validation of the "scoring power", the CASF benchmark can be used to estimate the "screening power" and the "ranking power" of the scoring function implemented in NAOMInext. The "ranking power" is important for the usability of a tool, since the molecules with the best binding affinities should be ranked first. The "screening power" on the other hand, allows to identify true binders within a set of random molecules.[215] In the context of fragment growing, extensions of the anchor fragment leading to random molecules could be detected.

8.3.4. Molecular Properties

NAOMInext does not filter any of the resulting molecules based on specific properties like Lipinski's Ro5[94] or other criteria like such described by Ghose *et al.* [333] Result ranking in NAOMInext is solely based on the empirical scoring function which is an estimate of protein ligand binding affinity (see Section 4.1.10). However, drug discovery is rather a multi-objective optimization

(MOOP) problem.[334] Hence, incorporating different pharmaceutically important properties, e.g. absorption, distribution, metabolism, excretion, and toxicity (ADMET) related properties, should be considered during the drug discovery process. Integration of physicochemical constraints would further improve the transparency of NAOMInext's results. Implementation of a filter, based on molecular properties, may be straightforward. Since NAOMInext uses a DB to store additional information, further molecular properties may be added and used to filter the results.

8.3.5. Additional Filters

Depending on the binding site, number of used reactions, BBs and available reaction vectors of the anchor fragment, the generated target focused library may be very large (several thousand compounds). Thus, it is nearly impossible for the user to investigate all compounds visually. One possibility would be the incorporation of a clustering strategy to generate a chemically diverse library. The user may then investigate the results based on a cluster representative (scaffold) and subsequently look into detailed results for a specific cluster. Furthermore, the tool Mona[252], [331] can be used for interactive compound library processing. However, NAOMInext was not designed for large scale compound enumeration. For this purpose other tools like FlexNovo[178] or FSees[182] exist.

8.3.6. Large-Scale Validation Data Set

The data set from Malhotra and Karanicolas[44] is a good starting point for a large-scale evaluation of a fragment growing approach. However, some incorporated structures (e.g. PDB id 2R3K[318]) exhibit structural issues (low EDIA[301], [302] score for specific structural motifs), which complicate the statements made by RMSD-based evaluation procedures. Hence, a validation data set should be filtered using the EDIA score as filter criteria.

Another issue are the ligand pairs themselves. Most of the provided ligand pairs in the data set are not related, for example 2XM2[296]/2WCA[297]. One ligand may be a substructure of the other ligand, but they do not necessarily originate from a common study or publication. A good source to find related ligand pairs is the PDB[1][217] itself. Some PDB files contain identifiers to related PDB entries in the *remark* section. Using this information, related ligand pairs can be easily identified. For example, the fragment-based study from Unzue and co-workers provides three related protein structures (PDB ids 3SVH, 3SVF,

[1] www.rcsb.org

and 3SVG).[335] Thus, two related ligand pairs can be derived without much effort.

8.3.7. Summary

As a bottom line, NAOMInext is a valuable contribution the the field of FBDD. Its interactive use may help medicinal chemists to generate new ideas for yet undrugged drug targets like protein-protein interfaces. The comprehensive validation implies NAOMInext a broad applicability domain. The newly developed conformational sampling algorithm provides sufficient results for possible extensions during the target-focused library design. The easy handling of the integrated reaction rules may broaden the users imagination of the available chemical space, and may lead to yet unexpected chemical extensions. Because each predicted ligand is provided with a possible synthesis route, users may focus on readily synthesizable compounds in subsequent steps. Thus, reducing time and costs. The opportunity to extend the provided reaction set by user-defined reaction rules further increases the value of NAOMInext.

Bibliography

[1] M. Palmer, "Phenotypic Screening," in *Small Molecule Medicinal Chemistry*, W. Czechtizky and P. Hamley, Eds., 1st ed., Hoboken, NJ: Wiley & Sons, 2016, ch. 10, pp. 281–304, ISBN: 978-1-118-77160-0.

[2] W. Friedrich, P. Knipping, and M. Laue, "Interferenzerscheinungen bei Röntgenstrahlen," *Annalen der Physik*, vol. 346, no. 10, pp. 971–988, 1913. DOI: 10.1002/andp.19133461004. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/andp.19133461004.

[3] M. M. Woolfson, "The development of structural x-ray crystallography," *Physica Scripta*, vol. 93, no. 3, p. 32 501, Jan. 2018. DOI: 10.1088/1402-4896/aa9c30. [Online]. Available: https://doi.org/10.1088%7B%5C%%7D2F1402-4896%7B%5C%%7D2Faa9c30.

[4] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. Parrish, H. Wyckoff, and D. C. Phillips, "A three-dimensional model of the myoglobin molecule obtained by X-ray analysis," *Nature*, vol. 181, no. 4610, pp. 662–666, 1958.

[5] J. C. Kendrew, R. E. Dickerson, B. E. Strandberg, R. G. Hart, D. R. Davies, D. C. Phillips, and V. C. Shore, "Structure of Myoglobin: A Three-Dimensional Fourier Synthesis at 2 Å. Resolution," *Nature*, vol. 185, no. 4711, pp. 422–427, 1960, ISSN: 1476-4687. DOI: 10.1038/185422a0. [Online]. Available: https://doi.org/10.1038/185422a0.

[6] G. Scapin, "Structural biology and drug discovery.," *Current Pharmaceutical Design*, vol. 12, no. 17, pp. 2087–2097, 2006, ISSN: 1359-6446. DOI: 10.1016/S1359-6446(05)03484-7.

[7] C. Tan, L. Wei, F. Ottensmeyer, I. Goldfine, B. A. Maddux, C. C. Yip, R. A. Batey, and L. P. Kotra, "Structure-based de novo design of ligands using a three-dimensional model of the insulin receptor," *Bioorganic & Medicinal Chemistry Letters*, vol. 14, no. 6, pp. 1407–1410, Mar. 2004, ISSN: 0960894X. DOI: 10.1016/j.bmcl.2004.01.064. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960894X04001441%20https://linkinghub.elsevier.com/retrieve/pii/S0960894X04001441.

[8] H.-J. Boehm, M. Boehringer, D. Bur, H. Gmuender, W. Huber, W. Klaus, D. Kostrewa, H. Kuehne, T. Luebbers, N. Meunier-Keller, and F. Mueller, "Novel Inhibitors of DNA Gyrase: 3D Structure Based Biased Needle Screening, Hit Validation by Biophysical Methods, and 3D Guided Optimization. A Promising Alternative to Random Screening," *Journal of Medicinal Chemistry*, vol. 43, no. 14, pp. 2664–2674, Jul. 2000, ISSN: 0022-2623. DOI: 10.1021/jm000017s.

[9] C. W. Murray and D. C. Rees, "The rise of fragment-based drug discovery," *Nature Chemistry*, vol. 1, no. 3, pp. 187–192, 2009, ISSN: 1755-4330. DOI: 10.1038/nchem.217. [Online]. Available: http://www.nature.com/doifinder/10.1038/nchem.217.

[10] W. B. Schwartz, "The Effect of Sulfanilamide on Salt and Water Excretion in Congestive Heart Failure," *New England Journal of Medicine*, vol. 240, no. 5, pp. 173–177, Feb. 1949, ISSN: 0028-4793. DOI: 10.1056/NEJM194902032400503. [Online]. Available: http://www.nejm.org/doi/abs/10.1056/NEJM194902032400503.

[11] W. Hunkeler, H. Möhler, L. Pieri, P. Polc, E. P. Bonetti, R. Cumin, R. Schaffner, and W. Haefely, "Selective antagonists of benzodiazepines," *Nature*, vol. 290, no. 5806, pp. 514–516, Apr. 1981, ISSN: 0028-0836. DOI: 10.1038/290514a0. [Online]. Available: http://www.nature.com/articles/290514a0.

[12] S. L. Schreiber, "Target-Oriented and Diversity-Oriented Organic Synthesis in Drug Discovery," *Science*, vol. 287, no. 5460, pp. 1964–1969, 2000, ISSN: 00368075, 10959203. DOI: 10.1126/science.287.5460.1964. [Online]. Available: http://www.sciencemag.org/cgi/doi/10.1126/science.287.5460.1964.

[13] J. Drews, "Drug Discovery : A Historical Perspective," *Science*, vol. 1960, no. 2000, pp. 1960–1965, 2008. DOI: 10.1126/science.287.5460.1960.

[14] S. Myers and A. Baker, "Drug discovery — an operating model for a new era," *Nature Biotechnology*, vol. 19, pp. 727–730, 2001. DOI: 10.1038/90765. [Online]. Available: https://doi.org/10.1038/90765.

[15] J. Drews, "Genomic sciences and the medicine of tomorrow," *Nature biotechnology*, vol. 14, no. 3, pp. 1516–1518, 1996, ISSN: 1087-0156 (Print) 1087-0156 (Linking). DOI: 10.1038/nbt0696-765.

[16] J. Drews, "Innovation deficit revisited: Reflections on the productivity of pharmaceutical R and D," *Drug Discovery Today*, vol. 3, no. 11, pp. 491–494, 1998, ISSN: 13596446. DOI: 10.1016/S1359-6446(98)01252-5.

[17] P. J. Hajduk and J. Greer, "A decade of fragment-based drug design: Strategic advances and lessons learned," *Nature Reviews Drug Discovery*, vol. 6, no. 3, pp. 211–219, 2007, ISSN: 14741776. DOI: 10.1038/nrd2220.

[18] W. P. Janzen, "Screening Technologies for Small Molecule Discovery: The State of the Art," *Chemistry & Biology*, vol. 21, no. 9, pp. 1162–1170, Sep. 2014, ISSN: 1074-5521. DOI: 10.1016/J.CHEMBIOL.2014.07.015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1074552114002440.

[19] A. Kumar, A. Voet, and K. Zhang, "Fragment Based Drug Design: From Experimental to Computational Approaches," *Current Medicinal Chemistry*, vol. 19, no. 30, pp. 5128–5147, 2012, ISSN: 09298673. DOI: 10.2174/092986712803530467. [Online]. Available: http://www.eurekaselect.com/openurl/content.php?genre=article%7B%5C&%7Dissn=0929-8673%7B%5C&%7Dvolume=19%7B%5C&%7Dissue=30%7B%5C&%7Dspage=5128.

[20] R. Macarron, M. N. Banks, D. Bojanic, D. J. Burns, D. A. Cirovic, T. Garyantes, D. V. S. Green, R. P. Hertzberg, W. P. Janzen, J. W. Paslay, and U. Schopfer, "Impact of high-throughput screening in biomedical research," *Nature*, vol. 10, no. March 2011, pp. 188–195, 2011, ISSN: 1474-1784. DOI: 10.1038/nrd3368. [Online]. Available: http://dx.doi.org/10.1038/nrd3368.

[21] Y. Bian and X.-Q. Xie, "Computational Fragment-Based Drug Design: Current Trends, Strategies, and Applications," *The AAPS Journal*, vol. 20, no. 3, p. 59, 2018, ISSN: 1550-7416. DOI: 10.1208/s12248-018-0216-7. [Online]. Available: http://link.springer.com/10.1208/s12248-018-0216-7.

[22] D. A. Erlanson, S. W. Fesik, R. E. Hubbard, W. Jahnke, and H. Jhoti, "Twenty years on: the impact of fragments on drug discovery," *Nature Reviews Drug Discovery*, vol. 15, no. 9, pp. 605–619, Jul. 2016, ISSN: 1474-1776. DOI: 10.1038/nrd.2016.109. [Online]. Available: http://www.nature.com/doifinder/10.1038/nrd.2016.109.

[23] W. A. Warr, "Fragment-based drug discovery," *Journal of Computer-Aided Molecular Design*, vol. 23, no. 8, pp. 453–458, 2009, ISSN: 0920654X. DOI: 10.1007/s10822-009-9292-1.

[24] P. G. Polishchuk, T. I. Madzhidov, and A. Varnek, "Estimation of the size of drug-like chemical space based on GDB-17 data," *Journal of Computer-Aided Molecular Design*, vol. 27, no. 8, pp. 675–679, 2013, ISSN: 0920654X. DOI: 10.1007/s10822-013-9672-4.

[25] S. B. Shuker, P. J. Hajduk, R. P. Meadows, and S. W. Fesik, "Discovering High-Affinity Ligands for Proteins: SAR by NMR," *Science*, vol. 274, no. 5292, pp. 1531–1534, 1996, ISSN: 0036-8075. DOI: 10.1126/science.274.5292.1531. [Online]. Available: http://science.sciencemag.org/content/274/5292/1531.abstract.

[26] A. Schuffenhauer, S. Ruedisser, A. Marzinzik, W. Jahnke, P. Selzer, and E. Jacoby, "Library Design for Fragment Based Screening," *Current Topics in Medicinal Chemistry*, vol. 5, no. 8, pp. 751–762, 2005, ISSN: 15680266. DOI: 10.2174/1568026054637700. [Online]. Available: http://www.eurekaselect.com/openurl/content.php?genre=article%7B%5C&%7Dissn=1568-0266%7B%5C&%7Dvolume=5%7B%5C&%7Dissue=8%7B%5C&%7Dspage=751.

[27] M. M. Hann, A. R. Leach, and G. Harper, "Molecular Complexity and Its Impact on the Probability of Finding Leads for Drug Discovery," *Journal of Chemical Information and Computer Sciences*, vol. 41, no. 3, pp. 856–864, 2001, ISSN: 00952338. DOI: 10.1021/ci000403i.

[28] D. G. Teotico, K. Babaoglu, G. J. Rocklin, R. S. Ferreira, A. M. Giannetti, and B. K. Shoichet, "Docking for fragment inhibitors of AmpC β-lactamase," *Proceedings of the National Academy of Sciences*, vol. 106, no. 18, pp. 7455–7460, 2009, ISSN: 0027-8424. DOI: 10.1073/pnas.0813029106. [Online]. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.0813029106.

[29] D. Fattori, "Molecular recognition: the fragment approach in lead generation," *Drug Discovery Today*, vol. 9, no. 5, pp. 229–238, Mar. 2004, ISSN: 13596446. DOI: 10.1016/S1359-6446(03)03007-1. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/14980541%20http://linkinghub.elsevier.com/retrieve/pii/S1359644603030071.

[30] M. Congreve, G. Chessari, D. Tisi, and A. J. Woodhead, "Recent Developments in Fragment-Based Drug Discovery," *Journal of Medicinal Chemistry*, vol. 51, no. 13, 2008.

[31] C. W. Murray, M. L. Verdonk, and D. C. Rees, "Experiences in fragment-based drug discovery," *Trends in Pharmacological Sciences*, vol. 33, no. 5, pp. 224–232, May 2012, ISSN: 01656147. DOI: 10.1016/j.tips.2012.02.006. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0165614712000326.

[32] G. M. Keserű and G. M. Makara, "Hit discovery and hit-to-lead approaches," *Drug Discovery Today*, vol. 11, no. 15-16, pp. 741–748, 2006, ISSN: 13596446. DOI: 10.1016/j.drudis.2006.06.016.

[33] R. F. Ludlow, M. L. Verdonk, H. K. Saini, I. J. Tickle, and H. Jhoti, "Detection of secondary binding sites in proteins using fragment screening.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, no. 52, pp. 15 910–15 915, Dec. 2015, ISSN: 1091-6490. DOI: 10.1073/pnas.1518946112. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26655740%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4703025.

[34] G. E. de Kloe, D. Bailey, R. Leurs, and I. J. P. de Esch, "Transforming fragments into candidates: small becomes big in medicinal chemistry," *Drug Discovery Today*, vol. 14, no. 13-14, pp. 630–646, 2009, ISSN: 13596446. DOI: 10.1016/j.drudis.2009.03.009.

[35] M. Baker, "Fragment-based lead discovery grows up," *Nat. Rev. Drug Discovery*, vol. 12, no. January, pp. 5–7, 2013, ISSN: 1474-1784. DOI: 10.1038/nrd3926. arXiv: NIHMS150003. [Online]. Available: http://dx.doi.org/10.1038/nrd3926%7B%5C%%7D5Cnhttp://dx.doi.org/10.1038/nrd3926%7B%5C%%7D5Cnhttp://www.nature.com/nrd/journal/v12/n1/full/nrd3926.html.

[36] K. H. Bleicher, H. J. Böhm, K. Müller, and A. I. Alanine, "Hit and lead generation: Beyond high-throughput screening," *Nature Reviews Drug Discovery*, vol. 2, no. 5, pp. 369–378, May 2003, ISSN: 14741776. DOI: 10.1038/nrd1086. arXiv: NIHMS150003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12750740%20http://www.nature.com/articles/nrd1086.

[37] L. Hoffer, C. Muller, P. Roche, and X. Morelli, "Chemistry-driven Hit-to-lead Optimization Guided by Structure-based Approaches," *Molecular Informatics*, vol. 37, no. 9-10, p. 1 800 059, Sep. 2018, ISSN: 18681743. DOI: 10.1002/minf.201800059. [Online]. Available: www.molinf.com%20http://doi.wiley.com/10.1002/minf.201800059.

[38] J. R. Huth, C. Park, A. M. Petros, A. R. Kunzer, M. D. Wendt, X. Wang, C. L. Lynch, J. C. Mack, K. M. Swift, R. A. Judge, J. Chen, P. L. Richardson, S. Jin, S. K. Tahir, E. D. Matayoshi, S. A. Dorwin, U. S. Ladror, J. M. Severin, K. A. Walter, D. M. Bartley, S. W. Fesik, S. W. Elmore, and P. J. Hajduk, "Discovery and Design of Novel HSP90 Inhibitors Using Multiple Fragment-based Design Strategies," *Chemical Biology & Drug Design*, vol. 70, no. 1, pp. 1–12, Jul. 2007, ISSN: 1747-0277. DOI: 10.1111/j.1747-0285.2007.00535.x. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17630989%20http://doi.wiley.com/10.1111/j.1747-0285.2007.00535.x.

[39] S. Chung, J. B. Parker, M. Bianchet, L. M. Amzel, and T. James, "Impact of linker strain and flexibility in the design of a fragment- based inhibitor," *Nature chemical biology*, vol. 5, no. 6, pp. 407–413, 2009. DOI: 10.1038/nchembio.163..

[40] J. C. Baber and M. Feher, "Predicting Synthetic Accessibility: Application in Drug Discovery and Development," *Mini-Reviews in Medicinal Chemistry*, vol. 4, no. 6, pp. 681–692, Aug. 2004, ISSN: 13895575. DOI: 10.2174/1389557043403765. [Online]. Available: http://www.eurekaselect.com/openurl/content.php?genre=article%7B%5C&%7Dissn=1389-5575%7B%5C&%7Dvolume=4%7B%5C&%7Dissue=6%7B%5C&%7Dspage=681.

[41] E. Vangrevelinghe, "Computational Approaches for Fragment Optimization," *Current Computer-Aided Drug Design*, vol. 3, pp. 69–83, 2007, ISSN: 15734099. DOI: 10.2174/157340907780058781.

[42] G. Schneider and U. Fechner, "Computer-based de novo design of drug-like molecules," *Nature Reviews Drug Discovery*, vol. 4, no. 8, pp. 649–663, Aug. 2005, ISSN: 1474-1776. DOI: 10.1038/nrd1799. [Online]. Available: http://www.nature.com/doifinder/10.1038/nrd1799.

[43] M. Hartenfeller, M. Eberle, P. Meier, C. Nieto-Oberhuber, K. H. Altmann, G. Schneider, E. Jacoby, and S. Renner, "A collection of robust organic synthesis reactions for in silico molecule design," *Journal of Chemical Information and Modeling*, vol. 51, no. 12, pp. 3093–3098, Dec. 2011, ISSN: 15499596. DOI: 10.1021/ci200379p. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/ci200379p.

[44] S. Malhotra and J. Karanicolas, "When Does Chemical Elaboration Induce a Ligand To Change Its Binding Mode?" *Journal of Medicinal Chemistry*, vol. 60, no. 1, pp. 128–145, Jan. 2017, ISSN: 15204804. DOI: 10.1021/acs.jmedchem.6b00725. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/acs.jmedchem.6b00725.

[45] Y. C. Martin, "Let's not forget tautomers," *Journal of Computer-Aided Molecular Design*, vol. 23, no. 10, pp. 693–704, 2009, ISSN: 0920654X. DOI: 10.1007/s10822-009-9303-2.

[46] M. Hartenfeller, M. Eberle, P. Meier, C. Nieto-Oberhuber, K. H. Altmann, G. Schneider, E. Jacoby, and S. Renner, "Probing the bioactivity-relevant chemical space of robust reactions and common molecular building blocks," *Journal of Chemical Information and Modeling*, vol. 52, no. 5, pp. 1167–1178, 2012, ISSN: 15499596. DOI: 10.1021/ci200618n.

[47] M. Hartenfeller, H. Zettl, M. Walter, M. Rupp, F. Reisen, E. Proschak, S. Weggen, H. Stark, and G. Schneider, "Dogs: Reaction-driven de novo design of bioactive compounds," *PLoS Computational Biology*, vol. 8, no. 2, 2012, ISSN: 1553734X. DOI: 10.1371/journal.pcbi.1002380.

[48] M. Hartenfeller, S. Renner, and E. Jacoby, "Reaction-Driven De Novo Design: a Keystone for Automated Design of Target Family-Oriented Libraries," in *De novo Molecular Design*, G. Schneider, Ed., Weinheim: WILEY-VCH Verlag, 2014, ch. 10, pp. 245–266, ISBN: 978-3-527-33461-2.

[49] K. Earm and Y. E. Earm, "Integrative approach in the era of failing drug discovery and development," *Integrative Medicine Research*, vol. 3, no. 4, pp. 211–216, 2014, ISSN: 22134220. DOI: 10.1016/j.imr.2014.09.002.

[50] M. L. Bolognesi, M. R. Popovic-Nikolic, R. R. Ramsay, E. Uliassi, and K. Nikolic, "A perspective on multi-target drug discovery and design for complex diseases," *Clinical and Translational Medicine*, vol. 7, no. 1, 2018, ISSN: 2001-1326. DOI: 10.1186/s40169-017-0181-2. [Online]. Available: https://doi.org/10.1186/s40169-017-0181-2.

[51] P. S. Kutchukian and E. I. Shakhnovich, "De novo design: balancing novelty and confined chemical space," *Expert Opinion on Drug Discovery*, vol. 5, no. 8, pp. 789–812, 2010. DOI: 10.1517/17460441.2010.497534. [Online]. Available: https://doi.org/10.1517/17460441.2010.497534.

[52] M. Hartenfeller and G. Schneider, "Enabling future drug discovery by de novo design," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 1, no. 5, pp. 742–759, 2011, ISSN: 17590876. DOI: 10.1002/wcms.49.

[53] B. C. Pearce, D. R. Langley, J. Kang, H. Huang, and A. Kulkarni, "E-Novo : An Automated Workflow for Efficient Structure-Based Lead Optimization," *Journal of Chemical Information and Modeling*, vol. 49, no. 7, pp. 1797–1809, Jul. 2009, ISSN: 1549-9596. DOI: 10.1021/ci900073k. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/ci900073k.

[54] J. D. Durrant, S. Lindert, and J. A. Mccammon, "AutoGrow 3.0 : An improved algorithm for chemically tractable , semi-automated protein inhibitor design," *Journal of Molecular Graphics and Modelling*, vol. 44, pp. 104–112, 2013, ISSN: 1093-3263. DOI: 10.1016/j.jmgm.2013.05.006. [Online]. Available: http://dx.doi.org/10.1016/j.jmgm.2013.05.006.

[55] N. Chéron, N. Jasty, and E. I. Shakhnovich, "OpenGrowth: An Automated and Rational Algorithm for Finding New Protein Ligands," *Journal of Medicinal Chemistry*, vol. 59, no. 9, pp. 4171–4188, 2016, ISSN: 15204804. DOI: 10.1021/acs.jmedchem.5b00886.

[56] J. Pottel and N. Moitessier, "Customizable Generation of Synthetically Accessible, Local Chemical Subspaces," *Journal of Chemical Information and Modeling*, vol. 57, no. 3, pp. 454–467, 2017, ISSN: 15205142. DOI: 10.1021/acs.jcim.6b00648.

[57] L. Batiste, A. Unzue, A. Dolbois, F. Hassler, X. Wang, N. Deerain, J. Zhu, D. Spiliotopoulos, C. Nevado, and A. Caflisch, "Chemical Space Expansion of Bromodomain Ligands Guided by in Silico Virtual Couplings (AutoCouple)," *ACS Central Science*, vol. 4, no. 2, pp. 180–188, 2018, ISSN: 23747951. DOI: 10.1021/acscentsci.7b00401.

[58] F. Chevillard, H. Rimmer, C. Betti, E. Pardon, S. Ballet, N. van Hilten, J. Steyaert, W. E. Diederich, and P. Kolb, "Binding-Site Compatible Fragment Growing Applied to the Design of β 2 -Adrenergic Receptor Ligands," *Journal of Medicinal Chemistry*, vol. 61, no. 3, pp. 1118–1129, 2018, ISSN: 0022-2623. DOI: 10.1021/acs.jmedchem.7b01558. [Online]. Available: http://pubs.acs.org/doi/10.1021/acs.jmedchem.7b01558.

[59] L. Hoffer, Y. V. Voitovich, B. Raux, K. Carrasco, C. Muller, A. Y. Fedorov, C. Derviaux, A. Amouric, S. Betzi, D. Horvath, A. Varnek, Y. Collette, S. Combes, P. Roche, and X. Morelli, "Integrated Strategy for Lead Optimization Based on Fragment Growing: The Diversity-Oriented-Target-Focused-Synthesis Approach," *Journal of Medicinal Chemistry*, vol. 61, no. 13, pp. 5719–5732, Jul. 2018, ISSN: 0022-2623. DOI: 10.1021/acs.jmedchem.8b00653. [Online]. Available: http://pubs.acs.org/doi/10.1021/acs.jmedchem.8b00653.

[60] D. Merk, L. Friedrich, F. Grisoni, and G. Schneider, "De Novo Design of Bioactive Small Molecules by Artificial Intelligence," *Molecular Informatics*, vol. 1700153, pp. 2–6, 2018, ISSN: 18681743. DOI: 10.1002/minf.201700153. [Online]. Available: http://doi.wiley.com/10.1002/minf.201700153.

[61] H. Patel, M. J. Bodkin, B. Chen, and V. J. Gillet, "Knowledge-Based Approach to ¡i¿de Novo¡/i¿ Design Using Reaction Vectors," *Journal of Chemical Information and Modeling*, vol. 49, no. 5, pp. 1163–1184, May 2009, ISSN: 1549-9596. DOI: 10.1021/ci800413m. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/ci800413m.

[62] V. J. Gillet, M. J. Bodkin, and D. Hristozov, "Multiobjective De Novo Design of Synthetically Accesible Compounds," in *De novo Molecular Design*, G. Schneider, Ed., 1st ed., Weinheim: WILEY-VCH Verlag, 2013, ch. 11, pp. 269–286, ISBN: 978-3-527-33461-2.

[63] V. J. Gillet, A. P. Johnson, P. Mata, S. Sike, and P. Williams, "SPROUT: A program for structure generation," *Journal of Computer-Aided Molecular Design*, vol. 7, no. 2, pp. 127–153, Apr. 1993, ISSN: 0920-654X. DOI: 10.1007/BF00126441. [Online]. Available: http://link.springer.com/10.1007/BF00126441.

