Knowledge-based Optimization of Protein-Ligand-Complex Geometries

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

dem
Fachbereich Pharmazie
der Philipps-Universität Marburg
vorgelegt von

Andreas Spitzmüller
aus Leonberg

Marburg/Lahn 2011

Vom Fachbereich Pharmazie der Philipps-Universität Marburg als Dissertation am 12.07.2011 angenommen.

Erstgutachter	Prof. Dr. Gerhard Klebe, Institut für Pharmazeutische Chemie, Philipps-Universität Marburg
Zweitgutachter	Prof. Dr. Eyke Hüllermeier, Institut für Mathematik und Informatik, Philipps-Universität Marburg

Tag der mündlichen Prüfung: 13.07.2011

Die Untersuchungen zur vorliegenden Arbeit wurden auf Anregung von Herrn Prof. Dr. Gerhard Klebe am Institut füur Pharmazeutische Chemie des Fachbereichs Pharmazie der Philipps-Universität Marburg in der Zeit von März 2007 bis Juni 2011 durchgeführt.

Meinen Eltern

Kurzfassung

Ziel dieser Arbeit war die Entwicklung eines Programms zur Optimierung *in-silico* generierter Protein-Ligand Komplexe auf Grund der DrugScoreX Potentiale. Die Scoringfunktion DrugScoreX wird typischerweise für die Nachbewertung gebundener Ligandgeometrien genutzt, die von Docking Programmen erzeugt wurden. Daher sind diese Geometrien zunächst für die intern verwendete Scoringfunktion des gewählten Docking Algorithmus optimiert. Wird DrugScore auf eine solche Geometrie angewandt, so ist nicht automatisch eine aussagekräftige Bewertung garantiert, selbst wenn die gedockte Pose sehr nahe an einem DrugScore Optimum liegt. Bedenkt man die Steilheit der DrugScoreX Potentiale, so können bereits kleine Abweichungen der Atompositionen zu großen Unterschieden in der Bewertung führen. Im allgemeinen gilt dies für jede Nachbewertung, so dass eine lokale Optimierung anhand der verwendeten Scoringfunktion ausdrücklich empfohlen wird.

2009 wurde von O'Boyle et al. ausgeführt, dass eine lokale Optimierung grundsätzlich auf das jeweilige Tal der Potentialoberfläche beschränkt ist, in dem sich die Ausgangspose bereits befindet. Das heißt, es könnte ein tieferes Tal in der Nähe geben, das bei einer lokalen Optimierung nicht berücksichtigt wird, obwohl es ebenso eine zulässige Lösung darstellen würde. Das in dieser Arbeit entwickelte Programm MiniMuDS soll diesem Problem gerecht werden. Andererseits soll aber keine globale Optimierung durchführen, da dies eine vollständige Suche nach dem insgesamt bestbewerteten Bindemodus bedeutet, und damit zu einem neuen Docking Algorithmus führen würde. Stattdessen soll der neue Algorithmus nahe an der ursprünglich erzeugten Pose bleiben und diese nur an die DrugScoreX Funktion anpassen.

Das in dieser Arbeit entwickelte Programm sollte beide Anforderungen erfüllen, indem es eine streng lokale Optimierung vermeidet, ohne dabei jedoch eine globale Suche durchzuführen. Hierzu wurde eine Suchstrategie implementiert, die Elemente einer globalen Suche enthält, sich aber dennoch auf einen abgegrenzten Teil des vollständigen Suchraums beschränkt. Einfach ausgedrückt kann der Algorithmus kleine Hürden auf der Potentialoberfläche überwinden, jedoch nur wenn sich direkt dahinter auch ein tieferes Tal befindet. Größere energetische Barrieren zwischen grundsätzlich unterschiedlichen Konformationen können so nicht passiert werden.

Durch die Validierung von MiniMuDS konnten verschiedene wichtige Eigenschaften gezeigt werden:

1. Die Optima der angewandten Zielfunktion stimmen beeindruckend genau mit experimentell bestimmten Komplexstrukturen überein. Dies wurde durch die Optimierung von original Kristallstrukturen gezeigt, die in einer mittleren Abweichung von etwa 0,5 Å resultierten. Dies sind deutlich kleinere Abweichungen als im Fall von *in-silico* generierten Geometrien. Darüber hinaus fallen diese Abweichungen etwa in den Bereich der geschätzten Genauigkeit experimenteller Strukturaufklärung.

2. Das Ziel den vorgegebenen Bindemodus beizubehalten wurde erreicht. Die vorgestellte Methode erlaubt Modifikationen bis zu 2 Å rmsd gegenüber der Ausgangspose. Bemerkenswerterweise nutzten nicht einmal 5 % der optimierten Docking Lösungen diesen Raum aus. Diese zeigten durchschnittliche Abweichungen von etwa 1 Å auf.

3. Bezogen auf den rmsd zur Kristallstruktur verbessert MiniMuDS eine gegebene Konformation um etwa 0,1 Å. Die besten Ergebnisse wurden für bereits gut gedockte Posen mit einem ursprünglichen rmsd zwischen 1 und 2 Å beobachtet, die im Mittel um bis zu 0,3 Å verbessert wurden.

4. Es konnte gezeigt werden, dass die Überwindung der Einschränkungen einer rein lokalen Suche die erzielte Rangliste verbessert wurde.

Im Vergleich zu einer lokalen Optimierung wurden bis zu 4,7 % bessere Erfolgsraten bei der Erkennung nativ-ähnlicher Posen unter 2 Å rmsd auf Rang 1 erzielt. Für Posen unter 1 Å lag die Verbesserung bei 9,3 %.

5. Betrachtet man nicht nur die bestbewertete Lösung, sonder die gesamte Rangliste, so konnte gezeigt werden, dass MiniMuDS die Trennung zwischen nativ-nahen und falsch platzierten Posen deutlich verbessert. Geometrien mit niedrigen rmsd Werten tauchen häufiger auf den vorderen Positionen der Rangliste auf.

6. Die Berücksichtigung zusätzlicher flexibler Komponenten in der Optimierung ist mit MiniMuDS leicht zu bewältigen, wodurch die erzielten Ergebnisse deutlich verbessert werden können. Dies wurde am Beispiel von flexiblen Protein Seitenketten und an der Bindung beteiligter Wasser Moleküle gezeigt.

7. Es wurde gezeigt, dass es ausreichend ist, die zehn besten Lösungen eines Docking Experiments zu Optimieren. Dadurch wurden durchgängig etwas bessere Ergebnisse erzielt als bei der Optimierung aller fünfzig erzeugten Lösungen. Bei 80 % geringerem Rechenaufwand wurden so bis zu 4,7 % bessere Erfolgsraten erzielt.

Besonders der letzte Punkt bestätigt, dass es empfehlenswert ist, sich auf solche Posen zu konzentrieren, die bereits von einer anderen Scoringfunktion gut bewertet wurden. So kann zusätzlich von einem Konsensus Effekt profitiert werden.

Mit Blick auf die erzielten Ergebnisse muss die Anwendung zumindest einer lokalen Optimierung dringend Empfohlen werden, bevor DrugScore für Nachbewertungen herangezogen wird. Allerdings wird die Verwendung einer darüber hinausgehenden Suchstrategie wie sie in MiniMuDS implementiert wurde nahe gelegt. Insbesondere bei kleineren Leitstrukturen verbesserte die vorgestellte Methode die Nachbewertungsergebnisse deutlich.

Contents

Chapter 1

Introduction

In the present thesis a new method for the knowledge-based optimization of protein-ligand complex structures was developed and validated. This chapter gives a general survey of the scientific background on which this thesis is based (section 1.1), followed by a short outline of the aims pursued with the presented work (section 1.2).

1.1 Scientific Background

In structure-based drug design protein-ligand docking is a well established and widely used technique. It is applied in lead identification as well as lead optimization to predict binding modes and affinities of compounds in complex with a target protein. Sousa et al. (2006) give a review of the current status of protein-ligand docking as well as available algorithms and software packages.

From a computational point of view docking can be divided into two equally important sub-problems. The first one is the generation of reasonable binding geometries inside the protein pocket. This is often called the *docking problem*. Current docking engines usually suggest multiple ligand poses. The second one is the detection of those poses that most closely resemble the biologically relevant situation out of the whole set of suggested solutions. Therefore, a scoring function is applied to rank all geometries

expecting the best one to appear on rank one. Thus, this is referred to as the *scoring problem*.

A community-wide accepted criterion for reasonable ligand geometries is an rmsd value below 2 Å compared to an experimentally determined structure. Modern docking programs are generally able to explore conformational space sufficiently to generate ligand geometries within this threshold. This was shown in various studies evaluating popular docking algorithms on independently compiled test sets (Cross et al., 2009; Warren et al., 2006; Kellenberger et al., 2004). In this light, the docking problem has often been claimed to be "sufficiently well solved".

In contrast, the problem of identifying the best pose among all generated solutions still remains to be satisfactorily resolved. Dozens of scoring functions have been reported in literature in the past. Research in this field is still going on as reflected in regular benchmark studies on scoring functions (Cheng et al., 2009; Ferrara et al., 2004; Wang et al., 2003). Current scoring functions can be roughly classified into three groups: force-field methods, empirical and knowledge-based functions.

The first ones employ molecular mechanics terms to estimate interaction energies between ligand and protein. Usually non-bonding terms like electrostatic and van-der-Waals interactions are applied. Internal energy contributions of the binding partners like torsional energies are also frequently included. In addition, desolvation energies can be considered through implicit solvent models. Several scoring function are based on the well established CHARMM (Brooks et al., 1983) and AMBER (Weiner et al., 1984) force fields like the one originally built into the docking program DOCK (Meng et al., 1992). Another representative is the GoldScore function from the program Gold (Jones et al., 1997).

Empirical scoring functions divide the total binding free energy into several additive terms that account for different interaction types, such as hydrogen bonds, hydrophobic or aromatic interactions. Entropic contributions like the number of immobilized rotatable bonds in the complexed state can also be included. The weighting coefficients for the individual terms are fitted by regression analysis on a set of protein-ligand complexes

with known binding affinity. One of the first functions of this class was the SCORE1 function used in the *de novo* design tool LUDI (Böhm, 1994). Other well known examples are the scoring function built into the docking program FlexX (Rarey et al., 1996) which is a variant of SCORE1, or the function ChemScore (Eldridge et al., 1997).

The last group is made up of knowledge-based scoring functions. These methods derive statistical potentials from intermolecular atomic distances as they are observed in 3D structure databases like the Protein Data Bank (PDB) (Berman et al., 2000) or the Cambridge Structural Database (CSD) (Allen, 2002). They are based on the idea that frequently occurring interaction distances are likely to be energetically more favorable than less frequently observed ones. Important representatives are the functions PMF (Muegge, 2006; Muegge and Martin, 1999), ASP (Mooij and Verdonk, 2005) and DrugScore (Velec et al., 2005; Gohlke et al., 2000).

1.1.1 Knowledge-Based Scoring: the DrugScore Function

In the method developed in the present thesis knowledge-based DrugScore potentials play a central role. Therefore, this section briefly describes the concept behind the DrugScore function.

Initially, the statistical potentials were derived from non-covalent atomic contacts in protein-ligand complexes as stored in the PDB (Gohlke et al., 2000). Structures were retrieved using the ReLiBase system (Hendlich et al., 2003; Hendlich, 1998) which contained 6 026 complexes in January 1998. These structures were applied to a set of filters like a minimum required resolution of 2.5 Å, no covalently bound ligands and a number of non-hydrogen ligand atoms between six and fifty. As DrugScore was designed to predict and evaluate protein-ligand interactions of drug-like compounds, this was a suitable choice for the underlying knowledge base.

However, the success of knowledge-based approaches strongly depends on the frequency of the events to be predicted in this knowledge base. Due to the limited number of structures used for potential extraction, a couple of PDB-based pair potentials actually showed minor statistical

significance.

In 2005, Velec et al. reported a new version, DrugScore^{CSD}, where the potentials were retrieved from the spacial distribution of non-covalent interactions in small molecule crystals as stored in the CSD. This was motivated by the assumption that atomic contacts in small molecule crystals are subject to the same physical principles as those in protein crystals. A total of 28 642 small molecule structures were selected for potential extraction. Considering this data base, the statistical basis for individual contact types was much broader as compared to the PDB-based potentials. This led to a highly increased statistical significance of many pair potentials along with an improved performance. In addition, the substantially higher resolution and experimental accuracy of small molecule crystal structures is related to lower positional uncertainties for individual atoms compared to the PDB.

On the other hand, these improvements introduced a remarkable side effect: Although the positions of the extrema are generally the same, the CSD-derived potentials show steeper energy wells and are more sharply bounded. They show more pronounced extrema as compared to the PDB-based version. Figure 1.1 illustrates this effect for the charged interaction between a deprotonated carboxylate oxygen and a protonated amidine nitrogen. This in turn made DrugScore^{CSD} more sensitive to small conformational and positional variations compared to the earlier PDB version.

Since 2005 again a new version of the DrugScore function was developed by Gerd Neudert in our group which is not yet published, however. This version will be called *DrugScoreX* in the following. Since only the final potentials are relevant for the presented work, the advancements in the derivation procedure are not explained here. Instead, only the major differences in the resulting potentials between DrugScoreX and the older versions are described.

DrugScoreX potentials are also derived from the CSD. A total of 326 685 structures have been selected to get the potential set employed in this work. Yet, a new set of atom types has been applied. Both DrugScore^{PDB} and DrugScore^{CSD} potentials were derived for atomic contacts according to a

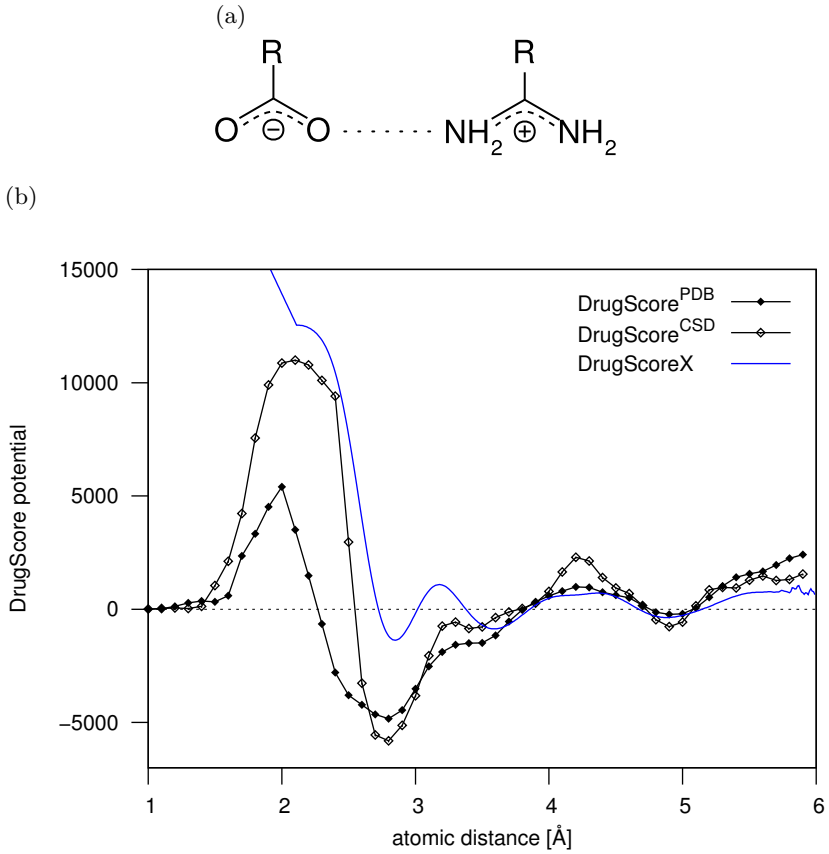


Figure 1.1: (a) Contact between a deprotonated carboxylate oxygen and a protonated amidine nitrogen. (b) Distance dependent pair potentials from DrugScore^{PDB} (filled squares), DrugScore^{CSD} (empty squares) and DrugScoreX (blue curve).

subset of the Sybyl atom types[1]. For the carboxylate-amidine example in Figure 1.1 these are the Sybyl types O.co2 and N.pl3, respectively. In contrast, DrugScoreX is based on the new fconv atom types (Neudert and Klebe, 2011) which are much more differentiated. 147 different types were discriminated for the potentials used in the presented work[2]. The appropriate types for the carboxylate-amidine example are O.co2 and N.mih, with the corresponding potential also given in Figure 1.1.

The bin width for sampling contact frequencies from the underlying knowledge-base was originally chosen to be 0.1 Å. As potential values for neighboring bins can still differ largely (cf. Figure 1.1), an interpolation was necessary for intermediate atomic distances. In DrugScoreX this interpolation is already included in the derivation procedure of the potentials. Discrete equidistant potential values are then recorded using a step size of 0.01 Å. This makes an ad-hoc interpolation unnecessary as instead a much more efficient look-up table of the discrete values can be used. In addition, the new potentials are extended by a linear term for very short atomic distances (arrow in Figure 1.1). Since obviously no observations can be made in the knowledge-base at these distances, potentials are artificially extended in this range to account for the strong repulsion that would occur between two non-bonded atoms. This term is attached at a distance corresponding to the larger one of either the first maximum of the derived potential or the sum of the van-der-Waals radii of the two involved atoms.

Figure 1.2 illustrates the previously described effects for an aromatic contact between a sp^2-carbon atom and a nitrogen like it occurs between a tryptophan side chain and a pyridine ring, for example.

Although the same tendencies are observed, they are not as prominent as in the case of the polar interaction shown in Figure 1.1. The final potentials appear much more similar. However, this is not necessarily the case for all apolar contacts. This gets obvious when looking at the hydrophobic interaction between a leucine side chain and an isobutane.

[1] See Table A.1 for a complete list of included types.
[2] A complete list is given in Table A.2.

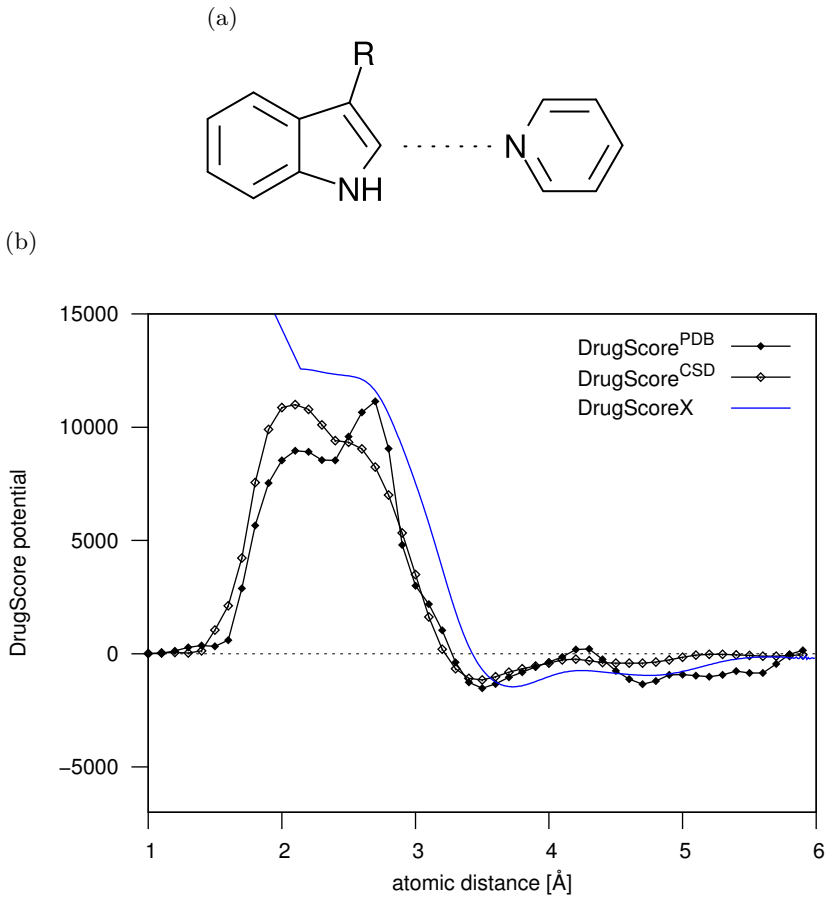


Figure 1.2: (a) Contact between an aromatic sp^2-carbon and an aromatic nitrogen. (b) Distance dependent pair potentials as shown in Figure 1.1.

Among others, two different contacts can be observed. The first one occurs between two CH_3 groups (called *primary* as they are bound to only one other heavy atom). The second one occurs between two CH groups (called *tertiary* as they are bound to three other heavy atoms): Both primary and tertiary carbon atoms are aliphatic sp^3-hybridized carbons. Thus, using the Sybyl atom type conventions, the type C.3 is assigned to all involved carbon atoms. Accordingly, both DrugScore^{PDB} and DrugScore^{CSD} apply the same potential to score the two different contacts. The respective graphs are given in Figure 1.3. In contrast, DrugScoreX discriminates between the types C.3p for a primary sp^3-carbon which is bound to only one other heavy atom and C.3t for a tertiary sp^3-carbon bound to three heavy atoms. Thus, two different potentials are applied in this case. Here the impact of the new set of atom types gets obvious. Whereas the potential used for the primary carbons shows a clear minimum around 4 Å, there is no minimum at all observed for the tertiary carbon contact. Instead, larger distances are clearly preferred. This is easy to understand since two tertiary carbons are likely to be sterically hindered from coming that close together, as opposed to the primary carbons. Nevertheless, it is impossible to distinguish these two cases when using Sybyl types. As a consequence, the Sybyl type potentials form a kind of "averaged" potential.

A central aspect common to all DrugScore versions is the exclusion of hydrogen atoms from the calculation of interaction potential values. It is assumed that the influence of a hydrogen on the interaction between two molecules is already implicitly regarded in the interaction potential of the corresponding heavy atoms interacting through that hydrogen. This entails the advantage of being independent of precise coordinates for hydrogen atoms that are usually not known even for experimentally solved protein structures.

1.1.2 Rescoring and Consensus Scoring

As mentioned above, there were plenty of scoring schemes proposed in the past that build on diverse principles. Different scoring functions focus on

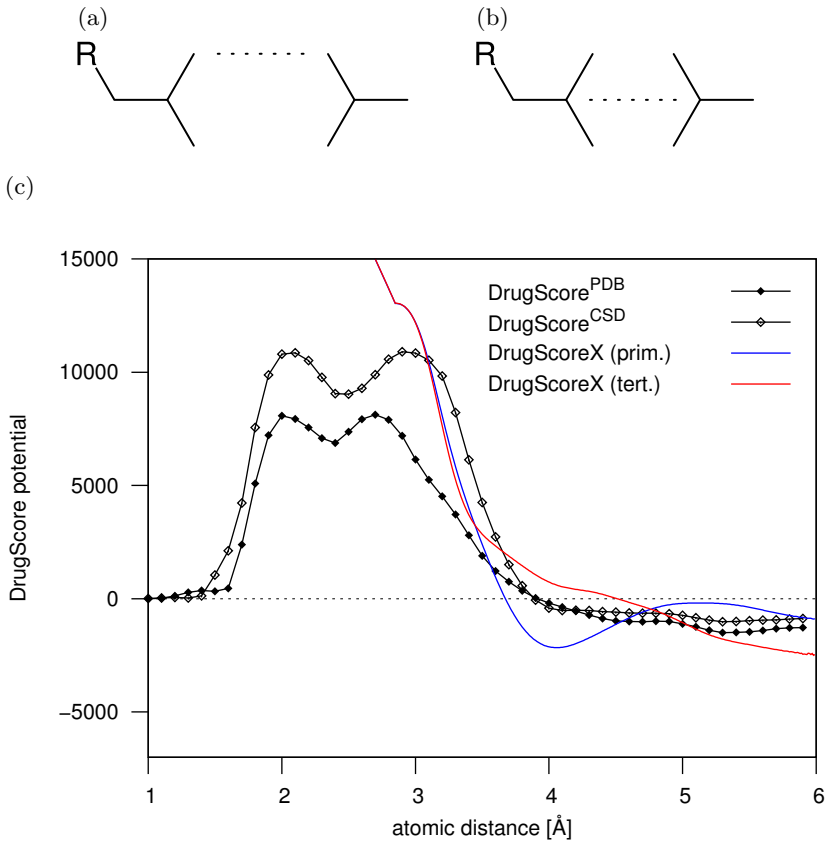


Figure 1.3: (a) Contact between two aliphatic *primary* sp^3-carbon atoms. (b) Contact between two aliphatic *tertiary* sp^3-carbon atoms. (c) Distance dependent pair potentials as shown in Figure 1.1.

different features and interaction types that are considered important for binding. The notion of *rescoring* tries to take advantage of these differences rather than considering them as contradicting.

Therefore, a rescoring strategy employs two different scoring functions. One is applied in the docking phase to generate reasonable binding modes, the other one to score and rank the solutions to identify the best ones. This was often reported to be more successful compared to the ranking obtained from the docking function alone (Stahl, 2000; Gohlke et al., 2000; Wallqvist and Covell, 1996).

Consensus scoring as a special case of rescoring goes one step further. It is based on the idea that the combination of results from different methods should lead to an improved performance and reliability. Thus, in order to making maximal use of the available scoring functions, consensus scoring integrates the results of several functions within one experiment. In the context of protein ligand docking, this idea was introduced by Charifson et al. (1999). Since than, a number of studies reported success stories of consensus scoring, particularly in the field of virtual screening (Oda et al., 2006; Clark et al., 2002; Bissantz et al., 2000). See Feher (2006) for a review on consensus scoring.

The DrugScore function is typically used for rescoring since it is provided as a stand alone function that cannot be incorporated directly into arbitrary docking programs without much additional effort. Bearing this in mind, the above described effects due to the sharpness of the DrugScore^{CSD} and the new DrugScoreX potentials become even more relevant. The poses, subjected to rescoring by DrugScoreX, have initially been generated and optimized under the influence of a different scoring function. However, local and global optima of the latter function and DrugScoreX will not be exactly the same. Thereby, a given ligand geometry positioned in a local optimum of the scoring function applied in docking can obtain a quite unfavourable DrugScoreX value even though a local DrugScoreX optimum may be in close vicinity.

Figure 1.4 illustrates this phenomenon for two imaginary scoring functions plotted against the conformational space. The same conformation

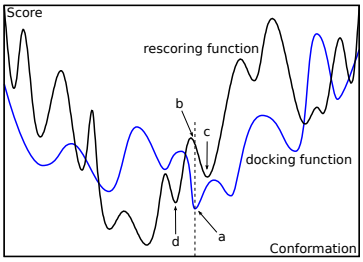


Figure 1.4: The discrepancy between docking score (blue) and rescoring function (black) makes a local optimization necessary to obtain valid rescore values.

that scored best according to the docking score (labeled *a*) can obtain a relatively poor value from the rescoring function (*b*). The corresponding local minimum on this function (*c*) would lead to a more meaningful rescore value, however.

This holds generally true for any attempt of rescoring. For this reason a local optimization of poses according to the applied rescoring function is essential in every rescoring scheme (Cole et al., 2005). O'Boyle et al. (2009) recently emphasized this point again, noting however, that a strictly local optimization cannot escape the energy well on the potential surface a pose resides in. Thus, a geometrically very close geometry from a neighboring well that would be better scored (Figure 1.4, label *d*) might not be considered.

This limitation of standard local optimization methods will be a central aspect in the presented thesis.

1.1.3 Current Obstacles in Protein-Ligand Docking and Scoring

From its early stages in the 1980s protein-ligand docking evolved to be routinely applied in today's structure based drug design projects. Nonetheless, a couple of problems are still not resolved satisfactorily. Two of the most prominent challenges comprise protein flexibility (B-Rao et al., 2009) and the participation of water molecules in ligand binding (de Beer et al., 2010).

It is known, that even small changes in the binding site conformation can strongly influence the performance of a docking experiment (McGovern and Shoichet, 2003). In another study Erickson et al. (2004) showed that an increasing degree of observed protein flexibility is directly mirrored in a decreasing docking accuracy. A detailed analysis of protein-ligand complexes from the PDB showed that at least 60 to 70 % of the binding pockets can be expected to undergo side chain rearrangements upon ligand binding (Najmanovich et al., 2000). At least, they show only a small number of flexible residues in general, with just about 10 % of the pockets exhibiting five or more such side chains. A fact that could help to restrict the problem complexity in docking applications.

Water molecules within the active site of a protein can play a key role during the process of ligand binding. Generally, water can stabilize a protein conformation due to its ability to act both as hydrogen bond donor and acceptor. This allows a water molecule to form multiple interactions, thereby mediating polar contacts between protein and ligand. On the other hand, a ligand that can mimic the structural role of a water by displacing it to bulk solvent may be energetically favoured (de Beer et al., 2010). Lu et al. (2007) found three bridging water molecules on average in every structure of a set of 392 examined protein-ligand complexes. More than 85 % of these structures contained at least one bridging water at the protein-ligand interface. A docking study by Roberts and Mancera (2008) on 240 protein-ligand complexes revealed a significant increase of docking accuracy when crystallographic water molecules were included.

Although the importance of both protein flexibility and water molecules for the prediction and evaluation of protein-ligand interactions is known, there is still no standard way of handling these effects. Both problems will be addressed in this thesis.

1.2 Previous Work and Aims of this Thesis

The most important aim of the presented work was the development of a computational method to optimize *in-silico* generated protein-ligand com-

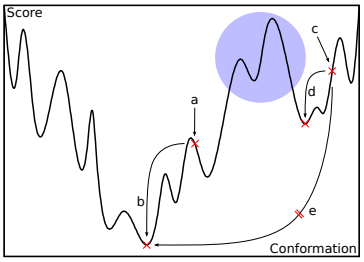


Figure 1.5: Semi-global search strategy as implemented in MiniMuDS.

plex geometries under the influence of the knowledge-based DrugScore potentials. This optimization approach should first of all account for and try to efficiently avoid the above-mentioned problem of strictly local optimizations. On the other hand, a true global optimization was also not desirable. This would be equivalent to the development of a new docking program. The idea of this work, though, is to adapt a given geometry to its surrounding energy landscape defined by the DrugScore function without generating completely new binding modes.

Therefore, a global minimization strategy based on local optima smoothing (Addis et al., 2005) was implemented into the program *MiniMuDS* (**Mini**mizing **M**olecules **u**sing **D**rug**S**core potentials). This algorithm does not perform a global search in the entire conformational and configurational space, but rather a *semi*-global one within a tightly restricted part of the search space. However, this restricted space is still significantly larger than the area covered by a strictly local search.

Generally, one can think of this optimization strategy to proceed with the search for better geometries in neighboring energy wells as long as they provide deeper minima. Major energy barriers, however, cannot be passed. Figure 1.5 illustrates this idea with the help of the previous example function. A given starting conformation a will be optimized to the global minimum b, which would not have been found by a strictly local optimization. In contrast, conformation c will end up in the local minimum d due to a large energy barrier (blue circle), while a true global optimization would try to find all the way down to the global optimum (e).

The basic idea for such an optimization method was introduced by Hans Velec in his PhD thesis (Velec, 2008). There, the scoring function DrugScore^{CSD} was developed. Based on these new potentials, the idea of a minimizer that takes further advantage of the improvements of the new DrugScore version to get enhanced geometries and scoring performance was born.

The initial implementation, written in Fortran 77, was based on the software tool *Torso* (Klebe and Mietzner, 1994) which was developed to generate biologically relevant molecular conformations. Therefore, internal van-der-Waals interactions and knowledge-based torsion potentials were used to find preferred conformations. Velec added the DrugScore potentials to this model to account for the interactions to the protein environment during the conformational search. A first evaluation of this method using a set of 63 protein-ligand complex structures yielded promising results.

For the presented work, the first task was to rewrite the minimizer to get a stand-alone software tool independent from the Torso program. The programming language Java 1.5 was chosen to accomplish this. The major part of this work then comprised an extensive validation of the minimization procedure including the search for a suitable parameter setup of the implemented algorithm.

The second step was to extend the optimization framework to enable multi-component optimizations. Thereby, the simultaneous consideration of additional flexible components beside the actual ligand influencing each other within one minimization run is facilitated. Various applications are imaginable for such a method. Two of the most important challenges in current structure-based drug design can be addressed this way: First, protein flexibility can be incorporated by treating one or more amino acid side chains as flexible to account for induced-fit adaptions. Second, water molecules can be included in the optimization as they are known to play a key role in ligand binding. Another scenario could be the optimization of several small fragments at the same time. Finally, the inclusion of cofactors as flexible parts in the optimization would be possible.

In the following, chapter 2 describes in detail the theory and implemen-

tation of the employed energy model and the selected optimization algorithm. This also includes the parametrization of this method. The data sets and methods used to validate MiniMuDS are introduced in chapter 3. The results of this validation are reported in chapter 4 which also contains selected case studies to exemplify the functionality of the developed procedure in more detail. Finally, chapter 5 completes this thesis with an outlook on possible further developments and research directions to further enhance MiniMuDS considering the quality of the results and the applicability of the software.

Chapter 2

2 MiniMuDS: Theory and Implementation

Whenever an optimization problem is to be solved, the decision on the applied optimization method is a crucial point. Two of the most important factors influencing this decision are the purpose of the optimization and the characteristics of the objective function. The better these factors are known a priori, the easier an appropriate algorithm can be chosen. Bearing in mind the aims of the presented work as described in section 1.2, a local optimization algorithm is ineligible. Instead, a global optimization algorithm that can be restricted to a clearly defined subdomain of the entire search space is needed.

The objective function is described in detail in section 2.1. Based on these considerations a suitable optimization algorithm was chosen for the implementation. This algorithm is presented in section 2.2. In the following, the corresponding parametrization (section 2.3) and some examined variations to this algorithm are introduced (section 2.4). Finally, the necessary modifications to enable multi-component optimizations and the corresponding pitfalls are explained in section 2.5.

2.1 The Energy Model

In the presented work protein-ligand interactions are to be optimized to improve bound ligand geometries and their scoring. The scoring function DrugScore represents the primary objective function on which the optimization procedure should operate. Therefore, DrugScoreX potentials account for all *inter*-molecular interactions between a ligand and its protein receptor in the applied energy model (see section 1.1 for an introduction to DrugScoreX potentials). However, to avoid a collapse of the ligand in the course of the optimization, *intra*-molecular interactions have to be modeled, too.

