

# **Application of coarse grained force fields for protein-protein docking**

Inaugural-Dissertation

zur Erlangung des Doktorgrades  
der Mathematisch-Naturwissenschaftlichen Fakultät  
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

**Philipp Kynast**  
aus Jena

Jülich, Mai 2016

aus dem Institute of Complex Systems (ICS-6): Strukturbiochemie  
des Forschungszentrums Jülich

Gedruckt mit der Genehmigung der  
Mathematisch-Naturwissenschaftlichen Fakultät der  
Heinrich-Heine-Universität Düsseldorf

Referent: Jun.-Prof. Dr. Birgit Strodel

Korreferent: Prof. Dr. Holger Gohlke

Tag der mündlichen Prüfung: 14. Juni 2016

# Erklärung

Ich versichere an Eides Statt, dass die Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich Heine Universität Düsseldorf“ erstellt worden ist. Die Arbeit wurde an keiner anderen Fakultät vorgelegt.



# Zusammenfassung

Protein-Protein-Interaktionen sind an beinahe jedem Prozess in lebenden Organismen beteiligt, zum Beispiel an der intrazellulären Kommunikation, der Immunantwort und der Enzymregulation. Es existieren einige experimentelle Methoden um Interaktionen von Proteinen zu bestimmen. Doch nicht alle diese Methoden können die Interaktionen auf atomarer Ebene aufklären. Die Aufklärung der Struktur von Protein-Protein-Komplexen ist relativ schwierig. Protein-Protein-Komplexe sind oft zu instabil für Röntgenkristallographie und zu gross für Kernspinresonanzspektroskopie.

Protein-Protein-Docking ist die Methode um die Struktur von Protein-Protein-Komplexen *in silico* aus den jeweiligen Strukturen der gefalteten Bindungspartner vorherzusagen. Diese Technik bietet so die Möglichkeit die Lücke der fehlenden Protein-Protein-Komplexe zu schliessen. Doch auch diese Technik hat ihre Herausforderungen, denn Proteine können theoretisch auf viele Arten miteinander interagieren. All diese Bindungsposen müssen produziert und nach ihrer Wahrscheinlichkeit bewertet werden. Wegen der Komplexität dieser Suche erlauben aktuelle Methoden keine vollständige Flexibilität der Proteine. Diese Arbeit untersucht die Möglichkeit der Benutzung von „coarse grained“-Kraftfeldern um Protein-Protein-Interaktionen vorherzusagen. Solche Kraftfelder könnten eine gute Wahl für die Simulation von grösseren Systemen sein, ohne die Flexibilität der Proteine einzuschränken.

Zuerst testen wir ein Kraftfeld welches von Bereau und Deserno entwickelt wurde. Konformationen die mit der grid-basierten Methode ZDOCK produziert werden, werden in die „coarse grained“-Konformation der Proteine transformiert. Ihre Energie wird minimiert und dann werden sie mit dem Kraftfeld von Bereau und Deserno bewertet. Es wurde herausgefunden, dass die Seitenketten-Darstellung des Kraftfeldes nicht ausreicht um erfolgreich zwischen Strukturen die nahe der gebundenen Komplexstruktur sind und welchen die falsch sind zu diskriminieren.

Deshalb wurde das „coarse grained“-Kraftfeld OPEP mit einer besseren Beschreibung der Seitenketten für den Rest der Arbeit benutzt. Wieder wurden die initialen Strukturen mit ZDOCK produziert und dann mit OPEP weiter

verarbeitet. Kleine Änderungen wurden am OPEP-Kraftfeld unternommen, es wurde ein weiches Potential für die Interaktionen zwischen den Seitenketten eingeführt und damit eine bessere Unterscheidung zwischen Strukturen nahe der nativen Lösung und falschen Strukturen, im Vergleich zum originalen OPEP, erreicht. Diese weiche Bewertungsfunktion wurde zusätzlich durch einen iterativen Lerner trainiert. Die sich daraus ergebene Bewertungsfunktion wurde auf einem weiteren Datensatz getestet und lieferte bessere Ergebnisse als die originale Bewertungsfunktion.

Schließlich wurde das OPEP-Kraftfeld dazu benutzt um Monte Carlo Simulationen, welche die Proteine als starre Körper behandelt, mit Strukturen in der Umgebung des gebundenen Komplexes durchzuführen. Mit dieser Methode ist es möglich Köderstrukturen nahe des nativen Komplexes zu erzeugen, welche eine gute Bewertung durch die weiche OPEP-basierte Bewertungsfunktion erhalten. Durch die Verwendung einer Starre-Körper-Minimierung können diese Köderstrukturen noch näher an die native Konformation gebracht werden. Clustering zeigte, dass 74 % der Köderstrukturen nahe des nativen Komplexes dem grösstem Cluster angehören, das heisst die OPEP-basierte Bewertungsfunktion ist in der Lage dem zum nativem Komplex führenden Energy-Trichter zu identifizieren.