[64] V. J. Gillet, W. Newell, P. Mata, G. J. Myatt, S. Sike, Z. Zsoldos, and A. P. Johnson, "SPROUT: Recent developments in the de novo design of molecules," *Journal of Chemical Information and Modeling*, vol. 34, no. 1, pp. 207–217, Aug. 1994, ISSN: 1549-9596. DOI: 10.1021/ci00017a027. [Online]. Available: http://pubs.acs.org/cgi-bin/doilookup/?10.1021/ci00017a027.

[65] P. Mata, V. J. Gillet, A. P. Johnson, J. Lampreia, G. J. Myatt, S. Sike, and A. L. Stebbings, "SPROUT: 3D Structure Generation Using Templates," *Journal of Chemical Information and Modeling*, vol. 35, no. 3, pp. 479–493, May 1995, ISSN: 1549-9596. DOI: 10.1021/ci00025a016. [Online]. Available: http://pubs.acs.org/cgi-bin/doilookup/?10.1021/ci00025a016.

[66] V. J. Gillet, G. J. Myatt, Z. Zsoldos, and A. P. Johnson, "SPROUT, HIPPO and CAESA: Tools for de novo structure generation and estimation of synthetic accessibility," *Perspectives in Drug Discovery and Design*, vol. 3, no. 1, pp. 34–50, Dec. 1995, ISSN: 0928-2866. DOI: 10.1007/BF02174466. [Online]. Available: http://link.springer.com/10.1007/BF02174466.

[67] Z. Szabó, M. Vargyas, and A. P. Johnson, "Novel Treatment of Conformational Flexibility Using Interval Analysis," *Journal of Chemical Information and Computer Sciences*, vol. 40, no. 2, pp. 339–346, 2000. DOI: 10.1021/ci990105p.

[68] K. Boda, "SynSPROUT : generating synthetically accessible ligands by de novo design," PhD thesis, University of Leeds, 2002.

[69] *SPROUT User Publications*, 2018. [Online]. Available: http://www.keymodule.co.uk/library/user-publications/sprout.html (visited on 10/24/2018).

[70] J. M. Law, D. Y. Fung, Z. Zsoldos, A. Simon, Z. Szabo, I. G. Csizmadia, and A. P. Johnson, "Validation of the SPROUT de novo design program," *Journal of Molecular Structure: THEOCHEM*, vol. 666-667, pp. 651–657, 2003, ISSN: 01661280. DOI: 10.1016/j.theochem.2003.08.104.

[71] I. V. Efremov and D. A. Erlanson, "Fragment-Based Lead Generation," in *Lead Generation*, J. Holenz, Ed., 1st ed., Weinheim: WILEY-VCH Verlag, 2016, ch. 6, pp. 133–158, ISBN: 9783527677047. DOI: 10.1002/

9783527677047.ch06. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9783527677047.ch06.

[72] D. Joseph-McCarthy, A. J. Campbell, G. Kern, and D. Moustakas, "Fragment-based lead discovery and design," *Journal of Chemical Information and Modeling*, vol. 54, no. 3, pp. 693–704, 2014, ISSN: 15205142. DOI: 10.1021/ci400731w.

[73] G. Klebe, *Wirkstoffdesign*, 2nd ed., G. Klebe, Ed. Heidelberg: Spektrum Akademischer Verlag, 2009, ISBN: 978-3-8274-2046-6.

[74] T. J. Ritchie and I. M. McLay, "Should medicinal chemists do molecular modelling?" *Drug Discovery Today*, vol. 17, no. 11-12, pp. 534–537, 2012, ISSN: 13596446. DOI: 10.1016/j.drudis.2012.01.005. [Online]. Available: http://dx.doi.org/10.1016/j.drudis.2012.01.005.

[75] 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*, vol. 51, no. 12, pp. 3199–3207, 2011, ISSN: 15499596. DOI: 10.1021/ci200324e.

[76] A. Kolodzik, S. Urbaczek, and M. Rarey, "Unique ring families: A chemically meaningful description of molecular ring topologies," *Journal of Chemical Information and Modeling*, vol. 52, no. 8, pp. 2013–2021, Aug. 2012, ISSN: 15499596. DOI: 10.1021/ci200629w. [Online]. Available: http://dx.doi.org/10.1021/ci200629w.

[77] S. Urbaczek, A. Kolodzik, I. Groth, S. Heuser, and M. Rarey, "Reading PDB: Perception of molecules from 3D atomic coordinates," *Journal of Chemical Information and Modeling*, vol. 53, no. 1, pp. 76–87, 2013, ISSN: 15499596. DOI: 10.1021/ci300358c.

[78] C. Schärfer, T. Schulz-Gasch, H.-C. Ehrlich, W. Guba, M. Rarey, and M. Stahl, "Torsion angle preferences in druglike chemical space: A comprehensive guide," *Journal of Medicinal Chemistry*, vol. 56, no. 5, pp. 2016–2028, Mar. 2013, ISSN: 00222623. DOI: 10.1021/jm3016816. [Online]. Available: http://dx.doi.org/10.1021/jm3016816.

[79] 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*, vol. 54, no. 3, pp. 756–766, 2014, ISSN: 15205142. DOI: 10.1021/ci400724v.

[80] 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*, vol. 6, no. 1, 2014, ISSN: 17582946. DOI: 10.1186/1758-2946-6-12.

[81] S. Urbaczek, "A consistent cheminformatics framework for automated virtual screening," PhD, Universität Hamburg, 2014.

[82] W. Guba, A. Meyder, M. Rarey, and J. Hert, "Torsion Library Reloaded: A New Version of Expert-Derived SMARTS Rules for Assessing Conformations of Small Molecules," *Journal of Chemical Information and Modeling*, vol. 56, no. 1, pp. 1–5, Jan. 2016, ISSN: 15205142. DOI: 10.1021/acs.jcim.5b00522. [Online]. Available: http://pubs.acs.org/doi/10.1021/acs.jcim.5b00522%20http://pubs.acs.org/doi/abs/10.1021/acs.jcim.5b00522.

[83] E. Nittinger, T. Inhester, S. Bietz, A. Meyder, K. T. Schomburg, G. Lange, R. Klein, and M. Rarey, "Large-Scale Analysis of Hydrogen Bond Interaction Patterns in Protein–Ligand Interfaces," *Journal of Medicinal Chemistry*, vol. 60, no. 10, pp. 4245–4257, May 2017, ISSN: 0022-2623. DOI: 10.1021/acs.jmedchem.7b00101. [Online]. Available: http://pubs.acs.org/doi/10.1021/acs.jmedchem.7b00101.

[84] F. Flachsenberg, N. Andresen, and M. Rarey, "RingDecomposerLib: An Open-Source Implementation of Unique Ring Families and Other Cycle Bases," *Journal of Chemical Information and Modeling*, vol. 57, no. 2, pp. 122–126, Feb. 2017, ISSN: 1549-9596. DOI: 10.1021/acs.jcim.6b00736.

[85] R. Lahana, "Who wants to be irrational?" *Drug Discovery Today*, vol. 8, no. 15, pp. 655–656, Aug. 2003, ISSN: 1359-6446. DOI: 10.1016/S1359-6446(03)02734-X. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S135964460302734X?via%7B%5C%%7D3Dihub.

[86] R. S. Bohacek, C. Mcmartin, and W. C. Guida, "The Art and Practice of Structure-Based Drug Design: A Molecular Modeling Perspective," *Medicinal Research Reviews*, vol. 16, no. 1, pp. 3–50, 1996.

[87] R. J. Bienstock, "Overview: Fragment-Based Drug Design," in *Library Design, Search Methods, and Applications of Fragment-Based Drug Design*, Washington, DC, 2011, ch. 1, pp. 1–26. DOI: 10.1021/bk-2011-1076.ch001. [Online]. Available: https://pubs.acs.org/doi/abs/10.1021/bk-2011-1076.ch001.

[88] D. A. Erlanson, "Introduction to Fragment-Based Drug Discovery," in *Fragment-Based Drug Discovery and X-Ray Crystallography*, T. G. Davies and M. Hyvönen, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 1–32, ISBN: 978-3-642-27540-1. DOI: 10.1007/128_2011_180. [Online]. Available: https://doi.org/10.1007/128%7B%5C_%7D2011%7B%5C_%7D180.

[89] T. Fink and J.-L. Reymond, "Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discov," *Journal of Chemical Information and Modeling*, vol. 47, no. 2, pp. 342–353, 2007. DOI: 10.1021/CI600423U. [Online]. Available: https://pubs.acs.org/doi/abs/10.1021/ci600423u.

[90] M. Congreve, R. Carr, C. W. Murray, and H. Jhoti, "A 'Rule of Three' for fragment-based lead discovery?" *Drug Discovery Today*, vol. 8, no. 19, pp. 876–877, 2003, ISSN: 13596446. DOI: 10.1016/S1359-6446(03)02831-9. arXiv: 03/ [1359-6446]. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1359644603028319.

[91] H. Jhoti, G. Williams, D. C. Rees, and C. W. Murray, "The 'rule of three' for fragment-based drug discovery: where are we now?" *Nature reviews. Drug discovery*, vol. 12, no. 8, pp. 644–5, 2013, ISSN: 1474-1784. DOI: 10.1038/nrd3926-c1. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23845999.

[92] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg, and A. L. Schacht, "How to improve RD productivity: The pharmaceutical industry's grand challenge," *Nature Reviews Drug Discovery*, vol. 9, no. 3, pp. 203–214, 2010, ISSN: 14741776. DOI: 10.1038/nrd3078. arXiv: /www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3006164{\&}tool=pmcentrez{\&}rendertype=abstract. [Figures, S., 2010. Supplementary information. Nature, 1(c), pp.1{7. Available at: http:].

[93] G. M. Keserű, D. A. Erlanson, G. G. Ferenczy, M. M. Hann, C. W. Murray, and S. D. Pickett, "Design Principles for Fragment Libraries: Maximizing the Value of Learnings from Pharma Fragment-Based Drug Discovery (FBDD) Programs for Use in Academia," *Journal of Medicinal Chemistry*, vol. 59, no. 18, pp. 8189–8206, 2016, ISSN: 15204804. DOI: 10.1021/acs.jmedchem.6b00197.

[94] 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," *Advanced drug delivery reviews*, vol. 46, pp. 3–26, 2001, ISSN: 0169409X. DOI: 10.1016/S0169-409X(00)00129-0. [Online]. Available: http://archive.org/details/a589638004lameuoft.

[95] J.-P. Renaud, T. Neumann, and L. Van Hijfte, "Fragment-Based Drug Discovery," in *Small Molecule Medicinal Chemistry*, W. Czechtizky and P. Hamley, Eds., 1st ed., Weinheim: Wiley & Sons, 2016, ch. 8, pp. 221–249.

[96] A. J. Woodhead, H. Angove, M. G. Carr, G. Chessari, M. Congreve, J. E. Coyle, J. Cosme, B. Graham, P. J. Day, R. Downham, L. Fazal, R. Feltell, E. Figueroa, M. Frederickson, J. Lewis, R. McMenamin, C. W. Murray, M. A. O'Brien, L. Parra, S. Patel, T. Phillips, D. C. Rees, S. Rich, D.-M. Smith, G. Trewartha, M. Vinkovic, B. Williams, and A. J.-A. Woolford, "Discovery of (2,4 - Dihydroxy - 5 -isopropyl - phenyl) - [5 - (4 - methylpiperazin - 1 -ylmethyl) - 1,3 - dihydroisoindol - 2 -yl]methanone (AT13387), a Novel Inhibitor of the Molecular Chaperone Hsp90 by Fragment Based Drug Design," *Journal of Medicinal Chemistry*, vol. 53, no. 16, pp. 5956–5969, Aug. 2010, ISSN: 0022-2623. DOI: 10.1021/jm100060b. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/jm100060b.

[97] H. Köster, T. Craan, S. Brass, C. Herhaus, M. Zentgraf, L. Neumann, A. Heine, and G. Klebe, "A small nonrule of 3 compatible fragment library provides high hit rate of endothiapepsin crystal structures with various fragment chemotypes," *Journal of Medicinal Chemistry*, vol. 54, no. 22, pp. 7784–7796, 2011, ISSN: 00222623. DOI: 10.1021/jm200642w.

[98] R. J. Hall, P. N. Mortenson, and C. W. Murray, "Efficient exploration of chemical space by fragment-based screening," *Progress in Biophysics and Molecular Biology*, vol. 116, no. 2-3, pp. 82–91, Nov. 2014, ISSN: 00796107. DOI: 10.1016/j.pbiomolbio.2014.09.007. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0079610714000960.

[99] G. Siegal, E. AB, and J. Schultz, "Integration of fragment screening and library design," *Drug Discovery Today*, vol. 12, no. 23-24, pp. 1032–1039, Dec. 2007, ISSN: 13596446. DOI: 10.1016/j.drudis.2007.08.005. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1359644607003108.

[100] C. Dalvit, "NMR methods in fragment screening: theory and a comparison with other biophysical techniques," *Drug Discovery Today*, vol. 14, no. 21-22, pp. 1051–1057, 2009, ISSN: 13596446. DOI: 10.1016/j.drudis.2009.07.013.

[101] M. Mazanetz, R. Law, and M. Whittaker, "Hit and Lead Identification from Fragments," in *De novo Molecular Design*. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA, Oct. 2013, ch. 6, pp. 143–200, ISBN: 9783527677016. DOI: 10.1002/9783527677016.ch6. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9783527677016.ch6%20http://doi.wiley.com/10.1002/9783527677016.ch6.

[102] M. Whittaker, R. J. Law, O. Ichihara, T. Hesterkamp, and D. Hallett, "Fragments: Past, present and future," *Drug Discovery Today: Technologies*, vol. 7, no. 3, pp. 163–171, 2010, ISSN: 17406749. DOI: 10.1016/j.ddtec.2010.11.007. [Online]. Available: http://dx.doi.org/10.1016/j.ddtec.2010.11.007.

[103] Y. Chen and B. K. Shoichet, "Molecular docking and ligand specificity in fragment-based inhibitor discovery," *Nat. Chem. Biol.*, vol. 5, no. 5, pp. 358–364, 2009. DOI: 10.1038/nchembio.155.

[104] B. Lamoree and R. E. Hubbard, "Current perspectives in fragment-based lead discovery (FBLD)," *Essays In Biochemistry*, vol. 61, no. 5, pp. 453–464, 2017, ISSN: 0071-1365. DOI: 10.1042/EBC20170028.

[105] W. P. Jencks, "On the attribution and additivity of binding energies.," *Proceedings of the National Academy of Sciences of the USA*, vol. 78, no. 7, pp. 4046–4050, 1981, ISSN: 1091-6490. DOI: 10.1073/pnas.78.7.4046. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16593049.

[106] N. Howard, C. Abell, W. Blakemore, G. Chessari, M. Congreve, S. Howard, H. Jhoti, C. W. Murray, L. C. Seavers, and R. L. Van Montfort, "Application of fragment screening and fragment linking to the discovery of novel thrombin inhibitors," *Journal of Medicinal Chemistry*, vol. 49, no. 4, pp. 1346–1355, 2006, ISSN: 00222623. DOI: 10.1021/jm050850v.

[107] A. Potter, V. Oldfield, C. Nunns, C. Fromont, S. Ray, C. J. Northfield, C. J. Bryant, S. F. Scrace, D. Robinson, N. Matossova, L. Baker, P. Dokurno, A. E. Surgenor, B. Davis, C. M. Richardson, J. B. Murray, and J. D. Moore, "Discovery of cell-active phenyl-imidazole Pin1 inhibitors by structure-guided fragment evolution," *Bioorganic and Medicinal Chemistry Letters*, vol. 20, no. 22, pp. 6483–6488, 2010, ISSN: 0960894X. DOI: 10.1016/j.bmcl.2010.09.063. [Online]. Available: http://dx.doi.org/10.1016/j.bmcl.2010.09.063.

[108] D. A. Erlanson, "Fragment-based lead discovery: a chemical update," *Current Opinion in Biotechnology*, vol. 17, no. 6, pp. 643–652, 2006, ISSN: 09581669. DOI: 10.1016/j.copbio.2006.10.007.

[109] E. Edink, P. Rucktooa, K. Retra, A. Akdemir, T. Nahar, O. P. Zuiderveld, R. Van Elk, E. Janssen, P. Van Nierop, J. Van Muijlwijk-Koezen, A. B. Smit, T. K. Sixma, R. Leurs, and I. J. P. De Esch, "Fragment growing induces conformational changes in acetylcholine-binding protein: A structural and thermodynamic analysis," *Journal of the American Chemical Society*, vol. 133, no. 14, pp. 5363–5371, 2011, ISSN: 00027863. DOI: 10.1021/ja110571r.

[110] W. Jahnke, J. M. Rondeau, S. Cotesta, A. Marzinzik, X. Pellé, M. Geiser, A. Strauss, M. Götte, F. Bitsch, R. Hemmig, C. Henry, S. Lehmann, J. F. Glickman, T. P. Roddy, S. J. Stout, and J. R. Green, "Allosteric non-bisphosphonate FPPS inhibitors identified by fragment-based discovery," *Nature Chemical Biology*, vol. 6, no. 9, pp. 660–666, 2010, ISSN: 15524469. DOI: 10.1038/nchembio.421.

[111] S. Roughley, L. Wright, P. Brough, A. Massey, and R. E. Hubbard, "Hsp90 inhibitors and drugs from fragment and virtual screening," *Topics in Current Chemistry*, 2012, ISSN: 03401022. DOI: 10.1007/128-2011-181.

[112] K. Loving, I. Alberts, and W. Sherman, "Computational Approaches for Fragment-Based and De Novo Design," *Current Topics in Medicinal Chemistry*, vol. 10, no. 1, pp. 14–32, 2010, ISSN: 15680266. DOI: 10.2174/156802610790232305. [Online]. Available: http://www.eurekaselect.com/openurl/content.php?genre=article%7B%5C&%7Dissn=1568-0266%7B%5C&%7Dvolume=10%7B%5C&%7Dissue=1%7B%5C&%7Dspage=14.

[113] G. Schneider, *De novo Molecular Design*, 1st ed., G. Schneider, Ed. Weimheim: WILEY-VCH Verlag GmbH & Co. KGaA, 2013, pp. 1–551, ISBN: 9783527333646. DOI: 10.1002/9783527677016.

[114] H. C. Kolb, M. G. Finn, and K. B. Sharpless, "Click Chemistry: Diverse Chemical Function from a Few Good Reactions," *Angewandte Chemie International Edition*, vol. 40, no. 11, pp. 2004–2021, Jun. 2001, ISSN: 1433-7851. DOI: 10.1002/1521-3773(20010601)40:11<2004::AID-ANIE2004>3.0.CO;2-5. arXiv: NIHMS150003. [Online]. Available: http://doi.wiley.com/10.1002/1521-3773%7B%5C%%7D2820010601%7B%5C%%7D2940%7B%5C%%7D3A11%7B%5C%%7D3C2004%7B%5C%%7D3A%7B%5C%%7D3AAID-ANIE2004%7B%5C%%7D3E3.0.CO%7B%5C%%7D3B2-5.

[115] S. C. Schürer, P. Tyagi, and S. M. Muskal, "Prospective exploration of synthetically feasible, medicinally relevant chemical space," *Journal of Chemical Information and Modeling*, vol. 45, no. 2, pp. 239–248, 2005, ISSN: 15499596. DOI: 10.1021/ci0496853.

[116] P. Ertl and A. Schuffenhauer, "Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions.," *Journal of cheminformatics*, vol. 1, no. 8, p. 8, Jun. 2009, ISSN: 1758-2946. DOI: 10.1186/1758-2946-1-8. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20298526%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3225829.

[117] F. Chevillard and P. Kolb, "SCUBIDOO: A Large yet Screenable and Easily Searchable Database of Computationally Created Chemical Compounds Optimized toward High Likelihood of Synthetic Tractability," *Journal of Chemical Information and Modeling*, vol. 55, no. 9, pp. 1824–1835, 2015, ISSN: 15205142. DOI: 10.1021/acs.jcim.5b00203.

[118] L. Friedrich, T. Rodrigues, C. S. Neuhaus, P. Schneider, and G. Schneider, "From Complex Natural Products to Simple Synthetic Mimetics by Computational de Novo Design," *Angewandte Chemie - International Edition*, vol. 55, no. 23, pp. 6789–6792, 2016, ISSN: 15213773. DOI: 10.1002/anie.201601941.

[119] F.-Y. Lin, E. X. Esposito, and Y. J. Tseng, "LeadOp+R: Structure-Based Lead Optimization With Synthetic Accessibility," *Frontiers in Pharmacology*, vol. 9, p. 96, Mar. 2018, ISSN: 1663-9812. DOI: 10.3389/fphar.2018.00096. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fphar.2018.00096/full.

[120] J. J. Irwin and B. K. Shoichet, *ZINC 15 - Catalogs*. [Online]. Available: http://zinc15.docking.org/catalogs (visited on 01/01/2019).

[121] R. J. Bienstock, *Fragment-Based Methods in Drug Discovery*, A. E. Klon, Ed., ser. Methods in Molecular Biology. New York, NY: Springer New York, 2015, vol. 1289, ch. 10, pp. 119–135, ISBN: 978-1-4939-2485-1. DOI: 10.1007/978-1-4939-2486-8. [Online]. Available: http://link.springer.com/10.1007/978-1-4939-2486-8.

[122] G. Schneider, M. L. Lee, M. Stahl, and P. Schneider, "De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks," *Journal of Computer-Aided Molecular Design*, vol. 14, no. 5, pp. 487–494, 2000, ISSN: 0920654X. DOI: 10.1023/A:1008184403558.

[123] R. C. Glen and A. Payne, "A genetic algorithm for the automated generation of molecules within constraints," *Journal of Computer-Aided Molecular Design*, vol. 9, no. 2, pp. 181–202, 1995, ISSN: 0920-654X. DOI: 10.1007/BF00124408. [Online]. Available: https://doi.org/10.1007/BF00124408.

[124] S. Makino, T. J. Ewing, and I. D. Kuntz, "DREAM++: flexible docking program for virtual combinatorial libraries.," *Journal of computer-aided molecular design*, vol. 13, no. 5, pp. 513–32, Sep. 1999, ISSN: 0920-654X. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10483532.

[125] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe, "A fast flexible docking method using an incremental construction algorithm.," *Journal of molecular biology*, vol. 261, no. 3, pp. 470–89, Aug. 1996, ISSN: 0022-2836. DOI: 10.1006/jmbi.1996.0477. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/8780787.

[126] Y. Nishibata and A. Itai, "Automatic creation of drug candidate structures based on receptor structure. Starting point for artificial lead generation.," *Tetrahedron*, vol. 47, no. 43, pp. 8985–8990, Nov. 1991, ISSN: 00404020. DOI: 10.1016/S0040-4020(01)86503-0. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0040402001865030.

[127] S. H. Rotstein and M. A. Murcko, "GenStar: A method for de novo drug design," *Journal of Computer-Aided Molecular Design*, vol. 7, no. 1, pp. 23–43, 1993, ISSN: 0920654X. DOI: 10.1007/BF00141573.

[128] D. A. Pearlman and M. A. Murcko, "CONCEPTS: New dynamic algorithm for de novo drug suggestion," *Journal of Computational Chemistry*, vol. 14, no. 10, pp. 1184–1193, Oct. 1993, ISSN: 1096987X. DOI: 10.1002/jcc.540141008. [Online]. Available: http://doi.wiley.com/10.1002/jcc.540141008.

[129] R. S. Bohacek and C. McMartin, "Multiple Highly Diverse Structures Complementary to Enzyme Binding Sites: Results of Extensive Application of a de Novo Design Method Incorporating Combinatorial Growth," *Journal of the American Chemical Society*, vol. 116, no. 13, pp. 5560–5571, Jun. 1994, ISSN: 0002-7863. DOI: 10.1021/ja00092a006. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/ja00092a006.

[130] I. D. Kuntz, E. C. Meng, and B. K. Shoichet, "Structure-Based Molecular Design," *Accounts of Chemical Research*, vol. 27, no. 5, pp. 117–123, 1994, ISSN: 15204898. DOI: 10.1021/ar00041a001.

[131] J. Robert Fischer, U. Lessel, and M. Rarey, "LoFT: Similarity-driven multiobjective focused library design," *Journal of Chemical Information and Modeling*, vol. 50, no. 1, pp. 1–21, 2010, ISSN: 15499596. DOI: 10.1021/ci900287p.

[132] J. B. Moon and W. J. Howe, "Computer Design of Bioactive Molecules - a Method for Receptor-Based de Novo Ligand Design," *Proteins-Structure Function and Genetics*, vol. 11, pp. 314–328, 1991, ISSN: 0887-3585. [Online]. Available: //a1991gv03200008.

[133] S. H. Rotstein and M. A. Murcko, "GroupBuild: a fragment-based method for de novo drug design.," *Journal of medicinal chemistry*, vol. 36, no. 12, pp. 1700–1710, 1993, ISSN: 02637855. DOI: 10.1016/0263-7855(94)80069-3.

[134] Z. Luo, R. Wang, and L. Lai, "RASSE: A new method for structure-based drug design," *Journal of Chemical Information and Computer Sciences*, vol. 36, no. 6, pp. 1187–1194, 1996, ISSN: 00952338. DOI: 10.1021/ci950277w.

[135] R. Wang, Y. Gao, and L. Lai, "LigBuilder: A Multi-Purpose Program for Structure-Based Drug Design," *Journal of Molecular Modeling*, vol. 6, no. 7-8, pp. 498–516, 2000, ISSN: 1610-2940. DOI: 10.1007/s0089400060498. [Online]. Available: http://link.springer.com/10.1007/s0089400060498.

[136] P. S. Kutchukian, D. Lou, and E. I. Shakhnovich, "FOG: Fragment Optimized Growth Algorithm for the ¡i¿de Novo¡/i¿ Generation of Molecules Occupying Druglike Chemical Space," *Journal of Chemical Information and Modeling*, vol. 49, no. 7, pp. 1630–1642, Jul. 2009, ISSN: 1549-9596. DOI: 10.1021/ci9000458. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/ci9000458.

[137] X. Q. Lewell, D. B. Judd, S. P. Watson, and M. M. Hann, "RECAP - Retrosynthetic Combinatorial Analysis Procedure: A powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry," *Journal of Chemical Information and Computer Sciences*, vol. 38, no. 3, pp. 511–522, 1998, ISSN: 00952338. DOI: 10.1021/ci970429i. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/ci970429i.

[138] A. P. Bento, A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Krüger, Y. Light, L. Mak, S. McGlinchey, M. Nowotka, G. Papadatos, R. Santos, and J. P. Overington, "The ChEMBL bioactivity database: An update," *Nucleic Acids Research*, 2014, ISSN: 03051048. DOI: 10.1093/nar/gkt1031. arXiv: arXiv:1011.1669v3.

[139] H. J. Böhm, "The computer program LUDI: A new method for the de novo design of enzyme inhibitors," *Journal of Computer-Aided Molecular Design*, vol. 6, no. 1, pp. 61–78, Feb. 1992, ISSN: 0920654X. DOI: 10.1007/BF00124387. [Online]. Available: http://link.springer.com/10.1007/BF00124387.

[140] G. Lauri and P. A. Bartlett, "CAVEAT: A program to facilitate the design of organic molecules," *Journal of Computer-Aided Molecular Design*, vol. 8, no. 1, pp. 51–66, Feb. 1994, ISSN: 0920654X. DOI: 10.1007/BF00124349. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/8035213.