The variables to be optimized comprise the ligand's torsion angles about all freely rotatable bonds as well as its free translation and rotation in space. In case of modifications at a torsion bond always the smaller of the two ends is rotated. Bond lengths and angles are fixed, assuming that the pose generating program already adjusted these parameters to reasonable values. The protein is also treated as a rigid body throughout the whole procedure. Thus, only torsion energies and internal van-der-Waals contacts are explicitly modeled during the optimization. The resulting objective function is given as

$$E = \alpha E_{DS} + \beta E_{tors} + \gamma E_{vdW}, \tag{2.1}$$

with E_{DS} = DrugScoreX potential, E_{tors} = torsion potential and E_{vdW} = van-der-Waals potential. The factors α, β and γ allow for different weighting of these three interaction types.

The van-der-Waals interactions are described using a standard 12-6-Lennard-Jones potential. The parameters are taken from the Tripos force field (Clark et al., 1989) as it is implemented in the Sybyl-X[1] software. The torsional degrees of freedom are modeled through knowledge-based torsion angle potentials. Like the DrugScoreX interactions they are derived from

[1] Tripos International, 1699 South Hanley Rd., St. Louis, Missouri, 63144, USA. `http://www.tripos.com`

the CSD as described by Klebe and Mietzner (1994).

To get a general idea of the shape of this objective function, it will be examined in more detail in the following. For the energy surface of a small molecule ligand binding to a target protein, an overall funnel shape can be expected (Tsai et al., 1999). However, this funnel can be locally strongly perturbed. A similar behavior for the objective function in the current optimization problem is assumed, since the non-covalent interactions between a small molecule ligand and its protein receptor in the complex state are to be optimized.

In the following, the validity of this assumption is illustrated with the help of the PDB structure of a thymidylate synthase in complex with Raltitrexed (PDB code 1hvy) as shown in Figure 2.1. One-dimensional cross sections of the high-dimensional energy surface were sampled for the translation of the rigid ligand along and its rotation around the three spatial axes as given in Figure 2.1 (a). The corresponding sections through the hypersurface are given in Figure 2.2 with interaction potentials shown on a relative scale.

All these transformations show a clearly pronounced optimum around zero which coincides with the configuration of the crystal structure. A broad region around this point is "falling down" onto the optimum in a funnel-like manner. Although there are additional local optima observed, none of them reaches the area of favourable interaction scores below zero (depicted in the dashed lines in Figure 2.2).

In fact, the observed rapid increase to highly unfavorable interaction potentials with increasing deviations from the input structure is not surprising. At least parts of the ligand are deeply buried inside the protein. This immediately causes heavy atomic clashes, no matter in which direction the ligand is translated or rotated. Extremely short contact distances as they are characteristic for such clashes yield exceptionally high potential values observed in these cases.

However, modifications on this scale, where large overlaps between ligand and protein are observed, are not part of the application domain of the presented method. When dealing with *in-silico* generated ligand geome-

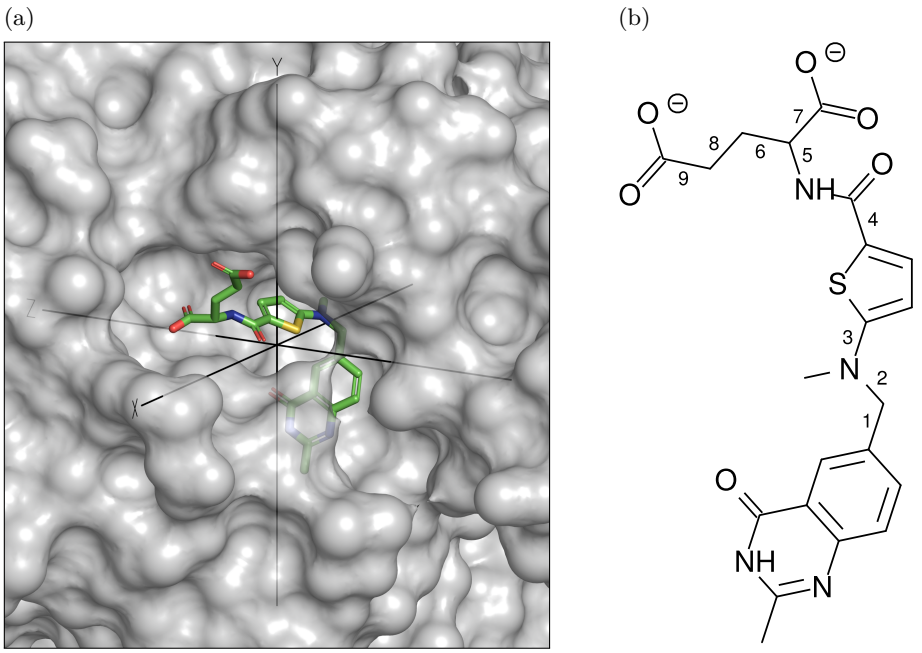


Figure 2.1: (a) The binding pocket of PDB complex 1hvy. The spatial axes correspond to the coordinate system in the PDB file and are centered at the geometric center of the ligand. (b) The chemical structure of the ligand with the freely rotatable bonds numbered from 1 to 9.

tries, one would expect them to be free of such overlaps. And of course, the here developed optimization method is also not intended to introduce atomic clashes into a complex structure.

Figure 2.3 shows the same interaction potentials as Figure 2.2, but focused on a 25-fold magnified section around the input structure. On this scale the funnels turn out to show extremely sharp and steep walls. This means that even smallest modifications that do not necessarily lead to atomic clashes still have a large effect on the score. These one-dimensional cross sections also suggest a similar funnel-like shape for higher dimensional combinations of the individual dimensions with an optimum close to the origin.

The picture changes, when examining the remaining degrees of freedom,

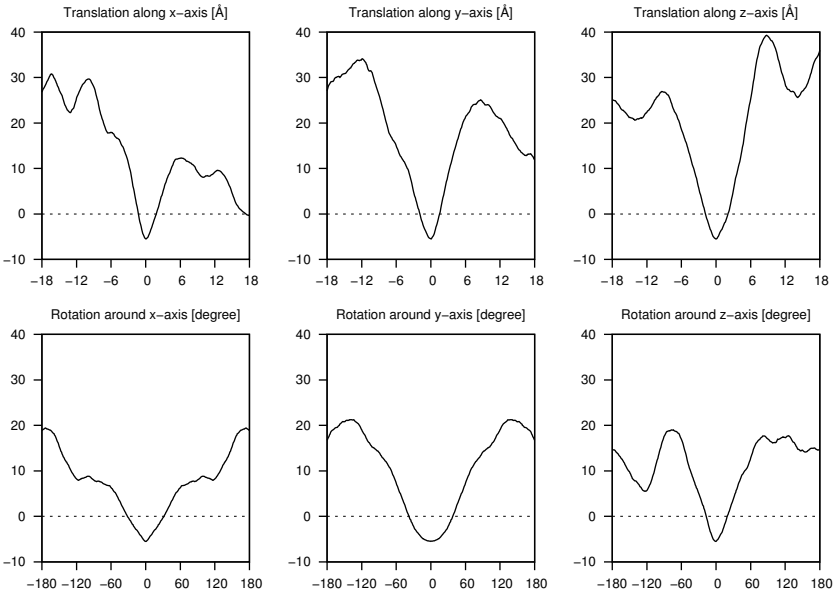


Figure 2.2: Relative interaction potentials according to Equation 2.1 for the translation (top) in steps of 0.1 Å and rotation (bottom) in steps of 1°. The axes correspond to Figure 2.1 (a).

namely the torsion angles at the nine freely rotatable bonds as labeled in Figure 2.1 (b). The potentials obtained through rotation about each of these bonds separately are given in Figure 2.4. Only one out of nine torsion angles (bond 2) shows a single sharp minimum across the entire definition range comparable to the previous ones. The others either comprise multiple favorable conformations (bonds 1, 3 and 5), exhibit a rather broad and flat funnel ground (bonds 4 and 8), or do not show a recognizable funnel shape at all (bonds 6, 7 and 9). Particularly variations of the torsion angles 7 and 9 seem to have nearly no influence on the total score.

This is a straight-forward consequence of the previous considerations. Modifications at bond 9 are limited to the rotation of a single carboxylate group, for example. Such modifications will never lead to atomic clashes with the protein. The rotated atoms are both most likely deprotonated carboxylate oxygens. These atoms will have a significant influence on the total interaction potential only if there is a suitable polar interaction

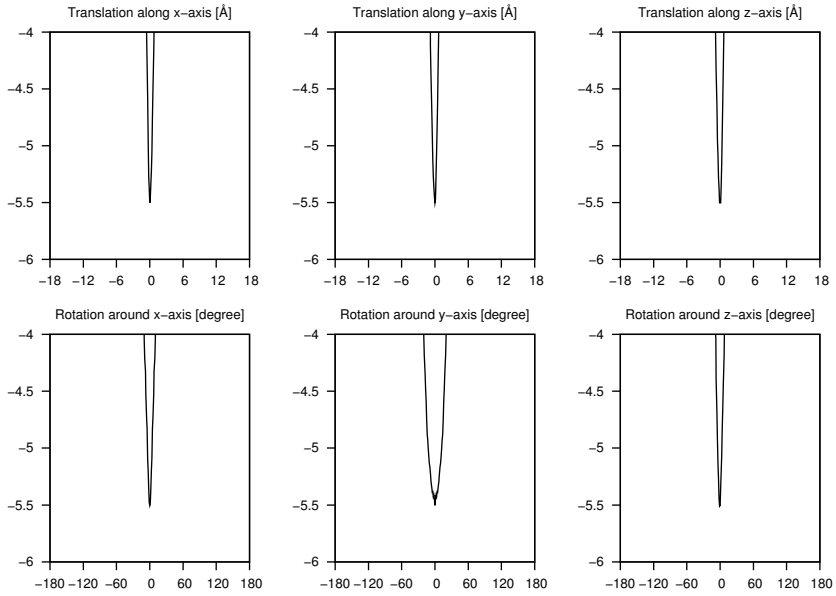


Figure 2.3: Relative interaction potentials as in Figure 2.2, but 25-fold magnified and focused on the region around the input structure.

partner available on the protein side. However, phenylalanine 80 is the only amino acid accessible within 5 Å (see Figure 2.5). This side chain cannot provide a hydrogen-bond donor functionality. Thus, the impact of modifications at this bond on the total score is rather small, independent of the actual configuration. In contrast, rotations around bond 2 lead again immediately to severe clashes in any direction, reflected in a pronounced and steep funnel in the corresponding energy surface.

Of course, it is impossible to depict the 15-dimensional hypersurface for the entire interaction potential of the example structure 1hvy. In combination with other modifications, the carboxylate group at bond 9 might well get into contact with polar protein atoms.

Altogether, the previous considerations about the one-dimensional subspaces still suggest a generally funnel-like shape. However, one has to be aware that this funnel can be distorted in a way to exhibit broad and flat grounds or to resemble narrow but long canyon-like structures. Additionally, the hypersurface can be strongly perturbed by local minima.

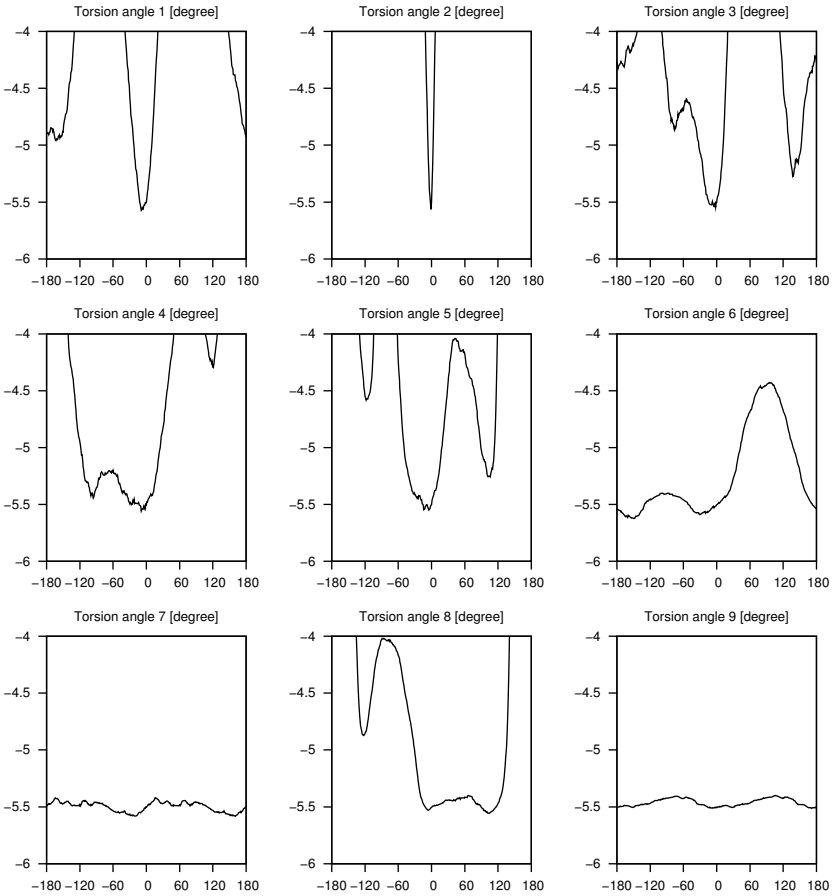


Figure 2.4: Relative interaction potentials for the torsion angles of ligand 1hvy sampled in steps of 1°, numbers according to Figure 2.1 (b).

This is sketched in Figure 2.6 which shows the interaction potential for the two-dimensional sub-space of torsion angle 4 and 5 as a contour map.

Considering the dimensions for these torsion angles separately as given in Figure 2.4, both adopt their global minimum close to the input structure. In the combined two-dimensional space a corresponding optimum also occurs (marked *a* in Figure 2.6). However, this optimum is no longer the global one. Instead it is part of a wide and rather flat plane (red contour line) that also comprises the global optimum *b* as well as two further local optima (marked *c*) with a lower potential compared to *a*. This plane

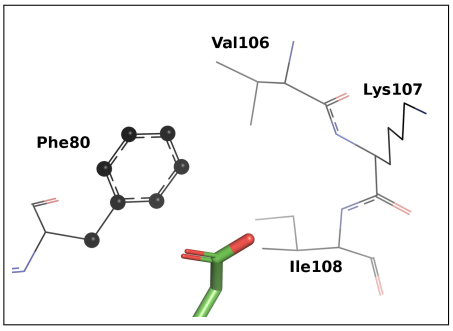


Figure 2.5: Local environment of the carboxylate group at rotatable bond 9. Protein atoms within 5 Å of the carboxylate oxygens are shown as spheres.

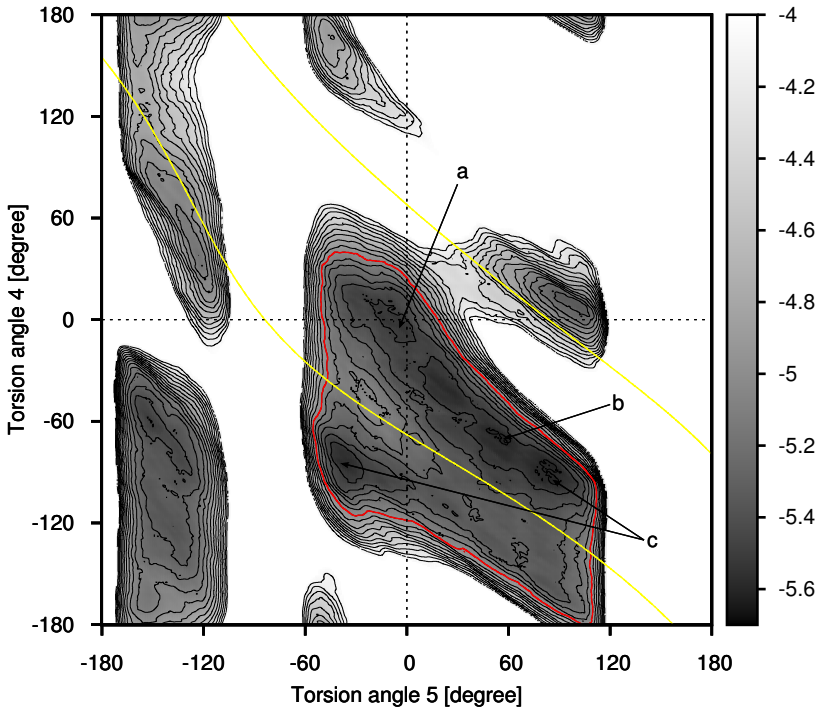


Figure 2.6: Relative interaction potentials according to Equation 2.1 for modifications at the torsion angles of bond 4 and 5, numbered as in Figure 2.1 (b). Values sampled in steps of 1° in each direction. Label *a* and *c* mark local minima, *b* the global minimum.

is surrounded by relatively steep walls indicated by the very dense contour lines. So there is still a funnel-like structure formed.

To reach the global minimum, starting from the input structure at the origin, rather strong modifications are necessary with rotations of -70° around bond 4 and 58° around bond 5. Yet, these modifications can partly compensate for each other due to the direct dependency between the torsion angles at the two bonds. Since they encompass a planar amide bond on both sides, bond 5 is part of the moiety affected by modifications at bond 4 and is oriented nearly in parallel with a deviation of only 6.8°. This results in only small conformational differences between the crystal conformation in the origin and the minimum conformation b in terms of an rmsd value of only 1.1 Å.

The dependency between these two variables is generally reflected in the observed rmsd. The region enclosed by the yellow lines in Figure 2.6 is characterized by rmsd values below 2 Å compared to the initial conformation at the origin. An almost linear dependency becomes apparent. Interestingly, the funnel containing the best scored conformations is distorted exactly in parallel to this region of mutually compensating modifications. Such effects are expected to get even stronger in higher-dimensional spaces where higher order dependencies can occur. This can yield to distortions of a funnel in multiple directions. Nevertheless, a general funnel-like shape with steep walls surrounding regions of favored conformations can be assumed.

2.2 The Minimization Strategy

2.2.1 Optimization Problems in General

Generally, an optimization problem can be formalized as

$$\begin{aligned} x^* &= \arg\min_{x \in S} f(x), \\ f^* &= \min_{x \in S} f(x) = f(x^*), \end{aligned} \tag{2.2}$$

where f is the objective function with a global minimum[2] at position x^* and value f^*. In this context f is also known as the *fitness function* of the given problem. The search space S is the domain of feasible positions x in which the minimum x^* is searched.

In case of the here developed work, the objective function f is described in Equation 2.1. Because of the use of discrete DrugScoreX potentials, there is no analytical and differentiable expression available for Equation 2.1.

In such a case, a numerical optimization is the method of choice to solve the minimization problem. An exhaustive search within S is then the only way to reliably find the global minimum. Nonetheless, an exhaustive search is often out of question for combinatorial optimization problems due to the size of S which increases exponentially with the number of dimensions. This phenomenon is known as the *combinatorial explosion* (Leach, 2001, pp. 460–461).

Regarding just the rotatable bonds of a ligand, the number of feasible conformations is given as

$$\text{Number of conformations} = \left(\frac{360}{\theta}\right)^N \tag{2.3}$$

where θ is the angle increment and N is the number of rotatable bonds. Table 2.1 lists the number of conformations for $N = 0, \ldots, 10$ and $\theta = 10°$. This example already shows how intractable an exhaustive search would be, even though the six dimensions for translating and rotating the molecule in space are still not considered. Beyond that, Figure 2.6 implies that an angle increment of 10° would probably be still too large to capture all important potential surface properties.

In a case like this stochastic search heuristics offer a practical alternative. Classical algorithms guarantee to find the optimal solution within optimal time for a given problem. If the optimal time is not acceptable though,

[2] Optimization can denote both the minimization and maximization of a function. However, both problems are generally equivalent with $\min f \equiv \max -f$. In the context of this work the terms *optimization* and *minimization* are used synonymously.

Table 2.1: Combinatorial explosion: number of feasible conformations for a molecule with N rotatable bonds and an angle increment of 10°.

N	conformations
0	1
1	36
2	1 296
3	46 656
4	1 679 616
5	60 466 176
6	2 176 782 336
7	78 364 164 096
8	2 821 109 907 456
9	101 559 956 668 416
10	3 656 158 440 062 976

heuristic procedures renounce to guarantee an optimal solution in order to achieve a "sufficiently good" solution within admissible time.

The most simple search heuristic is a random search. Until a predefined termination criterion is reached (maximum number of iterations, or adequate fitness, for example), f is evaluated at randomly chosen positions x. Since this method captures no information about the structure of f except the best position sampled so far, it is not only simple but also quite inefficient.

Beside the global minimum several local minima can exist. Finding a local minimum starting from a given position x is usually much easier than finding the global one. For a short introduction to local optimization methods, see Leach (2001, pp. 253–273), for example. Let

$$\bar{x} = LS_f(x) \tag{2.4}$$

denote the result of a local search on f starting at position x. In the ideal case, Equation 2.4 holds true if and only if x belongs to the region of attraction of $\bar{x}$. In practice, this is not necessarily true, since available local search algorithms usually depend on a given step width. Due to too large

steps, local minima can be omitted during the search. Nevertheless, an ideal local search is supposed to be available in the following considerations.

Obviously, finding the "best", meaning the lowest local minimum, is equivalent to finding the global minimum:

$$\min_{x \in S} f(x) = \min_{x \in S} f(LS_f(x)). \tag{2.5}$$

This knowledge leads to a straight-forward extension of the random search, the so-called *multistart* method, where every sampled point x_i is used as a starting point for a local search $LS_f(x_i)$. This way, instead of hitting the global minimum exactly, just any point within the region of attraction of the global minimum has to be sampled to solve the global optimization problem.

Based on this strategy, a broad range of enhanced optimization methods were proposed. Often, they try to improve the sampling of starting points for the local optimization by capturing and exploiting more information about the energy landscape or by restricting the search space.

Monte Carlo methods like *Simulated Annealing* (Kirkpatrick et al., 1983) perform a random walk through the search space by sampling random points within a defined neighborhood around the currently best point. Getting stuck within a local minimum is prevented by accepting points that decrease the fitness with a certain probability (Metropolis criterion). Dynamic adaption of this acceptance probability during optimization can further improve the efficiency.

Evolutionary algorithms like *Genetic Algorithms* (Goldberg, 1989) are inspired by biological strategies such as reproduction, mutation, recombination, and selection. Each single solution corresponds to an individual within a larger population of solutions. In analogy to evolution processes in Nature, the fitness of the solutions is iteratively increased.

Swarm intelligence algorithms like *Ant Colony Optimization* (Dorigo and Stützle, 2004) usually comprise a population of simple agents that locally interact with each other and with their environment. Individual agents follow simple rules while there is no central control unit influencing their

behavior. Still, local interactions between them result in the emergence of some kind of "intelligent" global behavior not observable for individual agents.

2.2.2 Potential Surface Smoothing

Prospects of success for optimization methods that combine global sampling with a local search strongly depend on the number of local minima. To find the global minimum it is essential to detect a starting point for the local search within its region of attraction. With an increasing number of local minima, this region will usually become smaller and therefore harder to be detected.

Often, a lot of these local optima from observed values can be considered as higher frequent perturbations, also called *noise*, on an underlying "true", yet unknown generating function. A common approach to reduce such noise is to apply any kind of smoothing technique (Moré and Wu, 1995).

The smoothing process transforms f into a smoother function $\langle f \rangle_\lambda$ with fewer local extrema. The parameter λ thereby controls the degree of smoothing: the original function is obtained for $\lambda \to 0$, while smoother functions are obtained with increasing λ. Generally, the value of the transformed function $\langle f \rangle_\lambda$ at a given point x is obtained by computing weighted averages of f in a defined neighborhood B around x.

To get the smoothed transform of a function f at position x, a so-called smoothing kernel g is applied to determine the weight of every point $y \in B$ depending on its distance to x:

$$\langle f \rangle_\lambda(x) = \frac{\int_B \; g_\lambda(\|x - y\|) \; f(y) \; \mathrm{d}y}{\int_B \; g_\lambda(\|x - y\|) \; \mathrm{d}y}. \tag{2.6}$$

The simplest case would be to have discrete, equidistant observations $f(x_1), \ldots, f(x_n)$. The smoothing parameter λ then defines the number of observations on each side of a current point x_i to be considered in averaging, thereby determining the radius of the neighborhood B. A weight of 1

is assigned to every observation:

$$\langle f \rangle_\lambda(x_i) = \frac{\sum_{j=i-\lambda}^{i+\lambda} f(x_j)}{2\lambda + 1}. \tag{2.7}$$

This is also known as the *unweighted sliding-average* method as a window of size $2\lambda+1$ centered at each x_i is moved over all observations. The smoothed value for an observation $\langle f \rangle_\lambda(x_i)$ is then the average of all observations within this window.

In case of continuous data, one of the most widely used kernel functions is the Gaussian kernel

$$g_\sigma(\|x - y\|) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-y}{\sigma}\right)^2\right). \tag{2.8}$$

Here, the smoothing parameter λ is proportional to the applied standard deviation σ of the Gaussian kernel.

The success of a smoothing approach, however, is dependent on both, the choice of λ as well as on the frequency and amplitude of the perturbations. In case of a strongly oscillating target function f, the smoothed function $\langle f \rangle_\lambda$ can emerge to be strongly perturbed itself, making an optimization as difficult as optimizing the original function f.

Addis et al. (2005) propose an approach to reduce this influence of the noise. In the following, their principle ideas are explained using an artificial, one-dimensional example (Figure 2.7). For a given objective function f (black curve) they try to reveal the underlying generating function (blue curve) which is free of any noise and therefore a much easier target for optimization.

To get there, however, not the original function f is smoothed. Instead the piecewise constant step function

$$L(x) = f(LS_f(x)) \tag{2.9}$$

is determined. In case of an ideal local search $LS_f(x)$, the constant segments of $L(x)$ correspond to the regions of attraction of the individual

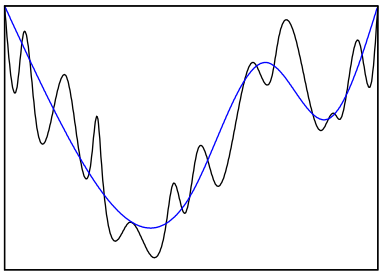


Figure 2.7: Schematic view of a noisy objective function $f(x)$ (black) and the assumed underlying, but unknown "true" function (blue).

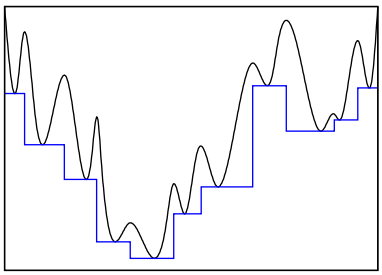


Figure 2.8: The original function $f(x)$ (black) and the result of the local search $L(x)$ (blue).

local minima of $f(x)$ (Figure 2.8, blue curve).

$L(x)$ clearly comprises the assumed underlying funnel shape, but in contrast to $f(x)$ it is free of any local perturbations. Obviously,

$$\min_{x \in S} f(x) = \min_{x \in S} L(x) \tag{2.10}$$

holds true. Yet, since a step function is discontinuous and exhibits a constant gradient of zero (if defined at all), it is not suitable for optimization although the number of local minima is already considerably reduced compared to $f(x)$. Applying $L(x)$ to a smoothing procedure according to Equation 2.6 using a Gaussian kernel overcomes this problem.

Since it is impossible to obtain an analytical expression for the objective function defined in Equation 2.1, such a form also lacks for both L and its smoothed form $\langle L \rangle_\sigma$. Moreover, it is even impossible to get a numerical

estimate for $\langle L \rangle_\sigma$ since its value at every single point x depends on values of L in the whole domain.

To resolve this problem, Addis et al. (2005) propose an approximation of $\langle L \rangle_\sigma$ within a delimited local environment based on a current point x_c. This local environment is a hypersphere $B(x_c, r)$ of predefined radius r centered at x_c. Restricting the smoothed step function $\langle L \rangle_\sigma$ to this neighborhood yields

$$\langle L \rangle_\sigma^B(x) = \frac{\int_{B(x_c,r)} g_\sigma(\|x - y\|)\, L(y)\, \mathrm{d}y}{\int_{B(x_c,r)} g_\sigma(\|x - y\|)\, \mathrm{d}y}. \tag{2.11}$$

To approximate $\langle L \rangle_\sigma^B(x)$ a uniform sample $y_1, \ldots, y_K$ of prefixed cardinality K is drawn from $L(x)$ within $B(x_c, r)$. Then the local approximation is given by

$$\hat{L}_\sigma^B(x) = \frac{\sum_{i=1}^K g_\sigma(\|x - y_i\|)\, L(y_i)}{\sum_{i=1}^K g_\sigma(\|x - y_i\|)}. \tag{2.12}$$

With K growing to infinity, this approximation converges to Equation 2.11.

Figure 2.9 shows the effect of this smoothing procedure on the example from Figure 2.7. The plotted area can be seen as a restricted environment B in which the step function L was sampled (black curve). Different standard deviations were used to obtain smoothed approximations $\hat{L}_\sigma^B$ (colored curves), with smaller σ resulting in transforms closer to the original function L. Like the step function itself, all of these transforms comprise the assumed funnel structure (dashed line). But in contrast to L, they are in addition suitable for an optimization, as they are continuous within B with a non-constant gradient.

2.2.3 The implemented Algorithm

The optimization procedure implemented in MiniMuDS should try to account for the characteristic shape of the the objective function as described in section 2.1. Essentially, a funnel-like, but distorted and locally perturbed shape is assumed. To optimize under these conditions, a modification of the *Algorithm based on Local Smoothing for Optimization* (ALSO)

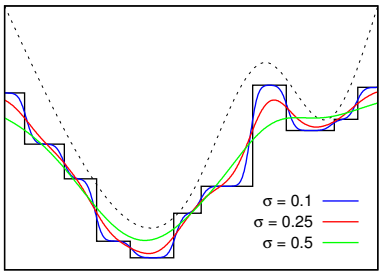


Figure 2.9: The step function L (black), the smoothed approximation $\hat{L}_\sigma^B$ obtained with different values of σ (colored curves), and the assumed funnel structure (dashed line).

as described by Addis et al. (2005) is applied.

This algorithm makes use of a local approximation of the smoothed step function $L(x)$ as given in Equation 2.12. The basic idea is to obtain the step function to reduce the number of local optima and then to approximate the smoothed form to predict a descent direction for $L(x)$ to enable further search steps based on information about the neighborhood of the currently best point.

Figure 2.10 depicts the implemented algorithm in pseudo-code. It consists of a sampling phase (line 16–29) and a smoothing phase (line 30–44) which are iteratively repeated. Within the first phase, a sample of K points $y_1, \ldots, y_K$ is drawn from a neighborhood $B(x_c, r)$ around the currently best point x_c (line 18). In each point y_i the respective value $L(y_i)$ on the step function is determined (line 19). During the smoothing phase, these values are used to approximate $\hat{L}_\sigma^B$ according to Equation 2.12, which is then locally minimized (line 32):

$$x_c^* = LS_{\hat{L}_\sigma^B}(x_c). \tag{2.13}$$

This minimum on $\hat{L}_\sigma^B$ in turn is applied to a local search on the original function f (line 33) to get

$$\bar{x}_c = LS_f(x_c^*). \tag{2.14}$$

```
1  MiniMuDS(x_0, r, K, MaxNoImprove, MaxNoDiff, MinDiff, MaxDiff)
2                                                  ▷ x_0: start configuration
3                                                  ▷ r: radius of sampling sphere
4                                                  ▷ K: number of samples in current sphere
5                                                  ▷ MaxNoImprove, MaxNoDiff,MinDiff: stopping criteria
6                                                  ▷ MaxDiff: acceptance criterion
7    NoImprove = 0
8    NoDiff = 0
9    x_c = LS_f(x_0)                               ▷ best position found so far
10   if (rmsd(x_c, x_0) ≤ MaxDiff)
11     x_c = x_0
12   record = f(x_c)                               ▷ best value found so far
13   while (NoImprove < MaxNoImprove and NoDiff < MaxNoDiff)
14     i = 0
15     current = record
16     while (i < K and record ≤ current)          ▷ sampling phase
17       i = i + 1
18       y_i = random uniform point in B(x_c, r)
19       y_i^* = LS_f(y_i)
20       if (rmsd(y_i^*, x_0) ≤ MaxDiff)
21         current = L(y_i)
22     if (current < record)
23       record = current                          ▷ new record found during sampling
24       x_c = y_i^*
25       NoImprove = 0
26       if (rmsd(current, record) < MinDiff)
27         NoDiff = NoDiff + 1
28       else
29         NoDiff = 0
30     else                                        ▷ smoothing phase based on y_1, ..., y_K
31       NoImprove = NoImprove + 1
32       x_c^* = LS_{L_σ^B}(x_c)
33       x̄_c = LS_f(x_c^*)
34       current = L(x_c^*)
35       if (current < record and rmsd(x̄_c, x_0) ≤ MaxDiff)
36         record = current                        ▷ new record found through smoothing
37         x_c = x̄_c
38         NoImprove = 0
39         if (rmsd(current, record) < MinDiff)
40           NoDiff = NoDiff + 1
41         else
42           NoDiff = 0
43       else if (rmsd(x_c^*, x_0) ≤ MaxDiff)
44         x_c = x_c^*
45   return x_c, record
```

Figure 2.10: The optimization procedure implemented in MiniMuDS. Differences compared to the ALSO algorithm as described by Addis et al. (2005) are marked in red.

In case $f(\bar{x}_c)$ is better than the best value found so far, $f(\bar{x}_c) < f(x_c)$, the respective optimum position $\bar{x}_c = x_{c+1}$ is used to define the new sampling sphere $B(x_{c+1}, r)$ for the next optimization cycle (line 37). If $f(\bar{x}_c) \geq f(x_c)$, the sampling sphere is at least moved to the minimum position on the approximated step function x_c^* (line 44).

This process is iterated until one of the predefined termination criteria is exceeded (line 13). To speed up the procedure, a new iteration is already started whenever a point y_i is sampled with $L(y_i) = f(LS_f(y_i)) < f(x_c)$, that is a point better than the currently best one (line 23). The sampling sphere is then moved to $LS_f(x_i)$ (line 24) and the sampling is restarted around this point.

All differences to the original ALSO procedure are shown in red in Figure 2.10. Three major modifications are introduced. First, MiniMuDS takes a given input geometry as starting point for the minimization whereas ALSO generates a random starting point (x_0, line 1).

Second, in addition to the number of iterations without improvement of the fitness value, a further termination criterion is established (line 13, 26, and 39). A minimum rmsd value *MinDiff* is defined to decide whether two optimum points found within two consecutive iterations differ significantly. A maximum number *MaxNoDiff* of consecutive iterations that fall below this threshold is allowed. The intention is to stop the optimization if virtually one and the same optimum is reached repeatedly.

Third, a new acceptance criterion is provided (line 10, 20, 35, and 43). It is used to restrict the search space to those geometries that show an rmsd value below the given threshold *MaxDiff* compared to the input structure. A solution with an rmsd above the allowed limit is not accepted, the search cannot be continued in this direction.

A minor modification concerns the optimization on the approximated smoothed step function $\hat{L}_\sigma^B$ (line 32). Addis et al. (2005) suggest a global minimization within the bounds of B. However, they just applied a single local search starting from the center of B to obtain their reported results. Arguing that this already yields adequate results they question whether a global search would be worthwhile considering the required additional

computational effort. Thus, this strategy was adopted in MiniMuDS.

At last, the termination criterion *MaxNoImprove* was slightly changed. While ALSO counts the number of consecutive samples drawn without improvement of the fitness value, MiniMuDS counts the number of consecutive unsuccessful iterations (line 31). Thereby this parameter is independent from the complexity of the given problem. In contrast, the number of samples K drawn within each iteration should well depend on the problem complexity (see section 2.3 for details).

As there is no analytical form of the implemented objective function f available, a gradient-free method is necessary to perform $LS_f(x)$ at the beginning of the optimization (line 9) as well as during the sampling phase (line 19) and at the end of the smoothing phase (line 33). This is done using a Downhill-Simplex algorithm (Nelder and Mead, 1965). A variant of this algorithm is applied to optimize on the approximated function $\hat{L}_\sigma^B$ (line 32). It is modified in a way to ensure that the search stays within the boundaries defined by B based on the ideas of the so-called *Complex* method described by Box (1965).

2.3 Parametrization

The algorithm introduced in section 2.2 needs a number of parameters to be predefined. In the following, the values chosen for the current implementation are given together with a short explanatory statement. This setup is taken as a reference when examining alternative parameter values. At the end of this section, Table 2.2 gives a summary of this reference parameter setup.

2.3.1 Algorithmic Parameters

The first parameter needed by the MiniMuDS procedure is the radius r of the sampling sphere. However, when optimizing molecular geometries, the individual dimensions are not equally scaled. Translation is measured in Å whereas rotation and dihedral angles are measured in degrees. Further-

more, one could even think about different radii for individual dimensions of the same type (to individually adjust the flexibility for different rotatable bonds, for example). On an absolute scale this would cause the sampling sphere to adopt an ellipsoidal shape, making uniform sampling difficult.

Therefore, random points are drawn within a hypersphere of radius 1. Afterward, every dimension is scaled according to the actual bounds of B. In the current implementation these bounds allow a maximal rotation of $\pm 5^\circ$ around each rotatable bond and a maximal translation of 0.1 Å in each spacial direction within a single search step. The total rotation around an arbitrary axis is limited to about 10°. This is realized through random rotation matrices generated according to the algorithm proposed by Arvo (1992). It guarantees uniformly distributed rotation matrices even within given limits. This would not be the case for randomly sampled Euler angles. To obtain such rotation matrices three random numbers are needed, corresponding to three dimensions in the optimization problem.

The number of samples drawn from $B(x_c, r)$ in each iteration is set to $K = 2n$, which is twice the dimensionality of the optimization problem, where n is the number of rotatable bonds plus 3 for the overall translation plus 3 for the rotation. These samples are used to approximate the smoothed step function as given in Equation 2.12. With increasing K the approximation gets closer to the real transform $\langle L \rangle_\sigma^B$, yet at the cost of higher computational efforts. However, to ensure a sufficient coverage of B, K should be increased with growing problem complexity. Addis et al. (2005) report results for $K = n$ and $K = 2n$ which are comparable although slightly better in the latter case for most of the examined test problems.

A newly found local optimum is accepted whenever it improves the fitness value compared to the best solution so far. In addition, the user can apply a further acceptance criterion, namely a maximal rmsd *MaxDiff* with respect to the input structure. This restricts the space to be searched, and thereby the computational complexity of the optimization problem. By default this parameter is set to 2 Å. This also guarantees that the overall

binding mode of the optimized geometry is conserved.

To determine convergence of the iterative procedure, MiniMuDS applies two independent convergence criteria. The number of successive iterations *MaxNoImprove* that do not improve the current optimum with respect to its fitness value is limited to 10. In addition, if a newly found optimum differs by less then 0.1 Å from the previously best point (*MinDiff* = 0.1) five times in a row (*MaxNoDiff* = 5) the optimization is also terminated.

The last remaining parameter is the standard deviation σ used in the smoothing kernel as given in Equation 2.8. According to Addis et al. (2005) this value is determined automatically depending on the radius r of the sampling sphere and the number of samples K drawn from it:

$$\sigma = rK^{-1/n}. \tag{2.15}$$

The rational behind this formula[3] is that given a n-dimensional sphere B of radius r, then K spheres of radius σ are necessary to cover the whole volume of B. In other words, to get enough information about the whole volume of B from K observations made within B, each observation has to cover a sphere of radius σ.

2.3.2 Weighting Factors for the Energy Model

To obtain suitable values for the weighting factors of the implemented energy model as described in section 2.1, α was set to 0.001 in order to transform DrugScoreX potentials roughly into the same numerical range as the other two potentials. Then, a systematic search to adjust β and γ was performed.

A training set consisting of 81 protein-ligand complexes, each one accompanied by 10 randomly deflected ligand conformations[4], was systematically optimized using all combinations of the values 0 and 10^i, $i = -3, \ldots, 3$ for β and γ. This range of values was selected to simulate situations from completely neglected to strongly dominating terms.

[3] See subsection A.3.1 for the derivation of Equation 2.15.
[4] See section 3.1 for details on the training data.

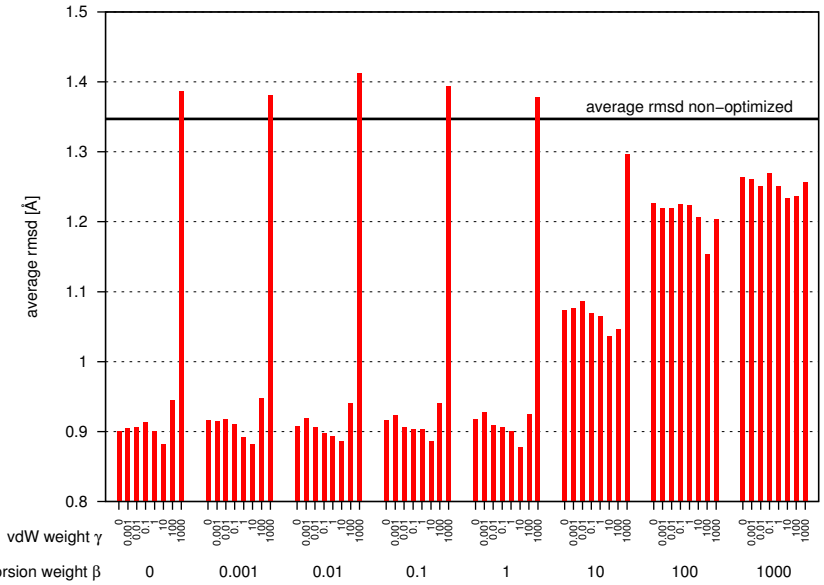


Figure 2.11: Average positional rmsd values of the training set before optimization (black line) and afterward (red bars), depending on the assigned weighting factors β for the torsion potentials and γ for the van-der-Waals potentials.

For each combination the average rmsd with respect to the atomic positions and with respect to the torsion angles between optimized and native state were calculated. Positional rmsd values are shown in Figure 2.11. Obviously, emphasized weighing of the van-der-Waals potentials is not beneficial for the optimization as they yield by far the highest rmsd values (last bar within each block). Similarly, strongly dominating contributions of the torsion potentials also result in rather large rmsd values (rightmost blocks). On the other hand, results seem to be quite robust with respect to torsion weights below 10, particularly if a van-der-Waals weight of 10 is applied. Best results are found for $\beta = 1$ and $\gamma = 10$. Rmsd values calculated with respect to deviation in the dihedral angles show exactly the same results.

To get a more fine-grained picture of the preferred values, all combinations of 0 and 2^i were assigned to β and γ in a second run. Figure 2.12

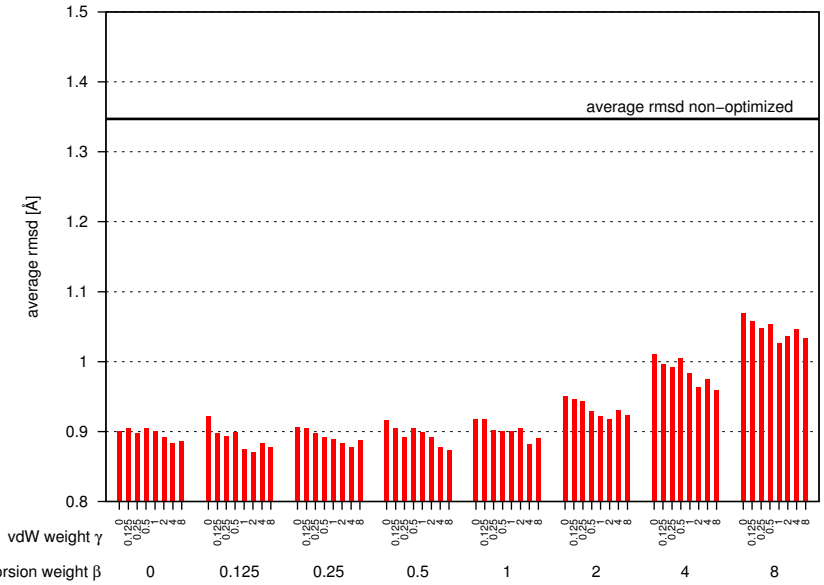


Figure 2.12: Rmsd values as in Figure 2.11, but for a smaller range of weighting factors.

shows the respective results for the positional rmsd values. Again, there is a tendency towards larger deviations in case bigger torsion weights between $\beta = 2$ and 8 are applied across the entire variation range of γ. Thus, it does not seem to be advisable to emphasize these interactions too much. Interestingly, although the higher torsion weights yield once more higher rmsd values (rightmost blocks), there is almost no difference observed for values of β between 0 and 1. The results calculated on dihedral angles again turned out to be nearly identical. As best performing values $\beta = 0.5$ and $\gamma = 8$ were selected for the current implementation.

Nonetheless, it seems a legitimate question whether the additional term for the torsion potentials E_{tors} in Equation 2.1 is necessary at all. A possible explanation for the observed results could be that the non-covalent interactions between protein and ligand dominate much stronger the native conformation compared to the dihedral angles. In this sense, the fact that rmsd values do not differ for torsion weights between 0 and 1, but increase from this point on, indicates that the impact of these potentials for $\beta \leq 1$ is

Table 2.2: Parameter setup of the reference implementation of MiniMuDS.

Parameter	Value	Explanation
r_{trans}	0.1 Å	maximal translation within one search step
r_{rot}	$\sim 10°$	maximal absolute rotation within one search step
r_{tors}	$\pm 5°$	maximal modification at dihedral angles within one search step
K	$2n$	number of samples to approximate smoothed step function
MaxNoImprove	10	stopping criterion based on success of the last iterations
MaxNoDiff	5	stopping criterion based on configuration of the last found optima
MinDiff	0.1 Å	minimum rmsd between two local optima to be "significantly" different
MaxDiff	2 Å	search space restriction around input geometry
σ	$K^{-1/n}$	standard deviation of Gaussian smoothing kernel
α	0.001	weighting factor for DrugScoreX interactions in Equation 2.1
β	0.5	weighting factor for torsion potentials in Equation 2.1
γ	8	weighting factor for van-der-Waals potentials in Equation 2.1

simply too small to show any effect on the total fitness value. But as soon as their weight becomes high enough to influence the total fitness value, results start to get worse. In this case, they could equally be omitted completely. This has to be examined separately, however.

The final reference parametrization implemented in MiniMuDS is given in Table 2.2.

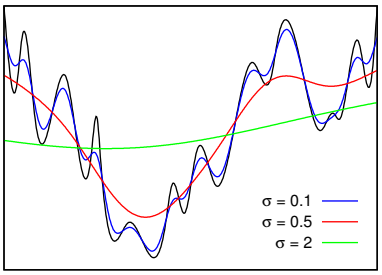


Figure 2.13: Smoothing of the original function $f(x)$ instead of the step function $L(x)$.

2.4 Algorithmic Variations

The algorithm and parametrization as introduced above are taken as a reference implementation in the following. In addition, two slight variations of the algorithm as well as the impact of the choices for the parameter K and the weighting factors for the objective function on the optimization performance will also be examined.

2.4.1 Alternative Smoothing Procedure

The first difference in the implementation concerns the function that is applied for the smoothing procedure. First of all, the success of the smoothing depends on the choice of the standard deviation σ, as illustrated in Figure 2.13. Too small values will not reduce the noise adequately (blue curve), whereas too large values may remove any structuring of the original objective function (green curve).

In the reference algorithm, the step function $L(x)$ is obtained and applied to the smoothing kernel in order to reduce the number of local minima prior to the actual smoothing. Thus, the performance should be increasingly independent of the choice of σ. This independence, however, is achieved at the expense of a highly increased number of cost function evaluations. For each sample point used to approximate the smoothed step function a local search has to be performed.

To assess whether the achieved results are worth this computational

effort, a variant of MiniMuDS was tested that applied the objective function $f(x)$ directly to the smoothing procedure instead of determining $L(x)$. Thus, the local searches on lines 9 and 19 of Figure 2.10 are omitted.

2.4.2 Alternative Parameter Setups

As mentioned above, Addis et al. (2005) report results for their optimization algorithm for two values of the parameter K, namely $K = n$ and $K = 2n$. In the latter case, their results show slightly better performance. Thus, this choice is adopted in the reference implementation of MiniMuDS. The value of K has two major effects. First, with increasing K the approximation $\hat{L}_\lambda^B$ of the smoothed function converges towards the real transform $\langle L \rangle_\lambda^B$. On the other hand, this value is directly correlated with the runtime of the optimization as it determines the number of samples and therefore the number of objective function evaluations. To examine the impact of this setting on the overall performance of MiniMuDS, the algorithm was also validated for a value of $K = n$.

The last variant concerns the weighting factors of the objective function as given in Equation 2.1. The search for suitable values raised the question whether the term accounting for internal torsion energies in the ligand is necessary at all. To answer this question, the optimization using an objective function without this contribution was also validated.

2.5 Multi-Component Optimization

To accomplish the second major task of this thesis, namely the simultaneous optimization of multiple components, the introduced algorithm does not have to be modified. In general, the simultaneous minimization of several components simply increases the dimensionality of the search problem. A single ligand optimization is represented by one dimension for each freely rotatable bond and six dimensions for translation and rotation in space. Any further optimization component adds the respective dimensions to the problem. So the minimization algorithm has to cope

with a larger search space. However, this does not imply any fundamental modification to the procedure as such.

Actually, it is sufficient to adapt the implementation of the objective function. In case of a single ligand, this function knows about a rigid set of atoms (the protein) and a flexible one (the ligand). To evaluate the DrugScore interaction potential, all pairwise contacts between protein and ligand have to be regarded. In addition, all internal ligand contacts have to be evaluated to get the van-der-Waals potential. The torsion potential affects all rotatable ligand bonds.

If any additional components should be considered, the implementation has to ensure that the van-der-Waals and torsion potentials for each component are evaluated separately. Furthermore, the inter-molecular interaction potentials have to be evaluated not only between the protein and each component but also between the different components. If this is guaranteed, generally an arbitrary number of components could be optimized simultaneously.

Of course, one has to be aware that each additional component increases the problem complexity and therefore the required runtime. However, this effect is not necessarily too dramatic. Including a water molecule, for example, adds a single atom to the system, since it is represented only by its central oxygen. There are no additional rotatable bonds to be considered. Furthermore, there is no need to rotate an isolated atom since this would have no effect on the interaction distances. Thus, each additional water molecule adds only three further dimensions for its translation to the problem.

Although the algorithm treats the protein as a rigid body, there is still a way to include individual side chains as flexible components. They have to be formally separated from the protein and can then be treated as if an additional covalently attached ligand would be minimized. This again keeps the number of additional dimensions low. A covalently bound ligand can only be modified at its rotatable bonds. As such a ligand cannot be moved independently from the protein, translation and rotation operations are permitted saving six respective dimensions. Thus, a flexible side chain

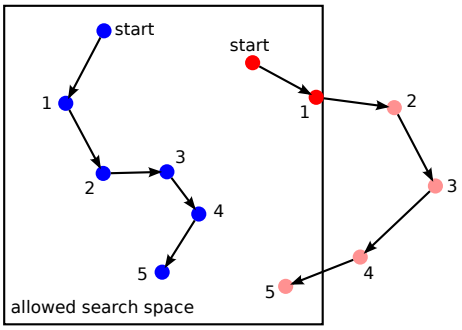


Figure 2.14: Possible course of a 2-component optimization raising the question of how to proceed if on component reaches the border of the available search space.

just adds as many dimensions as it contains rotatable bonds.

A more interesting issue is the restriction of the search space around the input structure given in the parameter *MaxDiff*. The reference implementation applies a default value of 2 Å. In case of a single ligand, the rmsd between input and output geometry is clearly defined. But how to calculate this value for a multi-component system?

There are two possible solutions to this problem. First, the rmsd for each individual component can be calculated and compared to the allowed threshold. As long as every component stays below this threshold a solution is accepted. If all components exceed this value, the solution is discarded. But this raises the question how to deal with a situation where only one or a few components fall beyond the allowed rmsd?

Discarding such a solution as a whole would prevent the other components from being further optimized although there is more search space available for them. This is illustrated in Figure 2.14. In position 1 the red component reaches the border of the search space. Stopping the minimization here would also cause the blue component to end in position 1, although there is much space left for optimization.

On the other hand, simply fixing the respective component at the border of the search space while the remaining ones are further minimized can lead to solutions that do not correspond to a local minimum on the po-

tential surface. This is related to a situation where the red component in Figure 2.14 is fixed in position 1 while the blue one ends up in position 5. Yet, the true minimum geometry for the red component corresponding to position 5 of the blue one could well be contained within the allowed search space.

The second possibility is not to calculate individual rmsd values for each component, but only one common value for the whole system. This way the calculation of the value and the behavior at the border of the search space is unambiguously defined again. This is the strategy used in the reference implementation of MiniMuDS.

However, since the rmsd is an averaged value, this can lead to large modifications of individual components as long as there are other ones compensating for this. In particular very small components consisting of only one atom (like water molecules, for example) can experience modifications far beyond the allowed threshold due to their small contributions to the total rmsd.

Although this is no problem *per se*, a major drawback of this way to calculate the rmsd concerns the representation of the protein. For each evaluation of the DrugScore interaction potential during minimization the distance for every pairwise protein-ligand contact has to be determined. Yet, since DrugScore only considers contacts up to 6 Å, large parts of the protein will be mostly irrelevant. Thus, a limited region around the ligand is extracted prior to the actual optimization for efficiency reasons. Of course, this region should be large enough to include all protein atoms that can possibly get into contact with the ligand within a 6 Å distance.

This is shown schematically in the left part of Figure 2.15. The input ligand geometry is depicted in dark blue and the black dashed line indicates the region relevant to a DrugScore evaluation for this geometry. Since the ligand can move upon optimization (denoted by the light blue ligand position), the actual region considered in MiniMuDS has to be increased appropriately (indicated in red).

Unfortunately, it is not trivial to exactly determine the region of all theoretically possible contacts already for a single ligand. Therefore, some

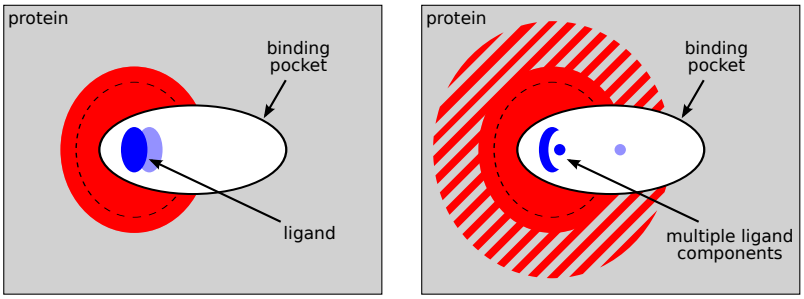


Figure 2.15: Definition of the protein region relevant for the evaluation of DrugScore interaction potentials.

simplifying assumptions are applied. Every protein atom within a given radius around any ligand atom is considered to define the relevant region. This assumes that the ligand can move equally far in any direction. Due to steric reasons this might not be true so that the relevant region is most likely overestimated by this simplification.

On the other hand it is not clear within which area a ligand can move based on arbitrary modifications and how this correlates with the resulting rmsd compared to the input geometry. An exception is the rigid translation. Translating the ligand by 2 Å in any direction results in an rmsd of 2 Å and cannot exceed the area of 2 Å around any ligand atom. But there is no way to formalize this relationship for arbitrary rotations, dihedral modifications or combinations of these operations. Thus, it is assumed that rigid translations approximate the area within which a ligand can move due to arbitrary modifications. The reference implementation thus considers every protein atom within a distance equal to the sum of the maximum DrugScore distance and the maximum allowed modification of the ligand around any ligand atom (8 Å by default). However, one has to be aware that this approximation can underestimate the relevant region in the worst case.

When dealing with multiple ligand components, these estimations become even more complicated (right part of Figure 2.15). A single small component can be transformed far beyond the allowed total rmsd thresh-

old (light blue position of the smaller component). Thus, the region of all theoretically possible contacts can get extremely large, indicated by the hatched area. Including all protein atoms that could get into the DrugScore distance in theory (although many will probably never do so in practise) would significantly slow down the optimization. On the other hand, omitting parts of this region can possibly neglect relevant contacts in the evaluation of the interaction potentials. In the worst case, a ligand component could be moved into some "empty" space from MiniMuDS' point of view that actually is occupied by protein atoms that were just not considered for optimization due to their large distance to the input geometry.

Since both variants of calculating the rmsd between the input an optimized configuration involve some drawbacks, this issue should be investigated in more detail in the future and definitely leaves room for further enhancements.

Chapter 3

3 Materials and Methods

3.1 Employed Data Sets

3.1.1 The Training Set

The Astex Diverse Set (Hartshorn et al., 2007) consisting of 85 high quality protein structures was chosen as a training set. It was originally compiled for the validation of protein-ligand docking algorithms. It is designed to show maximum diversity both with respect to protein targets and ligand compounds. Included complex structures represent pharmaceutically or agrochemically relevant proteins. They all contain drug-like ligands and have passed several quality filters.

This training set has been used to search for suitable weighting factors for the implemented energy model as described in section 2.1. During this search the weighting factor for the DrugScore potentials was kept fix so that only those for the dihedral and the van-der-Waals potentials had to be adjusted. Since both, E_{tors} and E_{vdW}, model internal ligand interactions, they are only affected by modifications of the dihedral angles of the optimized molecule. Thus, four complexes without any freely rotatable bond were excluded from the training set[1].

For each of the remaining 81 complexes, ten randomly deflected confor-

[1] PDB codes of excluded structures: 1gpk, 1hww, 1u4d, 1w1p.

mations were generated such that the torsion angle of every rotatable bond deviated between 10° and 60° from the native state. These conformations have been used for a systematic search for adequate weighting factors.

3.1.2 The Testing Set

The new optimization procedure introduced in chapter 2 is primarily validated on a data set compiled by Cheng et al. (2009) for a comparative assessment of scoring functions. It contains 65 diverse protein targets, each represented by three structures in complex with different ligands (195 complexes in total). All proteins and ligands are readily prepared for docking experiments concerning atom typing and protonation states. They are thus used without any modification.

Three complexes are shared with the above-mentioned Astex Diverse Set used to train the objective function parameters. To prevent any bias in the direction of the training data these structures are excluded from any further evaluation[2]. Two further complex structures are excluded due to structural deficiencies that make them inappropriate for the use in optimization studies[3]. Both of them contain short atomic contacts that cannot be reasonably explained from the structure. Details on the reasons that lead to the exclusion can be found in section A.2.

For each of the remaining 190 structures, two sets of docking solutions have been generated using two different docking engines:

1. FlexX (Rarey et al., 1996) in the LeadIT release 1.2[4] with default settings. Up to 50 solutions have been stored.

2. Gold version 4.1[5] (Verdonk et al., 2003; Jones et al., 1997) using default settings with 50 GA runs and early termination not allowed. To cover the available space within the binding pocket as broad as

[2]PDB codes of structures contained in the training set: 1l2s, 1n2v, 1v48.
[3]PDB codes of complexes excluded due to structural reasons: 1tyr, 2fzc.
[4]BioSolveIT GmbH, An der Ziegelei 79, 53757 Sankt Augustin, Germany.
`http://www.biosolveit.de`
[5]Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, UK.
`http://www.ccdc.cam.ac.uk`

possible, the generation of diverse solutions has been activated with an rmsd cutoff of 1 Å and a cluster size of 1.

In both cases, an area of 5 Å around any ligand atom in the crystal structure has been used to define the binding region. As input for the docking algorithms relaxed conformations of the ligands as included in the data set have been used.

Both FlexX and Gold are widely used programs that represent two principally different types of docking algorithms. Gold uses a Genetic Algorithm in combination with a force field based scoring function (GoldScore was chosen in the docking setup). In contrast, FlexX implements a deterministic incremental construction algorithm together with an empirical scoring scheme. See Sousa et al. (2006) for a review on different docking algorithms and further available programs.

The FlexX protocol did not generate 50 solutions in all cases due to the implemented algorithm. Therefore, the total number of poses to be examined is 9 036 instead of the theoretical number of 9 500 solutions for 190 complexes.

The employed test set contains a large range of compounds from small, fragment-like molecules up to oligo-peptides and -nucleotides. The latter ones are usually highly flexible due to their large number of rotatable bonds. In addition, they often bind to the surface of a protein receptor in rather flat binding regions. This makes them a difficult task for docking algorithms and also provides a challenge for the presented optimization method. On the other hand, such big ligands are usually not considered as drug-like or lead-like during the drug design process.

Therefore, the data set is divided into two subsets:

1. the *lead-like* subset containing 106 complexes that satisfy the definition by Oprea (2000);

2. the *non-lead* subset containing the remaining 84 structures.

Descriptors to determine lead-likeness have been calculated with the pro-

gram MOE[6] in version 2009.10. Using these subsets, the dependency of the optimization performance on the nature of the investigated molecules is examined.

3.2 Evaluation of the Pose Recognition Performance

The primary focus of the validation is put on the question whether an optimization of *in-silico* generated ligand geometries using the presented MiniMuDS method improves the capability of DrugScore to identify near native poses. Therefore, the pose recognition performance is evaluated in several steps.

3.2.1 Suitability of the MiniMuDS Energy Model for Geometry Optimization

First, it is examined whether the MiniMuDS fitness score is suitable at all as an objective function for a minimization according to DrugScore potentials. This involves two different aspects.

1. How well does the implemented objective function model the original DrugScore function?

2. How well does the implemented objective function reflect the true energy landscape of protein-ligand complexes?

The first question refers to the fact that the MiniMuDS fitness score as given in Equation 2.1 comprises two further terms in addition to the DrugScoreX potentials. Thus, the scores obtained by DrugScoreX and by MiniMuDS will differ in their absolute values. Generally, the optimization of one function, in this case the MiniMuDS energy model, does not necessarily lead to optimized results for another function, in this case DrugScoreX. Nonetheless, as long as the potential surfaces and in partic-

[6] Chemical Computing Group Inc., 1010 Sherbrooke St. W, Montreal, Canada. `http://www.chemcomp.com`

ular the positions of local minima are comparable, a MiniMuDS optimization still promises improved scoring results by DrugScoreX.

Thus, to answer the first question, the two scoring schemes have to be compared directly. In theory, MiniMuDS scores should be very similar to DrugScoreX. The inter-molecular interaction potentials are in common between the two functions. The first additional term in MiniMuDS accounts for intra-molecular dihedral energies. The applied potentials are defined in a way to show a minimum value of zero (Klebe and Mietzner, 1994). More unfavorable torsion angles yield more positive potential values. The second term considers ligand internal van-der-Waals energies. As these energies are modelled through a 12-6-Lennard-Jones potential, their contribution is slightly negative in a favorable case while it rapidly grows to large positive values for unfavorable conformations.

Hence, a ligand with ideal internal conformation will exhibit a torsion potential around zero together with a slightly negative van-der-Waals potential. In the end, this has to result in a MiniMuDS score very close to the corresponding DrugScore value. It is assumed that *in-silico* generated ligand geometries comprise reasonable torsion angles and avoid short atomic distances in order to keep intra-molecular interaction energies minimal. Thus, such conformations should get close to an ideal geometry concerning the internal score contributions.

To examine whether this actually holds true, all generated docking poses together with the native states are applied to both scoring schemes. The correlation is determined by means of the Pearson correlation coefficient[7]. One has to be aware that the MiniMuDS score directly depends on the DrugScore value by definition: $ms(x) = ds(x) + c(x)$ with $ms(x)$ the MiniMuDS score of conformation x, $ds(x)$ the DrugScore and $c(x)$ the sum of the ligand internal torsional and the van-der-Waals potentials. Thus, a high correlation is to be expected. However, this still allows to get a first impression of the difference between the two scoring schemes.

[7] See subsection A.3.2 for details on correlation measures and their statistical significance.

To get a more detailed picture of the compatibility of the two scores, every crystal structure is combined with its two corresponding sets of docking solutions to yield a total of up to 101 poses per complex. These are ranked according to both their DrugScoreX value and their MiniMuDS score. This time, the correlation of the resulting rankings is determined using Spearman's rank correlation coefficient.

Finally, the capability of the two scoring schemes to detect either the crystal structure itself or at least a geometry very close to the native state on rank 1 is compared. The rmsd values of the top ranked poses from the above mentioned rankings are calculated and the fraction of complexes with a rank 1 geometry within a given rmsd cutoff is determined.

The second question comprises an obvious requirement for any scoring function intended to recognize near native ligand geometries. It is addressed through the optimization of the crystal structures themselves. Thereby it is examined how well the optima of the MiniMuDS objective function are in agreement with crystallographically determined native states of the complexes.

It is assumed that crystal structures usually correspond to the global, at least however to a local minimum on the energy landscape of the corresponding protein-ligand complex. If the interaction potentials used by MiniMuDS adequately reflect this energy landscape, then the optimization of a geometry taken from the crystal should show only small effects. Both, the structures and the evaluated scores before and after the optimization are expected to be very similar.

The structural effects are measured in terms of rmsd values between input and output geometry. They reflect the degree of modification introduced through the minimization. To quantify the impact on the scoring of a given compound, the observed difference in the DrugScore values between the native and the optimized pose is calculated:

$$\Delta\mathrm{DS} = \frac{\mathrm{DS}_{in} - \mathrm{DS}_{out}}{|\mathrm{DS}_{in}|} \times 100. \tag{3.1}$$

The value is multiplied by a factor of 100 to get the difference as percentage of the native score DS_{in}.

3.2.2 Optimizing Docking Solutions

In the next step all generated docking solutions are optimized. Several parameters are determined to assess the influence of the optimization on the docking poses. As before, the rmsd values between input and output conformation are used to assess the structural effects, while ΔDS yields the impact on the scoring.

In addition, the effect on the quality of a docking pose is assessed. Therefore, the difference between the rmsd to the native state before and after minimization is calculated:

$$\Delta\text{rmsd} = \text{rmsd}_{in} - \text{rmsd}_{out}. \tag{3.2}$$

This difference indicates whether a geometry is moved towards the native state which is reflected in a positive Δrmsd, or away from it, resulting in a negative Δrmsd.

Finally, the effect on the ranking of the poses among each other is investigated. As described in section 1.2, the optimization is intended to adapt a given geometry to the DrugScore function without generating a completely new binding mode. In the ideal case, the structural changes will stay rather small, while the optimized scores strongly improve the detection of near native solutions from the whole set of suggested poses.

To evaluate the influence on the ranking of the solutions four different scenarios are examined:

1. the ranking directly obtained from the docking programs;
2. the ranking obtained from rescoring the unmodified docking solutions with DrugScoreX;
3. the ranking from rescoring poses strictly locally optimized according to the DrugScoreX potentials (named *Local* in the following);

4. the ranking obtained from rescoring the solutions optimized with MiniMuDS.

Furthermore, it is assessed whether it is actually necessary to optimize all fifty generated docking solutions. Therefore, a second set of rankings is determined. This time, only the ten poses scored best by the docking programs are considered for the following rescoring and optimization schemes (denoted by $^{\text{Top10}}$-labels added to the method names).

The ranking performance is given in terms of success rates for identifying near native poses on rank 1. The success rate is defined as the fraction of complexes, for which the top-ranked pose is found within a certain rmsd cutoff compared to the native state. This is evaluated for rmsd thresholds of 1 and 2 Å.

To compare the results obtained from MiniMuDS to a strictly local optimization according to the DrugScoreX potentials the docking poses are also minimized using a standard local optimization algorithm based on Powell's method (Powell, 1964) as it is implemented in the current development version of the DrugScore function[8].

In addition to the pose recognition performance on rank 1 also the complete rankings are evaluated. Thereby the general capability of the four scoring strategies to separate near native poses from misplaced geometries is examined. To this end, the median rmsd value across the test data set is calculated for each rank separately. These values are then compared to the rmsd values obtained from an ideal ranking.

In this context, the ideal ranking is that one generated by ordering the poses according to their rmsd value to the crystal structure. The median is preferred to the average of the rmsd values due to the fact that this statistic is more robust against outliers. In addition, when dealing with asymmetric distributions, the mean value can lead to a false impression of the "central" region of a distribution. Rmsd values are usually asymmetrically distributed since they are bounded to the lower side by zero, whereas there is no upper bound. In contrast, exactly 50 % of the distribution are

[8]Not yet published; personal communication with Gerd Neudert.

located below the median and 50 % are located above this value by definition. Thus, the median is usually much closer to the region where most of the observations are cumulating.

To quantify the similarity between the examined scoring strategies and the ideal ranking, the mean deviation between the median rmsd values of the compared rankings are calculated across all ranks.

All rmsd values used throughout the validation are calculated with fconv (Neudert and Klebe, 2011) regarding non-hydrogen atoms only and considering molecular symmetry.

Chapter 4

4 Results and Discussion

4.1 Suitability of the MiniMuDS Energy Model for the Optimization

MiniMuDS is supposed to improve DrugScore results through the minimization of a considered complex geometry prior to the actual scoring. However, the energy model of MiniMuDS as given in Equation 2.1 differs from the DrugScore function. Thus, it has to be clarified whether the MiniMuDS scoring scheme models the DrugScore function sufficiently accurate to be suitable for an optimization. In addition, this scoring scheme also has to reflect the true energy landscape of protein-ligand complexes adequately. Otherwise an optimization will not lead to a better detection of near native poses even though the mere DrugScore values are optimized.