[141] A. C. Pierce, G. Rao, and G. W. Bemis, "BREED: Generating novel inhibitors through hybridization of known ligands. Application to CDK2, P38, and HIV protease," *Journal of Medicinal Chemistry*, 2004, ISSN: 00222623. DOI: 10.1021/jm030543u.

[142] D. A. Pearlman and M. A. Murcko, "CONCERTS: Dynamic connection of fragments as an approach to de novo ligand design," *Journal of Medicinal Chemistry*, 1996, ISSN: 00222623. DOI: 10.1021/jm9507921.

[143] G. Schneider, W. Neidhart, T. Giller, and G. Schmid, ""SCAFFOLD-HOPPING" BY TOPOLOGICAL PHARMACOPHORE SEARCH: A CONTRIBUTION TO VIRTUAL SCREENING," *Angewandte Chemie International Edition*, vol. 38, no. 19, pp. 2894–2896, Oct. 1999, ISSN: 1433-7851. DOI: 10.1002/(SICI)1521-3773(19991004)38:19<2894::AID-ANIE2894>3.0.CO;2-F.

[144] H. J. Böhm, A. Flohr, and M. Stahl, "Scaffold hopping," *Drug Discovery Today: Technologies*, 2004, ISSN: 14652080. DOI: 10.1099/mic.0.075689-0.

[145] G. Schneider, P. Schneider, and S. Renner, *Scaffold-hopping: How far can you jump?* 2006. DOI: 10.1002/qsar.200610091.

[146] H. Mauser and W. Guba, "Recent developments in de novo design and scaffold hopping.," *Current Opinion in Drug Discovery & Development*, 2008, ISSN: 1367-6733. DOI: 10.1038/ejhg.2014.33.

[147] B. A. Krueger, A. Dietrich, K.-h. Baringhaus, and G. Schneider, "Scaffold-Hopping Potential of Fragment-Based De Novo Design: The Chances and Limits of Variations," *Combinatorial Chemistry*, 2009, ISSN: 1875-5402. DOI: 10.2174/138620709788167971.

[148] S. R. Langdon, P. Ertl, and N. Brown, *Bioisosteric replacement and scaffold hopping in lead generation and optimization*, 2010. DOI: 10.1002/minf.201000019.

[149] A. Schuffenhauer, "Computational methods for scaffold hopping," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2012, ISSN: 17590876. DOI: 10.1002/wcms.1106.

[150] N. Brown, *Scaffold Hopping in Medicinal Chemistry*, N. Brown, Ed., ser. Methods and Principles in Medicinal Chemistry. Wiley, Dec. 2013, ISBN: 9783527333646. DOI: 10.1002/9783527665143. [Online]. Available: https://onlinelibrary.wiley.com/doi/book/10.1002/9783527665143.

[151] Y. Hu, D. Stumpfe, and J. Bajorath, *Recent Advances in Scaffold Hopping*, 2017. DOI: 10.1021/acs.jmedchem.6b01437. arXiv: 1612.02189.

[152] P. Brear, C. De Fusco, K. Hadje Georgiou, N. J. Francis-Newton, C. J. Stubbs, H. F. Sore, A. R. Venkitaraman, C. Abell, D. R. Spring, and M. Hyvönen, "Specific inhibition of CK2α from an anchor outside the active site," *Chemical Science*, vol. 7, no. 11, pp. 6839–6845, 2016, ISSN: 20416539. DOI: 10.1039/c6sc02335e.

[153] C. De Fusco, P. Brear, J. Iegre, K. H. Georgiou, H. F. Sore, M. Hyvönen, and D. R. Spring, "A fragment-based approach leading to the discovery of a novel binding site and the selective CK2 inhibitor CAM4066," *Bioorganic and Medicinal Chemistry*, vol. 25, no. 13, pp. 3471–3482, 2017, ISSN: 14643391. DOI: 10.1016/j.bmc.2017.04.037. [Online]. Available: http://dx.doi.org/10.1016/j.bmc.2017.04.037.

[154] H. J. Böhm, "LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads," *Journal of Computer-Aided Molecular Design*, vol. 6, no. 6, pp. 593–606, 1992, ISSN: 0920654X. DOI: 10.1007/BF00126217.

[155] W. Kabsch, "A discussion of the solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A*, 1978, ISSN: 16005724. DOI: 10.1107/S0567739478001680. arXiv: 05677394.

[156] F. Dey and A. Caflisch, "Fragment-Based de Novo Ligand Design by Multiobjective Evolutionary Optimization," *Journal of Chemical Information and Modeling*, vol. 48, no. 3, pp. 679–690, 2008. DOI: 10.1021/CI700424B. [Online]. Available: https://pubs.acs.org/doi/abs/10.1021/ci700424b.

[157] F. Glover, "Tabu Search: A Tutorial," *Interfaces*, 1990, ISSN: 0092-2102. DOI: 10.1287/inte.20.4.74.

[158] F. Glover, "Tabu Search—Part II," *ORSA Journal on Computing*, 1990, ISSN: 0899-1499. DOI: 10.1287/ijoc.2.1.4. arXiv: arXiv:1011.1669v3.

[159] R. V. Devi, S. S. Sathya, and M. S. Coumar, "Evolutionary algorithms for de novo drug design - A survey," *Applied Soft Computing Journal*, vol. 27, pp. 543–552, 2015, ISSN: 15684946. DOI: 10.1016/j.asoc.2014.09.042.

[160] D. Douguet, E. Thoreau, and G. Grassy, "A genetic algorithm for the automated generation of small organic molecules: Drug design using an evolutionary algorithm," *Journal of Computer-Aided Molecular Design*, vol. 14, no. 5, pp. 449–466, 2000, ISSN: 0920654X. DOI: 10.1023/A:1008108423895.

[161] Y. Yuan, J. Pei, and L. Lai, "LigBuilder 2: A practical de novo drug design approach," *Journal of Chemical Information and Modeling*, vol. 51, no. 5, pp. 1083–1091, 2011, ISSN: 15499596. DOI: 10.1021/ci100350u.

[162] E. Shang, Y. Yuan, X. Chen, Y. Liu, J. Pei, and L. Lai, "De novo design of multitarget ligands with an iterative fragment-growing strategy," *Journal of Chemical Information and Modeling*, vol. 54, no. 4, pp. 1235–1241, 2014, ISSN: 15205142. DOI: 10.1021/ci500021v.

[163] H. M. Vinkers, M. R. de Jonge, F. F. D. Daeyaert, J. Heeres, L. M. H. Koymans, J. H. van Lenthe, P. J. Lewi, H. Timmerman, K. V. Aken, and P. A. J. Janssen, "SYNOPSIS: SYNthesize and OPtimize System in Silico," 2003. DOI: 10.1021/JM030809X. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/jm030809x.

[164] E.-W. Lameijer, J. N. Kok, T. Bäck, and A. P. IJzerman, "The Molecule Evoluator. An Interactive Evolutionary Algorithm for the Design of Drug-Like Molecules," *Journal of Chemical Information and Modeling*, vol. 46, no. 2, pp. 545–552, 2006. DOI: 10.1021/CI050369D. [Online]. Available: https://pubs.acs.org/doi/abs/10.1021/ci050369d.

[165] J. D. Durrant, R. E. Amaro, and J. A. McCammon, "AutoGrow: A Novel Algorithm for Protein Inhibitor Design," *Chemical Biology and Drug Design*, vol. 73, no. 2, pp. 168–178, 2009. DOI: doi:10.1111/j.1747-0285.2008.00761.x..

[166] N. Majeux, M. Scarsi, J. Apostolakis, C. Ehrhardt, and A. Caflisch, "Exhaustive docking of molecular fragments with electrostatic solvation," *Proteins: Structure, Function, and Genetics*, vol. 37, no. 1, pp. 88–105, Oct. 1999, ISSN: 0887-3585. DOI: 10.1002/(SICI)1097-0134(19991001)37:1<88::AID-PROT9>3.0.CO;2-0. [Online]. Available: http://doi.wiley.com/10.1002/%7B%5C%%7D28SICI%7B%5C%%7D291097-0134%7B%5C%%7D2819991001%7B%5C%%7D2937%7B%5C%%7D3A1%7B%5C%%7D3C88%7B%5C%%7D3A%7B%5C%%7D3AAID-PROT9%7B%5C%%7D3E3.0.CO%7B%5C%%7D3B2-0.

[167] N. Majeux, M. Scarsi, and A. Caflisch, "Efficient Electrostatic Solvation Model for Protein-Fragment Docking," *Proteins Structure Function And Bioinformatics*, vol. 42, no. 2, pp. 256–268, 2001.

[168] J. D. Durrant and J. A. McCammon, "AutoClickChem: Click Chemistry in Silico," *PLoS Computational Biology*, vol. 8, no. 3, H. Lapp, Ed., e1002397, Mar. 2012, ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1002397. arXiv: PMC3305364. [Online]. Available: https://dx.plos.org/10.1371/journal.pcbi.1002397.

[169] *Daylight Theory: SMARTS - A Language for Describing Molecular Patterns*, 2018. [Online]. Available: http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (visited on 10/05/2018).

[170] R. A. Lewis and P. M. Dean, "Automated site-directed drug design: the concept of spacer skeletons for primary structure generation," *Proceedings of the Royal Society B: Biological Sciences*, vol. 236, no. 1283, pp. 125–140, 1989, ISSN: 14712970. DOI: 10.1098/rspb.1989.0017.

[171] R. A. Lewis and P. M. Dean, "Automated Site-Directed Drug Design: The Formation of Molecular Templates in Primary Structure Generation," *Proceedings of the Royal Society B: Biological Sciences*, vol. 236, no. 1283, pp. 141–162, Mar. 1989, ISSN: 0962-8452. DOI: 10.1098/rspb.1989.0018. [Online]. Available: http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.1989.0018.

[172] R. A. Lewis, "Automated site-directed drug design: Approaches to the formation of 3D molecular graphs," *Journal of Computer-Aided Molecular Design*, vol. 4, no. 2, pp. 205–210, 1990, ISSN: 0920654X. DOI: 10.1007/BF00125319.

[173] N. P. Todorov and P. M. Dean, "Evaluation of a method for controlling molecular scaffold diversity in de novo ligand design," *Journal of Computer-Aided Molecular Design*, vol. 11, no. 2, pp. 175–192, 1997, ISSN: 0920654X. DOI: 10.1023/A:1008042711516.

[174] W.-D. Ihlenfeldt and J. Gasteiger, "Computer-Assisted Planning of Organic Syntheses: The Second Generation of Programs," *Angewandte Chemie International Edition in English*, vol. 34, no. 2324, pp. 2613–2633, 1996, ISSN: 0570-0833. DOI: 10.1002/anie.199526131. [Online]. Available: http://doi.wiley.com/10.1002/anie.199526131.

[175] K. Boda, T. Seidel, and J. Gasteiger, "Structure and reaction based evaluation of synthetic accessibility," *Journal of Computer-Aided Molecular Design*, vol. 21, no. 6, pp. 311–325, Jul. 2007, ISSN: 0920-654X. DOI: 10.1007/s10822-006-9099-2. [Online]. Available: http://link.springer.com/10.1007/s10822-006-9099-2.

[176] P. Bonnet, "Is chemical synthetic accessibility computationally predictable for drug and lead-like molecules? A comparative assessment between medicinal and computational chemists," *European Journal of Medicinal Chemistry*, vol. 54, pp. 679–689, Aug. 2012, ISSN: 02235234. DOI: 10.1016/j.ejmech.2012.06.024. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0223523412003765.

[177] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1988, ISBN: 0-934613-73-7.

[178] J. Degen, C. Wegscheid-Gerlach, A. Zaliani, and M. Rarey, "On the art of compiling and using 'drug-like' chemical fragment spaces," *ChemMedChem*, vol. 3, no. 10, pp. 1503–1507, 2008, ISSN: 18607179. DOI: 10.1002/cmdc.200800178.

[179] U. Fechner and G. Schneider, "Flux (1): A virtual synthesis scheme for fragment-based de novo design," *Journal of Chemical Information and Modeling*, vol. 46, no. 2, pp. 699–707, 2006, ISSN: 15499596. DOI: 10.1021/ci0503560.

[180] U. Fechner and G. Schneider, "Flux (2): Comparison of molecular mutation and crossover operators for ligand-based de novo design," *Journal of Chemical Information and Modeling*, vol. 47, no. 2, pp. 656–667, 2007, ISSN: 15499596. DOI: 10.1021/ci6005307.

[181] J. Degen and M. Rarey, "FLEXNOVO: Structure-based searching in large fragment spaces," *ChemMedChem*, vol. 1, no. 8, pp. 854–868, 2006, ISSN: 18607179. DOI: 10.1002/cmdc.200500102.

[182] F. Lauck and M. Rarey, "FSees: Customized Enumeration of Chemical Subspaces with Limited Main Memory Consumption," *Journal of Chemical Information and Modeling*, vol. 56, no. 9, pp. 1641–1653, 2016, ISSN: 15205142. DOI: 10.1021/acs.jcim.6b00117.

[183] G. van Rossum, "Python tutorial," Centrum voor Wiskunde en Informatica (CWI), Amsterdam, Tech. Rep. CS-R9526, May 1995.

[184] A. Massarotti, A. Brunco, G. Sorba, and G. C. Tron, "ZINClick: A database of 16 million novel, patentable, and readily synthesizable 1,4-disubstituted triazoles," *Journal of Chemical Information and Modeling*, vol. 54, no. 2, pp. 396–406, 2014, ISSN: 15499596. DOI: 10.1021/ci400529h.

[185] D. Levré, C. Arcisto, V. Mercalli, and A. Massarotti, "ZINClick v.18: Expanding Chemical Space of 1,2,3-Triazoles," *Journal of Chemical Information and Modeling*, acs.jcim.8b00615, 2018, ISSN: 1549-9596. DOI: 10.1021/acs.jcim.8b00615. [Online]. Available: http://pubs.acs.org/doi/10.1021/acs.jcim.8b00615.

[186] *Daylight Theory: SMIRKS - A Reaction Transform Language*. [Online]. Available: http://www.daylight.com/dayhtml/doc/theory/theory.smirks.html (visited on 10/05/2018).

[187] G. Landrum, *RDKit: Open-source Cheminformatics*, 2006. DOI: 10.2307/3592822. [Online]. Available: http://www.rdkit.org (visited on 07/15/2015).

[188] F. H. Reisen, G. Schneider, and E. Proschak, "Reaction-MQL: Line Notation for Functional Transformation," *Journal of Chemical Information and Modeling*, vol. 49, no. 1, pp. 6–12, Jan. 2009, ISSN: 1549-9596. DOI: 10.1021/ci800215t. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/ci800215t.

[189] G. L. Holliday, P. Murray-Rust, and H. S. Rzepa, "Chemical Markup, XML, and the World Wide Web. 6. CMLReact, an XML Vocabulary for Chemical Reactions," *Journal of Chemical Information and Modeling*, vol. 46, no. 1, pp. 145–157, 2005. DOI: 10.1021/CI0502698. [Online]. Available: https://pubs.acs.org/doi/abs/10.1021/ci0502698.

[190] N. Kochev, S. Avramova, and N. Jeliazkova, "Ambit-SMIRKS: a software module for reaction representation, reaction search and structure transformation," *Journal of Cheminformatics*, vol. 10, no. 1, p. 42, Dec. 2018, ISSN: 1758-2946. DOI: 10.1186/s13321-018-0295-6. [Online]. Available: https://jcheminf.springeropen.com/articles/10.1186/s13321-018-0295-6.

[191] A. Evers, G. Hessler, L.-h. Wang, S. Werrel, P. Monecke, and H. Matter, "CROSS: An Efficient Workflow for Reaction-Driven Rescaffolding and Side-Chain Optimization Using Robust Chemical Reactions and Available Reagents," *Journal of Medicinal Chemistry*, vol. 56, no. 11, pp. 4656–4670, Jun. 2013, ISSN: 0022-2623. DOI: 10.1021/jm400404v. [Online]. Available: http://pubs.acs.org/doi/10.1021/jm400404v.

[192] L. Hoffer, C. Chira, G. Marcou, A. Varnek, and D. Horvath, *S4MPLE-Sampler for multiple protein-ligand entities: Methodology and rigid-site docking benchmarking*, 5. 2015, vol. 20, pp. 8997–9028, ISBN: 3368793470. DOI: 10.3390/molecules20058997.

[193] A. R. Leach and I. D. Kuntz, "Conformational analysis of flexible ligands in macromolecular receptor sites," *Pharmaceutical Chemistry Journal*, vol. 24, no. 10, pp. 706–711, 1990. DOI: 10.1002/jcc.540130608. [Online]. Available: https://doi.org/10.1002/jcc.540130608.

[194] I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin, "A geometric approach to macromolecule-ligand interactions," *Journal of Molecular Biology*, vol. 161, no. 2, pp. 269–288, 1982, ISSN: 0022-2836. DOI: 10.1016/0022-2836(82)90153-X. [Online]. Available: https://www.sciencedirect.com/science/article/pii/002228368290153X?via%7B%5C%%7D3Dihub.

[195] P. C. Hawkins, "Conformation Generation: The State of the Art," *Journal of Chemical Information and Modeling*, vol. 57, no. 8, pp. 1747–1756, Aug. 2017, ISSN: 1549-9596. DOI: 10.1021/acs.jcim.7b00221. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/acs.jcim.7b00221.

[196] A. Griewel, O. Kayser, J. Schlosser, and M. Rarey, "Conformational sampling for large-scale virtual screening: accuracy versus ensemble size," *Journal of Chemical Information and Modeling*, vol. 49, no. 10, pp. 2303–2311, 2009, ISSN: 15499596. DOI: 10.1021/ci9002415.

[197] C. Schärfer, T. Schulz-Gasch, J. Hert, L. Heinzerling, B. Schulz, T. Inhester, M. Stahl, and M. Rarey, "CONFECT: Conformations from an expert collection of torsion patterns," *ChemMedChem*, vol. 8, no. 10, pp. 1690–1700, 2013, ISSN: 18607179. DOI: 10.1002/cmdc.201300242.

[198] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *The Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, 1953, ISSN: 00219606. DOI: 10.1063/1.1699114. arXiv: 5744249209.

[199] G. D. S. and O. A. J., "Automated docking of substrates to proteins by simulated annealing," *Proteins: Structure, Function, and Bioinformatics*, vol. 8, no. 3, pp. 195–202, 1990, ISSN: 0887-3585. DOI: 10.1002/prot.340080302. [Online]. Available: https://doi.org/10.1002/prot.340080302.

[200] J. Y. Trosset and H. a. Scheraga, "Reaching the global minimum in docking simulations: a Monte Carlo energy minimization approach using Bezier splines.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 14, pp. 8011–8015, 1998, ISSN: 0027-8424. DOI: 10.1073/pnas.95.14.8011.

[201] S. Kirkpatrick, C. J. Gelatt, and M. Vecchi, "Optimization by Simulated Annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983.

[202] R. S. DeWitte and E. I. Shakhnovich, "SMoG: De novo design method based on simple, fast, and accurate free energy estimates. 1. Methodology and supporting evidence," *Journal of the American Chemical Society*, vol. 118, no. 47, pp. 11 733–11 744, 1996, ISSN: 00027863. DOI: 10.1021/ja960751u.

[203] B. Parent, A. Kökösy, and D. Horvath, "Optimized Evolutionary Strategies in Conformational Sampling," *Soft Computing*, vol. 11, no. 1, pp. 63–79, Jan. 2007, ISSN: 1432-7643. DOI: 10.1007/s00500-006-0053-y. [Online]. Available: http://link.springer.com/10.1007/s00500-006-0053-y.

[204] G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor, "Development and validation of a genetic algorithm for flexible docking," *Journal of Molecular Biology*, vol. 267, no. 3, pp. 727–748, Apr. 1997, ISSN: 00222836. DOI: 10.1006/jmbi.1996.0897. arXiv: arXiv:1011.1669v3. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022283696908979?via%7B%5C%%7D3Dihub.

[205] 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 Lamarckian genetic algorithm and an empirical binding free energy function," *Journal of Computational Chemistry*, vol. 19, no. 14, pp. 1639–1662, 1998, ISSN: 01928651. DOI: 10.1002/(SICI)1096-987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B. arXiv: NIHMS150003.

[206] O. Trott and A. J. Olson, "AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading," *Journal of Computational Chemistry*, vol. 31, no. 2, pp. 455–461, 2010, ISSN: 01928651. DOI: 10.1002/jcc.21334. arXiv: NIHMS150003. [Online]. Available: http://doi.wiley.com/10.1002/jcc.21334.

[207] S. Ruiz-Carmona, D. Alvarez-Garcia, N. Foloppe, A. B. Garmendia-Doval, S. Juhos, P. Schmidtke, X. Barril, R. E. Hubbard, and S. D. Morley, "rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids," *PLoS Computational Biology*, vol. 10, no. 4, pp. 1–7, 2014, ISSN: 15537358. DOI: 10.1371/journal.pcbi.1003571.

[208] D. Horvath and C. Jeandenans, "Neighborhood Behavior of in Silico Structural Spaces with Respect to in Vitro Activity Spaces - A Novel Understanding of the Molecular Similarity Principle in the Context of Multiple Receptor Binding Profiles," *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 2, pp. 680–690, Mar. 2003, ISSN: 0095-2338. DOI: 10.1021/ci025634z. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/ci025634z.

[209] M. M. Mysinger and B. K. Shoichet, "Rapid Context-Dependent Ligand Desolvation in Molecular Docking," *Journal of Chemical Information and Modeling*, vol. 50, no. 9, pp. 1561–1573, Sep. 2010, ISSN: 1549-9596. DOI: 10.1021/ci100214a. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/ci100214a.

[210] B. D. Bursulaya, M. Totrov, R. Abagyan, and C. L. Brooks, "Comparative study of several algorithms for flexible ligand docking," *Journal of Computer-Aided Molecular Design*, vol. 17, no. 11, pp. 755–763, 2003, ISSN: 0920654X. DOI: 10.1023/B:JCAM.0000017496.76572.6f.

[211] R. Wang, Y. Lu, and S. Wang, "Comparative evaluation of 11 scoring functions for molecular docking," *Journal of Medicinal Chemistry*, vol. 46, no. 12, pp. 2287–2303, 2003, ISSN: 00222623. DOI: 10.1021/jm0203783.

[212] P. C. Hawkins, B. P. Kelley, and G. L. Warren, "The application of statistical methods to cognate docking: A path forward?" *Journal of Chemical Information and Modeling*, vol. 54, no. 5, pp. 1339–1355, 2014, ISSN: 15205142. DOI: 10.1021/ci5001086.

[213] J. W. M. Nissink, C. W. Murray, M. Hartshorn, M. L. Verdonk, J. C. Cole, and R. Taylor, "A new test set for validating predictions of protein-ligand interaction," *Proteins: Structure, Function and Genetics*, vol. 49, no. 4, pp. 457–471, 2002, ISSN: 08873585. DOI: 10.1002/prot.10232.

[214] M. J. Hartshorn, M. L. Verdonk, G. Chessari, S. C. Brewerton, W. T. M. Mooij, P. N. Mortenson, and C. W. Murray, "Diverse, high-quality test set for the validation of protein-ligand docking performance," *Journal of Medicinal Chemistry*, vol. 50, no. 4, pp. 726–741, 2007, ISSN: 00222623. DOI: 10.1021/jm061277y.

[215] Y. Li, M. Su, Z. Liu, J. Li, J. Liu, L. Han, and R. Wang, "Assessing protein–ligand interaction scoring functions with the CASF-2013 benchmark," *Nature Protocols*, vol. 13, pp. 666–680, Mar. 2018. [Online]. Available: https://doi.org/10.1038/nprot.2017.114%20http://10.0.4.14/nprot.2017.114%20https://www.nature.com/articles/nprot.2017.114%7B%5C#%7Dsupplementary-information.

[216] S. Malhotra and J. Karanicolas, "Correction to When Does Chemical Elaboration Induce a Ligand To Change Its Binding Mode?" *Journal of Medicinal Chemistry*, vol. 60, no. 13, pp. 5940–5940, Jul. 2017, ISSN: 0022-2623. DOI: 10.1021/acs.jmedchem.7b00868. [Online]. Available: http://pubs.acs.org/doi/10.1021/acs.jmedchem.7b00868.

[217] H. M. Berman, J. D. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic acids research*, vol. 28, no. 1, pp. 235–242, 2000, ISSN: 0305-1048. DOI: 10.1093/nar/28.1.235.

[218] M. N. Drwal, G. Bret, C. Perez, C. Jacquemard, J. Desaphy, and E. Kellenberger, "Structural Insights on Fragment Binding Mode Conservation," *Journal of Medicinal Chemistry*, vol. 61, no. 14, pp. 5963–5973, 2018, ISSN: 15204804. DOI: 10.1021/acs.jmedchem.8b00256.

[219] O. Korb, T. Stützle, and T. E. Exner, "Empirical scoring functions for advanced Protein-Ligand docking with PLANTS," *Journal of Chemical Information and Modeling*, 2009, ISSN: 15499596. DOI: 10.1021/ci800298z.

[220] L. Hoffer and D. Horvath, "S4MPLE - Sampler for multiple protein-ligand entities: Simultaneous docking of several entities," *Journal of Chemical Information and Modeling*, vol. 53, no. 1, pp. 88–102, 2013, ISSN: 15499596. DOI: 10.1021/ci300495r.

[221] B. Holzberger, S. Obeid, W. Welte, K. Diederichs, and A. Marx, "Structural insights into the potential of 4-fluoroproline to modulate biophysical properties of proteins," *Chemical Science*, vol. 3, no. 10, p. 2924, 2012, ISSN: 2041-6520. DOI: 10.1039/c2sc20545a. [Online]. Available: https://www.rcsb.org/structure/4dle%20http://xlink.rsc.org/?DOI=c2sc20545a.

[222] V. Cherezov, D. M. Rosenbaum, M. A. Hanson, S. G. F. Rasmussen, F. S. Thian, T. S. Kobilka, H. Choi, P. Kuhn, W. I. Weis, B. K. Kobilka, and R. C. Stevens, "High-Resolution Crystal Structure of an Engineered Human 2-Adrenergic G Protein-Coupled Receptor," *Science*, vol. 318, no. 5854, pp. 1258–1265, Nov. 2007, ISSN: 0036-8075. DOI: 10.1126/science.1150577. [Online]. Available: https://www.rcsb.org/structure/2rh1%20http://www.sciencemag.org/cgi/doi/10.1126/science.1150577.

[223] J. J. Irwin and B. K. Shoichet, "ZINC - A free database of commercially available compounds for virtual screening," *Journal of Chemical Information and Modeling*, vol. 45, no. 1, pp. 177–182, Jan. 2005, ISSN: 15499596. DOI: 10.1021/ci049714+.

[224] T. Sterling and J. J. Irwin, "ZINC 15 - Ligand Discovery for Everyone," *Journal of Chemical Information and Modeling*, vol. 55, no. 11, pp. 2324–2337, 2015, ISSN: 15205142. DOI: 10.1021/acs.jcim.5b00559. arXiv: 15334406.

[225] J. Erickson, D. Neidhart, J. VanDrie, D. Kempf, X. Wang, D. Norbeck, J. Plattner, J. Rittenhouse, M. Turon, and N. Wideburg, "Design, activity, and 2.8 A crystal structure of a C2 symmetric inhibitor complexed to HIV-1 protease.," *Science*, vol. 249, no. 4968, pp. 527–533, 1990. DOI: 10.2210/PDB9HVP/PDB. [Online]. Available: https://www.rcsb.org/structure/9hvp.