4.1.1 Comparability of the MiniMuDS Energy Model and the DrugScoreX Function

A total of 18 726 geometries including the crystal structures from the validation data set and all generated docking poses are scored using both scoring functions. The difference between the two scores is shown in Figure 4.1. Every point corresponds to one geometry with its DrugScore value given on the x-axis and the MiniMuDS score on the y-axis. The solid line

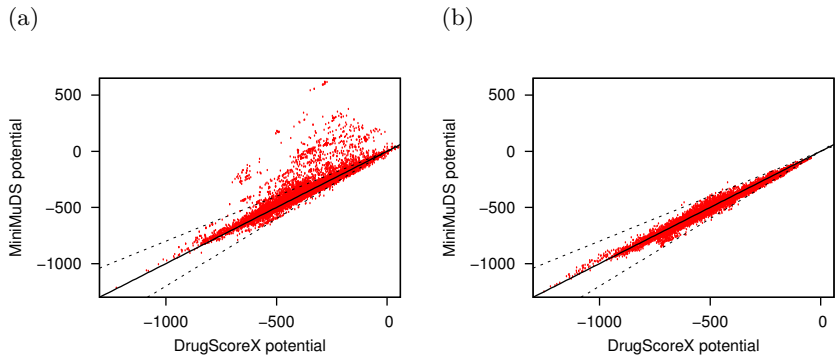


Figure 4.1: Relation between the absolute scores obtained by DrugScore and MiniMuDS for (a) non-optimized poses and (b) optimized poses.

indicates equality of the two scores, the dashed lines denote an absolute difference of 20 %.

Figure 4.1 (a) shows the scores for poses that were directly applied to the scoring functions without any modification. A majority of 89.8 % shows only minor differences with an absolute MiniMuDS score within 20 % of the respective DrugScore value. Only 1.1 % of all scored poses obtain a MiniMuDS score that is more than 20 % smaller than the corresponding DrugScore value (points below lower dashed line). This is due to the fact that the torsion potentials cannot yield negative values at all and the van-der-Waals potentials contribute notable negative values only in case of extremely well placed atomic distances.

In contrast, quite some poses obtain a MiniMuDS score much larger than the corresponding DrugScore value. In 9.1 % of all cases the MiniMuDS score is more than 20 % larger than the respective DrugScore value (points above the upper dashed line). These are nine times more poses than on the lower side.

Interestingly, almost all of these upper poses have been generated by FlexX (1 634 out of 1 700 cases). The reason for this phenomenon is found in the way FlexX generates the conformations. To prevent ligand internal clashes, a test for short atomic distances is performed. Distances below the

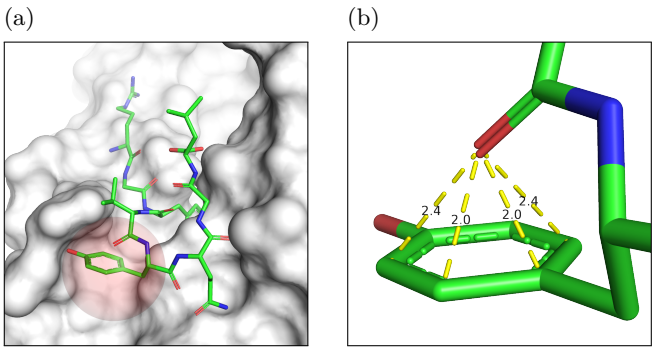


Figure 4.2: (a) FlexX docking solution for complex 1fzj. (b) Contact distances that are accepted by FlexX but penalized by MiniMuDS (in Å).

sum of the van-der-Waals radii of the involved atoms times a constant clash factor are not permitted (FlexX User Guide, 2010). This clash factor is set to 0.6 by default. The van-der-Waals radii used by FlexX and MiniMuDS are comparable. Yet, the applied clash factor can lead to atomic distances that are still accepted by FlexX while they are already penalized by large potential values in MiniMuDS.

This is illustrated using the example of FlexX solution 26 for complex 1fzj which shows the largest observed difference between the two scores. Figure 4.2 (a) shows the binding mode suggested by FlexX with close ligand internal contacts occurring in the moiety highlighted in red. This part is magnified in Figure 4.2 (b) with labeled contact distances. The van-der-Waals radii used by both FlexX and MiniMuDS for the involved atoms (1.7 Å for the carbons and 1.52 Å for the oxygen) add up to 3.22 Å. Thus, all distances depicted in Figure 4.2 (b) are far below this sum and are therefore penalized by the van-der-Waals potential in MiniMuDS. FlexX in contrast allows for distances as close as $3.22 \times 0.6 = 1.932$ Å without any penalty on the score.

Figure 4.1 (b) gives the scores of the two functions when the poses are optimized with MiniMuDS prior to their evaluation. There are almost no poses left (less than 0.4 %) for which the MiniMuDS scores deviate by more than 20 % from the corresponding DrugScore values. Also the

absolute differences, in particular in the upper direction, are much smaller compared to Figure 4.1 (a).

The reason for this is that both additional terms in MiniMuDS show minimum values close to zero. Thus, the contribution of these terms to the total score is reduced towards zero during the minimization. The optimized score is strongly dominated by the DrugScoreX potentials. This indicates already that a minimization with MiniMuDS can be used to optimize on the DrugScore function.

This is also reflected in the calculated Pearson correlation coefficient ρ between the two scores. For unmodified poses, ρ evaluates to 0.90. Such a high correlation was to be expected due to the direct dependency between the two scoring schemes by definition. However, through the optimization of the poses the correlation is even further increased to $\rho = 0.99$ which indicates a nearly perfect linear correlation between the two scoring schemes.

MiniMuDS is intended to minimize the DrugScore value of a given geometry although the implemented objective function is not exactly equivalent to the DrugScore function. The fact that the evaluated DrugScore value after the optimization is smaller than before in all but one out of 18 726 cases is further evidence for the suitability of the applied energy model.

4.1.2 Comparing the Ranking Capabilities of DrugScore and MiniMuDS

The results above show that MiniMuDS is suitable to optimize a single pose concerning its DrugScore value although the optimization itself evaluates a different objective function. In the following, it is investigated whether this observation also holds for the relative ranking of different conformations of the same molecule. Therefore, each crystal structure is combined with its corresponding non-optimized docking poses to a set of up to 101 geometries. These are ranked according to both, their DrugScoreX and their MiniMuDS score. Results on the ranking of optimized poses are presented in detail in section 4.3.

It turns out that around 60 % of the geometries are found among 5 positions in both rankings, almost 80 % are separated by 10 or less positions. For each of the 190 complexes, the correlation r_s between the two rankings is calculated according to Spearman's rank correlation coefficient. A majority shows high correlations with $r_s > 0.9$ in 72.1 % and $r_s > 0.8$ in 89.5 % of the cases. This is also reflected in a mean correlation of $\bar{r}_s = 0.92$.

There are also four cases observed with a lower r_s of about 0.5. Yet, all of them still suggest highly significant correlations. This is confirmed by a t-test which supports the found correlations at extremely high confidence levels far beyond 99.99 %. These high significance values can be assigned due to the large sample size of 101 observations in all cases except those where not a full set of 50 docking solutions was generated by FlexX.

Only one exception to this finding exists. For PDB complex 1fzk the calculated $r_s = 0.10$ indicates no correlation at all. Figure 4.3 shows the ranks according to both scoring schemes plotted against each other. Actually, there seems to exist a well defined correlations in the lower part of the plot. But beside this, there are some other poses ranked much lower by MiniMuDS compared to the DrugScore ranking.

Again, this can be assigned to short atom contacts as described above. All poses encircled in the upper left corner of Figure 4.3 are solutions generated by FlexX that comprise two oxygen atoms at a distance below 2 Å. Excluding these poses would yield a strong correlation of $r_s = 0.96$ for the remaining 67 geometries.

Finally, the capability of each scoring scheme to detect either the crystal structure itself or at least a near native geometry on rank one is examined. Therefore, the rmsd values of the top ranked poses from the above obtained rankings are calculated. Based on different thresholds it is decided whether the geometry is close enough to the native state. Results in Table 4.1 show that the two scoring functions achieve comparable success rates, no matter whether crystal structures are included into the set of ranked poses, Table 4.1 (a), or not, Table 4.1 (b).

The demonstrated similarities concerning both the absolute scores as well as the ranking capabilities suggest the MiniMuDS score to be a suf-

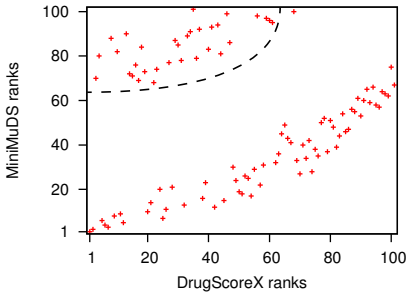


Figure 4.3: Comparison of DrugScoreX and MiniMuDS ranking of poses for PDB complex 1fzk.

Table 4.1: Recovery rate of DrugScoreX and the MiniMuDS score for the detection of near native ligand conformations from a set of docking solutions.

(a) Crystal structures included.

	success rate in % (absolute count)				
rmsd[1] ≤	0.0 Å	0.5 Å	1.0 Å	1.5 Å	2.0 Å
DrugScoreX	53.6 (103)	58.3 (112)	63.0 (121)	67.7 (130)	70.8 (136)
MiniMuDS	56.3 (108)	59.4 (114)	65.1 (125)	69.8 (134)	75.0 (144)
N[2]	190	190	190	190	190

1) Rmsd threshold for the top ranked geometry. 2) Number of complexes for which a geometry within the given rmsd threshold exists in the data set.

(b) Crystal structures excluded.[1]

	success rate in % (absolute count)				
rmsd ≤	0.0 Å	0.5 Å	1.0 Å	1.5 Å	2.0 Å
DrugScoreX	∼ (0)	44.4 (24)	40.7 (57)	53.1 (85)	56.3 (98)
MiniMuDS	∼ (0)	44.4 (24)	45.7 (64)	56.2 (90)	60.9 (106)
N	0	54	140	160	174

1) Rows corresponding to Table 4.1 (a).

Table 4.2: Impact of MiniMuDS on the DrugScore of crystal structures.

	all	lead-like	non-lead
average ΔDS [%][1)]	10.4 (± 7.7)	10.5 (± 8.1)	10.4 (± 7.1)
max ΔDS[2)]	39.8	39.8	32.7
ΔDS ≤ 10 %[3)]	61.1 (116)	61.3 (65)	60.7 (51)
ΔDS ≤ 20 %[3)]	88.4 (168)	87.7 (93)	89.3 (75)
N[4)]	190	106	84

1) Mean difference in DrugScore (ΔDS) between input and output conformation (standard deviation in parentheses). 2) Maximal ΔDS observed in the data set. 3) Percentage of optimized structures with ΔDS below the given threshold (absolute count in parentheses). 4) Total number of optimized geometries in the data set.

ficiently close model of the DrugScoreX function to allow for a successful optimization. This is supported by the fact that the impact of the additional terms in MiniMuDS is diminishing through the minimization.

4.1.3 Optimization of Native Geometries

The similarity between the implemented objective function and the original DrugScoreX function was shown above. Beside this, a sufficient modeling of the true energy landscape of protein-ligand complexes is required in order to be able to detect near native poses. Therefore, the given crystal structures from the validation data set are applied to the optimization in the following. Only minor effects with respect to both, the score and the geometry, are to be expected if the applied energy model adequately reflects the true energy surface.

Figure 4.4 (a) shows the impact of MiniMuDS in terms of ΔDS as given in Equation 3.1. Generally, only small changes in the DrugScore values around 10 % on average are observed. Most of the complexes show a stable DrugScore with a difference below 20 % in 88.4 % of the cases (see Table 4.2). Also the maximally observed difference of 39.8 % is still rather small. There is no noticeable difference between lead-like and non-lead compounds concerning the stability of the DrugScore values with respect to the evaluation of crystal structures.

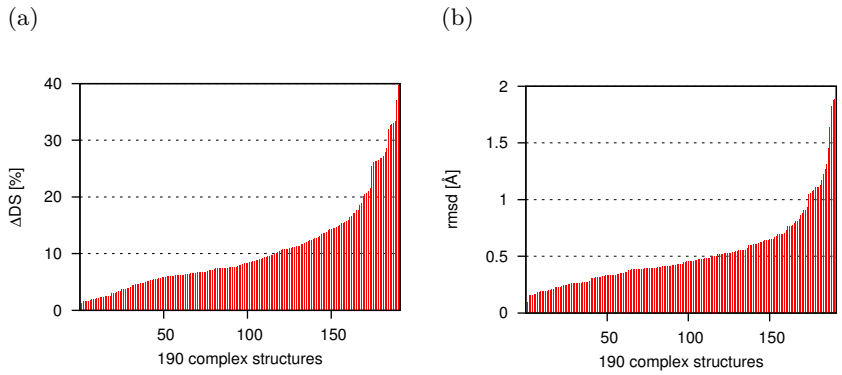


Figure 4.4: Impact of the optimization on native geometries. (a) Sorted DrugScore improvements ΔDS through the optimization. (b) Sorted rmsd values after the optimization.

Table 4.3 summarizes the impact of the optimization in terms of rmsd values between native and optimized ligand geometries which is also illustrated in Figure 4.4 (b). Generally, only small modifications are recorded (0.52 Å on average). Only very few structures show movements of more than 1 Å (8.9 %), although modifications up to 2 Å have been allowed during optimization.

In contrast to the DrugScore evaluations, the amount of structural modifications seems to depend on the nature of the ligands. Whereas lead-like structures show movements of 0.48 Å, the generally larger non-lead compounds are slightly stronger modified (around 0.56 Å). This is also reflected in the fraction of structures modified by more than 1 Å (7.5 % of the lead-like versus 10.7 % of the non-lead ligands). However, this difference is statistically not significant and can therefore only be viewed as a tendency.

None of the compounds has exploited the fully available search space as there is no structure with a final rmsd between 1.9 and 2 Å. This indicates that the examined space of 2 Å around the input structure is large enough to find a suitable optimum with respect to the DrugScoreX potentials, at least in the case crystal structures are used as input.

One question remains: Do the rather small shifts in the range of 0.5 Å

Table 4.3: Impact of MiniMuDS on the geometry crystal structures.

	all	lead-like	non-lead
average rmsd [Å][1)]	0.52 (± 0.32)	0.48 (± 0.28)	0.56 (± 0.35)
rmsd ≤ 0.5 Å[2)]	61.6 (117)	64.2 (68)	58.3 (49)
rmsd ≤ 1 Å[2)]	91.1 (173)	92.5 (98)	89.3 (75)
rmsd > 1.9 Å[3)]	0 (0)	0 (0)	0 (0)
N[4)]	190	106	84

1) Mean rmsd between input and output conformation (standard deviation in parentheses). 2) Percentage of optimized structures with rmsd below the given threshold (absolute count in parentheses). 3) Percentage of optimized structures with rmsd between 1.9 and 2 Å. 4) Total number of optimized geometries in the data set.

indeed reflect a good agreement between the implemented energy model and the true energy landscape? If the objective function models protein-ligand interactions too weakly, such small effects would be the general behavior of the optimization algorithm. Yet, this would be the case on any input data, no matter how far a given input geometry differs from the native state. To clarify this, all generated docking solutions have subsequently been optimized and evaluated.

4.2 Optimization of *in-silico* Generated Geometries

4.2.1 Effects on Score and Structure of Docking Poses

According to the previous section, the effects of MiniMuDS on the generated docking solutions for each protein-ligand complex are investigated. If the small modifications observed for the optimized crystal structures do not reflect the general behavior of the algorithm but rather the agreement between objective function and true energy landscape, then stronger modifications are to be expected in case of optimized docking solutions. Correspondingly, larger effects on the scores should be observed.

Figure 4.5 compares the influence of the optimization on native geometries (red curves) and on docking solutions (blue curves). Obviously, the effect of MiniMuDS on the DrugScore value of a docking solution is much

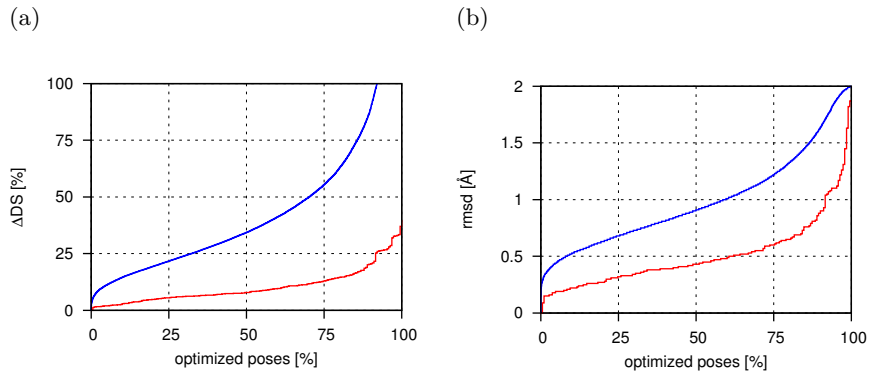


Figure 4.5: Impact of MiniMuDS on docking solutions (blue curves) and crystal structures (red curves, cf. Figure 4.4). (a) Sorted DrugScore differences ΔDS before and after optimization. (b) Sorted rmsd values between input and output structure.

stronger. While 61.1 % of the native conformations showed $\Delta DS \leq 10\,\%$, this is the case for only 3.6 % of the docking solutions (Table 4.4). On average, the scores of docking poses are improved by 64.1 %, which is six times more than the mean difference in the score of native geometries (10.4 %).

The remarkably high standard deviation of 664.7 % indicates a strong spreading of the observed scoring differences. This is also reflected in Figure 4.5 (a), which shows differences only up to 100 % or a 2-fold improvement. This covers 92.0 % of the optimized poses. The remaining 8 %, however, experience improvements up to 80-fold and there are even seven poses with scores improved between 130-fold and a maximum of about 475-fold.

Table 4.5 summarizes the structural effects on the docking poses, corresponding to Figure 4.5 (b). Note that the poses are modified twice as much as in the case of the crystal structures (about 1 Å average rmsd). Only 59.6 % of the optimized poses show movements below 1 Å, compared to more than 91.1 % of the crystal structures. However, there is still only a small fraction of optimization runs that fully exploit the available search space (4.1 % of the poses show modifications between 1.9 and 2 Å). This

Table 4.4: Impact of MiniMuDS on the DrugScore of docking poses.[1)]

	all	lead-like	non-lead
average ΔDS [%]	64.1 (± 664.7)	71.6 (± 718.4)	54.8 (± 591.5)
max ΔDS	47 546.7	47 248.2	47 546.7
ΔDS ≤ 10 %	3.6 (667)	3.5 (354)	3.8 (313)
ΔDS ≤ 20 %	21.3 (3 944)	18.7 (1 917)	24.4 (2 027)
N	18 536	10 241	8 295

1) Rows corresponding to Table 4.2.

Table 4.5: Impact of MiniMuDS on the geometry of docking poses.[1)]

	all	lead-like	non-lead
average rmsd [Å]	0.99 (± 0.41)	0.99 (± 0.43)	0.98 (± 0.40)
rmsd ≤ 0.5 Å	8.6 (1 598)	10.0 (1 023)	6.9 (575)
rmsd ≤ 1 Å	59.6 (11 046)	58.7 (6 013)	60.7 (5 033)
rmsd > 1.9 Å	4.1 (767)	4.6 (474)	3.5 (293)
N	18 536	10 241	8 295

1) Rows corresponding to Table 4.3.

supports the reasonable restriction of the search space to 2 Å rmsd with respect to the input geometry.

These findings confirm that the small modifications observed for the crystal structures concerning both score and rmsd are not the general behavior of the optimization algorithm, but indeed reflect a high agreement between the optima of the implemented objective function and the experimentally determined native states of the structures.

Whereas there is no difference in the degree of modification observed between lead-like and non-lead subsets, this time the improvements in the DrugScore values seem to be slightly larger in case of lead-like compounds (71.6 %) versus non-lead compounds (54.8 %), as shown in Table 4.5. Furthermore, 81.3 % of the lead-like subset show ΔDS > 20 %. On the other hand, this is the case for only 75.6 % of the non-lead subset. Yet, this is once again no significant difference.

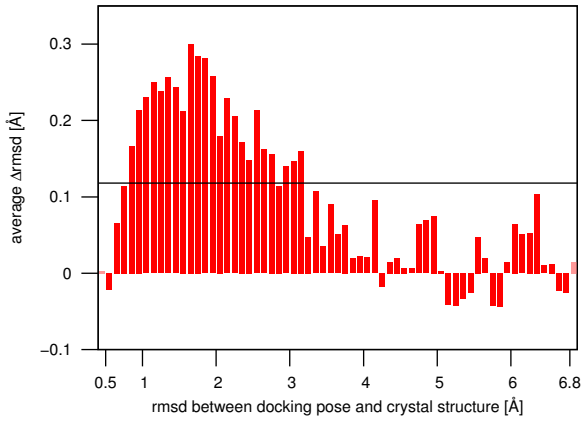


Figure 4.6: Average improvement for all 18 536 docking poses (black line) and for poses within a common range of rmsd to the crystal structure (red bars). The light bars pool all poses below 0.5 Å and above 6.8 Å rmsd, respectively.

4.2.2 Effects on the Quality of Docking Poses

Next, the influence of the optimization on the structural quality of the docked conformations is analyzed. This effect is expressed in terms of Δrmsd as given by Equation 3.2.

Figure 4.6 shows the results for all generated docking solutions. On average, a pose is improved by about 0.1 Å as depicted by the black line. This means that in general the optimization drives a docking pose towards the native state, although by only a marginal amount. However, given the large sample size this is still significantly different from 0 which would mean that on average the optimization neither lead to an improvement nor to a deterioration.

The bars in Figure 4.6 give the improvement for all docking poses that fall within the same range of rmsd to the crystal structure, averaged across bins of 0.1 Å. This allows for a more differentiated view. As the individual bins below 0.5 Å each contain less than 100 poses, they are summarized in the first bar (light red). This prevents the calculation of statistically meaningless average values based on only a few poses within one bin. The same has been applied for the last bin, which comprises all poses

above 6.8 Å.

The optimization of geometries between 1 and 2 Å rmsd performs well above average with an at least 2-fold better improvement. Notably, this is the most interesting range in structure-based drug design. Beyond 3 Å the improvement that can be expected from the optimization rapidly decreases with the increasing deviation of the input structure from the native state. Nevertheless, even in this range the docking solutions are rarely perturbed in a way to deviate further from the crystal structure.

Overall, rather small improvements are observed (generally below 0.3 Å). Together with the absolute modification of about 1Å of the docking poses, this reflects the intention not to generate new binding modes, but to preserve the given one. However, it was shown that these small modifications still have a rather strong influence on the scoring of the poses.

This leads to the most important question: How much do such small modifications finally affect the ranking of different poses for the same compound among each other? Considering the steepness of the applied potentials, these slight adaptions are assumed to be essential for scoring. In the ideal case, they should subsequently allow DrugScore to better discriminate between well docked and non-relevant poses.

4.3 Pose Recognition Performance

In this section, the success in identifying near native poses on rank 1 is validated for different ranking methods as described in subsection 3.2.2. Considering the aims of this work, two questions are of particular interest:

1. Does MiniMuDS improve the recognition rate of DrugScore?

2. Is there a benefit compared to a standard local optimization?

Figure 4.7 shows the success rates for all evaluated scenarios on the complete test set. The upper part gives the results for poses generated by FlexX, the lower part for the Gold solutions, respectively. FlexX yields a rank 1 solution within 2 Å to the crystal structure for 44.7 % of the

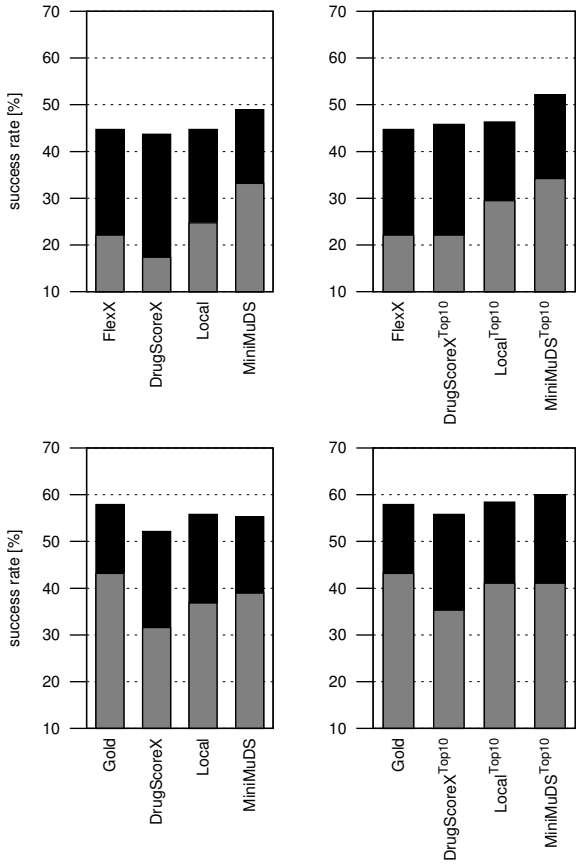


Figure 4.7: Success rates on the whole test set for top ranked solutions with an rmsd below 1 Å (gray) and 2 Å (black). Top: FlexX docking solutions; bottom: Gold docking solutions; left: optimization of all poses; right: optimization of top 10 docking poses.

cases (Figure 4.7 top left). In 22.1 % the top-ranked solution is found within 1 Å. When the unmodified poses are rescored with DrugScoreX these success rates drop by 1.1 % and 4.7 %, respectively. A slight increase of 2.6 % at the 1 Å cutoff compared to the original docking poses is obtained when a local optimization is performed before rescoring the solutions. At the 2 Å level no improvement is observed. In contrast, an optimization with MiniMuDS increases the success rates by 4.2 % at 2 Å and 11.1 % at 1 Å.

Slightly better results are obtained if only the ten top-ranked FlexX solutions are subjected to following rescoring procedures (Figure 4.7, top right). The success rates for direct rescoring do not drop any more. A local optimization increases success by 1.6 % at 2 Å and by 7.4 % at 1 Å. MiniMuDS$^{\text{Top10}}$ raises success by 7.4 % and 12.1 % to finally reach 52.1 % at 2 Å and 33.2 % at 1 Å, respectively.

In summary, a steadily increasing improvement of the results is observed starting from directly rescoring the poses using DrugScoreX, through the local optimization according to the rescoring function, to the point of minimizing with MiniMuDS. In particular poses very close to the native structure with rmsd values below 1 Å are considerably more often detected on rank 1 after the optimization. Notably, it seems to be sufficient to consider only the ten top-ranked solutions for the optimization as this strategy actually yields the best results.

In case of Gold docking, a rank 1 solution within 2 Å is received for 57.9 % of the complexes, within 1 Å for 43.2 % (Figure 4.7, down left). This time, a drop by 5.8 % and 11.6 %, respectively, is observed if DrugScoreX is directly used for rescoring. Neither a local minimization nor the use of MiniMuDS can fully compensate for this effect. Only if the optimization is limited to the top 10 Gold solutions comparable success rates to the original Gold ranking are obtained. A slight increase of 2.1 % at 2 Å and a decrease of 2.1 % at 1 Å (Figure 4.7, down right) are recorded in this case. The reason for the minor performance especially on the Gold solutions can be found in the composition of the data set.

Figure 4.8 shows the success rates only for the lead-like subset. The absolute level of success for both docking programs is about 5 to 10 % higher compared to the complete test set. And although the general tendency across the different methods is very much the same, the improvement from FlexX to MiniMuDS$^{\text{Top10}}$ raises from 7.4 % to 9.4 % at a 2 Å cutoff and from 12.1 % to 17.9 % at 1 Å. The results on the Gold solutions also raise from 2.1 % to 7.5 % at the 2 Å threshold, whereas almost no deterioration is observed any more at 1 Å.

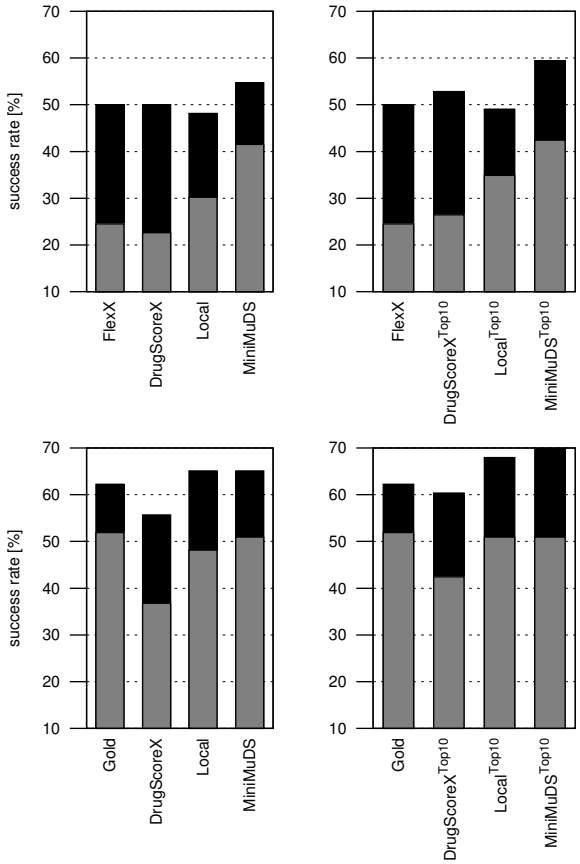


Figure 4.8: Success rates for the lead-like subset. Considered docking poses in the individual graphs correspond to Figure 4.7.

On the contrary, only minor improvements are observed for FlexX generated poses in the non-lead subset. For Gold solutions, even a decrease of success rates is found for these compounds throughout the different rescoring and optimization methods as shown in Figure 4.9.

The recorded difference discovered between the optimization of all or only the ten top-ranked solutions is worth closer inspection. For the FlexX poses of the complete test set, MiniMuDS$^{\text{Top10}}$ achieved 3.2 % better results at the 2 Å level compared to MiniMuDS. On the Gold poses the difference is 4.7 %.

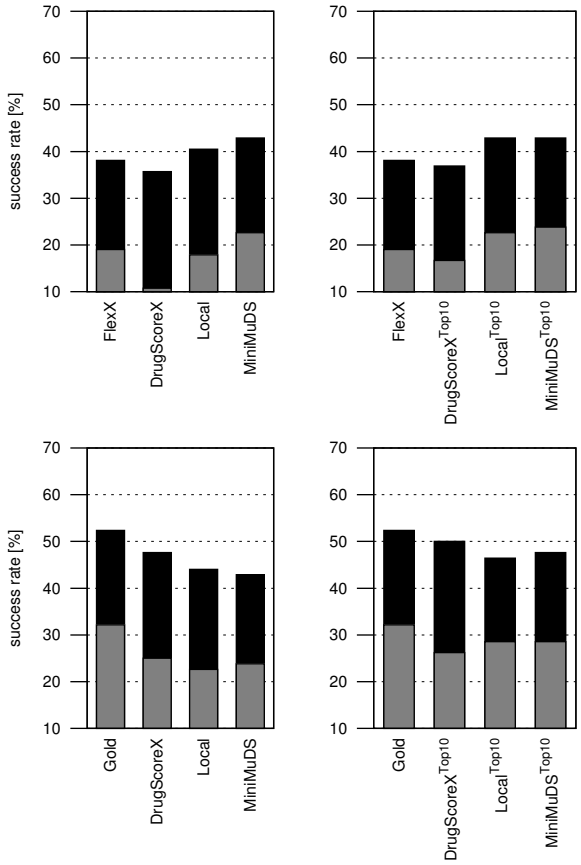


Figure 4.9: Success rates for the non-lead subset. Considered docking poses in the individual graphs correspond to Figure 4.7.

In particular the GoldScore function appears to be already quite successful in identifying the correct pose out of the fifty widely spread decoys. This obvious advantage is deteriorated if all solutions are subjected to MiniMuDS. In consequence, some of the actually less optimally placed solutions are artificially optimized into geometries that receive a superior DrugScoreX value compared to the better placed poses. In contrast, if we rely on GoldScore to detect the most promising solutions among the top ranked ones, we avoid these artifacts. This behavior observed for both docking programs can be seen as the result of a consensus effect. Only

the most relevant poses that have already scored high according to another function are considered for the optimization. This finally leads to the improved success rates.

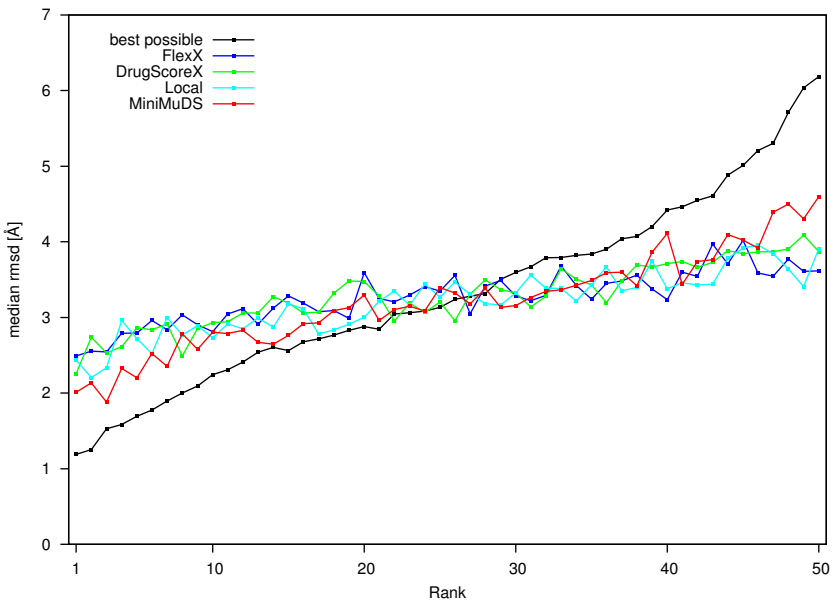
The previous analyses all considered the top ranked solutions only. In the following it is examined how MiniMuDS affects the ranking as a whole. Therefore, the rankings obtained by the different scoring methods are compared to the *best possible* way to rank the suggested poses for each complex. In this context, the "best possible way" is defined as a ranking according to the actual deviation between a given docking pose and the crystal structure in terms of rmsd values. Thus, it is evaluated whether a given scoring strategy is able to generally rank near native geometries higher than misplaced poses.

Figure 4.10 gives for each rank the median rmsd value of a given scoring scheme calculated across the complete test set. In addition, the best possible ranking as defined above is shown (black curve). Obviously, MiniMuDS (red curve) strongly improves the ranking capabilities of DrugScore. Top scored solutions on position one to about 20 show clearly lower rmsd values after the optimization compared to a direct scoring of the unmodified docking solutions (green curve). Accordingly, poses with higher rmsd values are more often ranked on the last positions.

All the other scoring schemes show a comparable ranking performance. The similarity between the curves of the different strategies and the best possible curve is quantified in Table 4.6 which gives the average difference between any scoring scheme and the best possible curve across all ranks. For all combinations of test sets and scoring strategies, MiniMuDS yields the smallest differences to the best possible ranking.

The previously observed performance difference between the two subsets of the test data set is confirmed by this analysis. Figure 4.11 shows the median rmsd values for the lead-like subset. For both docking protocols slightly better results compared to the complete test set are obtained. In contrast, there is almost no improvement recorded for the non-lead subset as shown in Figure 4.12.

(a)



(b)

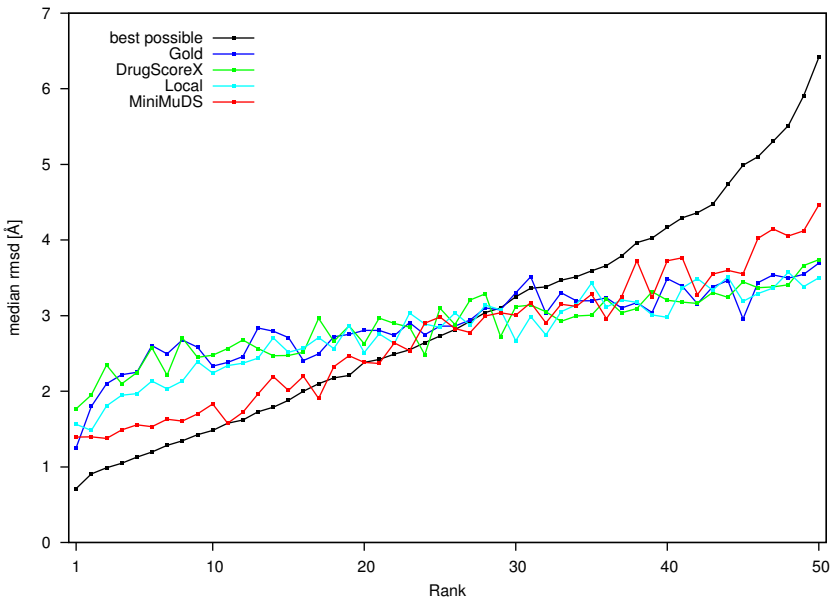
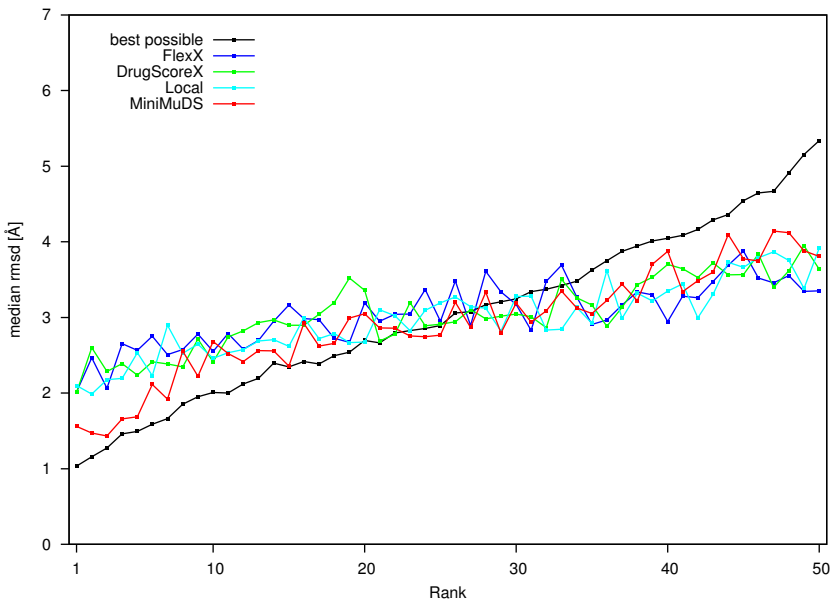


Figure 4.10: Median rmsd on each rank across the whole test set. (a) FlexX solutions; (b) Gold solutions.

(a)



(b)

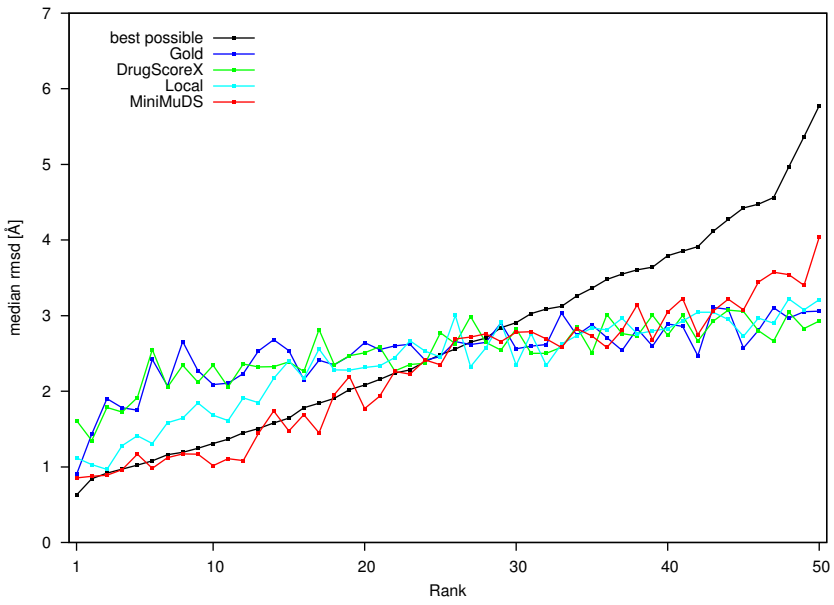
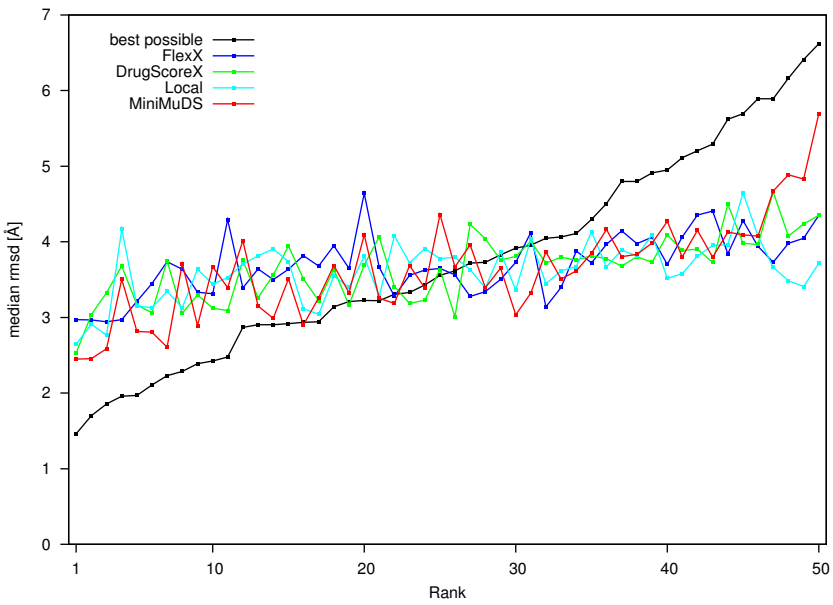


Figure 4.11: Median rmsd on each rank across the lead-like subset. (a) FlexX solutions; (b) Gold solutions.

(a)



(b)

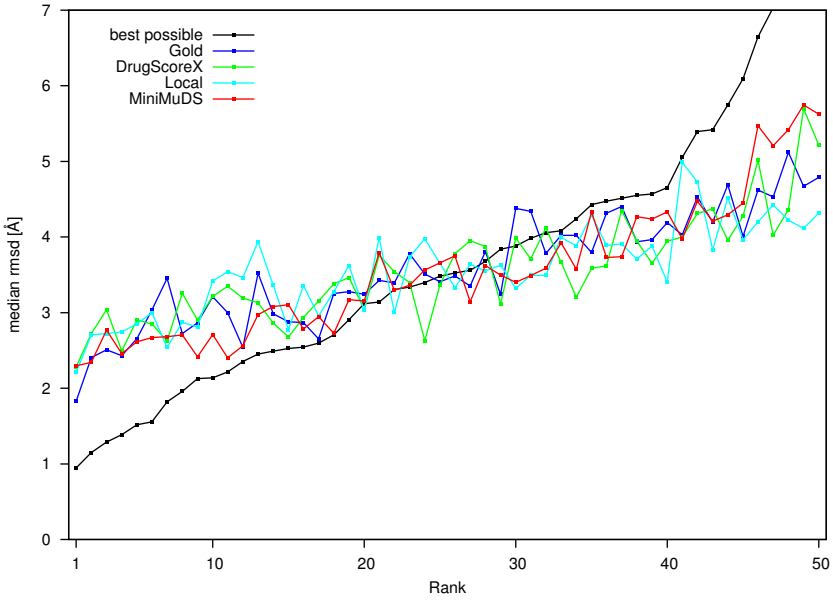


Figure 4.12: Median rmsd on each rank across the non-lead subset. (a) FlexX solutions; (b) Gold solutions.

Table 4.6: Mean difference between the different scoring strategies and the ideal ranking.

	FlexX				Gold			
	dock[1)]	ds[2)]	loc[3)]	opt[4)]	dock[1)]	ds[2)]	loc[3)]	opt[4)]
all	0.77	0.70	0.62	0.54	0.82	0.88	0.64	0.47
lead-like	0.73	0.70	0.61	0.51	0.81	0.81	0.55	0.44
non-lead	0.92	0.86	0.80	0.74	0.81	0.92	0.80	0.70
Top10 solutions[5)]:								
all	0.52	0.50	0.38	0.30	0.71	0.67	0.53	0.40
lead-like	0.48	0.44	0.36	0.30	0.62	0.66	0.41	0.35
non-lead	0.60	0.56	0.39	0.33	0.76	0.73	0.75	0.64

1) Original ranking from the docking program. 2) Ranking according to DrugScoreX. 3) Ranking according to DrugScoreX with preceding local minimization. 4) Ranking according to DrugScoreX with preceding MiniMuDS optimization. 5) Only top-10 docking poses are considered in the DrugScore rankings.

Considering only the ten top ranked docking solutions for the following DrugScore rankings reveals exactly the same results (Figure 4.13 and Table 4.6, lower part). The average difference of the median rmsd values in case of the original Gold ranking is almost two times higher than in case of the MiniMuDS ranking for both the complete test set and the lead-like subset. For the non-lead subset the improvement through MiniMuDS is smaller but still observable. This is the same, no matter whether all docking poses or only the top-10 are used for rescoring. Similar results are obtained for the FlexX generated poses.

The evaluation of the pose recognition performance on rank 1 solutions only as described above shows that MiniMuDS performs better on the FlexX poses compared to the Gold solutions. Interestingly, the analysis of the complete ranking revealed the opposite result. Larger improvements are obtained for the scoring of the Gold generated geometries.

While the success of GoldScore in identifying the best pose on rank 1 is confirmed, there are rapidly increasing rmsd values observed on the subsequent ranks. Therefore, no further discrimination between near native and

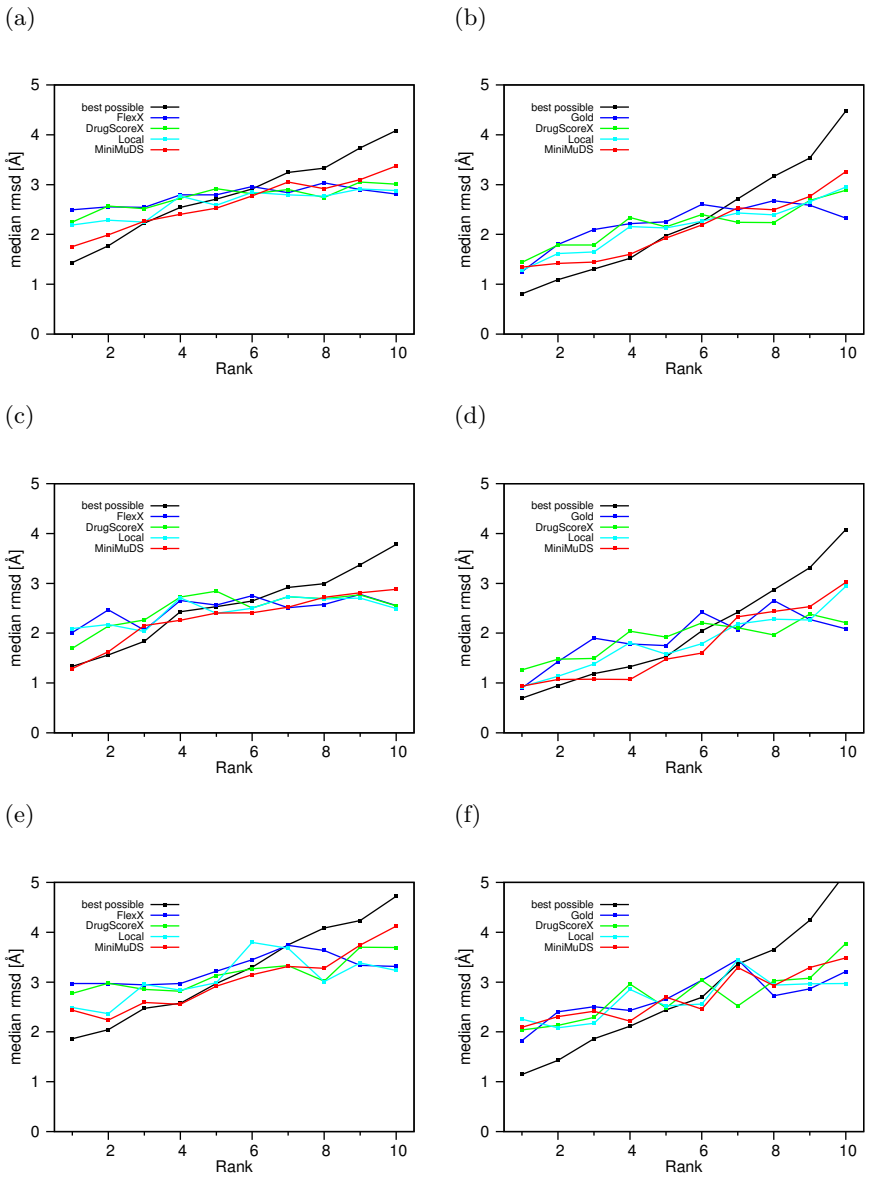


Figure 4.13: Median rmsd on each rank when only the top 10 docking solutions are considered in further DrugScore scoring. (a) FlexX poses, complete test set. (b) Gold poses, complete test set. (c) FlexX poses, lead-like subset. (d) Gold poses, lead-like subset. (e) FlexX poses, non-lead subset. (f) Gold poses, non-lead subset.

misplaced poses can be expected from this scoring scheme. In contrast, MiniMuDS places low-rmsd solutions on the first positions and high-rmsd solutions on the last positions much more reliably.

4.4 Results on Alternative Algorithm Setups

As described in section 2.4 several variations to the implemented algorithm and its parameter setup have been validated beside the reference implementation. Namely, these variations comprise:

1. the variation of the sample count K (denoted as *sample count*),
2. smoothing $f(x)$ instead of $L(x)$ (denoted as *no step*),
3. omitting the torsion potentials (denoted as *no torsion*).

In the following, these three alternatives are compared to the reference implementation. First of all, the impact of the modifications on the ranking performance is of interest. Figure 4.14 shows the mean difference to the ideal ranking for all variants, calculated for the complete test set. For comparison, the values for rescoring unmodified poses (DrugScoreX) and for the reference implementation (MiniMuDS) are also given.

At first glance, all variations seem to perform comparable to the reference, except for the *no step* variant that shows larger deviations particularly on the Gold poses. Yet, to gain deeper insights into the differences between the various implementations, all of them are examined separately in the following. Focus is put on two additional properties of the respective implementations that should help to discriminate between favorable and unfavorable variants.

First one is the convergence behavior. MiniMuDS uses a stochastic search algorithm and thus is a non-deterministic optimization method. Several applications on the same input data will usually yield different results. In the ideal case, however, the optimization should still converge to the same optimum, ending always up with virtually the same geometry.

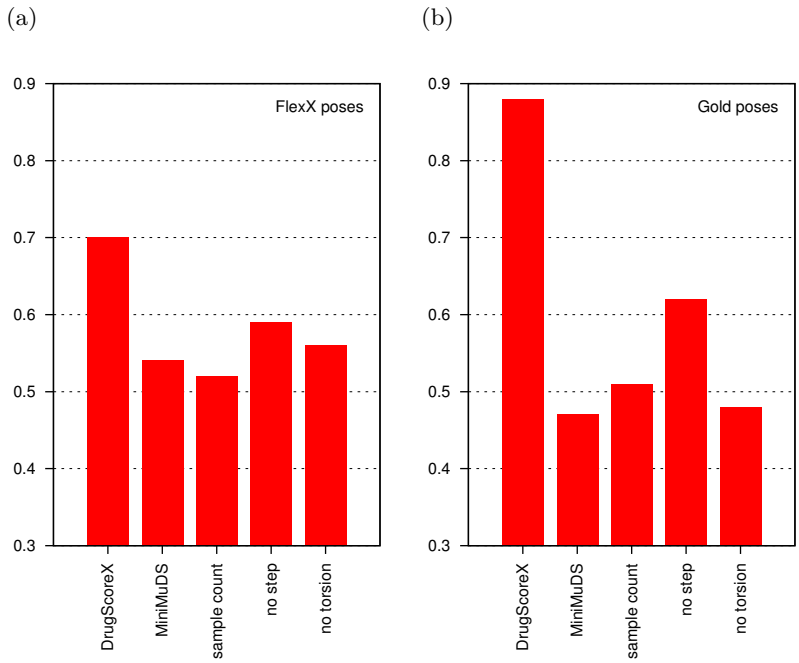


Figure 4.14: Mean difference in Å to the ideal ranking for the rankings obtained from the various variants of MiniMuDS as defined for Table 4.6 above, based on the complete test set.

Output structures should therefore be almost identical, no matter how often the optimization is actually performed.

To get statistics on the convergence, the first five docking solutions from each docking run have been selected. To keep the required computational effort tractable, this analysis has been limited to the lead-like subset, yielding a total of 1 060 geometries each of which has been optimized ten times. These ten corresponding optimization results have then been clustered according to there mutual rmsd values to find out how much the geometries obtained from the same input data differ. For this purpose, a hierarchical complete linkage clustering as provided by fconv (Neudert and Klebe, 2011) with a maximal rmsd of 0.5 Å within one cluster was used. To get an impression about the diversity within the results, the total number of resulting clusters is determined. This is shown in Figure 4.15 for the

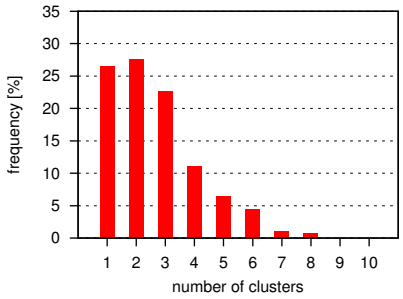


Figure 4.15: Distribution of the final number of clusters when clustering with an rmsd cutoff of 5 Å.

reference algorithm.

For 26.42 % of the 1 060 geometries all ten minimization runs end up within a single cluster. A total of 76.42 % group into three or less clusters whereas there is only one case or 0.09 % where solutions are divided into ten single-member clusters. This shows the general tendency of the algorithm to repeatedly converge into the same local optima. Yet, in case of more than one clusters, it is not visible from this distribution whether there is one large cluster accompanied by a few very small ones, or whether the optimization results are grouped into several medium-sized clusters.

To reveal this, the distribution of the cluster sizes is given in Figure 4.16. Again 26.42 % of the optimization results find themselves within a cluster of size ten. They correspond to those cases in Figure 4.15 that ended up in a single cluster. The remaining geometries however spread across all possible cluster sizes nearly uniformly. All of them are populated with 7.26 % to 9.51 % of the optimized geometries. This means that a geometry with two accessible local optima can end up nine times in the first one and only once in the second, or it can arrive five times in both of them with a similar likelihood, for example.

Second, the impact of the algorithmic variations on the search efficiency is examined. The computational time needed for a single optimization is basically determined by the number of objective function evaluations. Apart from this, only constant factors for pre- and post-processing steps

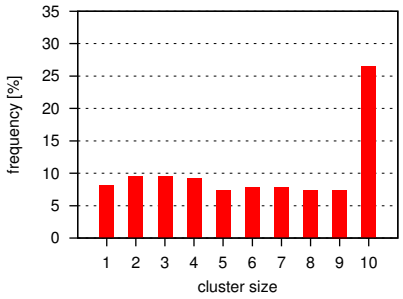


Figure 4.16: Distribution of the number of elements within the individual clusters.

add to the total runtime. Since these steps are not affected by the investigated modifications, they are not considered in the following analyses. Measuring the required computational effort in terms of the number of cost function evaluations makes this comparison independent from the specific hardware the optimization experiments have been performed on. For the following comparisons the average number of function calls across all 18 726 optimization runs performed on the complete test set is calculated for each algorithmic variant. In case of the reference implementation 254 096 cost function evaluations were performed on average to minimize a single geometry out of the test data set.

4.4.1 Variation of the Sample Count K

The reference implementation of MiniMuDS uses $K = 2n$ samples to approximate the smoothed step function. Addis et al. (2005) reported results on both, $K = n$ and $K = 2n$. They found that the latter choice usually yielded slightly better results so that this value has been adopted in MiniMuDS. Yet, for comparison, the *sample count* variation using only $K = n$ samples was also validated.

Looking at the pose recognition performance there is no significant difference observed between the two values (see Figure 4.14). Yet, a considerably reduced number of cost function evaluations can be expected, since the number of samples drawn in each cycle directly determines the number

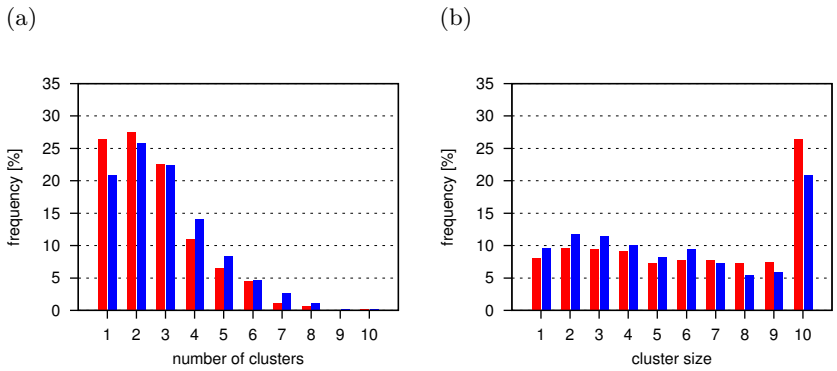


Figure 4.17: Convergence behavior of the reference algorithm (red bars) compared to the *sample count* variant (blue bars).

of local searches required to obtain the step function. Indeed, this variant required only 129 485 function calls per optimization which is a decrease of almost 50 % compared to the reference.

However, the smaller basis for the approximation in this case affects the convergence of the algorithm. As shown in Figure 4.17 (a), the distribution of the number of clusters obtained from the repeated optimization of the same input structure is shifted to the right. This means, that there is a tendency towards a higher number of local minima in which the algorithm can get stuck. Accordingly, the individual minima are less frequently populated so that the fraction of very small clusters between one and four members is larger compared to the reference, whereas clusters with eight or more members are less frequently observed, see Figure 4.17 (b).

The convergence behavior of the search is affected by the decrease of the sample count. On the other hand, the computational time required for the optimization is reduced two-fold while retaining the pose recognition and ranking performance. This suggests that the algorithm is relatively robust against this parameter, although a more comprehensive study covering a broader range of possible values would be necessary to confirm this finding.

Yet, to enable the user to take advantage of the shorter runtime in case of lower sample counts, this parameter was made adjustable via the

MiniMuDS user interface with a default value of $K = 2n$.

4.4.2 Smoothing $f(x)$ instead of $L(x)$

A central part of the implemented optimization algorithm is the derivation of the step function $L(x)$ which is then applied to the smoothing procedure (see section 2.2). The idea behind this is to filter all the local noise and to reveal the underlying funnel structure, while the smoothing is needed to make the resulting step function accessible to an optimization method.

The question is whether this is worth the effort of determining $L(x)$. It is also possible to apply the smoothing procedure directly to the objective function $f(x)$. Thereby, all the local searches that are performed during the sampling phase to determine the step function would no longer be necessary. In fact, the *no step* variant of MiniMuDS takes only 6 443 cost function evaluation per optimization run, which is a decrease by more than 97 % compared to the reference implementation.

If the direct smoothing could reliably filter the local perturbations, thereby revealing the actual funnel shape, than about the same number of local minima should remain on the potential surface as in the case of smoothing the step function. However, Figure 4.18 demonstrates that this is not the case. The range of solutions obtained from repeated optimization runs on the same input geometry is much more diverse.

The number of clusters that are to be expected from ten repetitions is increased, reflected in a strong shift of the corresponding distribution to the right, see Figure 4.18 (a). Accordingly, the number of sparsely populated clusters with only one to four members is substantially increased, Figure 4.18 (b). This indicates that there are more local optima left on the potential surface when smoothing $f(x)$ directly compared th the reference implementation.

Furthermore, a declined ranking performance is recorded for the *no step* variant. While the reference showed a mean difference to the ideal ranking of 0.54 and 0.47 Å for the FlexX and the Gold poses, respectively, omitting the step function leads to differences of 0.59 and 0.62 Å (see Figure 4.14).

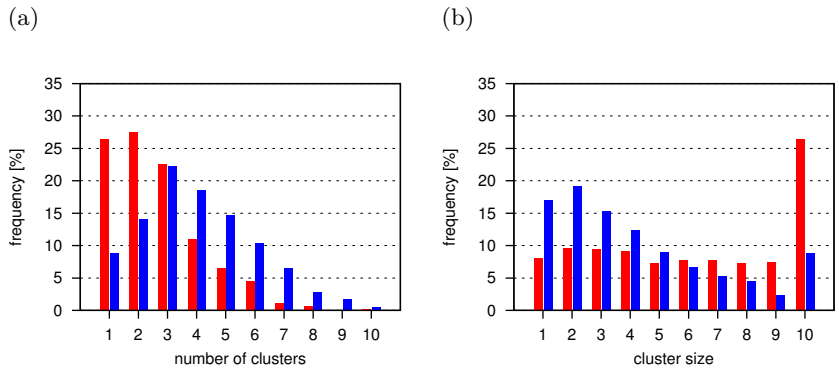


Figure 4.18: Convergence behavior of the reference algorithm (red bars) compared to the *no step* variant (blue bars).

The ranking results together with the worse convergence behavior of the *no step* variation indicate that MiniMuDS in fact benefits from the additional computational effort that is necessary to obtain the step function $L(x)$ before smoothing. However, the dramatically reduced number of cost function evaluations introduces the opportunity to increase the sample count. While the *no step* variant kept this number constant at $K = 2n$, a second experiment with $K = 5n$ was carried out.

Actually, this only led to slight improvements with both, the ranking results and the convergence still not comparable to the reference. More interesting, however, is the fact that the number of objective function calls raised to 7 169 which is only a 1.1-fold increase compared to the *no step* implementation, even though the sample count was increased 2.5-fold. This observation could be the basis for a further study to find out to what extend an increasing sample count can compensate for the loss of accuracy due to the direct smoothing of the objective function.

4.4.3 Omitting the Torsion Potentials

As a last algorithmic alternative the necessity to incorporate explicit torsion potentials in the MiniMuDS energy model is examined. This represents a special case. In contrast to the other variations described above,

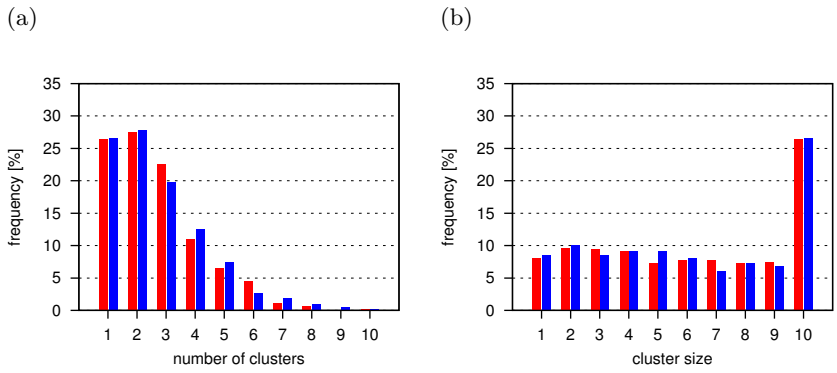


Figure 4.19: Convergence behavior of the reference algorithm (red bars) compared to the *no torsion* variant (blue bars).

disregarding torsion potentials does not alter the optimization algorithm as such, but only affects the applied objective function. Thus, it is not surprising that both the convergence behavior (see Figure 4.19) as well as the search efficiency with 261 715 function calls are nearly identical to the reference implementation which required 254 096 function calls. Even though omitting one term of the objective function obviously affects the absolute runtime of the optimization, this is almost negligible due to the small number of rotatable bonds compared to the number of inter-molecular atomic contacts that have to be evaluated during each function call.

Interestingly, the ranking performance seems to be only marginally influenced by this additional term (see Figure 4.14). These findings confirm once more the question raised in section 2.3 whether the torsion potentials are really necessary in the MiniMuDS energy model.

To clarify this, further properties of the resulting geometries have to be examined. Both the van-der-Waals and the torsion potentials are incorporated into the applied energy model to prevent unfavorable intra-molecular conformations. Most unfavorable ligand geometries would be those containing intra-molecular atomic clashes. Thus, the influence of these two potentials on the respective contact distances is examined first.

Figure 4.20 shows the distribution of short atom contacts within the

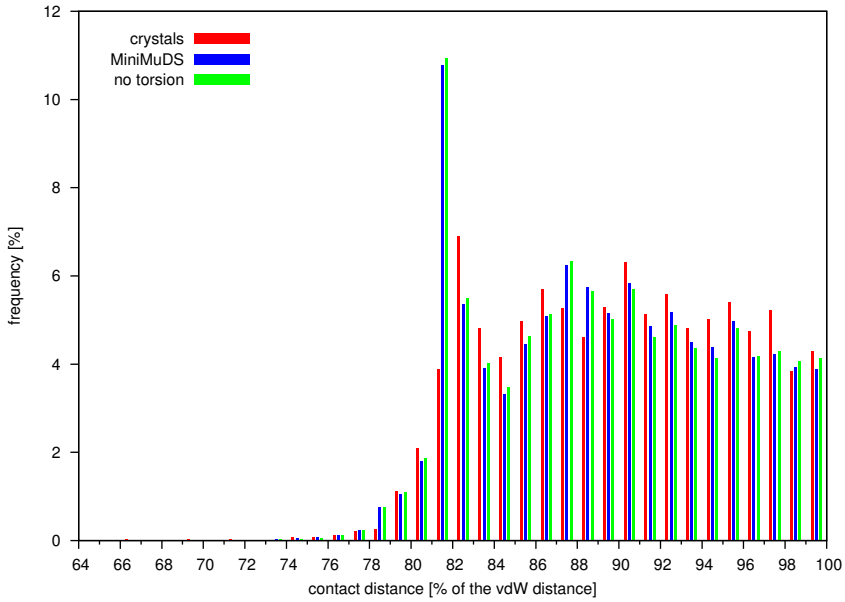


Figure 4.20: Distribution of short intra-molecular contact distances in crystal structures (red), geometries optimized with torsion potentials (blue), and geometries optimized without torsion potentials (green).

ligands of the test set after the optimization with torsion potentials (blue bars) and without them (green bars). Contacts are considered only if there are at least three bonds in between the two atoms. A contact is called short if the distance between the atoms is shorter than the sum of their corresponding van-der-Waals radii as defined in the Tripos force field (Clark et al., 1989). Contact lengths are depicted as the percentage of this van-der-Waals distance.

These contacts are distributed almost identically, no matter whether torsion potentials have been used or not. Moreover, the distribution is also very close to the one found in the original crystal structures (red bars in Figure 4.20). Shortest contacts are observed at distances of at least 70 % of the corresponding van-der-Waals distance which is in the range of the crystal structures.

This indicates that intra-molecular clashes are generally prevented by the van-der-Waals potentials, that have been present in both cases. To

confirm this, further experiments have been performed where the objective function missed the van-der-Waals term. In these cases contacts below 5 % of the corresponding van-der-Waals distance are observed, independently of the torsion potentials present or not.

Thus, it is clear that van-der-Waals potentials are essential in order to prevent intra-molecular clashes. In contrast, the torsion potentials are not necessary in this sense. This was expected.

There is only one striking difference between the crystal structures and the optimized geometries. This is the extreme peak for contacts between 81 and 82 % of the corresponding van-der-Waals distance in the latter case. All structures used as input for the optimization experiments had initially been subjected to the CHARMM force field (Brooks et al., 1983) by the authors of the test set (Cheng et al., 2009).

Contacts between two opposing atoms in a six-membered ring are considered in the presented statistics. For frequently represented phenyl rings the distance between two opposing carbon atoms in such minimized conformations usually falls exactly between 81 and 82 % of the applied van-der-Waals distance for two carbon atoms (3.4 Å). Hence, 55 % of all contacts between two aromatic carbon atoms fall in this range, in contrast to only 1 % of the contacts between any two other atoms. This yields the pronounced peak in this position.

On the other hand, the crystal structures usually have not been subjected to the same force field. Thus, the distances between aromatic carbon atoms are scattered in a broader range. Yet, still 16 % are falling into the aforementioned range of 81 and 82 % of the corresponding van-der-Waals distance, compared to again only 1 % of all other contacts.

Beside atomic clashes a molecular conformation can also be energetically inferior due to unfavorable dihedral angles. Considering only freely rotatable single bonds, the contribution of these angles to the internal energy of a molecule is known to be much smaller compared to bond stretching, angle bending or the rotation around multiple bonds, for example, which would require quite substantial energies to be altered from the equilibrium values (Leach, 2001, pp. 173–174). Therefore, most of the conformational

flexibility of a molecule arises from the combination of its torsional and non-bonded interactions and their energy contributions.