[226] B. L. Hodous, S. D. Geuns-Meyer, P. E. Hughes, B. K. Albrecht, S. Bellon, J. Bready, S. Caenepeel, V. J. Cee, S. C. Chaffee, A. Coxon, M. Emery, J. Fretland, P. Gallant, Y. Gu, D. Hoffman, R. E. Johnson, R. Kendall, J. L. Kim, A. M. Long, M. Morrison, P. R. Olivieri, V. F. Patel, A. Polverino, P. Rose, P. Tempest, L. Wang, D. A. Whittington, and H. Zhao, "Evolution of a highly selective and potent 2-(pyridin-2-yl)-1,3,5-triazine Tie-2 kinase inhibitor," *Journal of Medicinal Chemistry*, vol. 50, no. 4, pp. 611–626, Feb. 2007, ISSN: 00222623. DOI: 10.1021/jm0611071. arXiv: arXiv:1011.

1669v3. [Online]. Available: https://pubs.acs.org/doi/abs/10.1021/jm061107l%20http://pubs.acs.org/doi/abs/10.1021/jm061107l.

[227] A. Bateman, M. J. Martin, C. O'Donovan, M. Magrane, R. Apweiler, E. Alpi, R. Antunes, J. Arganiska, B. Bely, M. Bingley, C. Bonilla, R. Britto, B. Bursteinas, G. Chavali, E. Cibrian-Uhalte, A. Da Silva, M. De Giorgi, T. Dogan, F. Fazzini, P. Gane, L. G. Castro, P. Garmiri, E. Hatton-Ellis, R. Hieta, R. Huntley, D. Legge, W. Liu, J. Luo, A. Macdougall, P. Mutowo, A. Nightingale, S. Orchard, K. Pichler, D. Poggioli, S. Pundir, L. Pureza, G. Qi, S. Rosanoff, R. Saidi, T. Sawford, A. Shypitsyna, E. Turner, V. Volynkin, T. Wardell, X. Watkins, H. Zellner, A. Cowley, L. Figueira, W. Li, H. McWilliam, R. Lopez, I. Xenarios, L. Bougueleret, A. Bridge, S. Poux, N. Redaschi, L. Aimo, G. Argoud-Puy, A. Auchincloss, K. Axelsen, P. Bansal, D. Baratin, M. C. Blatter, B. Boeckmann, J. Bolleman, E. Boutet, L. Breuza, C. Casal-Casas, E. De Castro, E. Coudert, B. Cuche, M. Doche, D. Dornevil, S. Duvaud, A. Estreicher, L. Famiglietti, M. Feuermann, E. Gasteiger, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, F. Jungo, G. Keller, V. Lara, P. Lemercier, D. Lieberherr, T. Lombardot, X. Martin, P. Masson, A. Morgat, T. Neto, N. Nouspikel, S. Paesano, I. Pedruzzi, S. Pilbout, M. Pozzato, M. Pruess, C. Rivoire, B. Roechert, M. Schneider, C. Sigrist, K. Sonesson, S. Staehli, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, A. L. Veuthey, C. H. Wu, C. N. Arighi, L. Arminski, C. Chen, Y. Chen, J. S. Garavelli, H. Huang, K. Laiho, P. McGarvey, D. A. Natale, B. E. Suzek, C. R. Vinayaka, Q. Wang, Y. Wang, L. S. Yeh, M. S. Yerramalla, and J. Zhang, "UniProt: A hub for protein information," *Nucleic Acids Research*, 2015, ISSN: 13624962. DOI: 10.1093/nar/gku989. arXiv: NIHMS150003.

[228] R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, D. E. Shaw, P. Francis, and P. S. Shenkin, "Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy," *Journal of Medicinal Chemistry*, vol. 47, no. 7, pp. 1739–1749, 2004, ISSN: 00222623. DOI: 10.1021/jm0306430. arXiv: arXiv:1011.1669v3.

[229] T. A. Halgren, R. B. Murphy, R. A. Friesner, H. S. Beard, L. L. Frye, W. T. Pollard, and J. L. Banks, "Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening," *Journal of Medicinal Chemistry*, vol. 47, no. 7, pp. 1750–1759, Mar. 2004, ISSN: 0022-2623. DOI: 10.1021/jm030644s. [Online]. Available: https://doi.org/10.1021/jm030644s.

[230] R. A. Friesner, R. B. Murphy, M. P. Repasky, L. L. Frye, J. R. Greenwood, T. A. Halgren, P. C. Sanschagrin, and D. T. Mainz, "Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes," *Journal of Medicinal Chemistry*, vol. 49, no. 21, pp. 6177–6196, Oct. 2006, ISSN: 0022-2623. DOI: 10.1021/jm051256o. [Online]. Available: https://doi.org/10.1021/jm051256o.

[231] Biovia and Accelrys, *Biovia Pipeline Pilot Overview*, 2015. [Online]. Available: http://accelrys.com/products/pipeline-pilot/ (visited on 08/26/2015).

[232] 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," in *SIGKDD Explorations*, vol. 11, Springer, 2009, pp. 26–31, ISBN: 978-3-540-78239-1. DOI: 10.1145/1656274.1656280. [Online]. Available: http://doi.acm.org/10.1145/1656274.1656280%7B%5C%%7D5Cnhttp://centaur.reading.ac.uk/6139/.

[233] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open Babel: An open chemical toolbox," *Journal of Cheminformatics*, vol. 3, no. 1, p. 33, Jan. 2011, ISSN: 1758-2946. DOI: 10.1186/1758-2946-3-33. [Online]. Available: http://www.jcheminf.com/content/3/1/33%20http://jcheminf.springeropen.com/articles/10.1186/1758-2946-3-33.

[234] G. Madhavi Sastry, M. Adzhigirey, T. Day, R. Annabhimoju, and W. Sherman, "Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments," *Journal of Computer-Aided Molecular Design*, vol. 27, no. 3, pp. 221–234, 2013, ISSN: 0920654X. DOI: 10.1007/s10822-013-9644-8. arXiv: NIHMS150003.

[235] "Maestro," New York, NY, 2017.

[236] H. Ngo and L. Kang, "Structure of Xoo1075, a peptide deformylase from Xanthomonas oryzae pv oryze, in complex with fragment 571," *TO BE PUBLISHED*, DOI: 10.2210/PDB5CVP/PDB. [Online]. Available: https://www.rcsb.org/structure/5cvp.

[237] K. Sommer, N.-O. Friedrich, S. Bietz, M. Hilbig, T. Inhester, and M. Rarey, "UNICON: A Powerful and Easy-to-Use Compound Library Converter," *Journal of Chemical Information and Modeling*, vol. 56, no. 6, pp. 1105–1111, 2016, ISSN: 15205142. DOI: 10.1021/acs.jcim.6b00069.

[238] H.-C. Ehrlich and M. Rarey, "Systematic benchmark of substructure search in molecular graphs - From Ullmann to VF2," *Journal of Cheminformatics*, vol. 4, no. 7, pp. 1–17, 2012, ISSN: 17582946. DOI: 10.1186/1758-2946-4-13.

[239] H.-C. Ehrlich, A. M. Henzler, and M. Rarey, "Searching for Recursively Defined Generic Chemical Patterns in Nonenumerated Fragment Spaces," *Journal of Chemical Information and Modeling*, vol. 53, no. 7, pp. 1676–1688, Jul. 2013, ISSN: 1549-9596. DOI: 10.1021/ci400107k. [Online]. Available: http://pubs.acs.org/doi/10.1021/ci400107k.

[240] S. Bietz and M. Rarey, "ASCONA: Rapid Detection and Alignment of Protein Binding Site Conformations," *Journal of Chemical Information and Modeling*, vol. 55, no. 8, pp. 1747–1756, Aug. 2015, ISSN: 15205142. DOI: 10.1021/acs.jcim.5b00210. [Online]. Available: http://pubs.acs.org/doi/10.1021/acs.jcim.5b00210.

[241] G. Guennebaud, B. Jacob, *et al.*, *Eigen v3*, http://eigen.tuxfamily.org, 2010.

[242] J. L. Blanco and P. K. Rai, *nanoflann: a {C}++ header-only fork of {FLANN}, a library for Nearest Neighbor ({NN}) wih KD-trees*, 2014. [Online]. Available: https://github.com/jlblancoc/nanoflann.

[243] S. R. Heller, A. McNaught, I. Pletnev, S. Stein, and D. Tchekhovskoi, "InChI, the IUPAC International Chemical Identifier," *Journal of Cheminformatics*, vol. 7, no. 1, p. 23, Dec. 2015, ISSN: 17582946. DOI: 10.1186/s13321-015-0068-4. [Online]. Available: http://www.jcheminf.com/content/7/1/23.

[244] *Qt - a cross-platform application and ui framework.* (visited on 10/14/2018).

[245] J. D. Westbrook, C. Shao, Z. Feng, M. Zhuravleva, S. Velankar, and J. Young, "The chemical component dictionary: Complete descriptions of constituent molecules in experimentally determined 3D macromolecules in the Protein Data Bank," *Bioinformatics*, vol. 31, no. 8, pp. 1274–1278, Apr. 2015, ISSN: 14602059. DOI: 10.1093/bioinformatics/btu789. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/31/8/1274.long.

[246] T. Lippert and M. Rarey, "Fast automated placement of polar hydrogen atoms in protein-ligand complexes," *Journal of Cheminformatics*, vol. 1, no. 1, p. 13, 2009, ISSN: 1758-2946. DOI: 10.1186/1758-2946-1-13. [Online]. Available: https://doi.org/10.1186/1758-2946-1-13.

[247] T. Inhester, "Mining of Interaction Geometries in Collections of Protein Structures," PhD Thesis, Universität Hamburg, 2017.

[248] David Weininger, A. Weininger, and J. L. Weininger, "SMILES . 2 . Algorithm for Generation of Unique SMILES Notation," *Journal of chemical information and computer sciences*, vol. 29, no. 2, pp. 97–101, May 1989, ISSN: 1549-9596. DOI: 10.1021/ci00062a008. [Online]. Available: http://pubs.acs.org/cgi-bin/doilookup/?10.1021/ci00062a008.

[249] H. L. Morgan, "The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service.," *Journal of Chemical Documentation*, vol. 5, no. 2, pp. 107–113, 1965. DOI: 10.1021/c160017a018. [Online]. Available: https://doi.org/10.1021/c160017a018.

[250] S. Umeyama, "Least-Squares Estimation of Transformation Parameters Between Two Point Patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 4, pp. 376–380, 1991, ISSN: 01628828. DOI: 10.1109/34.88573.

[251] S. Bietz, "Methoden zur computergestützten Generierung und Aufbereitung von Strukturensembles für Proteinbindetaschen," PhD Thesis, Universität Hamburg, 2016.

[252] M. Hilbig, S. Urbaczek, I. Groth, S. Heuser, and M. Rarey, "MONA - Interactive manipulation of molecule collections," *Journal of Cheminformatics*, vol. 5, no. 8, pp. 1–10, 2013, ISSN: 17582946. DOI: 10.1186/1758-2946-5-38.

[253] L. Cordella, P. Foggia, C. Sansone, and M. Vento, "A (sub)graph isomorphism algorithm for matching large graphs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 10, pp. 1367–1372, Oct. 2004, ISSN: 0162-8828. DOI: 10.1109/TPAMI.2004.75. [Online]. Available: http://ieeexplore.ieee.org/document/1323804/.

[254] R. Schmidt, "Efficient incremental search of variable molecular patterns.," Master thesis in computer science, Universität Hamburg, 2017.

[255] T. Inhester, "Generation of Small-Molecule 3d Coordinates for High-Throughput Applications," Master thesis in computer science, Universität Hamburg, 2012.

[256] R. J. Gillespie, "The valence-shell electron-pair repulsion (VSEPR) theory of directed valency," *Journal of Chemical Education*, vol. 40, no. 6, p. 295, Jun. 1963, ISSN: 0021-9584. DOI: 10.1021/ed040p295. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/ed040p295.

[257] C. R. Groom, I. J. Bruno, M. P. Lightfoot, and S. C. Ward, "The Cambridge structural database," *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*, vol. 72, no. 2, pp. 171–179, 2016, ISSN: 20525206. DOI: 10.1107/S2052520616003954.

[258] M. D. Eldridge, C. W. Murray, T. R. Auton, G. V. Paolini, and R. P. Mee, "Empirical scoring functions .1. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes," *Journal of Computer-Aided Molecular Design*, vol. 11, no. 5, pp. 425–445, 1997, ISSN: 0920-654X. DOI: 10.1023/a:1007996124545.

[259] R. T. Marler and J. S. Arora, "The weighted sum method for multi-objective optimization: New insights," *Structural and Multidisciplinary Optimization*, vol. 41, no. 6, pp. 853–862, 2010, ISSN: 1615147X. DOI: 10.1007/s00158-009-0460-7.

[260] Y. V. Borodina, E. Bolton, F. Fontaine, and S. H. Bryant, "Assessment of conformational ensemble sizes necessary for specific resolutions of coverage of conformational space," *Journal of Chemical Information and Modeling*, vol. 47, no. 4, pp. 1428–1437, 2007, ISSN: 15499596. DOI: 10.1021/ci7000956. arXiv: NIHMS150003.

[261] D. M. Wong, H. M. Greenblatt, H. Dvir, P. R. Carlier, Y.-F. Han, Y.-P. Pang, I. Silman, and J. L. Sussman, "Acetylcholinesterase Complexed with Bivalent Ligands Related to Huperzine A: Experimental Evidence for Species-Dependent Protein-Ligand Complementarity," *Journal of the American Chemical Society*, vol. 125, no. 2, pp. 363–373, Jan. 2003, ISSN: 0002-7863. DOI: 10.1021/ja021111w. [Online]. Available: https://pubs.acs.org/doi/10.1021/ja021111w%20https://www.rcsb.org/structure/1h22.

[262] M. L. Verdonk, P. N. Mortenson, R. J. Hall, M. J. Hartshorn, and C. W. Murray, "Protein-ligand docking against non-native protein conformers," *Journal of Chemical Information and Modeling*, vol. 48, no. 11, pp. 2214–2225, 2008, ISSN: 15499596. DOI: 10.1021/ci8002254.

[263] F. Jiang and S. H. Kim, ""SOFT DOCKING": MATCHING OF MOLECULAR SURFACE CUBES," *Journal of Molecular Biology*, vol. 219, no. 1, pp. 79–102, 1991, ISSN: 00222836. DOI: 10.1016/0022-2836(91)90859-5.

[264] R. Abagyan, M. Totrov, and D. Kuznetsov, "ICM - A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation," *Journal of Computational Chemistry*, vol. 15, no. 5, pp. 488–506, 1994, ISSN: 1096987X. DOI: 10.1002/jcc.540150503. [Online]. Available: http://doi.wiley.com/10.1002/jcc.540150503.

[265] C. B. Barber, D. Dobkin, and H. Huhdanpaa, "The Quickhull Algorithm for Convex Hulls," *ACM Transactions on Mathematical Software*, vol. 22, no. 4, pp. 469–483, 1996. DOI: 10.1145/235815.235821.

[266] A. M. Henzler, "Modellierung molekularer und makromolekularer Zustände im geleiteten strukturbasierten virtuellen Screening Dissertation zur Erlangung des Doktorgrades," PhD thesis, University of Hamburg, 2015.

[267] B. D. Hudson, R. M. Hyde, E. Rahr, J. Wood, and J. Osman, "Parameter based methods for compound selection from chemical databases," *Quantitative Structure-Activity Relationships*, vol. 15, no. 4, pp. 285–289, 1996, ISSN: 09318771. DOI: 10.1002/qsar.19960150402.

[268] N.-O. Friedrich, F. Flachsenberg, A. Meyder, K. Sommer, J. Kirchmair, and M. Rarey, "Conformator: A Novel Method for the Generation of Conformer Ensembles," *Journal of Chemical Information and Modeling*, vol. 59, no. 2, acs.jcim.8b00704, Feb. 2019, ISSN: 1549-9596. DOI: 10.1021/acs.jcim.8b00704. [Online]. Available: http://pubs.acs.org/doi/10.1021/acs.jcim.8b00704.

[269] H.-J. Böhm, "Towards the automatic design of synthetically accessible protein ligands: Peptides, amides and peptidomimetics," *Journal of Computer-Aided Molecular Design*, vol. 10, no. 4, pp. 265–272, Aug. 1996, ISSN: 0920-654X. DOI: 10.1007/BF00124496. [Online]. Available: http://link.springer.com/10.1007/BF00124496.

[270] M. P. Repasky, R. B. Murphy, J. L. Banks, J. R. Greenwood, I. Tubert-Brohman, S. Bhat, and R. A. Friesner, "Docking performance of the glide program as evaluated on the Astex and DUD datasets: A complete set of glide SP results and selected results for a new scoring function integrating WaterMap and glide," *Journal of Computer-Aided Molecular Design*, vol. 26, no. 6, pp. 787–799, 2012, ISSN: 0920654X. DOI: 10.1007/s10822-012-9575-9.

[271] S. Bietz and M. Rarey, "SIENA: Efficient Compilation of Selective Protein Binding Site Ensembles," *Journal of Chemical Information and Modeling*, vol. 56, no. 1, pp. 248–259, Jan. 2016, ISSN: 15205142. DOI: 10.1021/acs.jcim.5b00588. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/acs.jcim.5b00588.

[272] R. Fährrolfes, S. Bietz, F. Flachsenberg, A. Meyder, E. Nittinger, T. Otto, A. Volkamer, and M. Rarey, "ProteinsPlus: a web portal for structure analysis of macromolecules," *Nucleic Acids Research*, vol. 45, no. W1, pp. 1–7, Jul. 2017, ISSN: 0305-1048. DOI: 10.1093/nar/gkx333. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkx333.

[273] *The PyMOL Molecular Graphics System, Version 2.2*, New York, NY, Nov. 2015.

[274] R. Powers, F. Morandi, and B. K. Shoichet, "Structure-based discovery of a novel, noncovalent inhibitor of AmpC beta-lactamase.," *Structure*, vol. 10, pp. 1013–1023, 2002. DOI: 10.2210/PDB1L2S/PDB. [Online]. Available: https://www.rcsb.org/structure/1l2s.

[275] K. Babaoglu and B. K. Shoichet, "Deconstructing fragment-based inhibitor discovery," *Nature Chemical Biology*, vol. 2, no. 12, pp. 720–723, Dec. 2006, ISSN: 1552-4450. DOI: 10.1038/nchembio831. [Online]. Available: http://www.nature.com/doifinder/10.1038/nchembio831.

[276] S. Maignan, J. Guilloteau, S. Pouzieux, Y. Choi-Sledeski, M. Becker, S. Klein, W. Ewing, H. Pauls, A. Spada, and V. Mikol, "Crystal structures of human factor Xa complexed with potent inhibitors.," *J.Med.Chem.*, vol. 43, pp. 3226–3232, 2000. DOI: 10.2210/PDB1F0T/PDB. [Online]. Available: https://www.rcsb.org/structure/1f0t.

[277] J. Yamane, M. Yao, Y. Zhou, Y. Hiramatsu, K. Fujiwara, T. Yamaguchi, H. Yamaguchi, H. Togame, H. Tsujishita, H. Takemoto, and I. Tanaka, "In-crystal affinity ranking of fragment hit compounds reveals a relationship with their inhibitory activities," *J.Appl.Crystallogr.*, vol. 44, pp. 798–804, 2011. DOI: 10.2210/PDB3RXG/PDB. [Online]. Available: https://www.rcsb.org/structure/3rxg%20https://www.rcsb.org/structure/3atk.

[278] Z.-K. Wan, J. Lee, W. Xu, D. V. Erbe, D. Joseph-McCarthy, B. C. Follows, and Y.-L. Zhang, "Monocyclic thiophenes as protein tyrosine phosphatase 1B inhibitors: Capturing interactions with Asp48," *Bioorganic & Medicinal Chemistry Letters*, vol. 16, no. 18, pp. 4941–4945, Sep. 2006, ISSN: 0960-894X. DOI: 10.1016/J.BMCL.2006.06.051. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960894X06007190?via%7B%5C%%7D3Dihub.

[279] D. P. Wilson, Z.-K. Wan, W.-X. Xu, S. J. Kirincich, B. C. Follows, D. Joseph-McCarthy, K. Foreman, A. Moretto, J. Wu, M. Zhu, E. Binnun, Y.-L. Zhang, M. Tam, D. V. Erbe, J. Tobin, X. Xu, L. Leung, A. Shilling, S. Y. Tam, T. S. Mansour, and J. Lee, "Structure-Based Optimization of Protein Tyrosine Phosphatase 1B Inhibitors: From the Active Site to the Second Phosphotyrosine Binding Site," *Journal of Medicinal Chemistry*, vol. 50, no. 19, pp. 4681–4698, Sep. 2007, ISSN: 0022-2623. DOI: 10.1021/jm0702478. [Online]. Available: https://pubs.acs.org/doi/abs/10.1021/jm0702478%20http://pubs.acs.org/doi/abs/10.1021/jm0702478.

[280] R. Morales, S. Perrier, J.-M. Florent, J. Beltra, S. Dufour, I. De Mendez, P. Manceau, A. Tertre, F. Moreau, D. Compere, A.-C. Dublanchet, and M. O'Gara, "Crystal Structures of Novel Non-peptidic, Non-zinc Chelating Inhibitors Bound to MMP-12," *Journal of Molecular Biology*, vol. 341, no. 4, pp. 1063–1076, Aug. 2004, ISSN: 00222836. DOI: 10.1016/j.jmb.2004.06.039. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022283604007302?via%7B%5C%%7D3Dihub%20https://linkinghub.elsevier.com/retrieve/pii/S0022283604007302.

[281] I. P. Holmes, S. Gaines, S. P. Watson, O. Lorthioir, A. Walker, S. J. Baddeley, S. Herbert, D. Egan, M. A. Convery, O. M. Singh, J. W. Gross, J. M. Strelow, R. H. Smith, A. J. Amour, D. G. Brown, and S. L. Martin, "The identification of β-hydroxy carboxylic acids as selective MMP-12 inhibitors," *Bioorganic & Medicinal Chemistry Letters*, vol. 19, no. 19, pp. 5760–5763, Oct. 2009, ISSN: 0960894X. DOI: 10.1016/j.bmcl.2009.07.155. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960894X09011238?via%7B%5C%%7D3Dihub%20https://linkinghub.elsevier.com/retrieve/pii/S0960894X09011238.

[282] E. Jones, T. Oliphant, P. Peterson, *et al.*, *SciPy: Open source scientific tools for Python*. [Online]. Available: http://www.scipy.org/.

[283] F. Perez and B. E. Granger, "IPython: A System for Interactive Scientific Computing," *Computing in Science Engineering*, vol. 9, no. 3, pp. 21–29, May 2007, ISSN: 1521-9615. DOI: 10.1109/MCSE.2007.53.

[284] R. Saito, J. M. Pruet, L. A. Manzano, K. Jasheway, A. F. Monzingo, P. A. Wiget, I. Kamat, E. V. Anslyn, and J. D. Robertus, "Peptide-Conjugated Pterins as Inhibitors of Ricin Toxin A," *Journal of Medicinal Chemistry*, vol. 56, no. 1, pp. 320–329, Jan. 2013, ISSN: 0022-2623. DOI: 10.1021/jm3016393. [Online]. Available: http://pubs.acs.org/doi/10.1021/jm3016393.

[285] H. C. A. Raaijmakers, J. E. Versteegh, and J. C. Uitdehaag, "The X-ray Structure of RU486 Bound to the Progesterone Receptor in a Destabilized Agonistic Conformation," *Journal of Biological Chemistry*, vol. 284, no. 29, pp. 19 572–19 579, Jul. 2009, ISSN: 0021-9258. DOI: 10.1074/jbc.M109.007872. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19372222%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2740583%20http://www.jbc.org/lookup/doi/10.1074/jbc.M109.007872.

[286] M. D. Jacobs, J. Black, O. Futer, L. Swenson, B. Hare, M. Fleming, and K. Saxena, "Pim-1 Ligand-bound Structures Reveal the Mechanism of Serine/Threonine Kinase Inhibition by LY294002," *Journal of Biological*

Chemistry, vol. 280, no. 14, pp. 13 728–13 734, Apr. 2005, ISSN: 0021-9258. DOI: 10.1074/jbc.M413155200. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15657054%20http://www.jbc.org/lookup/doi/10.1074/jbc.M413155200.

[287] B. A. Katz, C. Luong, J. D. Ho, J. R. Somoza, E. Gjerstad, J. Tang, S. R. Williams, E. Verner, R. L. Mackman, W. B. Young, P. A. Sprengeler, H. Chan, K. Mortara, J. W. Janc, and M. E. McGrath, "Dissecting and Designing Inhibitor Selectivity Determinants at the S1 Site Using an Artificial Ala190 Protease (Ala190 uPA)," *Journal of Molecular Biology*, vol. 344, no. 2, pp. 527–547, Nov. 2004, ISSN: 00222836. DOI: 10.1016/j.jmb.2004.09.032. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022283604011751?via%7B%5C%%7D3Dihub%20https://linkinghub.elsevier.com/retrieve/pii/S0022283604011751.

[288] M. Aguilar, T. M. Gloster, M. I. García-Moreno, C. Ortiz Mellet, G. J. Davies, A. Llebaria, J. Casas, M. Egido-Gabás, and J. M. García Fernandez, "Molecular Basis for β-Glucosidase Inhibition by Ring-Modified Calystegine Analogues," *ChemBioChem*, vol. 9, no. 16, pp. 2612–2618, Nov. 2008, ISSN: 14394227. DOI: 10.1002/cbic.200800451. [Online]. Available: http://doi.wiley.com/10.1002/cbic.200800451.

[289] T. M. Gloster, P. Meloncelli, R. V. Stick, D. Zechel, A. Vasella, and G. J. Davies, "Glycosidase inhibition: An assessment of the binding of 18 putative transition-state mimics," *Journal of the American Chemical Society*, vol. 129, no. 8, pp. 2345–2354, 2007, ISSN: 00027863. DOI: 10.1021/ja066961g.

[290] M. Aguilar, T. M. Gloster, M. I. García-Moreno, O. Mellet, G. J. Davies, A. Llebaria, J. Casas, M. Egido-Gabás, and J. M. García, "Molecular Basis for b-Glucosidase Inhibition by Ring-Modified Calystegine Analogues," *European Journal of Chemical Biology*, vol. 9, no. 16, pp. 2612–2618, 2008. DOI: 10.1002/cbic.200800451.

[291] L. E. Tailford, W. A. Offen, N. L. Smith, C. Dumon, C. Morland, J. Gratien, M. P. Heck, R. V. Stick, Y. Blériot, A. Vasella, H. J. Gilbert, and G. J. Davies, "Structural and biochemical evidence for a boat-like transition state in β-mannosidases," *Nature Chemical Biology*, vol. 4, no. 5, pp. 306–312, 2008, ISSN: 15524469. DOI: 10.1038/nchembio.81.

[292] D. A. Kuntz, C. A. Tarling, S. G. Withers, and D. R. Rose, "Structural Analysis of Golgi α-Mannosidase II Inhibitors Identified from a Focused Glycosidase Inhibitor Screen," *Biochemistry*, vol. 47, no. 38, pp. 10 058–10 068, Sep. 2008, ISSN: 0006-2960. DOI: 10.1021/bi8010785. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/bi8010785.