In this sense, the question raised in section 2.3 is whether the non-bonded contributions arising from the protein environment are strong enough to completely determine the bound ligand conformation on their own? Actually, this often seems to be the case. This is why ligands from protein-ligand complex structures frequently comprise dihedral angles that do not correspond to a minimum energy conformation as they are expected to be adopted in the isolated state. The energy gain due to favorable protein-ligand interactions compensates for the slightly higher torsional energy. In this case, the torsion potentials would indeed be unnecessary, assuming that the DrugScore potentials adequately represent the interactions to the environment. This is what the presented results suggest so far.

However, there exist special cases of torsion angles that need closer inspection. Bonds treated as rotatable by MiniMuDS also cover some single bonds comprising a partial double bond character. In principle, these bonds are rotatable, yet they exhibit an increased rotational barrier. Substituents at both sides of the bond tend to be planar as long as this is not prevented for sterical reasons.

An example is given in Figure 4.21. The formula shows the ligand bound to Factor Xa in PDB complex 1mq6 which is contained in the test data set. The atom types for the highlighted part of the molecule from top to bottom are given as C.3 – N.pl3 – C.2 – N.2. Relying upon a correct atom typing in the test data set, the central bond comprises a partial double bond character so that this fragment is expected to adopt a planar conformation. Nevertheless, this bond is regarded flexible during the optimization.

In the following, the torsion angle defined by the highlighted part of Figure 4.21 is regarded. All 50 docking solutions in the test set generated by Gold show a planar conformation, 26 out of them with an angle of 0° as depicted in Figure 4.21. 24 times the terminal ring is flipped by 180°. FlexX on the other hand suggests geometries in discrete steps of 45°. Each of the planar states is adopted 10 times, 21 cases show an angle of 45°/135° and a 90° angle is observed 9 times.

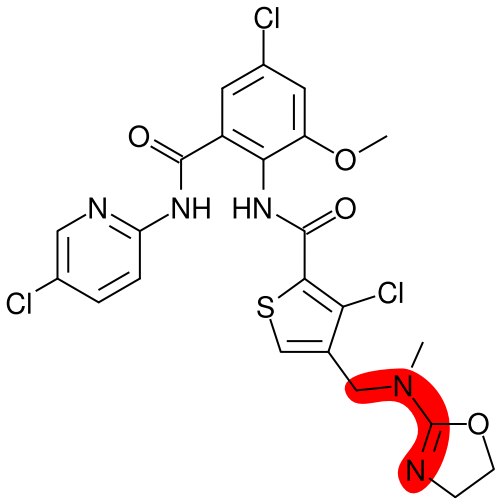


Figure 4.21: The ligand molecule from PDB complex 1mq6. The highlighted moiety is expected to be planar due to the partial double bond character of the central bond.

The corresponding angles after the optimization are given in Figure 4.22. This plot clearly shows that planar docking solutions are usually kept close to a planar conformation if an explicit torsion term is included in the energy model. Solutions with a 45°/135° angle are likely to be rotated towards the plane whereas perpendicular conformations mostly stay close to 90°, with the exception of two geometries that are rotated to be almost planar. This exactly reflects the knowledge-based potential applied to this torsion angle during the optimization, which is shown in Figure 4.23.

In contrast, geometries obtained from the optimization without torsion potentials do not show any preferred conformation for this angle. Input structures that have been planar are scattered up to about 60°. Geometries comprising a 45°/135° angle do not show any preference towards a planar conformation comparable to the reference implementation. Finally, the angles in the perpendicular docking solutions are widely spread across the whole range of possible angles after the optimization.

Interestingly, this difference between the geometries produced by the two variants of MiniMuDS is not reflected by the rmsd values of the solutions

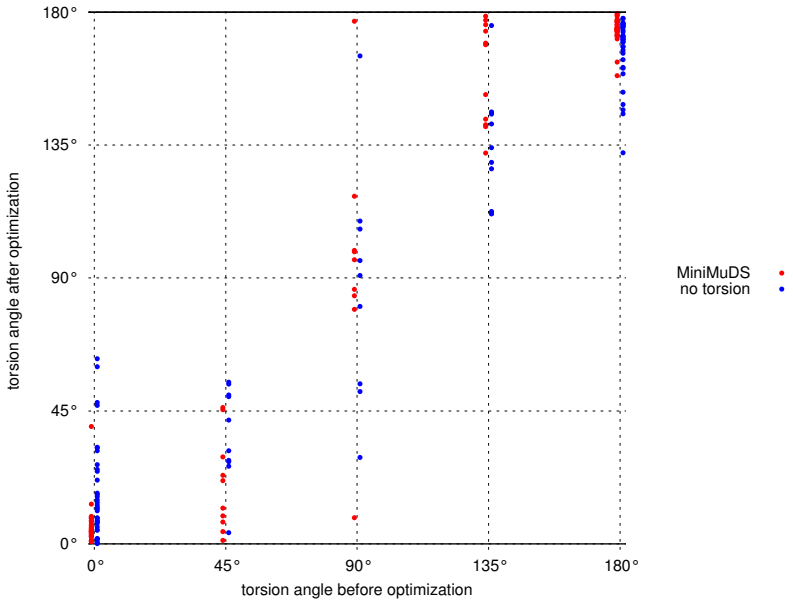


Figure 4.22: Dihedral angles for the torsion defined in Figure 4.21 before and after the optimization with (red) and without (blue) torsion potentials.

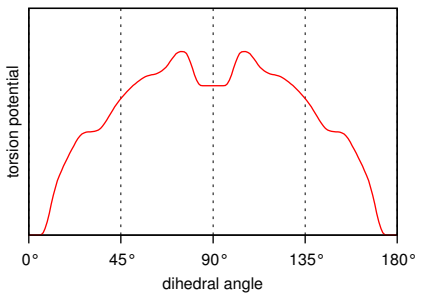


Figure 4.23: Torsion potential applied by MiniMuDS for the rotatable bond defined in Figure 4.21.

compared to the crystal structure of complex 1mq6. Since rotations about the considered bond affect only a relatively small part of the molecule, the final rmsd values are mainly determined by a correct placement of the remaining moieties. Thus, solutions of both variants cover the same range of rmsd values from 0.4 to 2.1 Å, no matter whether this explicit bond shows a planar conformation or not. The same observation is made for the resulting rankings. This can explain why there is no significant difference detectable in the pose recognition performance of the different implementations. Nonetheless, the behavior of the reference implementation, favoring planar conformations in agreement with chemical knowledge, is obviously the desired behavior from a structural point of view.

Keeping this in mind, the more reasonable results obtained from the explicit modeling of the torsional energy in the implemented objective function seems to be worth the little additional computational effort required to evaluate the corresponding potential values.

4.5 Case Studies

4.5.1 Single-Ligand Optimization

As a first example, the structure of human nuclear serine/threonine kinase Chk1 in complex with a pyrrolopyrimidine based inhibitor (Foloppe et al., 2005) is taken from the test data set (PDB code 2brm). The formula of the inhibitor is given in Figure 4.24 (a), its binding mode inside the receptor is shown in Figure 4.24 (b). The ligand is facing the so-called *hinge region* in the back and is covered by the glycine rich loop (G-loop) from above. In Figure 4.25 (a) the bound ligand is shown in more detail from above, together with the peptide backbone of the hinge region. Two hydrogen bonds are formed to Cys87.

The top ranked docking solution generated by FlexX for this complex is given in Figure 4.25 (b). Even though the two hydrogen bonds to the backbone are correctly predicted, the ligand is placed in the pocket in a completely different binding mode. With exception of pose number 14 the

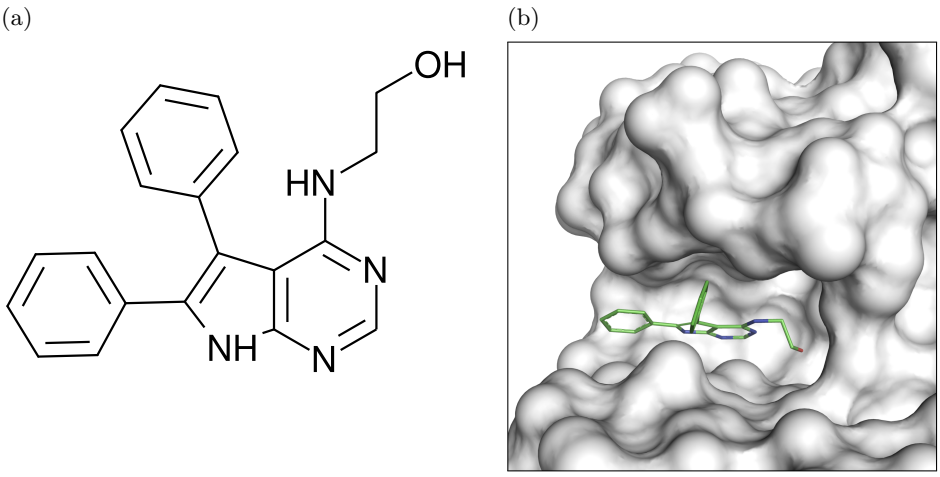


Figure 4.24: (a) Chemical structure and (b) binding mode of the ligand bound to the kinase Chk1 in the PDB complex 2brm.

first 16 docking solutions all comprise this binding mode with an rmsd of 5.4 Å to the crystal binding pose. Only three out of the fifty generated solutions show a binding mode similar to the crystal structure, see Figure 4.25 (c). Although they are shifted to the left by about 1.5 Å, they still constitute the best produced solutions with rmsd values between 1.7 and 2.1 Å. Nevertheless, they are only rank on positions 27, 28 and 30 by FlexX.

The application of DrugScoreX to rescore all solutions yielded exactly the same pose on first rank as in the original FlexX ranking. The three well placed poses from Figure 4.25 (c) are now found on positions 23, 25 and 21, respectively. Running a MiniMuDS optimization prior to rescoring the achieved poses considerably changes the picture. DrugScoreX is now able to recognize a pose very similar to the crystal structure on first rank. The optimized geometry of solution 28 is ranked best, followed by the optimized pose number 30 on rank two. As shown in Figure 4.25 (d), the optimized geometry of pose 28 is not only scored best, but its binding mode is additionally modified by MiniMuDS to exactly fit the crystal structure with a final rmsd of 0.5 Å.

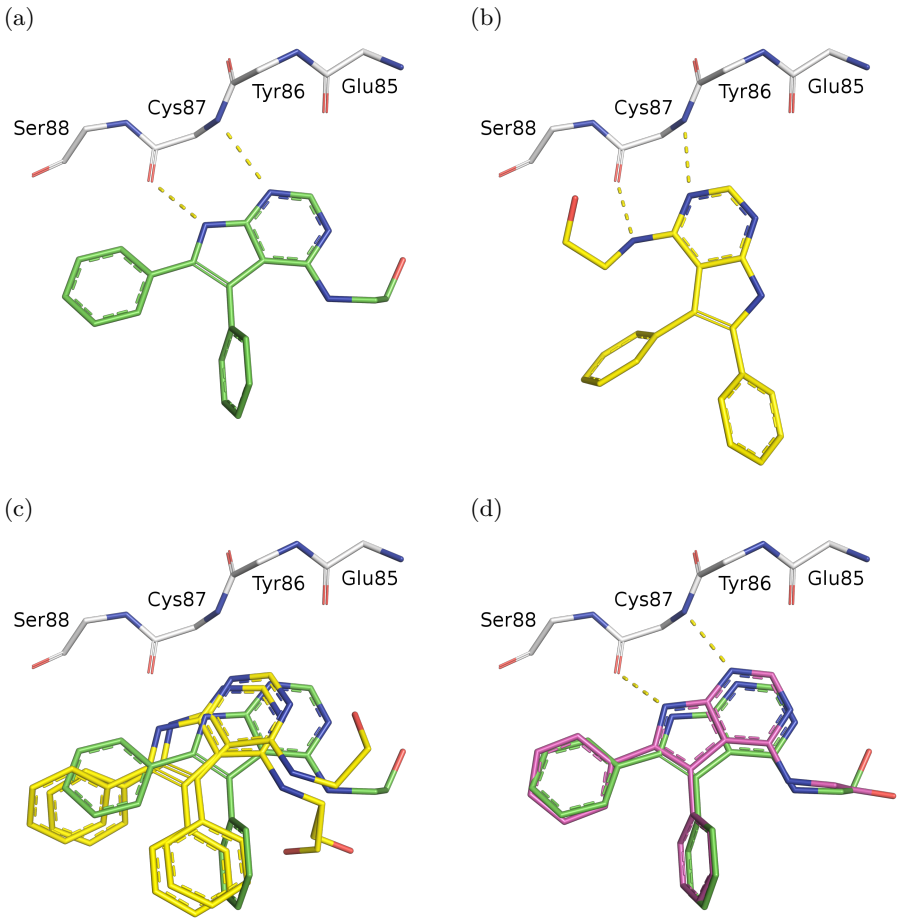
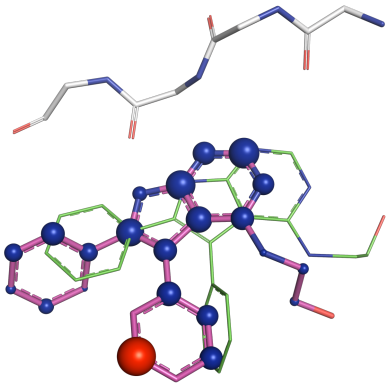


Figure 4.25: (a) The binding mode of the ligand in crystal structure 2brm interacting with the hinge backbone. Hydrogen bonds are shown with dashed lines. (b) Rank 1 docking solution generated by FlexX. This pose is also ranked best by DrugScoreX. (c) Best placed poses found within all docking solutions on positions 27, 28 and 30 (yellow) compared to the crystal structure (green). (d) Optimization result for pose 28 (purple), which is now placed on rank 1 by DrugScoreX, compared to the native state (green).

The modifications introduced to pose 28 in the course of the optimization are examined in detail in the following. The whole run comprised 50

optimization cycles from which a few snap shots representing significant intermediates are presented. The intermediate geometry is shown in purple, the crystal structure in green and the starting geometry in yellow are given for comparison. In addition, the contributions of individual ligand atoms to the total DrugScore value are visualized. Blue balls indicate favorable DrugScore interactions at the particular position whereas unfavorable contributions are depicted by red balls. The radius of a ball is related to the absolute value of the contribution. For each snap shot the current DrugScore potential and the rmsd to the native state are given.

Input:

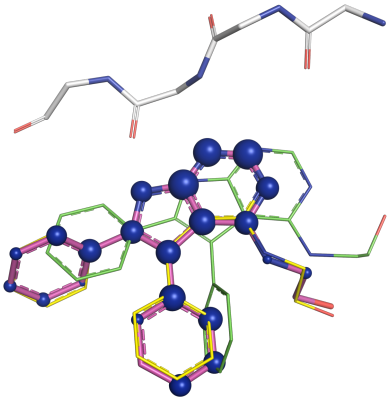


The pyrrolopyrimidine scaffold shows an off-set of 1.5 Å to the left; the aminoethanol moiety contributes almost nothing to the score; the phenyl ring in 2-position is inclined by 62° compared to the native state, the neighboring 3-phenyl substituent is displaced by 88°. Strongly unfavorable interaction distances occur between the latter phenyl residue (red ball) and the backbone atoms of Leu15 and Gly16 of the G-loop (not shown).

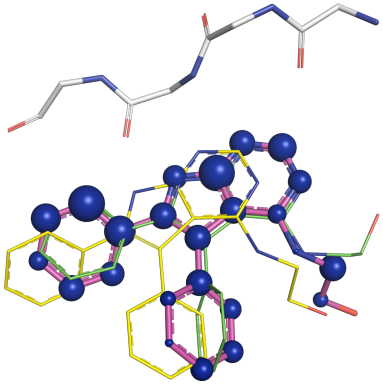
DS = -276 208

rmsd = 2.0 Å

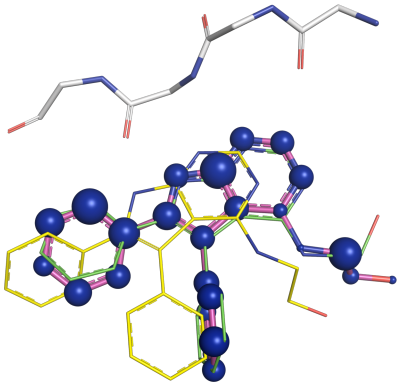
Cycle 1:
The unfavorable contact distances of the central phenyl group diminished due to a shift by 0.5 Å away from the G-loop.
DS = -337 455
rmsd = 1.9 Å



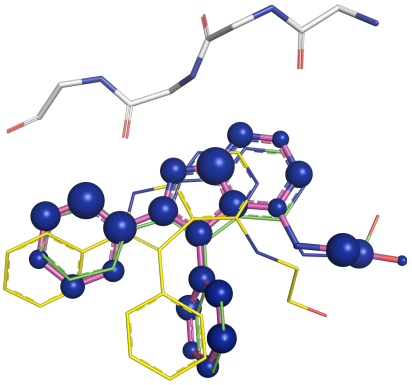
Cycle 21:
The scaffold is correctly placed at the position found in the crystal structure. The inclination of the 2-phenyl group is reduced to 36° while the central one is still rotated off by 73°.
DS = -419 490
rmsd = 1.1 Å



Cycle 39:
The central phenyl moiety nearly adopted the crystal conformation with a remaining deviation of 16°. The N-CH_2 fragment of the aminoethanol is placed correctly.
DS = -439 962
rmsd = 0.6 Å



Cycle 50:
The 2-phenyl ring is finally only inclined by 26° compared to the crystal structure, the central one is off by only 5°. Except the terminal hydroxyl group of the aminoethanol, every atom is placed almost exactly as in the native state. All parts of the ligand now yield substantial contributions to the total score.
DS = -462 088
rmsd = 0.5 Å



The next example is chosen to demonstrate the capability of MiniMuDS to discriminate between near-native and misplaced ligand poses. It concerns the RNA-dependent RNA polymerase (NS5B) from hepatitis C virus with a bound sulfonamide type inhibitor. The binding geometry is available from the PDB complex 2d3u (Biswal et al., 2006). The chemical structure of the inhibitor is shown in Figure 4.26 (a). It binds to the so-called *thumb* domain characteristic for RNA polymerases. Its binding mode is illustrated in Figure 4.26 (b).

The suggested poses obtained from the Gold docking protocol are shown in Figure 4.27 (a). 25 out of the 50 solutions resemble the correct binding mode (shown in yellow) with rmsd values to the crystal structure between 0.6 and 2.8 Å. The remaining 25 solutions are all placed in wrong orientations (shown in white). Their rmsd values range from 6.2 to 10.4 Å.

The question is whether it will be possible to reliably discriminate between the correct and the misplaced poses with the help of any scoring scheme? Figure 4.27 (b) shows the ten top ranked solutions according to their GoldScore. Although three of the correctly predicted poses appear on positions 3, 4 and 5 (yellow), the remaining ones still show great diversity. This indicates that the scoring scheme is not able to reliably retrieve near native solutions from the entire set of generated geometries.

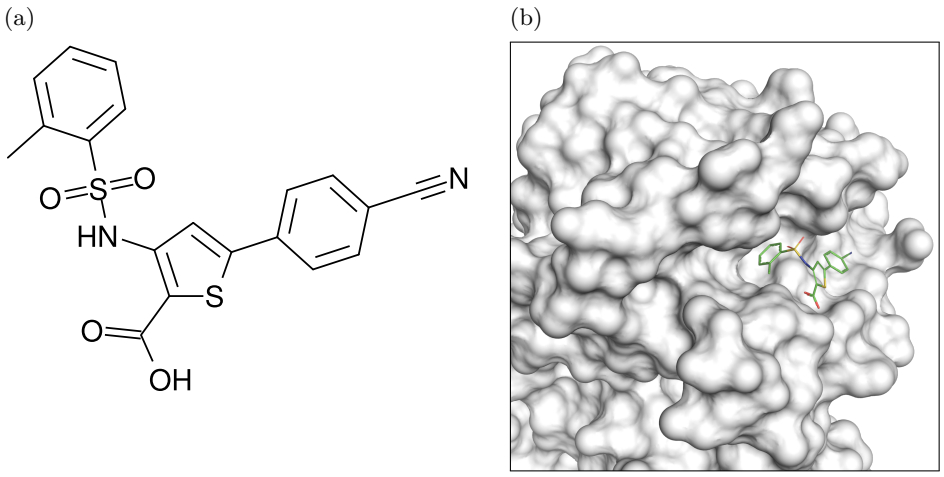


Figure 4.26: (a) Chemical structure and (b) binding mode of the inhibitor bound to the NS5B in the PDB complex 2d3u.

Rescoring the docking solutions directly with DrugScoreX without minimization yields the poses shown in Figure 4.27 (c) on the first ten ranks. This scoring scheme clearly favors two distinct orientations. The first cluster of binding modes with the cyano group pointing to the left (white conformations) is found four times among the top ten solutions. The second cluster contains the native binding mode with the cyano group oriented to the right (yellow conformations). It is represented by six geometries, among them the two best-scored solutions with rmsd values of 1.6 and 0.9 Å to the crystal structure (green conformation), respectively.

Finally, all poses have been optimized with MiniMuDS prior to rescoring. The ten geometries obtaining the best DrugScore values after the minimization are depicted in Figure 4.27 (d). This time, all top scored poses consistently represent the native binding mode (shown in purple). With rmsd values between 0.4 and 0.8 Å they all resemble the crystal structure (green) impressively accurate. There is only one case where the terminal phenyl ring attached to the sulfonamide is flipped by 180° so that the substituted methyl group points into the wrong direction. This leads to an rmsd of 1.5 Å for this geometry found on rank eight.

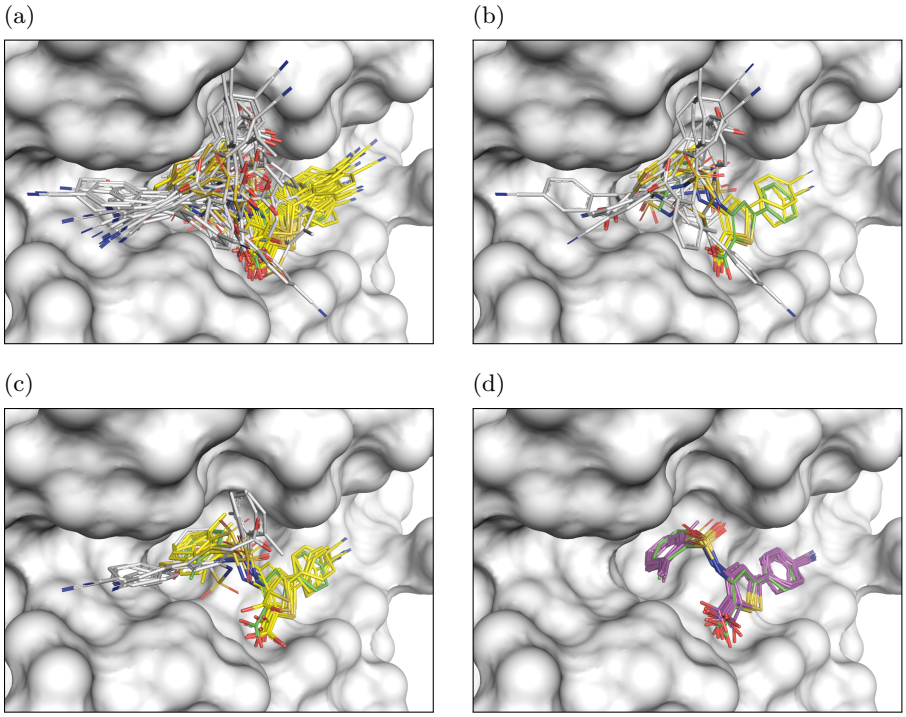


Figure 4.27: (a) 50 docking solutions generated by Gold for PDB complex 2d3u. The crystal structure is shown in green, correctly predicted docking poses in yellow and misplaced poses in white. (b) The top-10 Gold solutions. (c) The top-10 solutions according to DrugScoreX if all poses are directly rescored. (d) The top-10 solutions according to DrugScoreX after poses have been optimized with MiniMuDS.

The capability of MiniMuDS to enable DrugScoreX to reliably discriminate between near-native and misplaced poses becomes even more evident when considering the ranking of all 50 poses generated by Gold. Figure 4.28 shows the rmsd values of the individual geometries on each rank depending on the applied scoring scheme. The rmsd is encoded by the color starting with green for near-native geometries and turning to red with increasing deviation. The top row gives the perfect ordering according to the rmsd values, ignoring the actually achieved rankings.

In the Gold ranking, most of the well-docked solutions are placed on

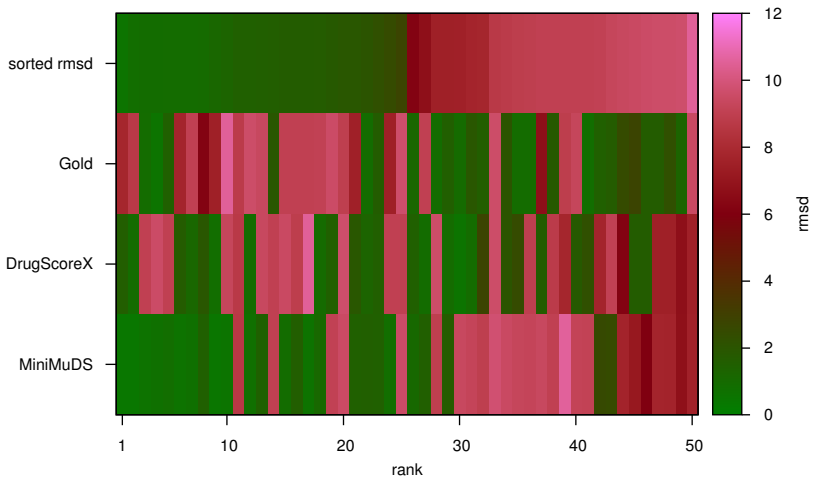


Figure 4.28: Rmsd values for all geometries generated for PDB complex 2d3u depending on the applied scoring scheme and the rank assigned to the geometry. Green colors depict low rmsd values, red colors high values.

high ranks beyond position 20, while almost all poses prior to this position show pronounced deviations. DrugScoreX places at least two near-native solutions on first ranks. Subsequently, a broad scatter of geometries with low and high rmsd follows. Only the optimization with MiniMuDS enables DrugScore to reliably identify near-native geometries on first ranks. They are now almost all placed within the first half of the ranking order. This example emphasizes once more the need for a subsequent minimization of docking solutions if they are supposed to be rescored using a scoring function different from the one used to generate the poses.

4.5.2 Protein Flexibility in MiniMuDS

As described in section 1.1, one of the major challenges in the prediction of protein-ligand interactions is the handling of protein flexibility and induced-fit effects that occur upon ligand binding. This section explains how to address this problem with MiniMuDS. In the original implementation, the protein is considered rigid during the optimization. Yet, a simple expansion has been introduced to also incorporate side chain movements.

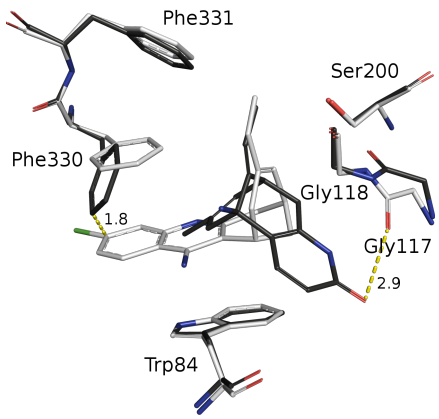


Figure 4.29: Superposition of the binding sites of 1e66 (white) and 1gpk (black). Distances are given in Å.

This is performed by treating the relevant side chains as additional flexible, but covalently bound portions during the minimization (see section 2.5 for details).

As an example, the *Torpedo californica* acetylcholinesterase (TcAChE) was selected. Figure 4.29 shows the superimposed binding pockets of two crystal structures with two different inhibitors. The first structure taken from PDB complex 1e66 (white, Dvir et al. 2002a) accommodates the ligand shown in Figure 4.30 (a). The second one originates from PDB complex 1gpk (black, Dvir et al. 2002b) containing the ligand depicted in Figure 4.30 (b). The ligands are pointing into opposite directions in the binding site.

In this study, a cross-docking experiment is performed. Thereby, the inhibitor from complex 1e66 is placed into the binding pocket observed in complex 1gpk.

Most residues align very closely among the two crystal structures, except for two amino acids. On the left hand side of the binding pocket Phe330 is rotated by about 53° in 1e66 compared to the conformation in 1gpk. This is obviously induced by the chlorine substituted aromatic moiety of the corresponding ligand. It would clearly clash with the phenyl ring of

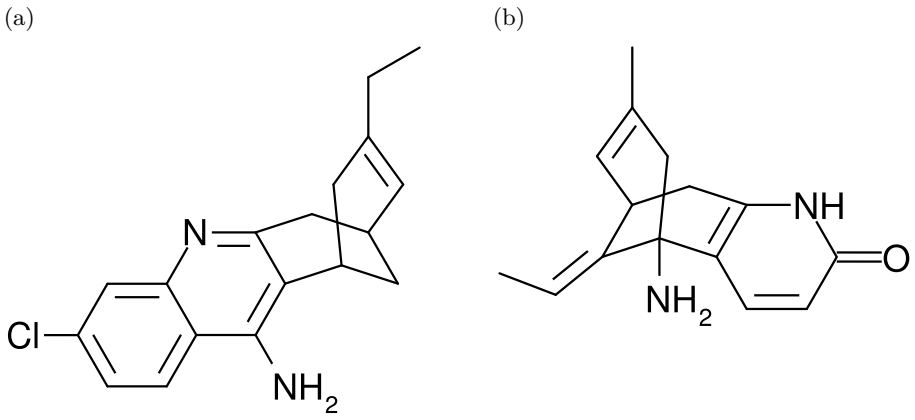


Figure 4.30: (a) The ligand bound in PDB structure 1e66. (b) The ligand bound in PDB structure 1gpk.

Phe330 in the crystal structure 1gpk. In contrast, the rotated conformation in 1e66 allows for a preferred π-stacking interaction between Phe330, Trp84 and the aromatic system of the inhibitor.

The second difference between the two structures concerns a backbone flip between Gly117 and Gly118. Yet, since it is not possible to simulate backbone movements with MiniMuDS, this difference is not regarded in the following considerations. In any case, there is no direct interaction observed between Gly117 and the ligand in complex 1e66. Thus, it seems to be unlikely that the conformation is of much relevance for the prediction of the inhibitor's native binding mode.

To begin with, ligand 1e66 was minimized together with Phe330 within its native pocket. This allows for an estimation of the stability of the correct binding mode applying DrugScoreX potentials. The result is shown in Figure 4.31 (a) with the optimized conformations depicted in purple. The inhibitor is slightly shifted resulting in an rmsd of 0.4 Å. The phenyl ring of the side chain is rotated by only 15°. This underlines that geometries optimized with MiniMuDS nicely correspond to near-native states.

In the next step, the crystal conformation of ligand 1e66 was placed into the binding site of complex 1gpk. This resulted in the clash between

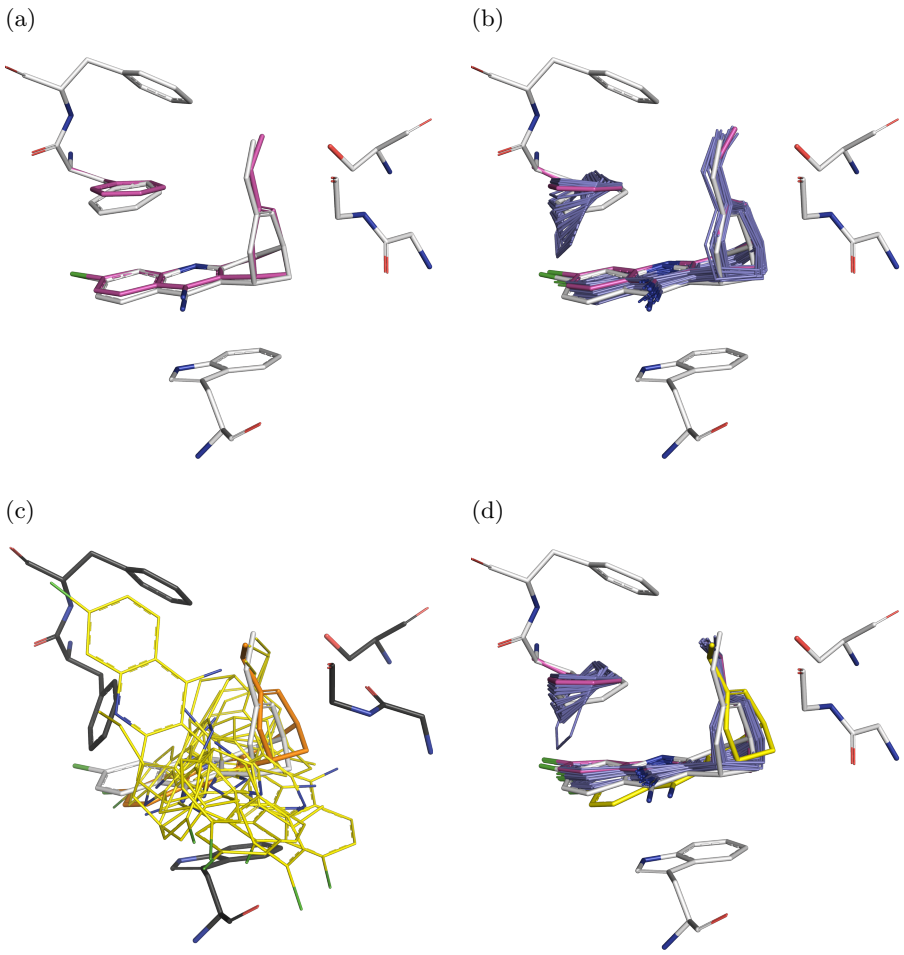


Figure 4.31: Optimization of ligand 1e66 and Phe330 from different starting conformations. The crystal structure of 1e66 is shown in white, 1gpk in black, final MiniMuDS results are depicted in purple, intermediate states in blue and docking solutions in yellow. (a) Optimization of the inhibitor and side chain conformation in the native crystal structure. (b) Optimization of the native inhibitor conformation together with the side chain adopted from 1gpk. (c) Ten diverse poses for the inhibitor docked into the 1gpk binding site. (d) Optimization of the best-placed docking solution inside the 1gpk binding pocket with a flexible Phe330 side chain.