[293] J. R. Somoza, J. Ho, C. Luong, M. Ghate, P. Sprengeler, K. Mortara, W. Shrader, D. Sperandio, H. Chan, M. McGrath, and B. A. Katz, "The Structure of the Extracellular Region of Human Hepsin Reveals a Serine Protease Domain and a Novel Scavenger Receptor Cysteine-Rich (SRCR) Domain," *Structure*, vol. 11, pp. 1123–1131, 2003. DOI: 10.2210/PDB1P57/PDB. [Online]. Available: https://www.rcsb.org/structure/1p57.

[294] C. W. Murray, M. G. Carr, O. Callaghan, G. Chessari, M. Congreve, S. Cowan, J. E. Coyle, R. Downham, E. Figueroa, M. Frederickson, B. Graham, R. McMenamin, M. A. O'Brien, S. Patel, T. R. Phillips, G. Williams, A. J. Woodhead, and A. J.-A. Woolford, "Fragment-based drug discovery applied to Hsp90. Discovery of two lead series with high ligand efficiency," *Journal of Medicinal Chemistry*, vol. 53, no. 16, pp. 5942–5955, 2010, ISSN: 00222623. DOI: 10.1021/jm100059d.

[295] J. Bussenius, C. M. Blazey, N. Aay, N. K. Anand, A. Arcalas, T. Baik, O. J. Bowles, C. A. Buhr, S. Costanzo, J. K. Curtis, S. C. DeFina, L. Dubenko, T. S. Heuer, P. Huang, C. Jaeger, A. Joshi, A. R. Kennedy, A. I. Kim, K. Lara, J. Lee, J. Li, J. C. Lougheed, S. Ma, S. Malek, J.-C. L. Manalo, J.-F. Martini, G. McGrath, M. Nicoll, J. M. Nuss, M. Pack, C. J. Peto, T. H. Tsang, L. Wang, S. W. Womble, M. Yakes, W. Zhang, and K. D. Rice, "Discovery of XL888: A novel tropane-derived small molecule inhibitor of HSP90," *Bioorganic & Medicinal Chemistry Letters*, vol. 22, no. 17, pp. 5396–5404, Sep. 2012, ISSN: 0960-894X. DOI: 10.1016/J.BMCL.2012.07.052. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960894X12009286?via%7B%5C%%7D3Dihub.

[296] Y. He, A. K. Bubb, K. A. Stubbs, T. M. Gloster, and G. J. Davies, "Inhibition of a bacterial O-GlcNAcase homologue by lactone and lactam derivatives: structural, kinetic and thermodynamic analyses," *Amino Acids*, vol. 40, no. 3, pp. 829–839, Mar. 2011, ISSN: 0939-4451. DOI: 10.1007/s00726-010-0700-6. [Online]. Available: http://link.springer.com/10.1007/s00726-010-0700-6.

[297] K. A. Stubbs, T. W. James, Y. He, D. J. Vocadlo, B. L. Mark, G. J. Davies, and M. D. Balcewich, "Insight into a strategy for attenuating AmpC-mediated β-lactam resistance: Structural basis for selective inhibition of the glycoside hydrolase NagZ," *Protein Science*, vol. 18, no. 7, pp. 1541–1551, 2009. DOI: 10.1002/pro.137.

[298] D. R. Davies, B. Mamat, O. T. Magnusson, J. Christensen, M. H. Haraldsson, R. Mishra, B. Pease, E. Hansen, J. Singh, D. Zembower, H. Kim, A. S. Kiselyov, A. B. Burgin, M. E. Gurney, and L. J. Stewart, "Discovery of Leukotriene A4 Hydrolase Inhibitors Using Metabolomics Biased Frag-

ment Crystallography †," *Journal of Medicinal Chemistry*, vol. 52, no. 15, pp. 4694–4715, Aug. 2009, ISSN: 0022-2623. DOI: 10.1021/jm900259h. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/jm900259h.

[299] M. Uttamchandani, H. Schuler, V. Suresh, T. Karlberg, C. Q. Pan, K. Thangavelu, B. C. Low, G. Balaji, and J. Sivaraman, "Structural basis for the allosteric inhibitory mechanism of human kidney-type glutaminase (KGA) and its regulation by Raf-Mek-Erk signaling in cancer cell metabolism," *Proceedings of the National Academy of Sciences*, vol. 109, no. 20, pp. 7705–7710, 2012, ISSN: 0027-8424. DOI: 10.1073/pnas.1116573109.

[300] B. DeLaBarre, S. Gross, C. Fang, Y. Gao, A. Jha, F. Jiang, J. Song J., W. Wei, and J. B. Hurov, "Full-Length Human Glutaminase in Complex with an Allosteric Inhibitor," *Biochemistry*, vol. 50, no. 50, pp. 10 764–10 770, Dec. 2011, ISSN: 0006-2960. DOI: 10.1021/bi201613d. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/bi201613d.

[301] E. Nittinger, N. Schneider, G. Lange, and M. Rarey, "Evidence of Water Molecules—A Statistical Evaluation of Water Molecules Based on Electron Density," *Journal of Chemical Information and Modeling*, vol. 55, no. 4, pp. 771–783, Apr. 2015, ISSN: 1549-9596. DOI: 10.1021/ci500662d. [Online]. Available: https://pubs.acs.org/doi/10.1021/ci500662d.

[302] A. Meyder, E. Nittinger, G. Lange, R. Klein, and M. Rarey, "Estimating Electron Density Support for Individual Atoms and Molecular Fragments in X-ray Structures," *Journal of Chemical Information and Modeling*, vol. 57, no. 10, pp. 2437–2447, Oct. 2017, ISSN: 1549-9596. DOI: 10.1021/acs.jcim.7b00391. [Online]. Available: http://pubs.acs.org/doi/10.1021/acs.jcim.7b00391.

[303] M. R. Groves, Z.-J. Yao, P. P. Roller, Burke,Terrence R. Jr., and D. Barford, "Structural Basis for Inhibition of the Protein Tyrosine Phosphatase 1B by Phosphotyrosine Peptide Mimetics," *Biochemistry*, vol. 37, no. 51, pp. 17 773–17 783, 1998. DOI: 10.1021/BI9816958. [Online]. Available: https://pubs.acs.org/doi/10.1021/bi9816958.

[304] E. Asante-Appiah, S. Patel, C. Desponts, J. M. Taylor, C. Lau, C. Dufresne, M. Therien, R. Friesen, J. W. Becker, Y. Leblanc, B. P. Kennedy, and G. Scapin, "Conformation-assisted inhibition of protein-tyrosine phosphatase-1B elicits inhibitor selectivity over T-cell protein-tyrosine phosphatase.," *The Journal of biological chemistry*, vol. 281, no. 12, pp. 8010–8015, Mar. 2006, ISSN: 0021-9258. DOI: 10.1074/jbc.M511827200. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16407290.

[305] T. A. Brandão, S. J. Johnson, and A. C. Hengge, "The Molecular Details of WPD-Loop Movement Differ in the Protein-Tyrosine Phosphatases YopH and PTP1B," *Archives of Biochemistry and Biophysics*, vol. 525, no. 1, pp. 53–59, 2012. DOI: 10.1016/j.abb.2012.06.002.The.

[306] R. M. Craparo, "Significance Level," in *Encyclopedia of Measurement and Statistics*, N. J. Salkind, Ed., Thousand Oaks, California: SAGE PublicationsSage CA: Los Angeles, CA, 2007, ch. 3, pp. 889–891, ISBN: 978-1-412-91611-0. DOI: 10.4135/9781412952644. [Online]. Available: https://methods.sagepub.com/reference/encyclopedia-of-measurement-and-statistics.

[307] Z. Zhu, Z. Y. Sun, Y. Ye, J. Voigt, C. Strickland, E. M. Smith, J. Cumming, L. Wang, J. Wong, Y. S. Wang, D. F. Wyss, X. Chen, R. Kuvelkar, M. E. Kennedy, L. Favreau, E. Parker, B. A. McKittrick, A. Stamford, M. Czarniecki, W. Greenlee, and J. C. Hunter, "Discovery of cyclic acylguanidines as highly potent and selective β-site amyloid cleaving enzyme (BACE) inhibitors: Part I - Inhibitor design and validation," *Journal of Medicinal Chemistry*, vol. 53, no. 3, pp. 951–965, 2010, ISSN: 00222623. DOI: 10.1021/jm901408p.

[308] M. D. Lloyd, N. Thiyagarajan, Y. T. Ho, L. W. L. Woo, O. B. Sutcliffe, A. Purohit, M. J. Reed, K. R. Acharya, and B. V. L. Potter, "First crystal structures of human carbonic anhydrase II in complex with dual aromatase-steroid sulfatase inhibitors.," *Biochemistry*, vol. 44, no. 18, pp. 6858–6866, 2005, ISSN: 0006-2960. DOI: 10.1021/bi047692e. [Online]. Available: http://opus.bath.ac.uk/7902/.

[309] E. Čapkauskaitė, A. Zubrienė, L. Baranauskienė, G. Tamulaitienė, E. Manakova, V. Kairys, S. Gražulis, S. Tumkevičius, and D. Matulis, "Design of [(2-pyrimidinylthio) acetyl] benzenesulfonamides as inhibitors of human carbonic anhydrases," *European Journal of Medicinal Chemistry*, vol. 51, pp. 259–270, May 2012, ISSN: 0223-5234. DOI: 10.1016/J.EJMECH.2012.02.050. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0223523412001419?via%7B%5C%%7D3Dihub.

[310] E. Čapkauskaitė, L. Baranauskienė, D. Golovenko, E. Manakova, S. Gražulis, S. Tumkevičius, and D. Matulis, "Indapamide-like benzenesulfonamides as inhibitors of carbonic anhydrases I, II, VII, and XIII," *Bioorganic & Medicinal Chemistry*, vol. 18, no. 21, pp. 7357–7364, Nov. 2010, ISSN: 0968-0896. DOI: 10.1016/J.BMC.2010.09.016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S096808961000831X?via%7B%5C%%7D3Dihub.

[311] H. Nar, M. Bauer, A. Schmid, J. Stassen, W. Wienen, H. Priepke, I. Kauffmann, U. Ries, and N. Hauel, "Structural Basis for Inhibition Promiscuity of Dual Specific Thrombin and Factor Xa Blood Coagulation Inhibitors," *Structure*, vol. 9, pp. 29–38, 2001. DOI: 10.2210/PDB1OYQ/PDB. [Online]. Available: https://www.rcsb.org/structure/1oyq.

[312] J. C. f. S. G. (JCSG), "Crystal structure of putative acetyltransferase (YP_001815201.1) from EXIGUOBACTERIUM SP. 255-15 at 1.62 A resolution," *TO BE PUBLISHED*, DOI: 10.2210/PDB3GYA/PDB. [Online]. Available: https://www.rcsb.org/structure/3gya.

[313] P. Bar-on, C. Millard, M. Harel, H. Dvir, A. Enz, J. L. Sussman, and I. Silman, "Kinetic and Structural Studies on the Interaction of Cholinesterases with the Anti-Alzheimer Drug Rivastigmine," *Biochemistry*, vol. 41, p. 3555, 2002. DOI: 10.2210/PDB1GQS/PDB. [Online]. Available: https://www.rcsb.org/structure/1gqs.

[314] H. M. Greenblatt, G. Kryger, T. Lewis, I. Silman, and J. L. Sussman, "Structure of Acetylcholinesterase Complexed with (-)-Galanthamine at 2.3A Resolution," *FEBS Lett.*, vol. 463, p. 321, 1999. DOI: 10.2210/PDB1DX6/PDB. [Online]. Available: https://www.rcsb.org/structure/1dx6.

[315] S. Howard, V. Berdini, J. A. Boulstridge, M. G. Carr, D. M. Cross, J. Curry, L. A. Devine, T. R. Early, L. Fazal, A. L. Gill, M. Heathcote, S. Maman, J. E. Matthews, R. L. McMenamin, E. F. Navarro, M. A. O'Brien, M. O'Reilly, D. C. Rees, M. Reule, D. Tisi, G. Williams, M. Vinković, and P. G. Wyatt, "Fragment-based discovery of the pyrazol-4-yl urea (AT9283), a multitargeted kinase inhibitor with potent aurora kinase activity," *Journal of Medicinal Chemistry*, vol. 52, no. 2, pp. 379–388, 2009, ISSN: 00222623. DOI: 10.1021/jm800984v.

[316] J. X. Qiao, C.-H. Chang, D. L. Cheney, P. E. Morin, G. Z. Wang, S. R. King, T. C. Wang, A. R. Rendina, J. M. Luettgen, R. M. Knabb, R. R. Wexler, and P. Y. Lam, "SAR and X-ray structures of enantiopure 1,2-cis-(1R,2S)-cyclopentyldiamine and cyclohexyldiamine derivatives as inhibitors of coagulation Factor Xa," *Bioorganic & Medicinal Chemistry Letters*, vol. 17, no. 16, pp. 4419–4427, Aug. 2007, ISSN: 0960-894X. DOI: 10.1016/J.BMCL.2007.06.029. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960894X0700710X?via%7B%5C%%7D3Dihub.

[317] A. Berndt, S. Miller, O. Williams, D. D. Le, B. T. Houseman, J. I. Pacold, F. Gorrec, W. C. Hon, Y. Liu, C. Rommel, P. Gaillard, T. Rückle, M. K. Schwarz, K. M. Shokat, J. P. Shaw, and R. L. Williams, "The p110δ

structure: Mechanisms for selectivity and potency of new PI(3)K inhibitors," *Nature Chemical Biology*, vol. 6, no. 2, pp. 117–124, Feb. 2010, ISSN: 15524469. DOI: 10.1038/nchembio.293. [Online]. Available: http://www.nature.com/articles/nchembio.293.

[318] T. O. Fischmann, A. Hruza, J. S. Duca, L. Ramanathan, T. Mayhood, W. T. Windsor, H. V. Le, T. J. Guzi, M. P. Dwyer, K. Paruch, R. J. Doll, E. Lees, D. Parry, W. Seghezzi, and V. Madison, "Structure-guided discovery of cyclin - dependent kinase inhibitors," *Biopolymers*, vol. 89, no. 5, pp. 372–379, May 2008, ISSN: 00063525. DOI: 10.1002/bip.20868. [Online]. Available: http://doi.wiley.com/10.1002/bip.20868.

[319] P. G. Wyatt, A. J. Woodhead, V. Berdini, J. A. Boulstridge, M. G. Carr, D. M. Cross, D. J. Davis, L. A. Devine, T. R. Early, R. E. Feltell, E. J. Lewis, R. L. McMenamin, E. F. Navarro, M. A. O'Brien, M. O'Reilly, M. Reule, G. Saxty, L. C. A. Seavers, D.-M. Smith, M. S. Squires, G. Trewartha, M. T. Walker, A. J.-A. Woolford, M. A. O'Brien, and M. O'Reilly, "Identification of N-(4-piperidinyl)-4-(2,6 - dichlorobenzoylamino) - 1H - pyrazole-3-carboxamide (AT7519), a novel cyclin dependent kinase inhibitor using fragment-based X-ray crystallography and structure based drug design," *Journal of Medicinal Chemistry*, vol. 51, no. 16, pp. 4986–4999, Aug. 2008, ISSN: 00222623. DOI: 10.1021/jm800382h. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/jm800382h.

[320] T. Waku, T. Shiraki, T. Oyama, K. Maebara, R. Nakamori, and K. Morikawa, "The nuclear receptor PPARγ individually responds to serotonin-and fatty acid-metabolites," *EMBO Journal*, vol. 29, no. 19, pp. 3395–3407, 2010, ISSN: 02614189. DOI: 10.1038/emboj.2010.197. [Online]. Available: http://dx.doi.org/10.1038/emboj.2010.197.

[321] T. Miura, T. A. Fukami, K. Hasegawa, N. Ono, A. Suda, H. Shindo, D. O. Yoon, S. J. Kim, Y. J. Na, Y. Aoki, N. Shimma, T. Tsukuda, and Y. Shiratori, "Lead generation of heat shock protein 90 inhibitors by a combination of fragment-based approach, virtual screening, and structure-based drug design," *Bioorganic and Medicinal Chemistry Letters*, vol. 21, no. 19, pp. 5778–5783, 2011, ISSN: 0960894X. DOI: 10.1016/j.bmcl.2011.08.001. [Online]. Available: http://dx.doi.org/10.1016/j.bmcl.2011.08.001.

[322] P. A. Brough, X. Barril, J. Borgognoni, P. Chene, N. G. M. Davies, B. Davis, M. J. Drysdale, B. Dymock, S. A. Eccles, C. Garcia-Echeverria, C. Fromont, A. Hayes, R. E. Hubbard, A. M. Jordan, M. R. Jensen, A. Massey, A. Merrett, A. Padfield, R. Parsons, T. Radimerski, F. I. Raynaud, A. Robertson, S. D. Roughley, J. Schoepfer, H. Simmonite, S. Y. Sharp, A. Surgenor, M. Valenti, S. Walls, P. Webb, M. Wood, P. Workman, and

L. Wright, "Combining Hit Identification Strategies: Fragment-Based and in Silico Approaches to Orally Active 2-Aminothieno[2,3- d]pyrimidine Inhibitors of the Hsp90 Molecular Chaperone," *Journal of Medicinal Chemistry*, vol. 52, no. 15, pp. 4794–4809, Aug. 2009, ISSN: 0022-2623. DOI: 10.1021/jm900357y. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/jm900357y.

[323] S. M. Saalau-Bethell, A. J. Woodhead, G. Chessari, M. G. Carr, J. Coyle, B. Graham, S. D. Hiscock, C. W. Murray, P. Pathuri, S. J. Rich, C. J. Richardson, P. A. Williams, and H. Jhoti, "Discovery of an allosteric mechanism for the regulation of HCV NS3 protein function," *Nature Chemical Biology*, vol. 8, no. 11, pp. 920–925, Nov. 2012, ISSN: 1552-4450. DOI: 10.1038/nchembio.1081. [Online]. Available: http://www.nature.com/articles/nchembio.1081.

[324] D. L. Cheney, J. M. Bozarth, W. J. Metzler, P. E. Morin, L. Mueller, J. A. Newitt, A. H. Nirschl, A. R. Rendina, J. K. Tamura, A. Wei, X. Wen, N. R. Wurtz, D. A. Seiffert, R. R. Wexler, and E. S. Priestley, "Discovery of novel P1 groups for coagulation factor VIIa inhibition using fragment-based screening," *Journal of Medicinal Chemistry*, vol. 58, no. 6, pp. 2799–2808, 2015, ISSN: 15204804. DOI: 10.1021/jm501982k.

[325] J. A. Borthwick, N. Ancellin, S. M. Bertrand, R. P. Bingham, P. S. Carter, C. W. Chung, I. Churcher, N. Dodic, C. Fournier, P. L. Francis, A. Hobbs, C. Jamieson, S. D. Pickett, S. E. Smith, D. O. Somers, C. Spitzfaden, C. J. Suckling, and R. J. Young, "Structurally Diverse Mitochondrial Branched Chain Aminotransferase (BCATm) Leads with Varying Binding Modes Identified by Fragment Screening," *Journal of Medicinal Chemistry*, vol. 59, no. 6, pp. 2452–2467, 2016, ISSN: 15204804. DOI: 10.1021/acs.jmedchem.5b01607.

[326] G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, and A. J. Olson, "AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility," *Journal of Computational Chemistry*, vol. 30, no. 16, pp. 2785–2791, Dec. 2009, ISSN: 01928651. DOI: 10.1002/jcc.21256. arXiv: NIHMS150003. [Online]. Available: http://doi.wiley.com/10.1002/jcc.21256.

[327] S. D. Roughley and A. M. Jordan, "The Medicinal Chemist's Toolbox: An Analysis of Reactions Used in the Pursuit of Drug Candidates," *Journal of Medicinal Chemistry*, vol. 54, no. 10, pp. 3451–3479, May 2011, ISSN: 0022-2623. DOI: 10.1021/jm200187y. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/jm200187y.

[328] C. A. Baxter, C. W. Murray, D. E. Clark, D. R. Westhead, and M. D. Eldridge, "Flexible docking using Tabu search and an empirical estimate of binding affinity," *Proteins: Structure, Function and Genetics*, vol. 33, no. 3, pp. 367–382, 1998, ISSN: 08873585. DOI: 10.1002/(SICI)1097-0134(19981115)33:3<367::AID-PROT6>3.0.CO;2-W.

[329] J. Kolb, B. Beck, M. Almstetter, S. Heck, E. Herdtweck, and A. Dömling, "New MCRs: The first 4-component reaction leading to 2,4-disubstituted thiazoles," *Molecular Diversity*, vol. 6, no. 3-4, pp. 297–313, 2003, ISSN: 13811991. DOI: 10.1023/B:MODI.0000006827.35029.e4.

[330] D. Weigelt and I. Dorange, "Lead Generation Based on Compound Collection Screening," in *Lead Generation: Methods, Strategies, and Case Studies*, J. Holenz, Ed., 1st ed., Weinheim: WILEY-VCH Verlag, 2016, ch. 5, pp. 95–132, ISBN: 9783527677047.

[331] M. Hilbig and M. Rarey, "MONA 2: A Light Cheminformatics Platform for Interactive Compound Library Processing," *Journal of Chemical Information and Modeling*, vol. 55, no. 10, pp. 2071–2078, Oct. 2015, ISSN: 15205142. DOI: 10.1021/acs.jcim.5b00292. [Online]. Available: http://dx.doi.org/10.1021/acs.jcim.5b00292.

[332] M. Su, Q. Yang, Y. Du, G. Feng, Z. Liu, Y. Li, and R. Wang, "Comparative Assessment of Scoring Functions: The CASF-2016 Update," *Journal of Chemical Information and Modeling*, vol. 59, no. 2, pp. 895–913, Feb. 2019, ISSN: 1549-9596. DOI: 10.1021/acs.jcim.8b00545. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/30481020%20http://pubs.acs.org/doi/10.1021/acs.jcim.8b00545.

[333] A. K. Ghose, V. N. Viswanadhan, and J. J. Wendoloski, "A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases.," *Journal of combinatorial chemistry*, vol. 1, no. 1, pp. 55–68, Jan. 1999, ISSN: 1520-4766. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10746014.

[334] C. A. Nicolaou and N. Brown, "Multi-objective optimization methods in drug design," *Drug Discovery Today: Technologies*, vol. 10, no. 3, pp. 1–9, 2013, ISSN: 17406749. DOI: 10.1016/j.ddtec.2013.02.001. [Online]. Available: http://dx.doi.org/10.1016/j.ddtec.2013.02.001.

[335] A. Unzue, M. Xu, J. Dong, L. Wiedmer, D. Spiliotopoulos, A. Caflisch, and C. Nevado, "Fragment-Based Design of Selective Nanomolar Ligands of the CREBBP Bromodomain," *Journal of Medicinal Chemistry*, vol. 59, no. 4, pp. 1350–1356, 2016, ISSN: 15204804. DOI: 10.1021/acs.jmedchem.5b00172.

[336] F. Morandi, E. Caselli, S. Morandi, P. J. Focia, J. Blázquez, B. K. Shoichet, and F. Prati, "Nanomolar Inhibitors of AmpC β-Lactamase," *Journal of the American Chemical Society*, vol. 125, no. 3, pp. 685–695, Jan. 2003, ISSN: 0002-7863. DOI: 10.1021/ja0288338. [Online]. Available: https://pubs.acs.org/doi/10.1021/ja0288338.

[337] E. Caselli, R. Powers, L. Blasczcak, C. Wu, F. Prati, and B. K. Shoichet, "Energetic, structural, and antimicrobial analyses of beta-lactam side chain recognition by beta-lactamases.," *Chem.Biol.*, vol. 8, pp. 17–31, 2001. DOI: 10.2210/PDB1FSW/PDB. [Online]. Available: https://www.rcsb.org/structure/1fsw.

[338] H. Deng, T. D. Bannister, L. Jin, R. E. Babine, J. Quinn, P. Nagafuji, C. A. Celatka, J. Lin, T. I. Lazarova, M. J. Rynkiewicz, F. Bibbins, P. Pandey, J. Gorga, H. V. Meyers, S. S. Abdel-Meguid, and J. E. Strickler, "Synthesis, SAR exploration, and X-ray crystal structures of factor XIa inhibitors containing an α-ketothiazole arginine," *Bioorganic & Medicinal Chemistry Letters*, vol. 16, no. 11, pp. 3049–3054, Jun. 2006, ISSN: 0960894X. DOI: 10.1016/j.bmcl.2006.02.052. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960894X06002356?via%7B%5C%%7D3Dihub%20https://linkinghub.elsevier.com/retrieve/pii/S0960894X06002356.

[339] M. S. Buchanan, A. R. Carroll, D. Wessling, M. Jobling, V. M. Avery, R. A. Davis, Y. Feng, Y. Xue, L. Öster, T. Fex, J. Deinum, J. N. A. Hooper, and R. J. Quinn, "Clavatadine A, A Natural Product with Selective Recognition and Irreversible Inhibition of Factor XIa †," *Journal of Medicinal Chemistry*, vol. 51, no. 12, pp. 3583–3587, Jun. 2008, ISSN: 0022-2623. DOI: 10.1021/jm800314b. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/jm800314b.

[340] O. Eidam, C. Romagnoli, E. Caselli, K. Babaoglu, D. T. Pohlhaus, J. Karpiak, R. Bonnet, B. K. Shoichet, and F. Prati, "Design, Synthesis, Crystal Structures, and Antimicrobial Activity of Sulfonamide Boronic Acids as β-Lactamase Inhibitors," *Journal of Medicinal Chemistry*, vol. 53, no. 21, pp. 7852–7863, Nov. 2010, ISSN: 0022-2623. DOI: 10.1021/jm101015z. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/jm101015z.

[341] O. Eidam, C. Romagnoli, G. Dalmasso, S. Barelier, E. Caselli, R. Bonnet, B. K. Shoichet, and F. Prati, "Fragment-guided design of subnanomolar -lactamase inhibitors active in vivo," *Proceedings of the National Academy of Sciences*, vol. 109, no. 43, pp. 17 448–17 453, Oct. 2012, ISSN: 0027-8424. DOI: 10.1073/pnas.1208337109. [Online]. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.1208337109.

[342] Z. Rankovic, J. Cai, J. Kerr, X. Fradera, J. Robinson, A. Mistry, W. Finlay, G. McGarry, F. Andrews, W. Caulfield, I. Cumming, M. Dempster, J. Waller, W. Arbuckle, M. Anderson, I. Martin, A. Mitchell, C. Long, M. Baugh, P. Westwood, E. Kinghorn, P. Jones, J. C. Uitdehaag, M. van Zeeland, D. Potin, L. Saniere, A. Fouquet, F. Chevallier, H. Deronzier, C. Dorleans, and E. Nicolai, "Optimisation of 2-cyano-pyrimidine inhibitors of cathepsin K: Improving selectivity over hERG," *Bioorganic & Medicinal Chemistry Letters*, vol. 20, no. 21, pp. 6237–6241, Nov. 2010, ISSN: 0960894X. DOI: 10.1016/j.bmcl.2010.08.101. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960894X10012370?via%7B%5C%%7D3Dihub%20https://linkinghub.elsevier.com/retrieve/pii/S0960894X10012370.

[343] D. Tondi, R. Powers, E. Caselli, M. Negri, J. Blazquez, M. Costi, and B. K. Shoichet, "Structure-based design and in-parallel synthesis of inhibitors of AmpC beta-lactamase.," *Chem.Biol.*, vol. 8, pp. 593–611, 2001. DOI: 10.2210/PDB1GA9/PDB. [Online]. Available: http://www.rcsb.org/structure/1GA9.

[344] K. C. Usher, L. C. Blaszczak, G. S. Weston, B. K. Shoichet, and S. J. Remington, "Three-Dimensional Structure of AmpC β-Lactamase from Escherichia c oli Bound to a Transition-State Analogue: Possible Implications for the Oxyanion Hypothesis and for Inhibitor Design," *Biochemistry*, vol. 37, no. 46, pp. 16 082–16 092, Nov. 1998, ISSN: 0006-2960. DOI: 10.1021/bi981210f. [Online]. Available: https://pubs.acs.org/doi/abs/10.1021/bi981210f%20http://pubs.acs.org/doi/abs/10.1021/bi981210f.

[345] M. Getlik, C. Grütter, J. R. Simard, S. Klüter, M. Rabiller, H. B. Rode, A. Robubi, and D. Rauh, "Hybrid Compound Design To Overcome the Gatekeeper T338M Mutation in cSrc #," *Journal of Medicinal Chemistry*, vol. 52, no. 13, pp. 3915–3926, Jul. 2009, ISSN: 0022-2623. DOI: 10.1021/jm9002928. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/jm9002928.

[346] J. A. Blair, D. Rauh, C. Kung, C.-H. Yun, Q.-W. Fan, H. Rode, C. Zhang, M. J. Eck, W. A. Weiss, and K. M. Shokat, "Structure-guided development of affinity probes for tyrosine kinases using chemical genetics," *Nature Chemical Biology*, vol. 3, no. 4, pp. 229–238, Apr. 2007, ISSN: 1552-4450. DOI: 10.1038/nchembio866. [Online]. Available: http://www.nature.com/articles/nchembio866.

[347] K. Ahn, D. S. Johnson, M. Mileni, D. Beidler, J. Z. Long, M. K. McKinney, E. Weerapana, N. Sadagopan, M. Liimatta, S. E. Smith, S. Lazerwith,

C. Stiff, S. Kamtekar, K. Bhattacharya, Y. Zhang, S. Swaney, K. Van Becelaere, R. C. Stevens, and B. F. Cravatt, "Discovery and Characterization of a Highly Selective FAAH Inhibitor that Reduces Inflammatory Pain," *Chemistry & Biology*, vol. 16, no. 4, pp. 411–420, Apr. 2009, ISSN: 10745521. DOI: 10.1016/j.chembiol.2009.02.013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1074552109000805?via%7B%5C%%7D3Dihub%20https://linkinghub.elsevier.com/retrieve/pii/S1074552109000805.

[348] M. Mileni, S. Kamtekar, D. C. Wood, T. E. Benson, B. F. Cravatt, and R. C. Stevens, "Crystal Structure of Fatty Acid Amide Hydrolase Bound to the Carbamate Inhibitor URB597: Discovery of a Deacylating Water Molecule and Insight into Enzyme Inactivation," *Journal of Molecular Biology*, vol. 400, no. 4, pp. 743–754, Jul. 2010, ISSN: 00222836. DOI: 10.1016/j.jmb.2010.05.034. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022283610005267?via%7B%5C%%7D3Dihub%20https://linkinghub.elsevier.com/retrieve/pii/S0022283610005267.

[349] A. Venturelli, D. Tondi, L. Cancian, F. Morandi, G. Cannazza, B. Segatore, F. Prati, G. Amicosante, B. K. Shoichet, and M. P. Costi, "Optimizing Cell Permeation of an Antibiotic Resistance Inhibitor for Improved Efficacy," *Journal of Medicinal Chemistry*, vol. 50, no. 23, pp. 5644–5654, Nov. 2007, ISSN: 0022-2623. DOI: 10.1021/jm070643q. [Online]. Available: https://pubs.acs.org/doi/abs/10.1021/jm070643q%20http://pubs.acs.org/doi/abs/10.1021/jm070643q.

[350] V. L. Thomas, A. C. McReynolds, and B. K. Shoichet, "Structural Bases for Stability–Function Tradeoffs in Antibiotic Resistance," *Journal of Molecular Biology*, vol. 396, no. 1, pp. 47–59, Feb. 2010, ISSN: 00222836. DOI: 10.1016/j.jmb.2009.11.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022283609013564?via%7B%5C%%7D3Dihub%20https://linkinghub.elsevier.com/retrieve/pii/S0022283609013564.

[351] L. Qiao, C. A. Baumann, C. S. Crysler, N. S. Ninan, M. C. Abad, J. C. Spurlino, R. L. DesJarlais, J. Kervinen, M. P. Neeper, S. S. Bayoumy, R. Williams, I. C. Deckman, M. Dasgupta, R. L. Reed, N. D. Huebert, B. E. Tomczuk, and K. J. Moriarty, "Discovery, SAR, and X-ray structure of novel biaryl-based dipeptidyl peptidase IV inhibitors," *Bioorganic & Medicinal Chemistry Letters*, vol. 16, no. 1, pp. 123–128, Jan. 2006, ISSN: 0960894X. DOI: 10.1016/j.bmcl.2005.09.037. [Online]. Available: https://www.sciencedirect.com/science/article/pii/

S0960894X05012059?via%7B%5C%%7D3Dihub%20https://linkinghub.elsevier.com/retrieve/pii/S0960894X05012059.

[352] H. B. Rasmussen, S. Branner, F. C. Wiberg, and N. Wagtmann, "Crystal structure of human dipeptidyl peptidase IV/CD26 in complex with a substrate analog," *Nature Structural Biology*, vol. 10, no. 1, pp. 19–25, Jan. 2003, ISSN: 10728368. DOI: 10.1038/nsb882. [Online]. Available: http://www.nature.com/doifinder/10.1038/nsb882.

[353] T. M. Gloster, R. Madsen, and G. J. Davies, "Structural basis for cyclophellitol inhibition of a β-glucosidase," *Org. Biomol. Chem.*, vol. 5, no. 3, pp. 444–446, Jan. 2007, ISSN: 1477-0520. DOI: 10.1039/B616590G. [Online]. Available: http://xlink.rsc.org/?DOI=B616590G.

[354] G. Scapin, S. B. Patel, J. W. Becker, Q. Wang, C. Desponts, D. Waddleton, K. Skorey, W. Cromlish, C. Bayly, M. Therien, J. Y. Gauthier, C. S. Li, C. K. Lau, C. Ramachandran, B. P. Kennedy, and E. Asante-Appiah, "The Structural Basis for the Selectivity of Benzotriazole Inhibitors of PTP1B †," *Biochemistry*, vol. 42, no. 39, pp. 11 451–11 459, Oct. 2003, ISSN: 0006-2960. DOI: 10.1021/bi035098j. [Online]. Available: https://pubs.acs.org/doi/abs/10.1021/bi035098j%20http://pubs.acs.org/doi/abs/10.1021/bi035098j.

[355] Z. Jia, Q. Ye, A. Dinaut, Q. Wang, D. Waddleton, P. Payette, C. Ramachandran, B. P. Kennedy, G. Hum, and S. Taylor, "Structure of protein tyrosine phosphatase 1B in complex with inhibitors bearing two phosphotyrosine mimetics.," *J.Med.Chem.*, vol. 44, pp. 4584–4594, 2001. DOI: 10.2210/PDB1KAK/PDB. [Online]. Available: https://www.rcsb.org/structure/1kak.

[356] R. Akué-Gédu, E. Rossignol, S. Azzaro, S. Knapp, P. Filippakopoulos, A. N. Bullock, J. Bain, P. Cohen, M. Prudhomme, F. Anizon, and P. Moreau, "Synthesis, Kinase Inhibitory Potencies, and in Vitro Antiproliferative Evaluation of New Pim Kinase Inhibitors," *Journal of Medicinal Chemistry*, vol. 52, no. 20, pp. 6369–6381, Oct. 2009, ISSN: 0022-2623. DOI: 10.1021/jm901018f. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/jm901018f.

[357] M. C. Abad, H. Askari, J. O'Neill, A. L. Klinger, C. Milligan, F. Lewandowski, B. Springer, J. C. Spurlino, and D. Rentzeperis, "Structural determination of estrogen-related receptor γ in the presence of phenol derivative compounds," *The Journal of Steroid Biochemistry and Molecular Biology*, vol. 108, no. 1-2, pp. 44–54, Jan. 2008, ISSN: 09600760. DOI: 10.1016/j.jsbmb.2007.06.006. [Online]. Available: https://www.sciencedirect.com/science/

article/abs/pii/S0960076007002361?via%7B%5C%%7D3Dihub%20https://linkinghub.elsevier.com/retrieve/pii/S0960076007002361.

[358] A. Matsushima, Y. Kakuta, T. Teramoto, T. Koshiba, X. Liu, H. Okada, T. Tokunaga, S.-i. Kawabata, M. Kimura, and Y. Shimohigashi, "Structural Evidence for Endocrine Disruptor Bisphenol A Binding to Human Nuclear Receptor ERR," *Journal of Biochemistry*, vol. 142, no. 4, pp. 517–524, Jul. 2007, ISSN: 0021-924X. DOI: 10.1093/jb/mvm158. [Online]. Available: https://academic.oup.com/jb/article-lookup/doi/10.1093/jb/mvm158.

[359] A. Blum, J. Böttcher, A. Heine, G. Klebe, and W. E. Diederich, "Structure-Guided Design of C 2 -Symmetric HIV-1 Protease Inhibitors Based on a Pyrrolidine Scaffold," *Journal of Medicinal Chemistry*, vol. 51, no. 7, pp. 2078–2087, Apr. 2008, ISSN: 0022-2623. DOI: 10.1021/jm701142s. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/jm701142s.

[360] B. A. Katz, J. M. Clark, J. S. Finer-Moore, T. E. Jenkins, C. R. Johnson, M. J. Ross, C. Luong, W. R. Moore, and R. M. Stroud, "Design of potent selective zinc-mediated serine protease inhibitors," *Nature*, vol. 391, no. 6667, pp. 608–612, Feb. 1998, ISSN: 0028-0836. DOI: 10.1038/35422. [Online]. Available: http://www.nature.com/doifinder/10.1038/35422.

[361] T. Brandt, N. Holzmann, L. Muley, M. Khayat, C. Wegscheid-Gerlach, B. Baum, A. Heine, D. Hangauer, and G. Klebe, "Congeneric but Still Distinct: How Closely Related Trypsin Ligands Exhibit Different Thermodynamic and Structural Properties," *Journal of Molecular Biology*, vol. 405, no. 5, pp. 1170–1187, Feb. 2011, ISSN: 00222836. DOI: 10.1016/j.jmb.2010.11.038. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022283610012684?via%7B%5C%%7D3Dihub%20https://linkinghub.elsevier.com/retrieve/pii/S0022283610012684.

[362] J. R. Pruitt, D. J. P. Pinto, R. A. Galemmo, R. S. Alexander, K. A. Rossi, B. L. Wells, S. Drummond, L. L. Bostrom, D. Burdick, R. Bruckner, H. Chen, A. Smallwood, P. C. Wong, M. R. Wright, S. Bai, J. M. Luettgen, R. M. Knabb, P. Y. S. Lam, and R. R. Wexler, "Discovery of 1 - (2 - Aminomethylphenyl) - 3 - trifluoromethyl - N - [3-fluoro-2'-(aminosulfonyl)[1,1'-biphenyl)]-4-yl]-1 H -pyrazole-5-carboxyamide (DPC602), a Potent, Selective, and Orally Bioavailable Factor Xa Inhibitor 1," *Journal of Medicinal Chemistry*, vol. 46, no. 25, pp. 5298–5315, Dec. 2003, ISSN: 0022-2623. DOI: 10.1021/jm030212h. [Online]. Available: https://pubs.acs.org/doi/abs/10.1021/jm030212h%20http://pubs.acs.org/doi/abs/10.1021/jm030212h.

[363] K. Sommer, F. Flachsenberg, and M. Rarey, "NAOMInext – Synthetically feasible fragment growing in a structure-based design context," *European Journal of Medicinal Chemistry*, vol. 163, pp. 747–762, Dec. 2018, ISSN: 0223-5234. DOI: 10.1016/J.EJMECH.2018.11.075. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0223523418310377?via%7B%5C%%7D3Dihub.

Appendix

Appendix A.

Invalid Ligand Pairs

PDB id larger ligand	Larger ligand	PDB id smaller ligand	Smaller ligand	Cause of exclusion
1MY8[336]	SM3	1FSW[337]	CTB	covalently bound
2FDA[338]	682	3BG8[339]	INH	covalently bound
3O87[340]	BSG	3O88[340]	BSH	covalently bound
3O88[340]	BSG	4E3I[341]	0N3	covalently bound
3O1G[342]	O75	3O0U[342]	O47	covalently bound
1GA9[343]	ETP	3BLS[344]	APB	covalently bound
3F3V[345]	1BU[a]	2HWO[346]	RBS	covalently bound
2WAP[347]	PIX	3LJ7[348]	OHO	covalently bound
2I72[349]	VA1	3IXG[350]	BZB[a]	covalently bound
1ZPC[338]	716	2FDA[338]	682	covalently bound
2AJL[351]	JNH	1N1M[352]	A3M[a]	covalently bound
2VRJ[288]	NCW[a]	2JAL[353]	YLL	covalently bound

PDB id larger ligand	Larger ligand	PDB id smaller ligand	Smaller ligand	Cause of exclusion
1GA9[343]	ETP	3BLS[344]	APB	covalently bound
1L2S[274]	STC	2HDQ[275]	C21	SIENA ligand extraction failed
1Q6S[354]	FNP	1KAK[355]	214	symmetric substructure
1ROS[280]	DEO	2WO9[281]	068	symmetric substructure
1ROS[280]	DEO	2WO8[281]	077	symmetric substructure
1YHS[286]	STU	3JPV[356]	1DR	symmetric substructure
2P7Z[357]	OHT	2E2R[358]	2OH	symmetric substructure
2QNN[359]	QN1	2PQZ[359]	G0G	symmetric substructure
1OYQ[311]	T87	1XUG[360]	BAB	symmetric substructure
2VTI[319]	LZ3	2VTH[319]	LZ2	invalid matching (spatial filter)
2ZFS[361]	12U	3ATK[277]	SZ1	invalid matching (no substructure)
3M35[362]	M35	3RXP[277]	SW3	invalid matching (no substructure)

Table (A.1) Invalid ligand pairs not used for evaluation. [a] Ligand not detected as covalently bound

Appendix B.

NAOMInext User Guide

In the following chapter the detailed usage of NAOMInext is outlined. Usage is described for the cmd-line as well as the interactive version. The NAOMInext graphical user interface (GUI) is organized in different views. The start window is the main view and incorporates all other views. For example, the available reaction schemes or loaded BBs are accessible via different windows. Last but not least, the user guide provides a detailed description on how to add your own SMIRKS rules to NAOMInext.

The minimal input to NAOMInext is a PDB file (*.pdb) or PDBx/mmCIF file (*.cif/*.mcif) including a protein and a co-crystallized ligand. Loading protein and ligand separately is also possible, but the ligand 3D - coordinates have to suit the given complex. Fetching PDB files from the PDB-server is possible via the shortcut `Ctrl+F` and `Command+F` on Linux/Windows and macOS, respectively.

B.1. Installation Guide

NAOMInext is available for Windows, macOS, and Linux operating systems and provided as installer package or compressed *.tar.gz and *.zip archive. Using the installer package is very convenient and installs NAOMInext in the system specific locations and can then be used like any other tool on the system (needs root privileges). If, for example, no root access is available, NAOMInext can be used via the provided archive and extracted to any user location on the system. Starting the application is then possible by clicking on the executable or via command line.

B.2. NAOMInext - Cmd-line Mode

NAOMInext is a tool to perform fragment growing within a protein binding site incorporating different types of constraints. The intended workflow should

be mainly interactive, but NAOMInext can also be used from the command line. To get a first impression about usage and available parameters just type the following on command line:

```
./NAOMInext --help
```

This starts the application and provides a list of available parameters together with a short description of each parameter (see Table B.1).

Using NAOMInext via command line may be useful if several unsupervised consecutive reaction steps should be performed in parallel. An example command for several iteration steps may look as follows:

```
./NAOMInext -i my/Protein.pdb -l my/ligand.sdf -o myResults.sdf -j
4 -b my/buildingBlock.smi --iterationCtr 2 --no-gui
```

This command performs two consecutive reaction steps starting from the given reference ligand. All reaction results from the first reaction are used as input for the second reaction step. This may lead to an exponential increase in runtime. The --no-gui parameter is necessary to start NAOMInext in headless mode, which is needed for calculations on high-performance clusters.

Another useful application would be performing the fragment growing step within an ensemble of proteins to incorporate protein flexibility. Therefore, NAOMInext should be called from a script (e.g. Python) using the same configuration with different (superposed) protein structures as input.

B.3. NAOMInext - GUI

The main purpose of NAOMInext is the interactive usage. Figure B.1 shows the initial start screen of NAOMInext. The `Toolbar` (see Subsection B.3.3) at the top provides most of the main functionality. The plus button, for example, provides convenient access to several different load functions. Using this button the users may load, proteins, ligands, and BB databases. Besides, clicking on the `File` menu in the `Menubar` enables detailed load functionalities for different input types. For example, loading ligands from a SD-file as BB instead of loading the molecules as input for the protein (3D - View).

The main view can be divided in three independent parts. First, the `Toolbar` at the top, which provides the main functionality via the image buttons. Hovering over a button shows a more detailed description. Second, the column on the left shows loaded input molecules as well as result molecules including the sampled poses.

Table (B.1) Command line options in conjunction with a short description

Command line option	Description
-h [--help]	Show command line options
--no-gui	Start NAOMInext in headless mode(no graphics) for cluster usage.
-v [--verbosity] arg (=3)	Set verbosity level (0 = Quiet, 1 = Errors, 2 = Warnings, 3 = Info)
-i [--input] arg	Input complex file, suffix is required.
-l [--ligand] arg	Input ligand file needed in cmd-line mode, suffix is required.
-o [--output] arg (=temp.sdf)	Output file, suffix is required.
-b [--buildingBlocks] arg	Input building block file, suffix is required.
-j [--jobs] arg	Number of parallel jobs allowed.
--reactionsFile arg	Reaction SMARTS/SMIRKS file name stored in HOME folder or relative to the binary. default: hartenfeller_reactions.csv
--iterationCtr arg (=0)	Configure the number of iterative reaction steps you want to perform automatically for a given anchor fragment
--reset	If you encounter any issues with the tool, use the '--reset' param to clear user specific settings.
--no-gui	Use NAOMInext in cmd-line mode (headless)
-n [--nofPosesToWrite] arg (=32)	The number of poses to write for each result molecule, default: 32
-p [--nofStartPoses] arg (=50)	Upper bound of used start poses, default: 50
--license arg	To reactivate the executable, please provide a new license key.

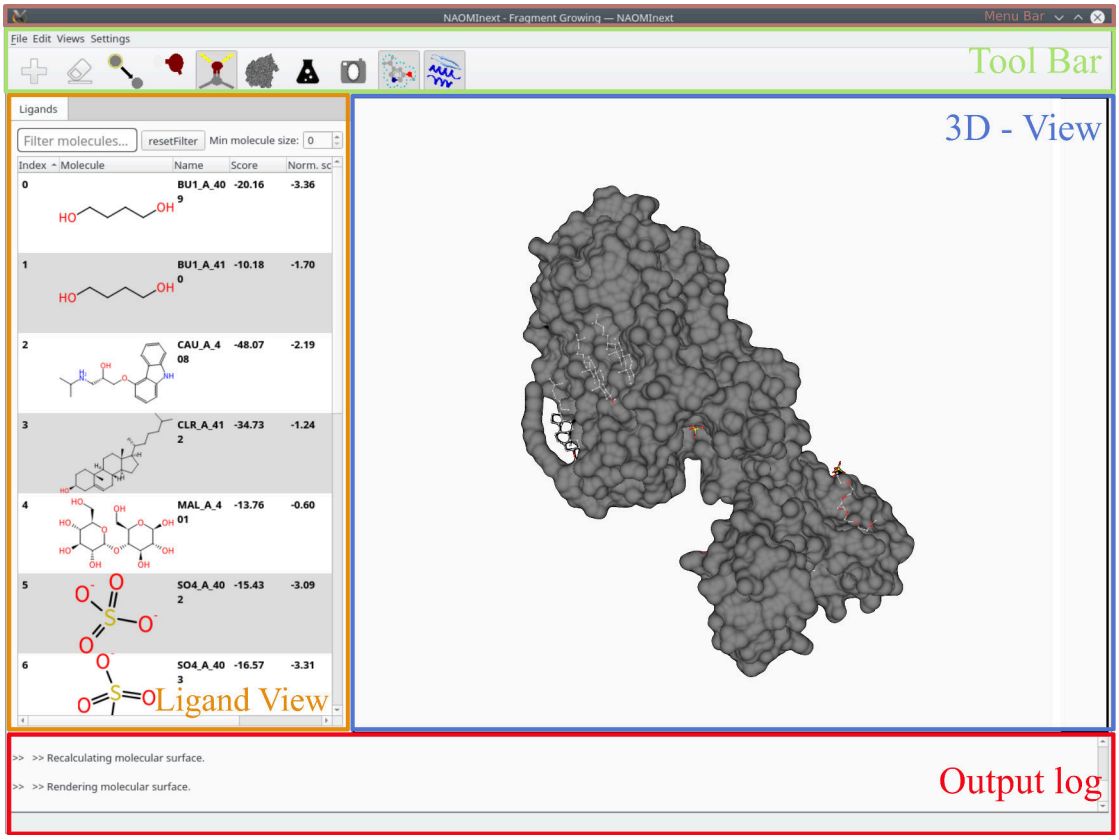


Figure (B.1) NAOMInext main view after loading a protein (PDB id: 2RH1[222])

B.3.1. User Settings

NAOMInext is based on the Qt-Framework and the user interface is implemented using QML. The overall platform support of Qt allows storage of any provided settings in a user-defined location. For example, dimensions and position of a window are stored and reloaded during the next application start. Moreover, important settings, affecting the algorithm, and usability settings are stored as well (see Section B.3.2).

B.3.2. Settings View

The settings view (accessible over the `Menubar`) provides access to most of the needed settings. All settings are initialized with default values but may be adapted to the users needs. All adapted settings are then saved and reloaded during the next program start. So, parametrization of NAOMInext does not have to be done all the time.

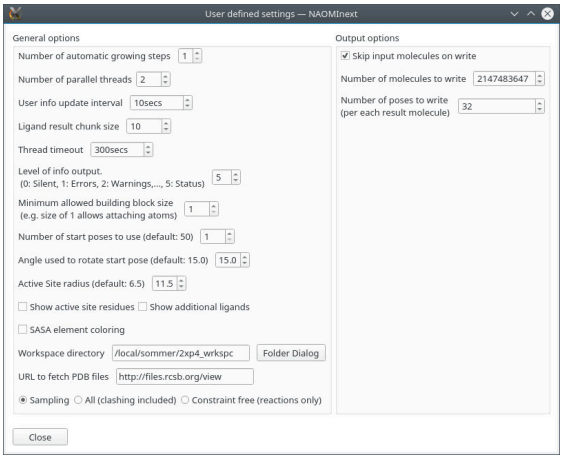


Figure (B.2) The Settings View provides easy access to important settings.

B.3.3. ToolBar

The `Toolbar` mainly provides access to input functions and different settings for the visualization view (see Section B.3.5). Each button (from left to right) is shortly described below.

1. Load PDB structures, ligands, and BB databases
2. Clear all loaded structural data (except BB database)
3. Clear marked/highlighted atoms and bonds
4. Show/hide water molecules
5. Show/hide molecule interactions to the protein
6. Show the whole protein instead of active site only
7. Start Fragment Growing
8. Make screenshot of the 3D-view
9. Show solvent accessible surface for the active site
10. Show/hide secondary structure

B.3.4. Ligand View

The ligand view shows all loaded ligands of the input protein structure as well as additionally loaded ligands. Generated result ligands including conformations are also provided here. The ligand view is organized as a table and provides a column for: 2D depiction of the molecule, name, score, normalized score, and an RMSD to the reference structure (if loaded). The table can be sorted by clicking on the respective column name. Clicking on the molecule column sorts the table according to the molecules heavy atom count. A double click on a row, loads the corresponding molecule into the 3D - View (section B.3.5).

A right-click shows additional features for the underlying molecule as well as other selected molecules.

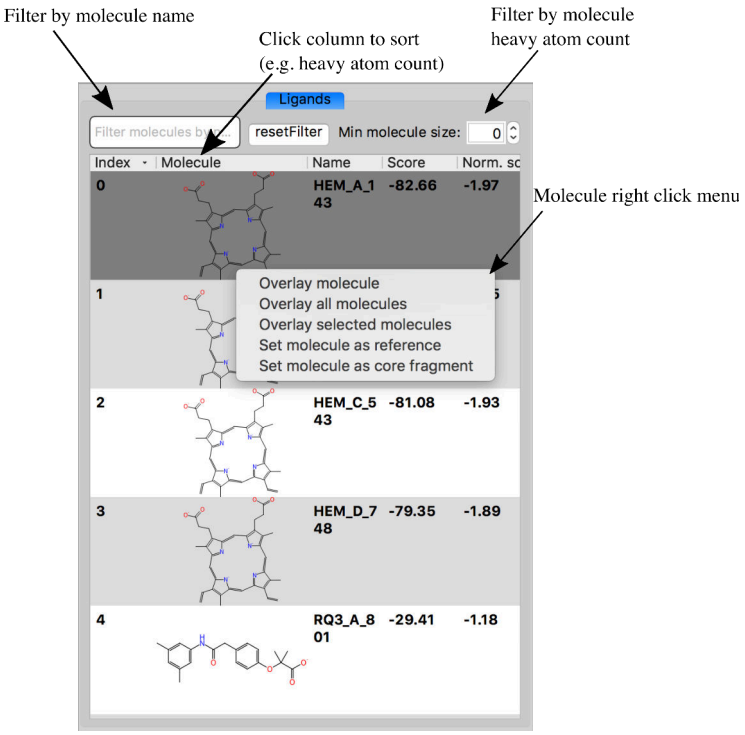


Figure (B.3) Ligand View of loaded input molecules. Filtering and sorting is provided. Additional features are available via the right click menu.

The ligand view can be filtered by name and molecule size to enhance clarity. Each column can be used to sort the ligands.

B.3.5. 3D - View

The 3D - View allows interactive inspection of the loaded protein structure including ligands. The initial view shows the whole protein including all extracted ligands regardless of the user settings. After double clicking on a molecule in the ligand view, the view is focused and shows interactions or active site residues (if enabled) around the molecule.

Navigation of the 3D scene is performed using the mouse. Following key bindings are defined for feasibility:

rotate - right mouse button
zoom - right mouse button + ctrl + shift or scroll wheel
translate - right mouse button + ctrl or middle mouse button

The 3D - View facilitates easy inspection of the loaded structure. Clicking on a ligand atom opens the tautomer and protonation state view. In this view, all available different states are shown and can be applied to the loaded ligand vie clicking on the particular 2D depiction.