Phe330 and the chloro-substituent of the inhibitor. The question was whether MiniMuDS first of all would be able to resolve this clash, and if so whether the amino acid or the inhibitor is shifted in space?

Figure 4.31 (b) illustrates the course of the optimization. The crystal conformation is shown in white, the final optimized geometry in purple. All intermediate solutions from the 29 minimization cycles are depicted in blue. This shows that it was not only possible to resolve the atomic clash between protein and ligand. Even more important is the fact that the inhibitor maintains its native position while the flexible side chain was moved upwards. This results in a geometry that nearly perfectly resembles the conformation obtained above from the optimization of the ligand in its native crystal structure, cf. Figure 4.31 (a). The inhibitor finally shows an rmsd of 0.6Å to its crystal conformation, the phenylalanine is off-set by 22°.

These are promising results. However, the initial setup comprised an inter-molecular atomic clash. Since the DrugScore potentials are distance dependent, such very short contact distances obtain substantially unfavorable scores. The minimization algorithm was kind of forced to remove this strong repulsion. Thus, the question remains, whether MiniMuDS is also able to find the native ligand pose along with the correct side chain conformation if the input geometry would not create a sterical clash making major structural modifications necessary?

To study this situation, a small but highly diverse set of ten docking solutions of the bound ligand inside the binding site of 1gpk was generated. They are shown in Figure 4.31 (c). The amino acids of 1gpk are depicted in black, the native ligand geometry in white and the docking poses in yellow.

Among the diverse docking solutions there is one that shows the correct orientation (orange lines), although it is inclined to the left by about 14° to avoid clashes with the side chain of Phe330. This pose was selected as a model-built input geometry lacking serious steric protein-ligand conflicts. Thus, steric force will not immediately modify the input structure.

The result of the optimization of this conformation is shown in Fig-

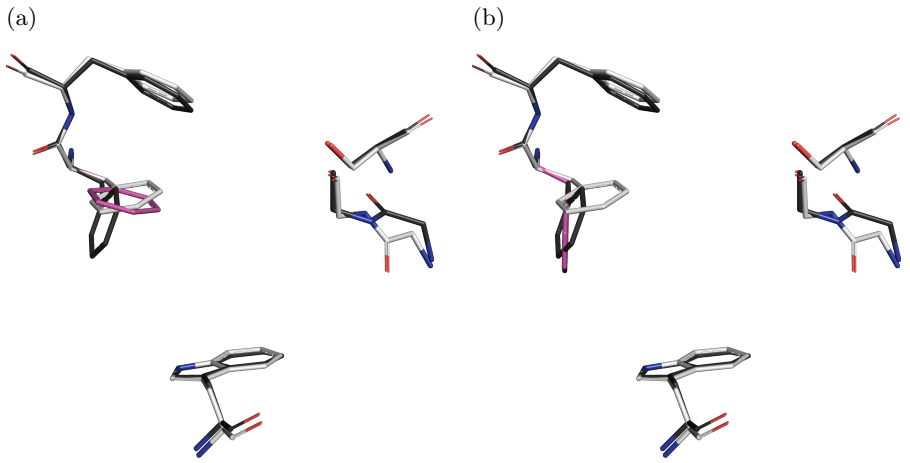


Figure 4.32: Optimization of the phenylalanine 330 in the absence of a ligand. PDB complex 1e66 is shown in white, 1gpk in black and MiniMuDS results in purple. (a) Input geometry from PDB complex 1e66. (b) Input geometry from PDB complex 1gpk.

ure 4.31 (d). Again, the side chain immediately moves upwards towards its native position. Interestingly, the inhibitor shows a comparable movement. This is indicated by the intermediate minimization steps between the docking solution (yellow) and the optimized conformation (purple). The final geometry exhibits an rmsd of 0.8Å to the crystal conformation.

To see whether the upwards rotation of the Phe330 side chain is actually induced by the placement of the inhibitor in the above described geometry, the uncomplexed pocket was also minimized considering both conformations of Phe330 as input geometry.

The optimization of the ligand bound state (1e66) resulted in a slightly rotated conformation if Phe330 that was already detected to be the optimum previously with an rmsd of 0.4 Å to the native state. This is shown in Figure 4.32 (a). In contrast, when starting from the alternative conformation (1gpk), the side chain shows a minimum that is slightly rotated into the opposite direction. The optimized geometry with an rmsd of 0.9 Å in this case is depicted in Figure 4.32 (b).

Table 4.7: Ranking of ten diverse docking solution for ligand 1e66 inside the 1gpk binding site according to different scoring schemes.

Gold		DrugScore[2)]	MiniMuDS		MiniMuDS/sc[3)]	
rank	rmsd[1)]	rank (score)	rank (score)	rmsd	rank (score)	rmsd
1	**6.1**	2 (-368028)	2 (-460909)	6.0	3 (-454394)	6.0
2	5.4	3 (-360616)	5 (-420297)	5.1	6 (-412294)	5.0
3	5.6	7 (-276706)	4 (-421803)	5.2	7 (-411483)	5.2
4	5.7	**1 (-417587)**	**1 (-470794)**	**5.5**	2 (-468052)	5.5
5	1.1	8 (-247368)	8 (-348680)	1.1	**1 (-526270)**	**0.8**
6	6.0	6 (-281407)	6 (-397941)	6.0	8 (-396123)	5.9
7	4.2	10 (-217684)	10 (-345049)	4.2	4 (-437377)	4.1
8	5.2	4 (-341340)	3 (-431097)	5.2	5 (-427305)	5.3
9	6.1	5 (-312152)	9 (-347549)	6.0	10 (-352023)	6.0
10	5.2	9 (-217970)	7 (-366793)	5.8	9 (-356136)	5.9

1) All rmsd values are given in Å. 2) Rmsd values are equal to those for the respective gold solutions. 3) Phenylalanine 330 side chain included as a flexible component in addition to the ligand.

Beside the finding that MiniMuDS is able to detect the native inhibitor placement along with a side chain conformation from a reasonable docking mode as starting point, the ranking of the individual docking solutions is of major interest, once again. As shown in Figure 4.31 (c), ten diverse docking solutions have been generated for the bound ligand taken from the crystal structure 1e66 inside the binding pocket with the conformation observed in 1gpk. Table 4.7 lists the ranking of these solutions according to different scoring schemes.

The first column gives the ranking according to the original docking score together with the rmsd of the respective solution to the native conformation of the ligand. The poses have been directly rescored with DrugScoreX, which is shown in the second column. Then, the ligand geometries have been optimized inside the rigid binding pocket before rescoring. The resulting ranking and rmsd values are given in the third column. Finally, the poses have been optimized simultaneously with the conformation of the side chain of Phe330 ("MiniMuDS/sc", last column). The best scored solutions are marked bold for each scoring scheme. In addition, the best

solution in terms of the rmsd to the correct binding mode is highlighted in green.

None of the scoring schemes, except for the last one including side chain movements was able to detect the correct binding mode on first rank. Moreover, in the latter case the best solution is not only scored on first rank, but also clearly distinguished from the remaining poses in terms of absolute scores. While in almost all other cases the inclusion of the side chain into the optimization had virtually no impact on the obtained DrugScore values compared to the rigid optimization, the overall score of the best solution has substantially improved.

4.5.3 Including Water Molecules in the Optimization

The last example illustrates the optimization of a ligand embedded into a surrounding water network. The protein chosen for this purpose is the carbonic anhydrase II (CAII) from PDB complex 3kig (Schulze Wischeler et al., 2011). In this case, an azide-containing ligand is covalently attached to the protein surface via a disulfide bridge. The chemical structure of this ligand is depicted in Figure 4.33 (a). The terminal thiol group is used to form a disulfide bridge to Cys64. The resulting binding mode is shown in Figure 4.33 (b). The disulfide anchor is placed in close vicinity to the active site of CAII.

The binding mode of the ligand is stabilized by two water molecules which mediate several contacts between protein and ligand. This is shown in detail in Figure 4.34 (a). The water molecule next to Lys169 is forming hydrogen-bonds to Gly63, Lys169 and the backbone of Phe231. In addition, it interacts with the nitrogen of the ligand's amide bond. The upper water molecule is bridging an interaction from the ligand's carbonyl oxygen to the backbone NH of Tyr7. Furthermore, this oxygen directly interacts with Asn11.

In the following, it will be investigated whether MiniMuDS is able to correctly reconstruct the given hydrogen-bonding network. Therefore, reasonable initial positions for the water molecules have to be generated. As

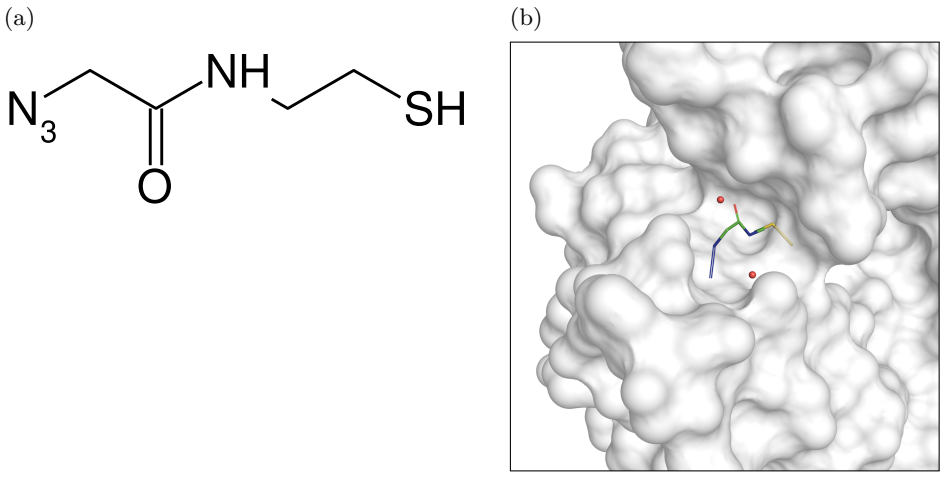


Figure 4.33: (a) Chemical structure and (b) binding mode of the covalently attached ligand to CAII in PDB complex 3kig.

initial model an uncomplexed binding site without attached ligand was assumed. Putative water sites were predicted using DrugScoreX with its implemented algorithm *dsx_wat*[1]. This tool produces a consistent network of water positions maximizing the DrugScore value when each placed water is evaluated against all remaining water molecules and against the given protein environment.

The water network obtained this way is shown in Figure 4.34 (b), the predicted oxygen positions are colored in yellow. For comparison, the two crystallographically observed water molecules are shown in red, however, they have not been considered in any of the calculations. Remarkably, the site of the water molecule next to Lys169 is almost exactly predicted with a deviation of only 0.5 Å.

This is not the case for the second water position. The network generated for the empty pocket suggests three water molecules that are nicely placed around the crystallographically observed one, as shown in Figure 4.34 (c). They are part of a larger network which is finally connected to Glu238, cf. Figure 4.34 (d).

[1]Unpublished results, personal communication by Gerd Neudert.

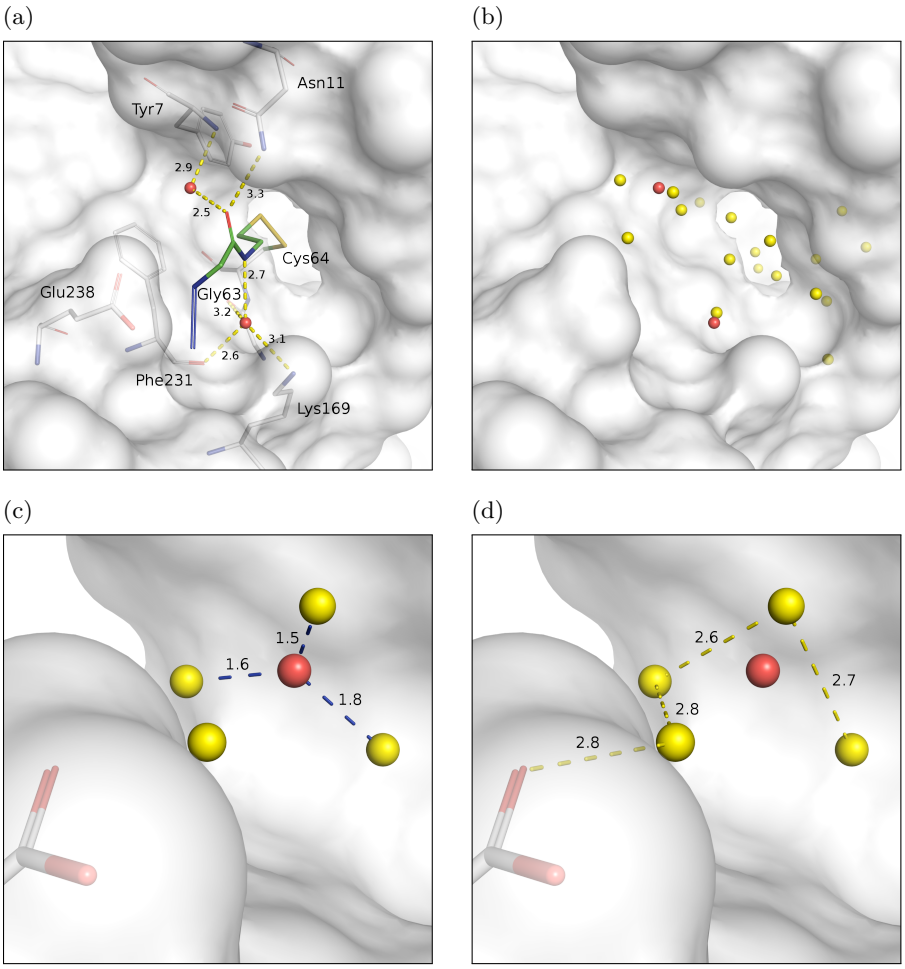


Figure 4.34: Predicted water positions in the uncomplexed CAII binding site. Crystallographically observed water molecules in red, predicted ones in yellow. Distances are shown in Å. (a) The hydrogen-bonding network in the crystal. (b) Predictions according to the DrugScoreX potentials. (c) Three predictions encompassing the crystallographically assigned water molecule. (d) They can form a consistent network among one another.

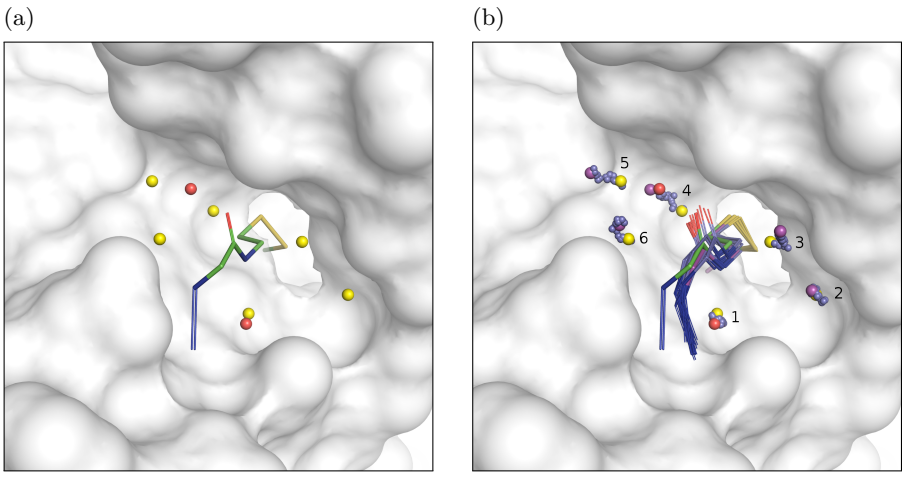


Figure 4.35: (a) MiniMuDS input geometries of the ligand and predicted water positions (yellow). (b) Optimization result (purple) with intermediate states (blue) and starting structure (yellow).

Subsequently, the crystal conformation of the ligand was placed onto the predicted water network. The ligand strongly modifies the local environment. Thus, the previously predicted water positions cannot be expected to match these new conditions. Generally, MiniMuDS is supposed to adapt the predicted water positions to the modified environment now accommodating the ligand. Of course, there are water sites predicted for the empty pocket in close vicinity to the native ligand position. Since these water molecules would cause heavy clashes with the overlaid ligand, all water molecules within 1.2 Å of the ligand's crystal conformation have been discarded due to their short distance. Additionally, also water molecules showing no direct contact to the ligand have been removed.

The resulting input geometry is shown in Figure 4.35 (a). Again, the predicted water molecules are shown in yellow whereas the crystallographically observed reference positions are depicted in red for comparison. The ligand conformation was than optimized together with the six remaining water positions.

Figure 4.35 (b) shows the corresponding optimization result. The final

ligand orientation is depicted in purple, the intermediate states in blue and the crystal structure in green. The ligand immediately moves to its preferred orientation which then experiences no substantial further modifications. The final rmsd with respect to the crystal structure amounts to 0.5 Å.

The well-predicted position of the water molecule next to Lys169 is only slightly modified (number 1 in Figure 4.35 (b)) . It shows a final deviation of 0.4 Å to the crystallographically observed position. Also the water molecule 2 is nearly not modified with a final deviation of 0.3 Å from its starting position. Water molecule 3 is slightly moved out of the binding site. This presumably results from its initial short distance to the ligand of 1.9 Å.

In contrast, all waters on the left hand side are strongly shifted away from the ligand. In case of water molecule 4, this is due to its initial distance of 1.6 Å to the ligand yielding strongly repulsive potentials. Interestingly, it is shifted almost exactly to the position of the second crystallographically observed water molecule, deviating by 0.8 Å. The finally adopted position would clash with the predicted position of water molecule 5, however. In consequence, also this one is translated, parallel to the first one, in order to retain favorable mutual distance. Water molecule 6 is also shifted to the right, correlated the other two molecules. Thereby, the network of favorable interactions among the water molecules is conserved.

Finally, the MiniMuDS solution resembles nicely the hydrogen-bonding network given in the crystal structure, now predicting two additional water molecules which expand the network to the right. This is illustrated in Figure 4.36 (a). This demonstrates MiniMuDS' capability to detect and optimize important contacts not only between protein and ligand, but to include also small and highly flexible solvent molecules.

As promising as the obtained results are, they also make some current drawbacks of MiniMuDS apparent. The current objective function focuses the DrugScore pair potentials which are only distance-dependent between two atoms (cf. concerted movement of water molecules in the network

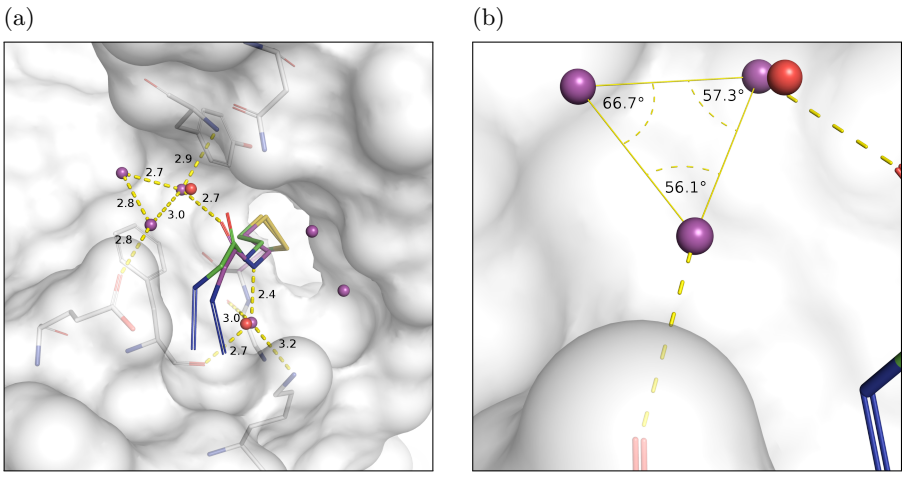


Figure 4.36: (a) Final water and hydrogen-bonding network predicted by MiniMuDS (purple), crystal waters in red, crystal ligand in green. (b) Unfavorable interaction angles predicted by MiniMuDS. Distances are shown in Å.

on the left-hand side). They comprise six directly connected hydrogen bonded contacts, all exhibiting distances from 2.7 to 3.0 Å. These are ideal distances for these types of interaction.

However, it is also known that these interactions are directional and will show angle dependence. In ideal case, a water molecule providing two hydrogen bond donor and two hydrogen bond acceptor functionalities should produce an almost tetrahedral coordination geometry.

Considering again the example above, the angles between the individual interactions forming the central triangle of this hydrogen-bonding network are between 56° and 67°, as shown in Figure 4.36 (b). This, however, is far from the tetrahedral angle of 109.5°. This clearly provides a broad field for future development and investigation.

Chapter 5

5 Summary and Outlook

5.1 Summary

The aim of this work was to develop a tool for the optimization of *in-silico* generated protein-ligand complexes according to the DrugScoreX potentials. The scoring function DrugScoreX is typically used to rescore bound ligand geometries that were generated by any docking program. Thus, these ligand poses are optimized according to the internally implemented scoring function used by the selected docking algorithm. Applying DrugScore to such a geometry does not necessarily guarantee reliable and relevant scoring although the docking poses may be geometrically very close to a DrugScore optimum. Considering the steepness of the DrugScoreX potentials, even small variations in the atomic positions can lead to large differences in the resulting scores. In general, this holds true for every rescoring scenario, so that a local optimization with respect to the subsequently used scoring scheme is strongly recommended in these cases.

In 2009, O'Boyle et al. stated, that a local optimization is always constrained to the energy well on the potential surface in which the original pose already resides. So there may be an even deeper well nearby which will not be considered in the local optimization, but would be equally valid. The new tool MiniMuDS, developed in this thesis, should account for this

problem. On the other hand, MiniMuDS is not intended to perform a global optimization since this would require an exhaustive search for the overall best-scoring binding mode and, at the end, would result in a new docking algorithm. Instead, the new optimization algorithm is supposed to stay close to the pose generated by the original docking engine and simply adapts this solution to the DrugScoreX function, a task typically addressed by local search methods.

The tool developed in this thesis was to combine these two tasks by avoiding a strictly local optimization without extending to a fully global search at the same time. To this end, an optimization strategy was implemented, that contains elements of a global optimization, but is still restricted to a local part of the search space. Simply speaking, the applied algorithm can overcome small hills on the potential surface, but only if the following valley is deeper than the current one. Thus, major energetic barriers between basically different conformations will not be passed.

In the validation of MiniMuDS several important properties were shown:

1. The optima of the applied energy model correspond impressively well to the experimentally determined native states of the evaluated complexes. This was shown by the optimization of the original crystal structures, which resulted in an average rmsd of about 0.5 Å, a value much smaller than the one observed in case of *in-silico* generated geometries. This deviation has also to be seen in light of the positional accuracy estimated for experimental structure determination. The observed deviations virtually fall into the same range.

2. The aim of conserving the given binding modes was achieved. The presented method allows for modifications up to 2 Å rmsd compared to the input geometry. Remarkably, not even 5 % of the optimized docking poses fully exploited this available space. On average a modified geometry shows an rmsd of about 1 Å to the input structure.

3. MiniMuDS improves a given docking solution by about 0.1 Å on average considering its rmsd to the native state. The best performance

was observed for well docked poses between 1 and 2 Å rmsd which could be improved by up to 0.3 Å on average.

4. It could be shown that an optimization exceeding the restrictions of a strictly local search can improve the resulting ranking. Up to 4.7 % better success rates at a 2 Å cutoff and an improvement of up to 9.3 % at the 1 Å level were received when comparing MiniMuDS to a local optimization.

5. Taking into account not only the top ranked solution but the whole ranking, it could be shown that MiniMuDS strongly improves the discrimination between near-native and misplaced poses. Geometries with lower rmsd values to the crystal structure are more likely to be placed within the first positions of the ranking.

6. The inclusion of additional flexible components into the optimization is easy to manage using MiniMuDS while results may strongly benefit. This was shown using the example of protein side chain flexibility and binding relevant water molecules.

7. Considering computational efforts, it was shown that it is sufficient to only subject the 10 top-ranked docking solutions to an optimization. This consistently yielded slightly better ranking results for all applied protocols compared to an optimization of all generated docking solutions. At 80 % less computational effort, up to 4.7 % higher success rates at 2 Å and 2.1 % higher once at a 1 Å cutoff were recorded.

Especially the last aspect mentioned above confirms that it is advisable to focus only on those docking poses for optimization that were already ranked high by another scoring function. This way, only poses that score well with respect to two different scoring functions are considered, taking thereby advantage of some kind of consensus effect.

In light of these findings, the usage of at least a local optimization has to be strongly recommend before applying DrugScore for rescoring purposes. Beyond that, the application of a more sophisticated search strategy

like the one implemented in MiniMuDS is suggested. In particular, when dealing with small, lead-like structures, the presented method showed to substantially improve the rescoring results.

5.2 Outlook

Usually, it is not possible to define an explicit point of completion for a software tool as well as for a research project in general. This is also true in the case of the presented thesis. In the following, this is exemplified through a few ideas for the further development of MiniMuDS.

5.2.1 Algorithmic Enhancements

To begin with, individual parts of the central search algorithm could be subject to improvements. The first two suggestions regard mainly the runtime, whereas the last one concerns the artificial restriction of the search space.

Random Sampling within an n-dimensional Hypersphere

During the sampling phase of the MiniMuDS algorithm, an n-dimensional hypersphere of radius one is centered on the currently best point. Random points are sampled from within this sphere and scaled to the actually applied limits in each dimension afterward. Currently, this is implemented in a "naive" way: Random points are uniformly drawn from an n-dimensional hypercube that is enclosing the hypersphere. This requires the sampling of n uniform random values. Such a point is rejected, however, if its distance to the center is larger than one, meaning that it is placed inside the cube but outside the sphere.

The drawback of this method is that the ratio of the volume of an n-dimensional sphere compared to the corresponding n-dimensional cube is rapidly decreasing with increasing n. Thus, for increasing problem sizes n the time to sample a valid point within the sphere will be increased accordingly.

To overcome this drawback, a sampling method that requires exactly $n+1$ random values to guarantee a uniformly distributed point within the hypersphere could be used. An example is described by Press et al. (2007, p. 1129).

Local Search Method

One of the central step in the MiniMuDS optimization algorithm is the determination of the step function $L(x)$. This is done using the downhill simplex method as described by Nelder and Mead (1965). This is a very simple and intuitive local search algorithm. Yet, it is known to require many cost function evaluations.

An alternative could be given by the direction set methods based on Powell's method (Powell, 1964). These methods are often much faster than the downhill simplex while they still do not require the computation of gradients. See Press et al. (2007, pp. 509–515) for an introduction.

Determination of Protein-Ligand Contacts

Whenever the MiniMuDS cost function is evaluated, the contact distance for each pair of a protein and a ligand atom has to be determined. Yet, since proteins can get rather large while the DrugScore interactions are limited to a distance of 6 Å, large parts of the protein will usually never be considered. To reduce the number of atom pairs that need to be evaluated, a binding pocket is extracted from the protein in a pre-processing step. The determination of the radius of the binding pocket is not trivial, however, especially in the case of multiple flexible components that are to be optimized (see section 2.5 for details).

A solution to this problem would be the implementation of a special data structure supporting efficient nearest-neighbor searches. This could be realized by so-called *KD Trees*, for example. In the current implementation of MiniMuDS, all possible protein-ligand atom pairs have to be evaluated. In contrast, a KD Tree needs to compare much less atom pairs in order to find all protein atoms within the radius of interest.

This helps not only in reducing the number of atomic distances that have to be calculated explicitly, thereby reducing the required runtime. In addition, this supersedes the artificial determination of the pocket radius, thus preventing the related difficulties. The desired conservation of a given binding mode should not be affected, since it was shown that more than 95 % of the optimization runs did not exploit the allowed range of 2 Å, so that they would still conserve their general binding mode, even if the search space was not restricted to this range. An introduction to KD Trees and related data structures can be found in Press et al. (2007, pp. 1101–1110).

5.2.2 Methodology

Beside the possible improvements on the algorithmic level, the following gives some ideas for further developments on a higher level.

The currently developed version of DrugScoreX also provides a set of torsion potentials derived from the CSD in an analog way to the distance dependent pair potentials. These potentials could replace the currently applied set of potentials taken from the MIMUMBA program (Klebe and Mietzner, 1994). This would lead to a more consistent treatment of pairwise protein-ligand interactions and intra-molecular torsions, as both potential types would be based on the same atom-type model and the same statistical evaluation.

Concerning the optimization of protein side chains, the incorporation of rotamer libraries could lead to a reduced problem complexity. Currently, flexible side chains are treated the same way as any other ligand, namely to be fully flexible during the optimization. However, it is known that amino-acid side chains possess distinct preferred conformations. Using this information could be used to reduce the number of possible states that a given side chain can adopt during the optimization process.

With an increasing number of components to be optimized simultaneously, possible solutions and thus the difficulty to find a suitable optimum increases. In particular, if multiple water molecules are included in the

optimization, it can get quite difficult to find an optimal position for each of them. Yet, often only the geometry and score of the actual ligand are of major interest. To reflect this in the optimization process, one could think about a special weighting factor that lowers the influence of water molecules, for example. This could reduce the penalty for slightly shifting them from their theoretical optimum position.

Finally, the question arises what would be missing to make a full docking program out of MiniMuDS? A major step into this direction would be the combination of the implemented optimization algorithm with an algorithm yielding a set of initial placements inside the binding pocket for a given ligand. For sure it would be interesting to examine the capabilities of such a combination in the future.

Appendix A

A Appendix

A.1 Atom Types used for Pair-Potential Derivation

Table A.1: Sybyl Atom Types used by DrugScore^{PDB} and DrugScore^{CSD}.

atom type[1)]	description
C.3	sp^3-hybridized carbon
C.2 (C.1)	sp^2- and sp-hybridized carbon
C.ar	carbon in aromatic ring systems
C.cat	carbon in amidino or guanidino groups
N.3 (N.4)	sp^3-hybridized nitrogen
N.ar (N.2)	nitrogen in aromatic ring systems and sp^2-hybridized nitrogen
N.am	nitrogen in amide bonds
N.pl3	nitrogen in amidino or guanidino groups
O.3	sp^3-hybridized oxygen
O.2	sp^2-hybridized oxygen
O.co2	oxygen in carboxylate groups
S.3 (S.2)	sp^3 and sp^2-hybridized sulfur
P.3	sp^3-hybridized phosphor
F	fluorine
Cl	chlorine
Br	bromine
I[2)]	iodine
Met	calcium, iron, zinc, nickel[3)]

1) Types in brackets are merged into the aforementioned type due to insufficient observations of the original type in the PDB structures. 2) Iodine was only considered in DrugScore^{CSD}. 3) Nickel was only considered in DrugScore^{PDB}.