B.3.6. Building Blocks

Building blocks can be loaded in every supported chemical file format using the `plus` button in the `Toolbar` or using the shortcut `Ctrl+B`. Loading via the `plus` button only allows for building blocks including linker atoms to be loaded correctly (see Section B.7.1 for an example SD file). A valid building block in SMILES format may look as follows: `[?linkerName?]CCC=O`. Invalid molecules are discarded and required 3D coordinates are calculated on the fly. Via the `Building Block View` (`Ctrl+Shift+B`), the loaded molecules can be investigated visually by provided 2D depictions. Compiled BB libraries can be saved and reloaded using the `File menu` via the `Menubar`.

B.3.7. Reaction View

In the reaction view all incorporated reactions are listed (figure B.4). The view can be opened via the `View menu` in the `Menubar`.

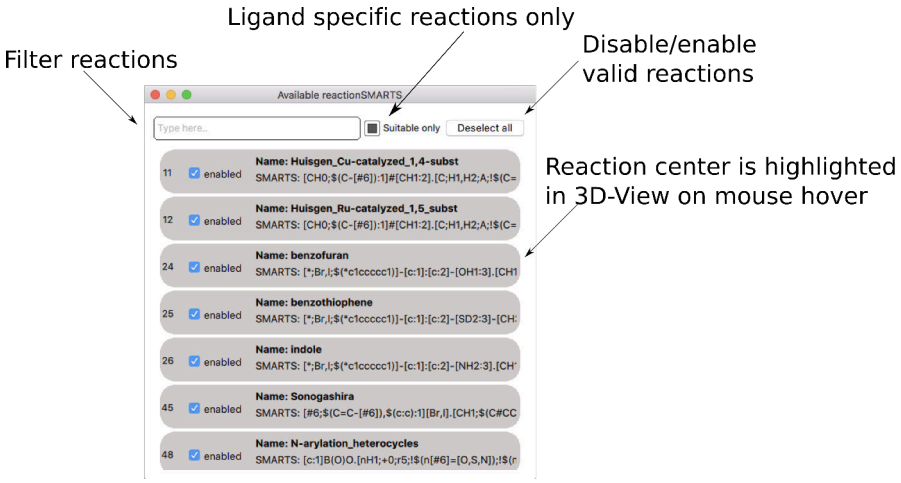


Figure (B.4) NAOMInext reaction view
Reaction view showing only ligand specific reactions

Initially, all reactions are disabled, hence, no ligand is loaded. After loading a protein structure including a ligand, suitable reactions are enabled. To list only suitable reactions, click the given checkbox. If a different ligand is clicked

in the Ligand View, suitable reactions are updated. Using the upper text field, reactions can be filtered based on a given text string. Suitable extension vectors of the molecule, i.e. bonds, are highlighted in green and can be seen in the 3D - View. Using the mouse and hovering over a specific reaction, additionally highlights the corresponding reaction center using yellow spheres.

B.4. Fragment Growing

Fragment growing can be performed in two different ways. First, using a defined exit vector at the anchor fragment (clicked bond highlighted in magenta see Figure B.5 for an example) and a pre-processed BB library containing linker atoms defining the connection vector. Second, using a BB library, e.g. a vendor catalogue and incorporated reaction rules.

B.4.1. Single Bond Fragment Growing

Clicking on a hydrogen bond, allows fragment growing on user defined exit vectors via single bond formation (see magenta marked bond in Figure B.5). All hydrogen atoms can be shown in the 3D - View via the Settings menu in the Menubar. If the desired hydrogen atom is not shown, please click on the corresponding heavy atom and select the tautomer providing the desired hydrogen atom. Additionally, users have to provide a pre-processed BB library. Each BB in the library needs a linker atom which defines the connection point. If both requirements are met, clicking on the fragment growing button in the ToolBar starts the growing process for the whole BB library.

B.4.2. Reaction based Fragment Growing

The incorporated reaction schemes[47] allow for synthetic feasible fragment growing. The reaction schemes are based on Reaction SMARTS. Hence, the BB library does not need to be pre-processed and allows loading any supported library, e.g. vendor catalogues or in-house databases. The used Reaction SMARTS implicitly selects suitable BBs based on compatibility to the anchor fragment via SMARTS matching. Suitable reactions can be investigated in the Reaction View (section B.3.7) and filtered based on specific needs. After loading a protein target, a fragment to elaborate, and a BB library, parallelized fragment growing can be started by clicking the reaction flask button in the toolbar (see Section B.3.3). Available reaction centers are visualized with a yellow sphere in the 3D – View after hovering over the reaction entry in the Reaction View (see Figure B.6 for an example).

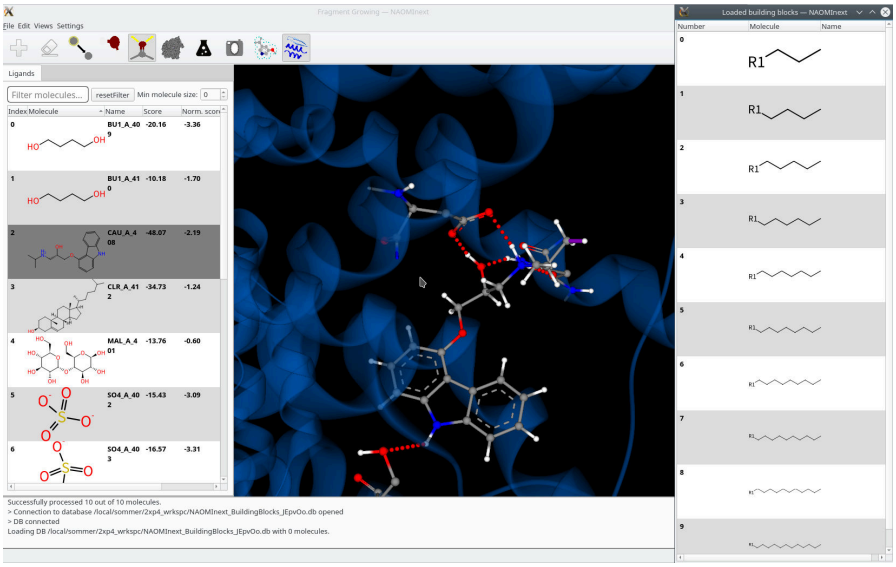


Figure (B.5) Exit vector selection for a user defined growing vector. Please note that using the manual extension mode requires a pre-processed building block library with provided attachment points (linker atoms)

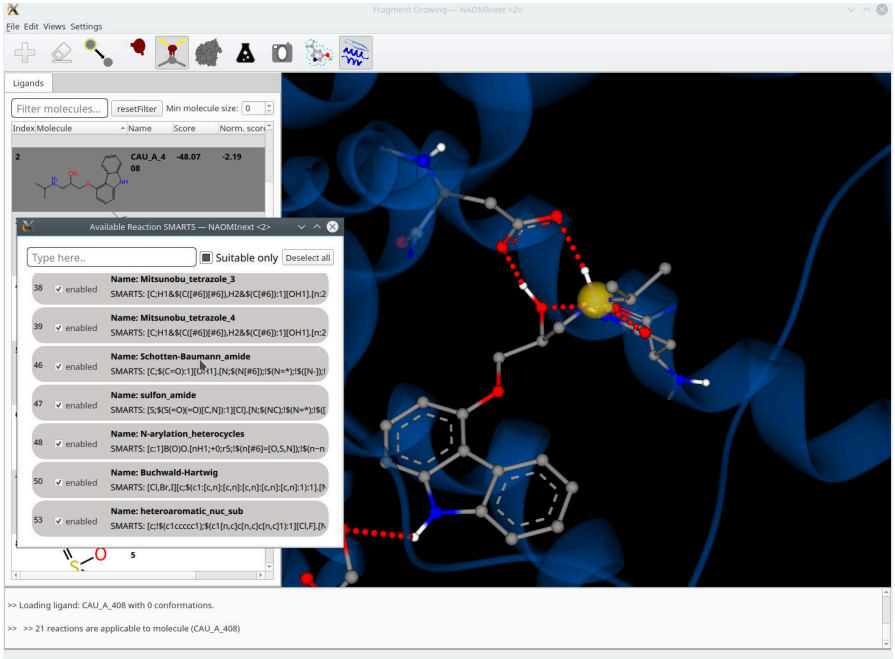


Figure (B.6) Reaction center depiction for individual highlighted reactions (mouse over).

B.5. Providing User-Defined Reactions

NAOMInext incorporates published reaction rules from Hartenfeller and co-workers. [43] These reactions are provided as *.csv file. To provide user defined reactions the supplied file in the install package needs to be modified. A new reactions file (as comma separated *.csv file) may be also provided via command line as `./NAOMInext --reactionsFile myOwnReactions.csv` or provided as `hartenfeller_reactions.csv` stored in the users home folder. Subsequently, the file needs to be reloaded via the `Menu Bar`.

B.6. Known issues

1. LigandView - Sorting columns for several hundreds of ligands is very slow
2. LigandView - Highlighting selected ligand (single click) does not work if any filter is used
3. LigandView - 2D depiction is not updated when sorting is updated (resize ligand view to force update)
4. 3dView - Rotation may not work. Click “Ctrl+Shift” or “Ctrl+Tab” to re-enable rotation.

B.7. Additional Data

B.7.1. Example SD File Including Linker

data.txt

```
          12011811072D 1   1.00000     0.00000     0
unicon 1.2 kaisommer
 11 10  0  0  0  0            999 V2000
    0.0000    0.0000    0.0000 R#  0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.0000    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.0000    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.0000    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.0000    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.0000    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.0000    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.0000    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
```

```
  1  2  1  0  0  0  0
  2  3  1  0  0  0  0
  3  4  1  0  0  0  0
  2  5  1  0  0  0  0
  2  6  1  0  0  0  0
  3  7  1  0  0  0  0
  3  8  1  0  0  0  0
  4  9  1  0  0  0  0
  4 10  1  0  0  0  0
  4 11  1  0  0  0  0
M  RGP  1    1    1
M  END
$$$$
```

Appendix C.

Automated GLIDE Docking Workflow

This chapter comprises the used commands for the evaluation performed using Glide.

C.1. Glide Commands

Here, commands of the Glide docking workflow are listed in the applied order. Input parameters for each script are denoted in curly braces. The scripts extract_ligands.py and site_by_reflig.py can be found in the Supporting Information provided by Sommer *et al.* [363] All other scripts are provided by the Schrödinger Maestro Suite[235].

prepwizard call=$SCHRODINGER/utilities/prepwizard -disulfides -fillsidechains -watdist 0.0 -propka_pH 7 - NOJOBID {input} {output}

extract_ligand_schrodinger call=$SCHRODINGER/run ~/extractLigands.py -reference_ligand {reflig:} - output_overlapping_ligands {output_reflig} {target}

prepare_docking_schrodinger call=$SCHRODINGER/run ~/extractLigands.py -output_structure {output_target:} -reference_ligand {reflig} -delete_overlapping_ligands {target}

ref_lig_site_schrodinger call=$SCHRODINGER/run ~/siteByReflig.py -reference_ligand {reflig} -input_structure {target} – size {size}

glide_grid_input call=$SCHRODINGER/run $SCHRODINGER/mmshare-v3.7/python/common/glide_sif.py - grid_center "{grid_center}" -innerbox {innerbox}

-outerbox {outerbox} -gridfile {gridfile} - recep_file {recep_file} {output_-file}

glide_docking_input call=$SCHRODINGER/run $SCHRODINGER/mmshare-v3.7/python/common/glide_sif.py - pose_outtype ligandlib_sd -poses_-per_lig {poses_per_lig} -precision SP -gridfile {gridfile} - ligandfile {ligandfile} {output_file}

glide call=$SCHRODINGER/glide -NOJOBID {input_file}

Appendix D.

Scientific Contributions

Scientific contributions during the time of my PhD are listed in chronological order. Contributions relating to this work are printed in bold.

D.1. Publications

Friedrich, N.-O.; Flachsenberg, F.; Meyder, A.; Sommer, K.; Kirchmair, J.; Rarey, M. Conformator: A Novel Method for the Generation of Conformer Ensembles. J. Chem. Inf. Model. 2019, acs.jcim.8b00704.

Sommer, K.; Flachsenberg, F.; Rarey, M. NAOMInext – Synthetically Feasible Fragment Growing in a Structure-Based Design Context. Eur. J. Med. Chem. 2018, 163, 747–762.

Bietz, S.; Inhester, T.; Lauck, F.; Sommer, K.; von Behren, M. M.; Fährrolfes, R.; Flachsenberg, F.; Meyder, A.; Nittinger, E.; Otto, T.; *et al.* From Cheminformatics to Structure-Based Design: Web Services and Desktop Applications Based on the NAOMI Library. J. Biotechnol. 2017, 261, 207–214.

Friedrich, N.-O.; de Bruyn Kops, C.; Flachsenberg, F.; Sommer, K.; Rarey, M.; Kirchmair, J. Benchmarking Commercial Conformer Ensemble Generators. J. Chem. Inf. Model. 2017, 57 (11), 2719–2728.

Friedrich, N.-O.; Meyder, A.; de Bruyn Kops, C.; Sommer, K.; Flachsenberg, F.; Rarey, M.; Kirchmair, J. High-Quality Dataset of Protein-Bound Ligand Conformations and Its Application to Benchmarking Conformer Ensemble Generators. J. Chem. Inf. Model. 2017, 57 (3), 529–539.

Inhester, T.; Nittinger, E.; Sommer, K.; Schmidt, P.; Bietz, S.; Rarey, M. NAOMInova: Interactive Geometric Analysis of Noncovalent Interactions in Macromolecular Structures. J. Chem. Inf. Model. 2017, 57 (9), 2132–2142.

Wagner, V.; Jantz, L.; Briem, H.; Sommer, K.; Rarey, M.; Christ, C. D. Computational Macrocyclization: From de Novo Macrocycle Generation to Binding Affinity Estimation. ChemMedChem 2017, 12 (22), 1866–1872.

Sommer, K.; Friedrich, N.-O.; Bietz, S.; Hilbig, M.; Inhester, T.; Rarey, M. UNICON: A Powerful and Easy-to-Use Compound Library Converter. J. Chem. Inf. Model. 2016, 56 (6), 1105–1111.

Paulke, A.; Proschak, E.; Sommer, K.; Achenbach, J.; Wunder, C.; Toennes, S. W. Synthetic Cannabinoids: In Silico Prediction of the Cannabinoid Receptor 1 Affinity by a Quantitative Structure-Activity Relationship Model. Toxicol. Lett. 2016, 245, 1–6.

D.2. Talks

Sommer, K., M. Rarey; Towards Interactive Fragment Growing, Gordon Research Seminar on Computer Aided Drug Design, 2017, Mount Snow, USA

D.3. Poster Presentations

Flachsenberg, F., Sommer K., Rarey, M.; JAMDA – Just Another Molecular Docking Algorithm, EuroQSAR, 2018, Thessaloniki, GRC

Sommer, K.; Flachsenberg, F.; Rarey, M.; NAOMInext – Reaction-Driven Probing of Protein Binding Sites, 2018, ICCS – International Conference on Chemical Structures, Noordwijkerhout, NL

Sommer, K.; Rarey, M.; Fragment Growing Linked with User Constraints Through Intuitive Usage, Gordon Research Conferences on Computer Aided Drug Design, 2017, Mount Snow, USA

Sommer, K.; Friedrich, N. O.; Bietz, S.; Hilbig, M.; Inhester, T.; Rarey, M.; An easy-to-use software tool for compound library conversion and isomeric enumeration, 2016, Seventh Joint Sheffield Conference on Chemoinformatics, Sheffield, GB

Appendix E.

Hartenfeller Reactions Adapted

#{name}	smirks	educt1 smiles [optional]	educt2 smiles [optional]	additional Reactant [optional]	adaptation
{Pictet-Spengler}	[cH1:1]1:[c:2](-[CH2:7]-[CH2:8]-[NH2:9]):[c:3]:[c:4]:[c:5]:[c:6]:1.[#6:11]-[CH1;R0:10]=[OD1]>>[c:1]12:[c:2](-[CH2:7]-[CH2:8]-[NH1:9]-[C:10]-2(-[#6:11])):[c:3]:[c:4]:[c:5]:[c:6]:1	c1cc(CCN)ccc1	CC(=O)		
{benzimidazole_derivatives_carboxylic-acid/ester}	[c;r6:1](-[NH1;$(N-[#6]):2]):[c;r6:3](-[NH2:4]).[#6:6]-[C;R0:5](=[OD1])-[#8;H1,$(O-[CH3])]>>[c:3]2:[c:1]:[n:2]:[c:5](-[#6:6]):[n:4]@2	c1c(NC)c(N)ccc1	CC(=O)O		
{benzimidazole_derivatives_aldehyde}	[c;r6:1](-[NH1;$(N-[#6]):2]):[c;r6:3](-[NH2:4]).[#6:6]-[CH1;R0:5](=[OD1])>>[c:3]2:[c:1]:[n:2]:[c:5](-[#6:6]):[n:4]@2	c1c(NC)c(N)ccc1	CC(=O)		
{benzothiazole}	[c;r6:1](-[SH1:2]):[c;r6:3](-[NH2:4]).[#6:6]-[CH1;R0:5](=[OD1])>>[c:3]2:[c:1]:[s:2]:[c:5](-[#6:6]):[n:4]@2	c1c(S)c(N)ccc1	CC(=O)		
{benzoxazole_arom-aldehyde}	[c:1](-[OH1;$(Oc1ccccc1):2]):[c;r6:3](-[NH2:4]).[c:6]-[CH1;R0:5](=[OD1])>>[c:3]2:[c:1]:[o:2]:[c:5](-[c:6]):[n:4]@2	c1cc(O)c(N)cc1	c1ccccc1C(=O)		
{benzoxazole_carboxylic-acid}	[c;r6:1](-[OH1:2]):[c;r6:3](-[NH2:4]).[#6:6]-[C;R0:5](=[OD1])-[OH1]>>[c:3]2:[c:1]:[o:2]:[c:5](-[#6:6]):[n:4]@2	c1cc(O)c(N)cc1	CC(=O)O		
{thiazole}	[#6:6]-[C;R0:1](=[OD1])-[CH1;R0:5](-[#6:7])-[*;#17,#35,#53].[NH2:2]-[C:3]=[SD1:4]>>[c:1]1(-[#6:6]):[n:2]:[c:3]:[s:4]:[c:5]:1(-[#6:7])	CC(=O)C(I)C	NC(=S)C		ring numbering

{Niementowski_quinazoline}	[c:1](-[C;$(C-c1ccccc1):2](=[OD1:3])-[OH1]):[c:4](-[NH2:5]).[N;!H0;!$(N-N);!$(N-C=N);!$(N(-C=O)-C=O):6]-[C;H1,$(C-[#6]):7]=[OD1]>>[c:4]2:[c:1][c:2](=[O:3])[n:6][c:7][n:5]2	c1c(C(=O)O)c(N)ccc1	C(=O)N	aromaticity and bond specification
{tetrazole_terminal}	[CH0;$(C-[#6]):1]#[NH0:2]>>[c:1]1[n:2]n[nH]n1	CC#N		aromaticity
{tetrazole_connect_regioisomere_1}	[CH0;$(C-[#6]):1]#[NH0:2].[C;A;!$(C=O):3]-[#17,#35,#53]>>[c:1]1[n:2]n(-[C:3])nn1	CC#N	CBr	aromaticity
{tetrazole_connect_regioisomere_2}	[CH0;$(C-[#6]):1]#[NH0:2].[C;A;!$(C=O):3]-[#17,#35,#53]>>[c:1]1[n:2]nnn1(-[C:3])	CC#N	CBr	aromaticity
{Huisgen_Cu-catalyzed_1,4-subst}	[CH0;$(C-[#6]):1]#[CH1:2].[C;H1,H2;A;!$(C=O):3]-[#17,#35,#53,OH1]>>[c:1]1[c:2]n(-[C:3])nn1	CC#C	CCBr	aromaticity
{Huisgen_Ru-catalyzed_1,5_subst}	[CH0;$(C-[#6]):1]#[CH1:2].[C;H1,H2;A;!$(C=O):3]-[#17,#35,#53,OH1]>>[c:1]1[c:2]nnn1(-[C:3])	CC#C	CCBr	aromaticity
{Huisgen_disubst-alkyne}	[#6]):2].[C;H1,H2;A;!$(C=O):3]-[#17,#35,#53,OH1]>>[c:1]1[c:2]nnn1(-[C:3])	CC#CC	CCBr	aromaticity
{1,2,4-triazole_acetohydrazide}	[CH0;$(C-[#6]):1]#[NH0:2].[NH2:3]-[NH1:4]-[CH0;$(C-[#6]);R0:5]=[OD1]>>[n:2]1[c:1][n:3][nH:4][c:5]1	CC#N	NNC(=O)C	aromaticity
{1,2,4-triazole_carboxylic-acid/ester}	[CH0;$(C-[#6]):1]#[NH0:2].[CH0;$(C-[#6]);R0:5](=[OD1])-[#8;H1,$(O-[CH3]),$(O-[CH2]-[CH3])]>>[n:2]1[c:1]n[nH][c:5]1	CC#N	OC(=O)C	aromaticity

{3-nitrile-pyridine}	[#6;!$([#6](-C=O)-C=O):4]-[CH0:1](=[OD1])-[C;H1&!$(C-[*;!#6])&!$(C-C(=O)O),H2:2]-[CH0;R0:3](=[OD1])-[#6;!$([#6](-C=O)-C=O):5]>>[c:1]1(-[#6:4]):[c:2]:[c:3](-[#6:5]):n:c(-O):c:1(-C#N)	CC(=O)CC(=O)C		[?3?]N=C(-O)-C([?1?])-C#N	aromaticity (and additional reactant added)
{spiro-chromanone}	[c:1](-[C;$(C-c1ccccc1):2](=[OD1:3])-[CH3:4]):[c:5](-[OH1:6]).[C;$(C1-[CH2]-[CH2]-[N,C]-[CH2]-[CH2]-1):7](=[OD1])>>[O:6]1-[c:5]:[c:1]-[C:2](=[OD1:3])-[C:4]-[C:7]-1	c1cc(C(=O)C)c(O)cc1	C1(=O)CCNCC1		
{pyrazole}	[#6;!$([#6](-C=O)-C=O):4]-[CH0:1](=[OD1])-[C;H1&!$(C-[*;!#6])&!$(C-C(=O)O),H2:2]-[CH0;R0:3](=[OD1])-[#6;!$([#6](-C=O)-C=O):5].[NH2:6]-[N;!H0;$(N-[#6]),H2:7]>>[c:1]1(-[#6:4])[c:2][c:3](-[#6:5])[n:7][n:6]1	CC(=O)CC(=O)C	NNC		aromaticity
{phthalazinone}	[c;r6:1](-[C;$(C=O):6]-[OH1]):[c;r6:2]-[C;H1,$(C-C):3]=[OD1].[NH2:4]-[NH1;$(N-[#6]);!$(NC=[O,S,N]):5]>>[c:1]1[c:2][c:3][n:4][n:5][c:6]1	c1cc(C(=O)O)c(C(=O)C)cc1	NNC		aromaticity
{Paal-Knorr_pyrrole}	[#6:5]-[C;R0:1](=[OD1])-[C;H1,H2:2]-[C;H1,H2:3]-[C:4](=[OD1])-[#6:6].[NH2;$(N-[C,N]);!$(NC=[O,S,N]);!$(N([#6])[#6]);!$(N~N~N):7]>>[c:1]1(-[#6:5])[c:2][c:3][c:4](-[#6:6])[n:7]1	CC(=O)CCC(=O)C	NC		aromaticity
{triaryl-imidazole}	[C;$(C-c1ccccc1):1](=[OD1])-[C;D3;$(C-c1ccccc1):2]~[O;D1,H1].[CH1;$(C-c):3]=[OD1]>>[c:1]1n[c:3][nH1][c:2]1	c1ccccc1C(=O)C(=O)c1ccccc1	c1ccccc1C(=O)		aromaticity

{Fischer_indole}	[NH1;$(N-c1ccccc1):1](-[NH2])-[c:5]:[cH1:4].[C;$(C([#6])[#6]):2](=[OD1])-[CH2;$(C([#6])[#6]);!$(C(C=O)C=O):3]>>[c:5]1[n:1][c:2][c:3][c:4]1	c1ccccc1NN	CCC(=O)C	
{Friedlaender_chinoline}	[NH2;$(N-c1ccccc1):1]-[c:2]:[c:3]-[CH1:4]=[OD1].[C;$(C([#6])[#6]):6](=[OD1])-[CH2;$(C([#6])[#6]);!$(C(C=O)C=O):5]>>[n:1]1:[c:2]:[c:3]:[c:4]:[c:5]:[c:6]:1	c1cccc(C=O)c1N	CCC(=O)C	aromaticity
{benzofuran}	[*;Br,I;$(*c1ccccc1)]-[c:1]:[c:2]-[OH1:3].[CH1:5]#[C;$(C-[#6]):4]>>[c:1]1[c:2][o:3][c:4][c:5]1	c1cc(I)c(O)cc1	CC#C	aromaticity
{benzothiophene}	[*;Br,I;$(*c1ccccc1)]-[c:1]:[c:2]-[SD2:3]-[CH3].[CH1:5]#[C;$(C-[#6]):4]>>[c:1]1[c:2][s:3][c:4][c:5]1	c1cc(I)c(SC)cc1	CC#C	aromaticity
{indole}	[*;Br,I;$(*c1ccccc1)]-[c:1]:[c:2]-[NH2:3].[CH1:5]#[C;$(C-[#6]):4]>>[c:1]1[c:2][n:3][c:4][c:5]1	c1cc(I)c(N)cc1	CC#C	aromaticity
{oxadiazole}	[#6:6][C:5]#[#7;D1:4].[#6:1][C:2](=[OD1:3])[OH1]>>[#6:6][c:5]1[n:4][o:3][c:2]([#6:1])n1	CC#N	CC(=O)O	
{Williamson_ether}	[#6;$([#6]~[#6]);!$([#6]=O):2][#8;H1:3].[Cl,Br,I][#6;H2;$([#6]~[#6]):4]>>[CH2:4][O:3][#6:2]	CCO	CCBr	
{reductive_amination}	[#6:4]-[C;H1,$([CH0](-[#6])[#6]):1]=[OD1].[N;H2,$([NH1;D2](C)C);!$(N-[#6]=[*]):3]-[C:5]>>[#6:4][C:1]-[N:3]-[C:5]	CC(=O)	NC	
{Suzuki}	[#6;H0;D3;$([#6](~[#6])~[#6]):1]B(O)O.[#6;H0;D3;$([#6](~[#6])~[#6]):2][Cl,Br,I]>>[#6:2][#6:1]	c1ccccc1B(O)O	c1ccccc1Br	

{piperidine_indole}	[c;H1:3]1:[c:4]:[c:5]:[c;H1:6]:[c:7]2:[nH:8]:[c:9]:[c;H1:1]:[c:2]:1:2.O=[C:10]1[#6;H2:11][#6;H2:12][N:13][#6;H2:14][#6;H2:15]1>>[#6;H2:12]3[#6;H1:11]=[C:10]([c:1]1:[c:9]:[n:8]:[c:7]2:[c:6]:[c:5]:[c:4]:[c:3]:[c:2]:1:2)[#6;H2:15][#6;H2:14][N:13]3	c1cccc2c1C=CN2	C1CC(=O)CCN1	
{Negishi}	[#6;$([#6]~[#6]);!$([#6]~[S,N,O,P]):1][Cl,Br,I].[Cl,Br,I][#6;$([#6]~[#6]);!$([#6]~[S,N,O,P]):2]>>[#6:2][#6:1]	CCBr	CCBr	
{Mitsunobu_imide}	[C;H1&$(C([#6])[#6]),H2&$(C[#6]):1][OH1].[NH1;$(N(C=O)C=O):2]>>[C:1][N:2]	CC(O)C	CC(=O)NC(=O)C	
{Mitsunobu_phenole}	[C;H1&$(C([#6])[#6]),H2&$(C[#6]):1][OH1].[OH1;$(Oc1ccccc1):2]>>[C:1][O:2]	CC(O)C	c1ccccc1O	
{Mitsunobu_sulfonamide}	[C;H1&$(C([#6])[#6]),H2&$(C[#6]):1][OH1].[NH1;$(N([#6])S(=O)=O):2]>>[C:1][N:2]	CC(O)C	CNS(=O)(=O)C	
{Mitsunobu_tetrazole_1}	[C;H1&$(C([#6])[#6]),H2&$(C[#6]):1][OH1].[nH1:2]1[n:3][n:4][n:5][c:6]1>>[C:1][n:2]1[n:3][n:4][n:5][c:6]1	CC(O)C	N1=N[NH1]C=N1	aromaticity
{Mitsunobu_tetrazole_2}	[C;H1&$(C([#6])[#6]),H2&$(C[#6]):1][OH1].[nH1:2]1[n:3][n:4][n:5][c:6]1>>[n:2]1[n:3]([C:1])[n:4][n:5][c:6]1	CC(O)C	N1=N[NH1]C=N1	aromaticity
{Mitsunobu_tetrazole_3}	[C;H1&$(C([#6])[#6]),H2&$(C[#6]):1][OH1].[n:2]1[nH1:3][n:4][n:5][c:6]1>>[C:1][n:2]1[n:3][n:4][n:5][c:6]1	CC(O)C	[NH1]1N=NC(C)=N1	aromaticity
{Mitsunobu_tetrazole_4}	[C;H1&$(C([#6])[#6]),H2&$(C[#6]):1][OH1].[n:2]1[nH1:3][n:4][n:5][c:6]1>>[n:2]1[n:3]([C:1])[n:4][n:5][c:6]:1	CC(O)C	[NH1]1N=NC(C)=N1	aromaticity

{Heck_terminal_vinyl}	[#6;c,$(C(=O)O),$(C#N):3][#6;H1:2]=[#6;H2:1].[#6;$([#6]=[#6]),$(c:c):4][Cl,Br,I]>>[#6:4]/[#6:1]=[#6:2]/[#6:3]	c1ccccc1C=C	c1ccccc1Br	
{Heck_non-terminal_vinyl}	[#6;c,$(C(=O)O),$(C#N):3][#6:2]([#6:5])=[#6;H1;$([#6][#6]):1].[#6;$([#6]=[#6]),$(c:c):4][Cl,Br,I]>>[#6:4][#6;H0:1]=[#6:2]([#6:5])[#6:3]	c1ccccc1C(C)=CC	c1ccccc1Br	
{Stille}	[#6;$(C=C-[#6]),$(c:c):1][Br,I].[Cl,Br,I][c:2]>>[c:2][#6:1]	c1ccccc1Br	c1ccccc1Br	
{Grignard_carbonyl}	[#6:1][C:2]#[#7;D1].[Cl,Br,I][#6;$([#6]~[#6]);!$([#6]([Cl,Br,I])[Cl,Br,I]);!$([#6]=O):3]>>[#6:1][C:2](=O)[#6:3]	CC#N	CCBr	
{Grignard_alcohol}	[#6:1][C;H1,$([C]([#6])[#6]):2]=[OD1:3].[Cl,Br,I][#6;$([#6]~[#6]);!$([#6]([Cl,Br,I])[Cl,Br,I]);!$([#6]=O):4]>>[#6:1][#6:2]([OH1:3])[#6:4]	CC(=O)C	CCBr	
{Sonogashira}	[#6;$(C=C-[#6]),$(c:c):1][Br,I].[CH1;$(C#CC):2]>>[#6:1][C:2]	c1cc(Br)ccc1	CC#C	
{Schotten-Baumann_amide}	[C;$(C=O):1][OH1].[N;$(N[#6]);!$(N=*);!$([N-]);!$(N#*);!$([ND3]);!$([ND4]);!$(N[O,N]);!$(N[C,S]=[S,O,N]):2]>>[C:1][N+0:2]	CC(=O)O	NCC	
{sulfon_amide}	[S;$(S(=O)(=O)[C,N]):1][Cl].[N;$(NC);!$(N=*);!$([N-]);!$(N#*);!$([ND3]);!$([ND4]);!$(N[c,O]);!$(N[C,S]=[S,O,N]):2]>>[S:1][N+0:2]	CS(=O)(=O)Cl	NCC	
{N-arylation_heterocycles}	[c:1]B(O)O.[nH1;+0;r5;!$(n[#6]=[O,S,N]);!$(n~n~n);!$(n~n~c~n);!$(n~c~n~n):2]>>[c:1]-[n:2]	c1ccccc1B(O)O	N1C=NC=C1	bond specification

{Wittig}	[#6:3]-[C;H1,$([CH0](-[#6])[#6]);!$(CC=O):1]=[OD1].[Cl,Br,I][C;H2;$(C-[#6]);!$(CC[I,Br]);!$(CCO[CH3]):2]>>[C:3][C:1]=[C:2]	CC(=O)C	BrCC	
{Buchwald-Hartwig}	[Cl,Br,I][c;$(c1:[c,n]:[c,n]:[c,n]:[c,n]:[c,n]:1):1].[N;$(NC)&!$(N=*)&!$([N-])&!$(N#*)&!$([ND3])&!$([ND4])&!$(N[c,O])&!$(N[C,S]=[S,O,N]),H2&$(Nc1:[c,n]:[c,n]:[c,n]:[c,n]:[c,n]:1):2]>>[c:1][N:2]	c1ccccc1Br	CNC	
{imidazole}	[C;$(C([#6])[#6;!$([#6]Br)]):4](=[OD1])[CH;$(C([#6])[#6]):5]Br.[#7;H2:3][C;$(C(=N)(N)[c,#7]):2]=[#7;H1;D1:1]>>[c:4]1[cH0:5][nH:3][c:2][n:1]1	CC(=O)C(Br)C	N=C(N)NC	aromaticity
{decarboxylative_coupling}	[c;$(c1[c;$(c[C,S,N](=[OD1])[*;R0;!$([OH1])])]cccc1):1][C;$(C(=O)[O;H1])].[c;$(c1aaccc1):2][Cl,Br,I]>>[c:1]-[c:2]	c1c(C(=O)O)c([N+](=O)[O-])ccc1	c1ccccc1Br	bond specification
{heteroaromatic_nuc_sub}	[c;!$(c1ccccc1);$(c1[n,c]c[n,c]c[n,c]1):1][Cl,F].[N;$(NC);!$(N=*);!$([N-]);!$(N#*);!$([ND3]);!$([ND4]);!$(N[c,O]);!$(N[C,S]=[S,O,N]):2]>>[c:1][N:2]	c1cnc(F)cc1	CN	
{nucl_sub_aromatic_ortho_nitro}	[c;$(c1c(N(~O)~O)cccc1):1][Cl,F].[N;$(NC);!$(N=*);!$([N-]);!$(N#*);!$([ND3]);!$([ND4]);!$(N[c,O]);!$(N[C,S]=[S,O,N]):2]>>[c:1]-[N:2]	c1c([N+](=O)[O-])c(F)ccc1	CN	
{nucl_sub_aromatic_para_nitro}	[c;$(c1ccc(N(~O)~O)cc1):1][Cl,F].[N;$(NC);!$(N=*);!$([N-]);!$(N#*);!$([ND3]);!$([ND4]);!$(N[c,O]);!$(N[C,S]=[S,O,N]):2]>>[c:1]-[N:2]	c1c(F)ccc([N+](=O)[O-])c1	CN	

{urea}	[N;$(N-[#6]):3]=[C;$(C=O):1].[N;$(N[#6]);!$(N=*);!$([N-]);!$(N#*);!$([ND3]);!$([ND4]);!$(N[O,N]);!$(N[C,S]=[S,O,N]):2]>>[N:3]-[C:1]-[N+0:2]	CN=C=O	CN
{thiourea}	[N;$(N-[#6]):3]=[C;$(C=S):1].[N;$(N[#6]);!$(N=*);!$([N-]);!$(N#*);!$([ND3]);!$([ND4]);!$(N[O,N]);!$(N[C,S]=[S,O,N]):2]>>[N:3]-[C:1]-[N+0:2]	CN=C=S	CN

Appendix F.

Ligand Pairs Evaluation Data Set (corrected)

Binding mode changed/ preserved	UniProt ID	PDB id larger ligand	Larger ligand	PDB id smaller ligand	Smaller ligand
1	P00811	3o88	BSH	3o86	BSF
1	Q7D785	3rv6	VAE	3st6	RVE
1	P18031	1q6s	214	1kak	FNP
1	P24941	2r3k	SCQ	2vta	LZ1
1	P04058	1dx6	GNT	1gqs	SAF
1	P00760	1yp9	UIZ	3ati	SZ4
1	P00918	2gd8	PO1	3ibn	O60
1	P00918	2gd8	PO1	3ibl	O59
1	O96017	2xbj	XBJ	2xm8	B4W
1	P00811	1ga9	ETP	3bls	APB
1	P00811	1my8	SM3	1fsw	CTB
1	Q8WSF8	2y58	V38	2y54	V63
1	P00811	1xgj	HTC	2hds	4MB
1	P00811	1l2s	STC	2hdq	C21
1	P42574	3dek	RXD	3deh	RXA
1	P56817	2ohu	IP7	2ohm	8AP
1	Q9WYE2	2zxb	ZXB	2zxd	ZXD
1	P56817	3rvi	RVI	3rtm	RTM
1	P00749	3kid	2BS	3mhw	ABV
1	P24941	1pxm	CK5	1pxj	CK2
1	P37231	3ads	IMN	3adt	HID
1	P00918	3f8e	TE1	4e49	RCO
1	P07900	3b27	B2T	2wi3	ZZ3
1	P21836	2ha2	SCK	2ha3	CHT

Binding mode changed/ preserved	UniProt ID	PDB id larger ligand	Larger ligand	PDB id smaller ligand	Smaller ligand
1	P62508	2p7z	OHT	2e2r	2OH
1	P00918	4e3d	GTQ	4e3h	HQE
1	P09960	3fuk	58Z	3fuh	5H1
1	P37231	3ads	IMN	3adu	MYI
1	P00811	3088	BSH	3087	BSG
1	P37231	3ads	IMN	3adv	SRO
1	P21836	2xuf	TZ4	2hao	CHH
1	P24941	2r3g	SC9	2r3h	SCE
1	P00523	3f3v	1BU	2hwo	RBS
1	P00918	3m96	E38	4e4a	JKE
1	P00760	1oyq	T87	1c1r	BAI
1	Q16539	3u8w	09J	3hvc	GG5
1	P00811	3088	BSH	4e3i	0N3
1	P07900	4awq	592	2xdl	2DL
1	P02879	3ej5	EJ5	1il5	DDP
1	P00760	1oyq	T87	1xug	BAB
1	P02766	3cfn	2AN	3cft	5NS
-1	P39900	3f15	HS1	3f1a	HS7
-1	P39900	3f15	HS1	3lka	M4S
-1	P39900	1jiz	CGS	3f19	HS6
-1	P07900	2uwd	2GG	2yjw	YJW
-1	O14965	3hoz	45B	3unz	0BZ
-1	P56817	3l5d	BDV	3l5b	BDO
-1	P39900	1jiz	CGS	3ehy	TBL
-1	P39900	1ros	DEO	2wo8	077
-1	O35904	2wxi	S30	2wxn	DLN
-1	P04058	1h22	E10	1gpn	HUB
-1	P00760	1lqe	IMA	3rx0	SW2
-1	P00760	102p	972	1032	801
-1	Q8A0N1	2wvz	KIF	2wwo	SWA
-1	Q02750	3orn	30R	4an9	2P7
-1	P66992	3r6c	17N	3twp	SAL
-1	P00749	1owi	426	1gja	135
-1	P11309	1yhs	STO	3jpv	1DR
-1	P39900	1ros	DEO	2wo9	068
-1	P07900	4fcr	0TM	4fcq	2N6
-1	Q16539	2yis	YIS	2baj	1PP

Binding mode changed/ preserved	UniProt ID	PDB id larger ligand	Larger ligand	PDB id smaller ligand	Smaller ligand
-1	O14757	2ym4	4YM	2wmx	ZY6
-1	P97612	2wap	PIX	3lj7	OHO
-1	P56817	3qbh	QBH	3pi5	3P5
-1	P00760	1oyq	T87	2fx6	270
-1	P00918	3r17	5UM	2weg	FBV
-1	P26663	4b74	1LH	4b6f	20L
-1	P24941	1y91	CT9	2vts	LZC
-1	P23458	4e5w	0NT	4ei4	0Q2
-1	P00523	3el7	PD3	3uqg	B5A
-1	P07900	3hek	BD0	3ekr	PY9
-1	P07900	3r4p	FU7	2xdx	WOE
-1	P20701	1xdd	AAY	1cqp	803
-1	P66992	3qqs	17C	3uu1	14B
-1	Q16539	3uvq	FS8	3uvp	048
-1	O94925	3u09	04A	3voz	04A
-1	P08473	2qpj	I20	1r1h	BIR
-1	Q16539	2yis	YIS	3nnv	MOL
-1	P24941	1pxp	CK8	1pxk	CK3
-1	Q8A3I4	2xii	TA9	2wvt	FHN
-1	P56817	4h3g	10Q	3rsx	RSV
-1	P00749	2viw	D56	2vin	505
-1	P26663	4b74	1LH	4b76	PW1
-1	P00918	3k2f	NKX	1bcd	FMS
-1	P07900	3hek	BD0	3k97	4CD
-1	P07900	2vci	2GJ	2xht	C0Y
-1	P23470	3qcj	NX4	3qck	NX5
-1	P0A5R0	3ivc	FG4	3ime	BZ2
-1	P23470	3qcj	NX4	3qce	NXY
-1	P02879	4hv7	19J	3px8	JP2
-1	P00918	3myq	E27	3sbi	E90
-1	O14965	3hoz	45B	3u06	0BY
-1	P09955	2pja	33Z	2piz	606
-1	P96222	308h	O8H	308g	O8G
-1	P45452	3i7i	518	3i7g	732
-1	P00918	2f14	FL1	4e3g	PHB
-1	O14965	3hoz	45B	3uok	0C6
-1	P00749	3ig6	438	3kgp	4AZ

Binding mode changed/ preserved	UniProt ID	PDB id larger ligand	Larger ligand	PDB id smaller ligand	Smaller ligand
-1	Q08638	2wc3	AM3	2cbv	CGB
-1	Q08638	2vrj	NCW	2jal	YLL
-1	P28720	3gc4	AAQ	3eou	PK3
-1	P56817	3msk	EV4	3msj	EV3
-1	Q16539	3bv3	P39	2rg5	279
-1	O92972	3u4r	08F	3u40	08E
-1	P24941	2vti	LZ3	2vth	LZ2
-1	P28720	3gc4	AAQ	3tll	62D
-1	P00760	1030	693	103e	696
-1	P56817	3lnk	74A	3ivh	1LI
-1	P00523	3f3v	1BU	3f3u	1AW
-1	P04062	2v3e	NND	3rik	3RI
-1	P09955	2pjc	343	2pjo	922
-1	O60674	4e6q	0NV	4fo9	JAK
-1	P08581	3c1x	CKK	3cth	319
-1	P04637	4agq	P96	4agm	P86
-1	Q08638	2vrj	NCW	2cbu	CTS
-1	P00760	102r	CR9	1035	802
-1	P09955	2pjb	983	2pj6	059
-1	P07900	2ykc	YKC	2yk9	YK9
-1	O14965	3m11	AKI	3k5u	PFQ
-1	P27487	3eio	AJH	2iiv	565
-1	P00760	3m35	M35	3rxp	SW3
-1	Q76353	3zso	O2N	3zsy	OM3
-1	P00742	2vwm	LZI	2p93	ME1
-1	Q16539	3bv3	P39	3mvl	38P
-1	O14965	3hoz	45B	4dea	NHI
-1	O14965	3hoz	45B	3u05	0BX
-1	P00749	3ig6	438	3khv	4AL
-1	P23458	4fk6	4AL	4ehz	JAK
-1	Q70I53	1zz1	SHH	1zz3	3YP
-1	Q16539	3hv5	R24	3hv7	1AU
-1	P18031	2qbp	527	2qbr	910
-1	Q81R22	3fl8	RAR	3fl9	TOP
-1	P07900	3r40	FU3	3r4n	FU5
-1	P56658	1qxl	FR8	2e1w	FR6
-1	Q16539	3hll	I45	3hp2	P36

Binding mode changed/ preserved	UniProt ID	PDB id larger ligand	Larger ligand	PDB id smaller ligand	Smaller ligand
-1	Q16539	3bv2	P38	4eh6	0ON
-1	O15530	3qcy	3Q3	3qcx	3Q2
-1	O15530	2pe2	464	2pe1	517
-1	P07900	2wi7	2KL	3rlp	3RP
-1	Q08638	2vrj	NCW	1oif	IFM
-1	P18031	2qbp	527	2qbs	024
-1	Q08638	2vrj	NCW	2j75	NOY
-1	P0A4Z6	3n86	RJP	3n7a	FA1
-1	Q04609	2c6c	24I	2jbj	G88
-1	P00749	1o5c	CR9	1gi9	123
-1	P04058	3i6z	G6X	3i6m	G3X
-1	O14965	3vap	0FY	3myg	EML
-1	P18031	2fjm	073	1bzj	PIC
-1	P28720	3ge7	AFQ	3rr4	HRD
-1	Q13526	2xpb	4GE	2xp7	4F8
-1	P03951	1zpc	716	2fda	682
-1	Q16539	2yiw	YIW	2yix	YIX
-1	P18031	1t4j	FRJ	1t48	BB3
-1	P00918	2x7s	WZC	3ibu	O48
-1	P00811	2i72	VA1	3ixg	BZB
-1	P56109	3c56	PH4	3c52	PGH
-1	Q76353	3zso	O2N	3zsz	OM2
-1	Q76353	3zso	O2N	3zt1	OM1
-1	A4GRE3	3rf4	FUN	3rf5	FUZ
-1	Q9F4L3	3iae	D7K	3iaf	TPP
-1	O94925	3u09	04A	3vp4	BP9
-1	O94925	3voz	04A	3vp2	BP0
-1	P03367	2qnn	QN1	2pqz	G0G
-1	P56817	3l5e	BDW	3l5d	BDV
-1	Q13526	2xpb	4GE	2xp6	4G2
-1	P00918	3ryx	RYX	3ryv	RYV
-1	P00749	1owk	303	1owe	675
-1	O14757	2ym4	4YM	2ym3	YM3
-1	P19491	2p2a	MP9	1p1q	AMQ
-1	P00760	2zfs	12U	3atk	SZ1
-1	P28720	2qzr	S79	3voy	SQO
-1	P03951	2fda	682	3bg8	INH

Binding mode changed/ preserved	UniProt ID	PDB id larger ligand	Larger ligand	PDB id smaller ligand	Smaller ligand
-1	P23470	3qcj	NX4	3qch	NX2
-1	P00734	3dhk	23U	2zfp	19U
-1	P07900	2wi7	2KL	2wi6	ZZ6
-1	Q16539	4aa5	NQB	4eh2	0OK
-1	P28720	1k4g	AIQ	1q4w	DQU
-1	Q9BJF5	3v51	I76	3i7c	BK2
-1	P00760	1y3w	UIP	1utn	ABN
-1	P21836	2ha2	SCK	2ha4	ACH
-1	P02829	2xx5	13N	2xx2	13C
-1	O60674	4e6q	0NV	4fo8	1RS
-1	P56817	3vv8	B02	3vv6	B00
-1	P00760	102n	762	1gi4	122
-1	P09955	2pjb	983	2piy	528
-1	P08311	1t32	OHH	1kyn	KTP
-1	P00918	3ni5	C1H	3d8w	D8W
-1	P62508	2p7z	OHT	2zas	1OH
-1	P43235	301g	O75	300u	O47
-1	Q8AAK6	2vot	NHV	2vmf	MVL
-1	P00760	102u	847	1gi1	BMZ
-1	P26663	3cj5	SX6	3cj2	SX3
-1	P00760	102u	847	102s	CR4
-1	P00760	102q	991	1gi6	124
-1	P00918	3rz7	RZ7	3rz8	RZ8
-1	P00760	1036	607	1039	780
-1	P00760	102u	847	103j	334
-1	P00918	3rz7	RZ7	3ryx	RYX
-1	P56817	2ohu	IP7	2oht	IP6
-1	P05981	105f	CR9	1p57	CR4
-1	P00918	3ryz	RYZ	3ryj	RYJ
-1	P00749	103p	655	1gi8	BMZ
-1	Q08638	2vrj	NCW	2j77	NOJ
-1	P00918	2q08	3CC	2nns	M25
-1	Q9WYE2	2zx7	ZX7	2zxa	ZXA
-1	P00760	1y3x	UIB	1y3y	UIR
-1	P00918	3ml2	SU0	3oys	OYS
-1	P18031	2qbp	527	2h4k	509
-1	P27487	3ccc	7AC	3ccb	B2Y

Binding mode changed/ preserved	UniProt ID	PDB id larger ligand	Larger ligand	PDB id smaller ligand	Smaller ligand
-1	P00918	1ze8	PIU	2nng	ZYX
-1	Q9WYE2	2zx9	ZX9	2zwz	ZWZ
-1	Q54276	1w1p	GIO	106i	0HZ
-1	P00918	2gd8	PO1	3ibi	BOW
-1	P00918	3rz1	RZ1	3ryy	RYY
-1	Q24451	2f7p	2SK	2f70	MSN
-1	P00918	3ryz	RYZ	3rz0	RZ0
-1	P04637	4agm	P86	4agl	P84
-1	Q24451	3d52	GHR	1hxk	DMJ
-1	Q24451	3ejt	HN6	1hww	SWA
-1	Q54276	1w1y	TYP	1w1p	GIO
-1	P78536	3lgp	50X	3kme	Z59
-1	Q89ZI2	2wca	NP6	2xm2	LOG
-1	Q13526	3kag	4D7	3kac	4BX
-1	P22498	2cer	PGI	2ceq	GIM
-1	Q13526	2xp8	4FY	2xp5	4FF
-1	P00918	3m67	E36	2weh	FB1
-1	P56817	4h3g	10Q	4ha5	13W
-1	P24941	2vtr	LZB	2vtm	LZM
-1	Q08638	2j7e	GI2	2ces	GIM
-1	Q13526	2xp8	4FY	2xp4	G14
-1	Q24451	2f7p	2SK	3dx3	YTB
-1	P04062	2v3e	NND	2v3d	NBV
-1	P24941	2vtj	LZ4	1wcc	CIG
-1	P0A5R0	3isj	A8D	3imc	BZ3
-1	P26663	3h5s	H5S	3d28	B34
-1	Q8A3I4	2xii	TA9	2xib	DFU
-1	O14757	2c3k	ABO	2c3l	IDZ
-1	Q24451	3dx2	MZB	3dx1	YHO
-1	P28720	3gc4	AAQ	3s1g	ITE
-1	P00918	3k2f	NKX	3s78	EVJ
-1	P18031	2qbs	024	2hb1	512
-1	P00918	3s72	EVE	3s76	EVH
-1	P27487	2qtb	474	20ph	277
-1	P27487	2ajl	JNH	1n1m	A3M
-1	P09955	2pjc	343	2pj3	86A
-1	P00760	1g36	R11	3gy4	PBZ

Binding mode changed/ preserved	UniProt ID	PDB id larger ligand	Larger ligand	PDB id smaller ligand	Smaller ligand
-1	P56817	4djv	0KM	4dju	0KK
-1	P28720	1k4g	AIQ	1s39	AQO
-1	P00918	3r17	5UM	2weo	FBW
-1	P28720	3gc4	AAQ	3gc5	2MQ
-1	P28720	3gc4	AAQ	3gev	SAQ
-1	P00918	3d8w	D8W	2wej	FB2
-1	P28720	1k4g	AIQ	1s38	MAQ
-1	Q873X9	3chc	ZRG	3ch9	XRG
-1	O15530	3qcy	3Q3	3qcq	3Q0
-1	O15530	3qcy	3Q3	3qcs	3Q1
-1	P56817	3rvi	RVI	3ru1	3RU
-1	O14965	3vap	0FY	3nrm	NRM
-1	P07900	3b28	B2X	3b27	B2T
-1	P07900	2xjx	XJX	2xab	VHD
-1	P00760	1fot	PR1	3rxf	4AP
-1	P00760	1g3c	109	1c5p	BAM
-1	Q08499	1y2d	4DE	1y2b	DEE
-1	P56817	3vv8	B02	3hvg	EV0
-1	P24941	2vti	LZ3	2vtl	LZ5
-1	P11309	3r02	UNM	3r00	UNJ
-1	P00918	3s71	EVD	3s75	EVG
-1	P26663	4b74	1LH	4b71	DJL
-1	P39900	1jiz	CGS	3lk8	Z79
-1	P07900	30wd	MEY	30w6	MEX
-1	P00918	3caj	EZL	3s77	EVI
-1	P00760	1g3c	109	1ce5	BEN
-1	P11086	2g71	FTS	1hnn	SKF
-1	P0A5R0	3iub	FG2	3isj	A8D
-1	O14965	2w1c	L0C	2w1f	L0F
-1	P00749	1sqa	UI1	1sq0	UI2
-1	P07900	2xjx	XJX	3k99	PFT
-1	P56817	3l5e	BDW	3l5f	BDX
-1	P0A6D3	1x8t	RC1	2aa9	SKM
-1	P22734	30zs	OZS	30zr	OZR
-1	P24941	2vtp	LZ9	2vtn	LZ7
-1	P07900	3b27	B2T	2wi2	ZZ3
-1	P00749	1gi9	123	1c5z	BAM

Binding mode changed/ preserved	UniProt ID	PDB id larger ligand	Larger ligand	PDB id smaller ligand	Smaller ligand
-1	P31749	3mvh	WFE	3mv5	XFE
-1	P07900	2xjx	XJX	3eko	PYU
-1	P06401	2w8y	486	1sqn	NDR
-1	O15530	3qd3	3Q5	3nus	JNZ
-1	Q16539	3iw7	IPK	4eh4	0OL
-1	Q9QYJ6	3qpp	PFW	3qpn	PFK
-1	O14965	2w1c	L0C	2w1d	L0D
-1	O60674	3e64	5B3	3e63	5B2
-1	P21836	4ara	C56	4a23	C56
-1	P04058	3i6m	G3X	1dx6	GNT
-1	P28482	2ojj	82A	2ojg	19A
-1	P04642	4al4	W7E	4ajk	88S
-1	P04642	4al4	W7E	4ajl	88W
-1	P00918	3ml2	SU0	1okl	MNS

Table (F.1) Complete and corrected list of related ligand pairs from Malhotra and Karanicolas. Only relevant data for this thesis is shown. Additional information is provided by the authors in their Supporting Information.[44], [216] A binding mode change in the table is indicated with 1 (unchanged -1).