Table A.2: Fconv Atom Types considered by DrugScoreX.

atom type	description
C.ar6p	sp^2-hybridized carbon with a positive charged resonance structure in a protonated 6-membered heteroaromatic ring
C.ar6x	sp^2-hybridized carbon in a 6-membered heteroaromatic ring
C.ar6	sp^2-hybridized carbon in a benzene ring
C.arp	sp^2-hybridized carbon with a positive charged resonance structure in other protonated heteroaromatic rings
C.arx	sp^2-hybridized carbon in other heteroaromatics
C.ar	sp^2-hybridized carbon in other aromatics
C.2r3o	carbonyl carbon in cyclopropanone or cyclopropenone
C.2r3x	sp^2-hybridized carbon in heterocyclic 3-membered rings
C.2r3	sp^2-hybridized carbon in 3-membered rings
C.3r3x	sp^3-hybridized carbon in heterocyclic 3-membered rings
C.3r3	sp^3-hybridized carbon in 3-membered rings
C.1n	sp-hybridized carbon in cyano groups
C.1p	sp-hybridized carbon with one heavy atom bonded
C.1s	sp-hybridized carbon with two heavy atoms bonded
C.co2h	sp^2-hybridized carbon in explicitly protonated COOH groups
C.co2	sp^2-hybridized carbon in COO^- groups (also set if protonation state is unknown)
C.es	carbonyl carbon in ester groups or anhydrides
C.hal	carbonyl carbon in acidhalogenides
C.am	carbonyl carbon in amides
C.o	other carbonyl carbon
C.s	thionyl carbon
C.gu	sp^2-hybridized carbon in unprotonated guanidino groups
C.guh	sp^2-hybridized carbon in protonated guanidino groups (also set if protonation state is unknown)
C.mi	sp^2-hybridized carbon in unprotonated amidino groups
C.mih	sp^2-hybridized carbon in protonated amidino groups (also set if protonation state is unknown)
C.n	sp^2-hybridized carbon in imines
C.2p	other sp^2-hybridized carbon with one heavy atom bonded
C.2s	other sp^2-hybridized carbon with two heavy atoms bonded
C.2t	other sp^2-hybridized carbon with three heavy atoms bonded
C.et	sp^3-hybridized carbon in ethers
C.ohp	sp^3-hybridized carbon in primary alcoholes

Table A.2: Fconv Atom Types *(continued)*.

atom type	description
C.ohs	sp^3-hybridized carbon in secondary alcoholes
C.oht	sp^3-hybridized carbon in tertiary alcoholes
C.3n	other sp^3-hybridized carbon bonded to nitrogen
C.3p	other sp^3-hybridized carbon with one heavy atom bonded
C.3s	other sp^3-hybridized carbon with two heavy atoms bonded
C.3t	other sp^3-hybridized carbon with three heavy atoms bonded
C.3q	other sp^3-hybridized carbon with four heavy atoms bonded
N.ar6p	positive charged nitrogen in 6-membered aromatics (e.g. pyridinium or NAD+)
N.ar6	sp^2-hybridized nitrogen in 6-membered aromatics
N.arp	sp^2-hybridized nitrogen in protonated aromatics (e.g. both nitrogens in protonated imidazole)
N.ar2	sp^2-hybridized nitrogen in aromatics with two bonded atoms (corresponding to sybyl type N.2)
N.ar3	sp^2-hybridized nitrogen in aromatics with three heavy atoms (corresponding to sybyl type N.pl3)
N.ar3h	sp^2-hybridized nitrogen in aromatics with two heavy atoms and one hydrogen (corresponding to sybyl type N.pl3)
N.r3	sp^3-hybridized in aziridine or azirene rings
N.az	middle nitrogen in azides
N.1	other sp nitrogen
N.o2	nitrogen in nitro groups
N.ohac	nitrogen in hydroxamic acids
N.oh	nitrogen in hydroxylamines
N.ims	imide nitrogen with two heavy atoms bonded
N.imt	imide nitrogen with three heavy atoms bonded
N.amp	carbon- or thionamide with one heavy atom bonded
N.ams	carbon- or thionamide with two heavy atoms bonded
N.amt	carbon- or thionamide with three heavy atoms bonded
N.samp	sulfonamide with one heavy atom bonded
N.sams	sulfonamide with two heavy atoms bonded
N.samt	sulfonamide with three heavy atoms bonded
N.gu1	NH nitrogen in unprotonated guanidino group (only if explicitly protonated)
N.gu2	NH_2 nitrogen in unprotonated guanidino group (only if explicitly protonated)

Table A.2: Fconv Atom Types *(continued)*.

atom type	description
N.guh	nitrogen in protonated guanidino group (also set if protonation state is unknown)
N.mi1	NH in unprotonated amidino group (only if explicitly protonated)
N.mi2	NH_2 in unprotonated amidino group (only if explicitly protonated)
N.mih	nitrogen in protonated amidino group (also set if protonation state is unknown)
N.aap	primary aromatic amine (hybridization can't be determined exactly)
N.aas	sp^2- or sp^3-hybridized secondary aromatic amine
N.aat2	sp^2-hybridized tertiary aromatic amine
N.aat3	sp^3-hybridized tertiary aromatic amine
N.2n	sp^2-hybridized nitrogen bonded to another nitrogen
N.2p	other sp^2-hybridized nitrogen with one heavy atom
N.2s	other sp^2-hybridized nitrogen with two heavy atoms
N.3n	sp^3-hybridized nitrogen bonded to another nitrogen
N.3p	sp^3-hybridized nitrogen with one heavy atom bonded
N.3s	sp^3-hybridized nitrogen with two heavy atoms bonded
N.3t	sp^3-hybridized nitrogen with three heavy atoms bonded
N.4	sp^3-hybridized nitrogen with four bonded atoms
O.ar	aromatic oxygen
O.r3	oxygen in oxiran ring
O.h2o	water oxygen
O.n	oxygen in nitro groups
O.noh	sp^3-hybridized oxygen in hydroxylamine or hydroxamic acid
O.2co2	sp^2-hybridized oxygen in COOH (sp^2-hybridized bonded to C.co2h)
O.2es	sp^2-hybridized oxygen in esters or anhydrids
O.2hal	sp^2-hybridized oxygen in acidhalogenides
O.am	oxygen in carbonamides
O.co2	oxygen in COO^- or CSO^-
O.2po	sp^2-hybridized oxygen in P=O (non deprotonated groups)
O.2so	sp^2-hybridized oxygen in S=O (non deprotonated groups)
O.2p	sp^2-hybridized oxygen in OPO_3H^- or PO_3H^- or POO^-
O.2s	sp^2-hybridized oxygen in OSO_3^- or SO_3^- or deprotonated sulfonamides
O.3po	sp^3-hybridized oxygen with two heavy atoms bonded to at least one phosphor
O.3so	sp^3-hybridized oxygen with two heavy atoms bonded to at least one sulfur

Table A.2: Fconv Atom Types *(continued)*.

atom type	description
O.carb	oxygen in other carbonyl groups
O.o	oxygen in peroxo groups
O.3ac	OH oxygen in COOH, CSOH, $PO(OH)_2$, POOH or SO_2OH
O.ph	oxygen in phenolic hydroxyl group
O.3oh	oxygen in hydroxyl group
O.3es	sp^3-hybridized oxygen in esters or anhydrids
O.3eta	oxygen in aromatic ether
O.3et	oxygen in aliphatic ether
S.ar	aromatic sulfur
S.r3	sulfur in thiiran ring
S.thi	sulfur in thionyl group
S.o	sulfur in SO
S.o2h	sulfur in protonated sulfonamide or other SO_2
S.o3h	sulfur in SO_3
S.o4h	sulfur in OSO_3
S.o2	sulfur in SO_2 or deprotonated sulfonamides (or unknown protonation state)
S.o3	sulfur in SO_3^- (or unknown protonation state)
S.o4	sulfur in OSO_3^- (or unknown protonation state)
S.2	sulfur in CSO^-, COS^- or other sp^2-hybridized sulfur
S.sh	sulfur in SH groups
S.s	suflur in S-S bonds
S.3	other sp^3-hybridized sulfur
P.r3	phosphor in phosphiran rings
P.o	phosphor in PO groups
P.o2h	phosphor in not deprotonated PO_2 groups
P.o3h	phosphor in not deprotonated PO_3 groups
P.o4h	phosphor in not deprotonated PO_4 groups
P.o2	phosphor in deprotonated PO_2 groups (or unknown protonation state)
P.o3	phosphor in deprotonated PO_3 groups (or unknown protonation state)
P.o4	phosphor in deprotonated PO_4 groups (or unknown protonation state)
P.3	other sp^3-hybridized phosphor
F	bonded fluorine or fluoride ion
Cl	bonded chlorine or chloride ion
Br	bonded bromine or bromide ion

Table A.2: Fconv Atom Types *(continued)*.

atom type	description
I	bonded iodine or iodide ion
Li	lithium
Na	sodium
Mg	magnesium
Al	aluminium
Si	silicon
K	potassium
Ca	calcium
Cr.th	chromium (tetrahedral)
Cr.oh	chroumium (octahedral)
Mn	manganese
Fe	iron
Co	cobalt
Cu	copper
Zn	zinc
Se	selenium
Mo	molybdenum
Sn	tin
Ni	nickel
Hg	mercury
B	boron

A.2 The Validation Data Set

The PDB codes of all complex structures in the validation data set as compiled by Cheng et al. (2009) are listed in Table A.3. Three structures (1l2s, 1n2v and 1v48) are already part of the training data set used to parametrize the energy model of the implemented optimization tool (see section 2.3 for details). To prevent any bias, these structures were not considered in the validation of the method.

Another two structures are excluded due to structural deficiencies. The first one is PDB complex 2fzc with the chemical structure of the ligand shown in Figure A.1 (a). Two oxygen atoms, one from each of the two

Table A.3: PDB codes of all complexes in the validation data set.

subset	members
lead-like	1a69, 1ai5, 1ajp, 1ajq, 1avn, 1b8o, 1bcu, 1bra, 1d7j, 1df8, 1e1v, 1e5a, 1e66, 1ela, 1f4e, 1f5k, 1fcx, 1fcz, 1fd0, 1fki, 1flr, 1ftm, 1gpk, 1ha2, 1hk4, 1if7, 1j16, 1jys, 1k4g, 1kv1, 1kv5, 1m0n, 1m2q, 1nc1, 1ndw, 1ndy, 1nfy, 1nja, 1nje, 1o3f, 1o3p, 1ols, 1olu, 1om1, 1p1q, 1pb9, 1pbq, 1pr5, 1pxo, 1q7a, 1q8t, 1re8, 1s39, 1sqa, 1sv3, 1syh, 1toi, 1toj, 1tok, 1trd, 1ttm, 1u2y, 1utp, 1uwt, 1v16, 1vfn, 1xgj, 1y1m, 1y6q, 1ydt, 1zc9, 1zs0, 1zvx, 2aou, 2aov, 2azr, 2b1v, 2baj, 2bok, 2brb, 2brm, 2bz6, 2ceq, 2cer, 2cet, 2cgr, 2ctc, 2d0k, 2d3u, 2d3z, 2f01, 2fai, 2flb, 2g5u, 2gss, 2hdq, 2j78, 2std, 2usn, 3pce, 3pch, 3pcj, 3std, 4tim, 4tln, 6std
non-lead	10gs, 1a08, 1a1b, 1a30, 1apw, 1b39, 1b7h, 1b9j, 1bma, 1bxo, 1d09, 1det, 1dhi, 1elb, 1f4f, 1f4g, 1fh7, 1fh8, 1fh9, 1fkb, 1fkn, 1fzj, 1fzk, 1g7f, 1g7q, 1gni, 1h23, 1hfs, 1hi4, 1is0, 1j17, 1jaq, 1jq9, 1jqd, 1k9s, 1lol, 1loq, 1m0q, 1mq6, 1ndz, 1nhu, 1nny, 1nvq, 1o0h, 1ppm, 1pz5, 1rnt, 1sl3, 1slg, 1tmn, 1tsy, 1u1b, 1u33, 1v2o, 1vzq, 1x1z, 1xd1, 1zoe, 2ayr, 2b7d, 2bak, 2bzz, 2c02, 2d1o, 2drc, 2er9, 2f80, 2fdp, 2g8r, 2g94, 2h3e, 2i0d, 2j77, 2qwb, 2qwd, 2qwe, 2rkm, 3gss, 4er2, 4tmn, 5er1, 6rnt, 7cpa, 8cpa
excluded	1l2s[1)], 1n2v[1)], 1tyr[2)], 1v48[1)], 2fzc[2)]

1) Structure already part of the training data set. 2) Structure contains atomic clashes.

phosphate groups, occur within a distance of 2.2 Å in the crystal structure. Therefore, the corresponding electron density as retrieved from the Electron Density Server[1] (Kleywegt et al., 2004) was inspected. Figure A.1 (b) displays the $2F_o - F_c$ map at a σ-level of 1.0. No density is visible for the left phosphate group.

This is confirmed by the $F_0 - F_c$ difference map as given in Figure A.1 (c) at a σ-level of 3.0 (green map) and -3.0 (red map), respectively. The respective phosphate is surrounded by negative density, implicating that there is no experimental evidence for this group to be positioned in this

[1] `http://eds.bmc.uu.se/`

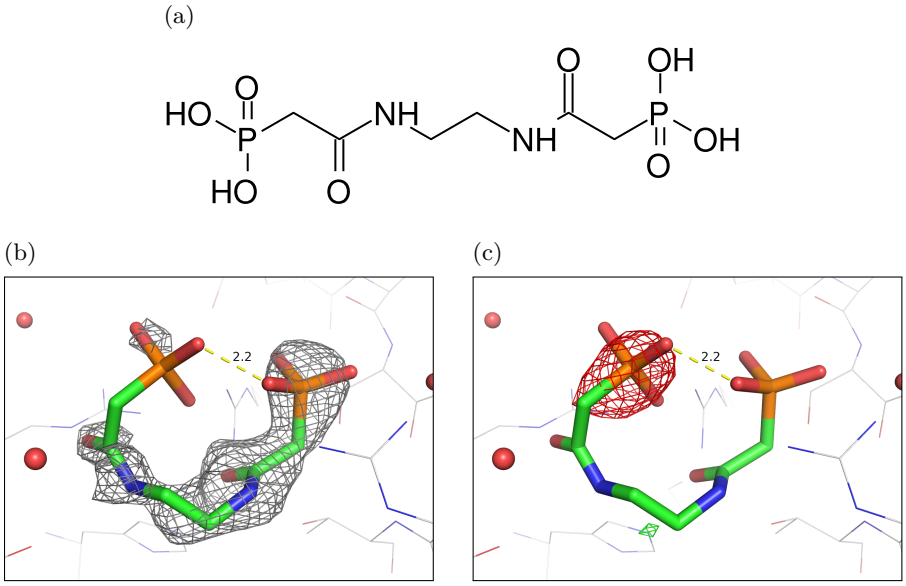


Figure A.1: Structure (a) and electron density maps of the ligand in PDB complex 2fzc: (b) $2F_o - F_c$ map $\sigma = 1.0$, (c) $F_0 - F_c$ map $sigma = -3.0$ (red).

area. This is further reflected in a rather high B-value for this group (70.1 Å^2 for the phosphor atom). Since there is no positive density (green map) observed in this region, it is not clear how to correctly place the phosphate group.

The short atomic distance together with insufficient experimental evidence for such an unusual conformation make this complex ineligible for the validation of a structure optimization method.

In case of PDB complex 1tyr, a clash between ligand and protein was found in the crystal structure. One of the carboxylate oxygens of the ligand, a retinoic acid molecule, falls within 2.0 Å of a protein oxygen atom. This is illustrated in Figure A.2 (a). As this distance is definitely too short for an inter-molecular contact, the structure was checked in more detail. Unfortunately, there is no electron density available for this complex.

Yet, from the PDB structure it was found that the protein atom clashing with the ligand, an oxygen of the C-terminal carboxylate group of the

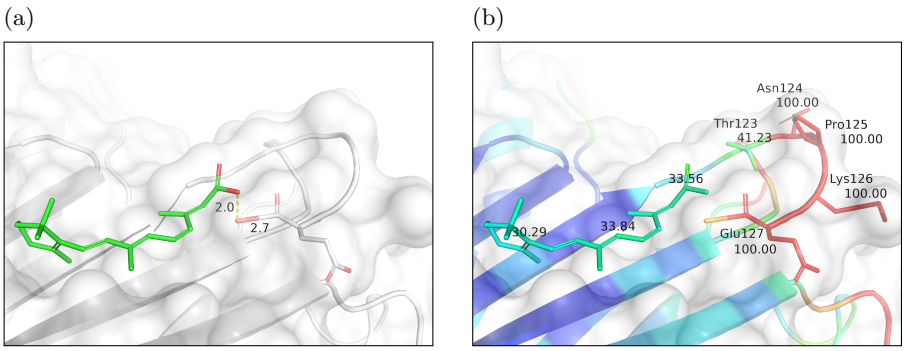


Figure A.2: Binding mode of PDB complex 1tyr; (a) clashing atoms and (b) binding site labeled and colored according to B-values (low B-values: blue; high B-values: red).

respective peptide chain, shows a bond length of 2.7 Å to the neighboring carboxylate carbon. This indicates that the protein structure was not properly minimized during the refinement process.

Figure A.2 (b) shows the questionable part of the protein colored according to B-values. Most of the amino acids below the ligand show values of less than 20 $Å^2$, while the ligand itself reaches about 33 $Å^2$. In contrast, the last four amino acids of the peptide chain (Asn124 – Glu127) turned out to exhibit B-values of 100.0 $Å^2$, suggesting that these residues were not observed in the electron density at all.

Since there is no reliable information available on how much the terminal loop actually affects the binding mode of the retinoic acid, this complex is also excluded from the validation data set.

A.3 Mathematical Definitions

A.3.1 Definition of σ for the Gaussian Smoothing Kernel

Section 2.3 states that, given Equation 2.15, K n-dimensional balls of radius σ have the same volume as a single ball of radius r. Generally, the volume of a n-dimensional ball with radius r is given as

$$V_n = \frac{\pi^{\frac{n}{2}} r^n}{\Gamma\left(\frac{n}{2}+1\right)}. \tag{A.1}$$

This reduces to

$$V_n = C_n r^n \tag{A.2}$$

where C_n is a proportionality factor independent of r. Thus to determine the radius σ needed for K balls to cover the same volume as a single ball with radius r, the following equation has to be solved:

$$\begin{aligned} & K \times C_n\, \sigma^n = C_n\, r^n \\ \Rightarrow \quad & K\, \sigma^n = r^n \\ \Rightarrow \quad & \sigma^n = r^n\, K^{-1} \\ \Rightarrow \quad & \sigma = r\, K^{-1/n}. \end{aligned} \tag{A.3}$$

A.3.2 Correlation as a Measure of Dependence among Observed Variables

Correlations are often used as a measure of a predictive relationship between two observable quantities. There are several correlation coefficients defined which usually yield values between 1 (perfect correlation) and -1 (perfect anti-correlation). A correlation coefficient of 0 then denotes uncorrelated or independent quantities. It is crucial, however, to consider the type of correlation for which a given coefficient is sensitive in order to be able to interpret a calculated correlation value.

Pearson's Correlation Coefficient

The most commonly used measure of correlation is *Pearson's correlation coefficient* ρ which is sensitive only to a linear relationship between two quantities. Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a set of n observations for two quantities X and Y for which a linear relationship can be expected. Then Pearson's correlation coefficient is given as

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{A.4}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of X and Y, and s_x and s_y are the sample standard deviations of X and Y.

Spearman's Rank Correlation Coefficient

If the functional relationship between two quantities is not known, *Spearman's Rank Correlation Coefficient* r_s is often used as an alternative to Pearson's ρ. It assesses how well the relationship can be described using any monotonic function. That is, with increasing X also Y is expected to increase although this has not necessarily to be a linear increase. If Y decreases with increasing X, an anti-correlation can be measured. Thereby, Spearman's Rank Correlation Coefficient is more robust against outliers compared to Pearson's Correlation Coefficient.

Let $(x_1, y_1), \ldots, (x_n, y_n)$ again be a set of n observations for two quantities X and Y. The raw values (x_i, y_i) are converted to the corresponding ranks $(r(x_i), r(y_i))$ among all observations, and the differences $d_i = r(x_i) - r(y_i)$ between the ranks of each observation on the two variables are calculated.

If there are no tied ranks, then Spearman's rank correlation coefficient is given as

$$r_s = 1 - \frac{6\sum_{i=2}^{n} d_i^2}{n(n^2-1)}. \tag{A.5}$$

If the same value exists more than once, then the same rank has to be assigned to all of them. It is defined as the average of their positions among all ordered values. In this case, Spearman's r_s is equivalent to Pearson's ρ calculated on the ranks of the observations:

$$r_s = \frac{\sum_{i=1}^{n} \left(r(x_i) - \bar{r}(x)\right)\left(r(y_i) - \bar{r}(y)\right)}{(n-1)s_{r(x)}s_{r(y)}}. \tag{A.6}$$

Significance of Correlation Coefficients

A correlation is called *statistically significant* if it is unlikely to occur just by chance. This is expressed in the p-value. It gives the probability of observing a correlation coefficient greater than or equal to the one calculated from the measured data, although the measured quantities actually are not correlated (this is called the *null hypothesis*).

A common approach to calculate this probability is to use a t-test. The test statistic

$$t = r\sqrt{\frac{n-2}{1-r^2}}, \tag{A.7}$$

where r is the correlation coefficient and n is the sample size, approximately follows a Student's t-distribution with $n-2$ degrees of freedom if the null hypothesis is supported. The p-value is then given as

$$p = \int_t^\infty f_{n-2}(x)\,dx \tag{A.8}$$

where f_{n-2} is the probability density function of Student's t-distribution with $n-2$ degrees of freedom.

Usually the significance level α is predefined and the null hypothesis is rejected if the calculated p-value is lower than or equal to $1-\alpha$. Critical values of the test statistic t for various confidence levels and degrees of freedom are given in statistical tables. If t is greater than or equal to the critical value at a given confidence level, the null hypothesis can be rejected. This way, the exact p-value does not need to be calculated explicitly.

Bibliography

Bernardetta Addis, Marco Locatelli, and Fabio Schoen. Local optima smoothing for global optimization. *Optim Method Softw*, 20(4):417–437, 2005.

Frank H. Allen. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr*, B58(Pt 3 Pt 1):380–388, Jun 2002.

James Arvo. Fast random rotation matrices. In David Kirk, editor, *Graphics Gems III*, pages 117–120, San Diego, CA, USA, 1992. Academic Press Professional, Inc. ISBN 0-12-409671-9.

Chandrika B-Rao, Jyothi Subramanian, and Somesh D. Sharma. Managing protein flexibility in docking and its applications. *Drug Discovery Today*, 14(7-8):394–400, 2009.

Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Res*, 28(1):235–242, Jan 2000.

Caterina Bissantz, Gerd Folkers, and Didier Rognan. Protein-Based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations. *Journal of Medicinal Chemistry*, 43(25): 4759–4767, 2000.

Bichitra K. Biswal, Meitian Wang, Maia M. Cherney, Laval Chan, Constantin G. Yannopoulos, Darius Bilimoria, Jean Bedard, and Michael N. G. James. Non-nucleoside Inhibitors Binding to Hepatitis C Virus

NS5B Polymerase Reveal a Novel Mechanism of Inhibition. *Journal of Molecular Biology*, 361(1):33–45, 2006.

Hans-Joachim Böhm. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J Comput Aided Mol Des*, 8(3): 243–256, Jun 1994.

M. J. Box. A New Method of Constrained Optimization and a Comparison With Other Methods. *The Computer Journal*, 8(1):42–52, 1965.

Bernard R. Brooks, Robert E. Bruccoleri, Barry D. Olafson, David J. States, S. Swaminathan, and Martin Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4(2):187–217, 1983.

Paul S. Charifson, Joseph J. Corkery, Mark A. Murcko, and W. Patrick Walters. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J Med Chem*, 42(25):5100–5109, Dec 1999.

Tiejun Cheng, Xun Li, Yan Li, Zhihai Liu, and Renxiao Wang. Comparative assessment of scoring functions on a diverse test set. *J Chem Inf Model*, 49(4):1079–1093, Apr 2009.

Matthew Clark, Richard D. Cramer, and Nicole Van Opdenbosch. Validation of the general purpose tripos 5.2 force field. *J Comput Chem*, 10 (8):982–1012, 1989.

Robert D. Clark, Alexander Strizhev, Joseph M. Leonard, James F. Blake, and James B. Matthew. Consensus scoring for ligand/protein interactions. *Journal of Molecular Graphics and Modelling*, 20(4):281–295, 2002.

Jason C. Cole, Christopher W. Murray, J. Willem M. Nissink, Richard D. Taylor, and Robin Taylor. Comparing protein-ligand docking programs is difficult. *Proteins*, 60(3):325–332, Aug 2005.

Jason B. Cross, David C. Thompson, Brajesh K. Rai, J. Christian Baber, Kristi Yi Fan, Yongbo Hu, and Christine Humblet. Comparison of several molecular docking programs: pose prediction and virtual screening accuracy. *J Chem Inf Model*, 49(6):1455–1474, Jun 2009.

Stephanie B. de Beer, Nico P. Vermeulen, and Chris Oostenbrink. The role of water molecules in computational drug design. *Current topics in medicinal chemistry*, 10(1):55–66, 2010.

Marco Dorigo and Thomas Stützle. *Ant Colony Optimization*. The MIT Press, Cambridge, MA, USA, 1st edition, July 2004. ISBN 0262042193.

Hay Dvir, Hua Liang Jiang, Dawn M. Wong, Michal Harel, M. Chetrit, Xu Chang He, Gui Yu Jin, G. L. Yu, X. C. Tang, Israel Silman, Dina L. Bai, and Joel L. Sussman. X-ray Structures of Torpedo californica Acetylcholinesterase Complexed with (+)-Huperzine A and (–)-Huperzine B: Structural Evidence for an Active Site Rearrangement. *Biochemistry*, 41(35):10810–10818, 2002a.

Hay Dvir, Dawn M. Wong, Michal Harel, Xavier Barril, Modesto Orozco, F. Javier Luque, Diego Mu noz Torrero, Pelayo Camps, Terrone L. Rosenberry, Israel Silman, and Joel L. Sussman. 3D Structure of Torpedo californica Acetylcholinesterase Complexed with Huprine X at 2.1 Å Resolution: Kinetic and Molecular Dynamic Correlates. *Biochemistry*, 41(9):2970–2981, 2002b.

Matthew D. Eldridge, Christopher W. Murray, Timothy R. Auton, Gaia V. Paolini, and Roger P. Mee. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J Comput Aided Mol Des*, 11(5):425–445, Sep 1997.

Jon A. Erickson, Mehran Jalaie, Daniel H. Robertson, Richard A. Lewis, and Michal Vieth. Lessons in Molecular Recognition: The Effects of Ligand and Protein Flexibility on Molecular Docking Accuracy. *Journal of Medicinal Chemistry*, 47(1):45–55, 2004.

Miklos Feher. Consensus scoring for protein-ligand interactions. *Drug Discovery Today*, 11(9-10):421–428, May 2006.

Philippe Ferrara, Holger Gohlke, Daniel J. Price, Gerhard Klebe, and Charles L. Brooks. Assessing scoring functions for protein-ligand interactions. *J Med Chem*, 47(12):3032–3047, Jun 2004.

FlexX User Guide. *FlexX Protein-Ligand Docker – User & Technical Reference as Part of LeadIT 1.2*. BioSolveIT GmbH, An der Ziegelei 79, 53757 St. Augustin, Germany, 2010. URL `http://www.biosolveit.de`.

Nicolas Foloppe, Lisa M. Fisher, Rob Howes, Peter Kierstan, Andrew Potter, Alan G. S. Robertson, and Allan E. Surgenor. Structure-Based Design of Novel Chk1 Inhibitors: Insights into Hydrogen Bonding and Protein–Ligand Affinity. *J Med Chem*, 48(13):4332–4345, 2005.

Holger Gohlke, Manfred Hendlich, and Gerhard Klebe. Knowledge-based scoring function to predict protein-ligand interactions. *J Mol Biol*, 295 (2):337–356, Jan 2000.

David E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Professional, Boston, MA, USA, 1st edition, January 1989. ISBN 0201157675.

Michael J. Hartshorn, Marcel L. Verdonk, Gianni Chessari, Suzanne C. Brewerton, Wijnand T. M. Mooij, Paul N. Mortenson, and Christopher W. Murray. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J Med Chem*, 50(4):726–741, Feb 2007.

Manfred Hendlich. Databases for protein-ligand complexes. *Acta Crystallogr*, D54(Pt 6 Pt 1):1178–1182, Nov 1998.

Manfred Hendlich, Andreas Bergner, Judith Günther, and Gerhard Klebe. Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions. *J Mol Biol*, 326(2):607–620, Feb 2003.

Gareth Jones, Peter Willett, Robert C. Glen, Andrew R. Leach, and Robin Taylor. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol*, 267(3):727–748, Apr 1997.

Esther Kellenberger, Jordi Rodrigo, Pascal Muller, and Didier Rognan. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins*, 57(2):225–242, Nov 2004.

Scott Kirkpatrick, C. D. Gelatt, and Mario P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, May 1983.

Gerard Klebe and Thomas Mietzner. A fast and efficient method to generate biologically relevant conformations. *J Comput-Aided Mol Des*, 8 (5):583–606, Oct 1994.

Gerard J. Kleywegt, Mark R. Harris, Jin-yu Zou, Thomas C. Taylor, Anders Wählby, and T. Alwyn Jones. The Uppsala Electron-Density Server. *Acta Crystallographica*, D60(12 Part 1):2240–2249, Dec 2004.

Andrew R. Leach. *Molecular Modelling: Principles and Applications*. Prentice Hall, 2 edition, April 2001. ISBN 0582382106.

Yipin Lu, Renxiao Wang, Chao-Yie Yang, and Shaomeng Wang. Analysis of Ligand-Bound Water Molecules in High-Resolution Crystal Structures of Protein-Ligand Complexes. *Journal of Chemical Information and Modeling*, 47(2):668–675, 2007.

Susan L. McGovern and Brian K. Shoichet. Information Decay in Molecular Docking Screens against Holo, Apo, and Modeled Conformations of Enzymes. *Journal of Medicinal Chemistry*, 46(14):2895–2907, 2003.

Elaine C. Meng, Brian K. Shoichet, and Irwin D. Kuntz. Automated docking with grid-based energy evaluation. *Journal of Computational Chemistry*, 13(4):505–524, 1992.

Wijnand T. M. Mooij and Marcel L. Verdonk. General and targeted statistical potentials for protein-ligand interactions. *Proteins*, 61(2):272–287, Nov 2005.

Jorge J. Moré and Zhijun Wu. Smoothing Techniques for Macromolecular Global Optimization. Technical Report MCS-P542-0995, Argonne National Laboratory, September 1995.

Ingo Muegge. PMF scoring revisited. *J Med Chem*, 49(20):5895–5902, Oct 2006.

Ingo Muegge and Yvonne C. Martin. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J Med Chem*, 42(5):791–804, Mar 1999.

Rafael Najmanovich, Josef Kuttner, Vladimir Sobolev, and Marvin Edelman. Side-chain flexibility in proteins upon ligand binding. *Proteins: Structure, Function, and Bioinformatics*, 39(3):261–268, 2000.

John A. Nelder and Roger Mead. A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313, 1965.

Gerd Neudert and Gerhard Klebe. fconv: format conversion, manipulation and feature computation of molecular data. *Bioinformatics*, 27(7):1021–1022, 2011.

Noel M. O'Boyle, John W. Liebeschuetz, and Jason C. Cole. Testing assumptions and hypotheses for rescoring success in protein-ligand docking. *J Chem Inf Model*, 49(8):1871–1878, Aug 2009.

Akifumi Oda, Keiichi Tsuchida, Tadakazu Takakura, Noriyuki Yamaotsu, and Shuichi Hirono. Comparison of Consensus Scoring Strategies for Evaluating Computational Models of Protein-Ligand Complexes. *Journal of Chemical Information and Modeling*, 46(1):380–391, 2006.

Tudor I. Oprea. Property distribution of drug-related chemical databases. *J Comput-Aided Mol Des*, 14(3):251–264, Mar 2000.

Michael J. D. Powell. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The Computer Journal*, 7(2):155–162, 1964.

William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, 3 edition, September 2007. ISBN 0521880688.

Matthias Rarey, Bernd Kramer, Thomas Lengauer, and Gerard Klebe. A fast flexible docking method using an incremental construction algorithm. *J Mol Biol*, 261(3):470–489, Aug 1996.

Benjamin C. Roberts and Ricardo L. Mancera. Ligand-Protein Docking with Water Molecules. *Journal of Chemical Information and Modeling*, 48(2):397–408, 2008.

Johannes Schulze Wischeler, Dong Sun, Nicola U. Sandner, Uwe Linne, Andreas Heine, Ulrich Koert, and Gerhard Klebe. Stereo- and Regioselective Azide/Alkyne Cycloadditions in Carbonic Anhydrase II via Tethering, Monitored by Crystallography and Mass Spectrometry. *Chemistry – A European Journal*, 17(21):5842–5851, 2011.

Sérgio Filipe Sousa, Pedro Alexandrino Fernandes, and Maria João Ramos. Protein-ligand docking: Current status and future challenges. *Proteins: Structure, Function, and Bioinformatics*, 65(1):15–26, 2006.

Martin Stahl. Modifications of the scoring function in FlexX for virtual screening applications. *Perspectives in Drug Discovery and Design*, 20: 83–98, 2000.

Chung-Jung Tsai, Sandeep Kumar, Buyong Ma, and Ruth Nussinov. Folding funnels, binding funnels, and protein function. *Protein Sci*, 8(6): 1181–1190, Jun 1999.

Hans F. G. Velec. *Neue Methoden in der Computerchemie zur Bewertung und Optimierung von Protein-Ligand-Komplexen*. PhD thesis, Philipps-Universität Marburg, 2008.

Hans F. G. Velec, Holger Gohlke, and Gerhard Klebe. DrugScore(CSD)–knowledge-based scoring function derived from small molecule crystal

data with superior recognition rate of near-native ligand poses and better affinity prediction. *J Med Chem*, 48(20):6296–6303, Oct 2005.

Marcel L. Verdonk, Jason C. Cole, Michael J. Hartshorn, Christopher W. Murray, and Richard D. Taylor. Improved protein-ligand docking using GOLD. *Proteins*, 52(4):609–623, Sep 2003.

Anders Wallqvist and David G. Covell. Docking enzyme-inhibitor complexes using a preference-based free-energy surface. *Proteins: Structure, Function, and Bioinformatics*, 25(4):403–419, 1996.

Renxiao Wang, Yipin Lu, and Shaomeng Wang. Comparative evaluation of 11 scoring functions for molecular docking. *J Med Chem*, 46(12): 2287–2303, Jun 2003.

Gregory L. Warren, C. Webster Andrews, Anna-Maria Capelli, Brian Clarke, Judith LaLonde, Millard H. Lambert, Mika Lindvall, Neysa Nevins, Simon F. Semus, Stefan Senger, Giovanna Tedesco, Ian D. Wall, James M. Woolven, Catherine E. Peishoff, and Martha S. Head. A critical assessment of docking programs and scoring functions. *J Med Chem*, 49(20):5912–5931, Oct 2006.

Scott J. Weiner, Peter A. Kollman, David A. Case, U. Chandra Singh, Caterina Ghio, Guliano Alagona, Salvatore Profeta, and Paul Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society*, 106(3):765–784, 1984.

Danksagung

Mein besonderer Dank gilt *Herrn Prof. Dr. Gerhard Klebe* für die Möglichkeit und das Vertrauen, diese Arbeit in seiner Arbeitsgruppe anzufertigen. Er gewährte große Freiheiten bei der Ausgestaltung des Themas und nahm sich stets Zeit für Diskussionen, Unterstützung und Hilfestellung.

Herrn Thomas Mietzner danke ich für die große Hilfsbereitschaft bei der Einführung in das Programm Torso und die geduldige Beantwortung zahlreicher Fragen.

Ebenso bedanke ich mich bei *Gerd Neudert* für die umfassende Hilfe und Unterstützung bei allen Fragen zu DrugScore und fconv. Auch meine vielen Wünsche nach zusätzlichen Funktionen wurde immer erfüllt.

Ich danke allen Anwendern von MiniMuDS in der Arbeitsgruppe, insbesondere *Tobias Craan*, für die zahlreichen wertvollen Rückmeldungen und Verbesserungsvorschläge.

Bei *Frau Lydia Hartleben* bedanke ich mich für die Unterstützung in allen organisatorischen und „bürokratischen“ Angelegenheiten.

Bei allen Kollegen der Arbeitsgruppe Klebe bedanke ich mich für die gute Zusammenarbeit und angenehme Arbeitsatmosphäre. Besonderer Dank gilt hier *Patrick, Tina, Jürgen, Johannes, Tobias, Florian, Michael* und *Chris* für vier wunderbare Jahre in Marburg.

Besonders herzlich möchte ich *meinen Eltern, Geschwistern* und *Lucia* danken für ihr unerschütterliches Vertauen in mich, ihren Rückhalt und ihre Unterstützung, die ich immer erfahren habe.

Erklärung

Ich versichere, dass ich meine Dissertation

„Knowledge-based Optimization of Protein-Ligand-Complex Geometries“

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

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

Marburg, den

.....

(Andreas Spitzmüller)

Lebenslauf

Persönliche Daten

Name	Andreas Spitzmüller
Geburtsdatum	22. Februar 1981
Geburtsort	Leonberg
Nationalität	Deutsch

Bildungsweg

1987 – 1991	Grundschule Bildungszentrum St. Konrad, Ravensburg
1991 – 2000	Gymnasium Bildungszentrum St. Konrad, Ravensburg
Juni 2000	**Abitur**
2000 – 2001	Studium der Bioinformatik, Eberhard-Karls-Universität Tübingen
2001 – 2004	Studium der Bioinformatik, Technische Universität München und Ludwig-Maximilians Universität München
Juni 2004	**Bachelor of Science (B.Sc)**
2004 – 2006	Studium der Bioinformatik, Technische Universität München und Ludwig-Maximilians Universität München
Mai 2006	**Master of Science (M.Sc)**
2007 – 2011	Promotion am Institut für Pharmazeutische Chemie, Philipps-Universität Marburg, Arbeitskreis Prof. Dr. Gerhard Klebe