In dieser Arbeit wurde das OPEP-Kraftfeld das erste Mal auf seine Anwendbarkeit für Protein-Protein-Docking getestet. Die Ergebnisse zeigen, dass das OPEP-Kraftfeld, mit kleinen Änderungen, eine vielversprechende neue Methode für dieses Problem ist.

# Abstract

Protein-protein interactions are involved in nearly every process in living organisms, such as intracellular signaling, immune response and enzyme regulation. Several experimental methods exist that allow to determine the interaction of proteins. However, not all of these methods reveal the interactions on an atomic level. Moreover, determining the structure of protein-protein complexes is an experimental challenging task. Protein-protein complexes are often too transient for X-Ray crystallography and too big for nuclear magnetic resonance.

Protein-protein docking is the method to predict the interaction of protein-protein complexes *in silico* from the individual structures of the folded binding partners. Thus, this technique offers the possibility to fill the gap of missing protein-protein complexes. Nonetheless, also this technique comes with its own challenges given that, in theory, proteins could bind to each other in different ways. All of these different binding poses have to be produced and scored for their likelihood. Because of the complexity of this search, current methods for global search do not allow the proteins to be flexible. This work examines the possibility to use coarse grained force fields to predict protein-protein interactions. Such force fields may be a good choice for simulating bigger systems without restricting the flexibility of the proteins.

First the force field developed by Bereau and Deserno is tested for this purpose. The initial decoys are produced with ZDOCK, that provides a grid representation of the proteins, which are transformed to the coarse grained presentation of the proteins. Their energy is minimized and then scored using the force field by Bereau and Deserno. However, it is found that the side chain description in this force field is not sufficient to allow a reliable discrimination between near-native and misdocked decoys.

Therefore, the coarse-grained OPEP force field with a better amino acid dependent representation of the side chain is used in the remainder of this work. Again, the initial decoys are produced with ZDOCK and then further treated with OPEP. Slight modifications to the OPEP force field are made, revealing a softer potential for the interactions between the side chain beads and therefore leading to a better discrimination between conformations close to the bound

conformation and mis-docked conformations compared to the original OPEP force field. The soft rescoring function was further trained by an iterative learning procedure. The resulting function was then tested on an independent data set, revealing that this trained rescoring function performs better than the original rescoring function.

Finally, the OPEP force field is used to perform Monte Carlo simulations treating the proteins as rigid bodies to produce docking decoys in the vicinity of the native complex. With this approach it is possible to produce more near-native decoys, which obtain a good score with the soft OPEP-based rescoring function. Using rigid-body minimization, these decoys can be brought even closer to the native complex structure. Clustering revealed that 74 % of the near-native decoys are members of the biggest cluster, i.e., that the OPEP based scoring function is able to identify the energy funnel leading to the native conformation.

In summary, in this work OPEP was tested for the very first time for its applicability to protein-protein docking. The results show that a slight modification of OPEP is indeed a promising new rescoring function for this problem.

# Acknowledgement

First I want to thank Jun.-Prof. Dr. Birgit Strodel for giving me the opportunity to join her Computational Biochemistry Group and supporting me, being open for discussions and providing me useful advice throughout my Ph.D. studies. Thank you Birgit! Thanks to Prof. Dr. Holger Gohlke and his group for giving me the opportunity to present my work and for fruitful discussions. I want to thank Falk for explaining to me physical concepts. I want to thank Gunnar for his advice and helpful discussions over lunch and for always being willing to answer all of my questions no matter how stupid. I want to thank Dusan for the discussions and nice conversations about plotting data. I want to thank Michael Owen for teaching me that glycine in a peptide chain is not called glycine. I want to thank Qinghua for helping me with gromacs and OPLS. I want to thank Bogdan for sharing his Tcl and force field knowledge with me. I want to thank my colleagues: Marianne, Martin, Matthias, Jide, Ken, Chetan, Oliver, André, Tatjana, Michaela, Amudha, Dennis and Kateryna for the discussions and nice company over the years.

I also want to thank my friends and family: Bettina, Jürgen, Clemens, Claudia, Adele, Marga and especially Hannah. My friends in Jülich: Volkmar, Viktor, Stephan, Matthias, Maren, Kerstin, Aleksey, Dima, Kevin, Keyvan, Kristina, Peter and Vineet. You have all made my time in Jülich unforgettable.







# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Proteins . . . . .	1
1.1.1	Building blocks of proteins . . . . .	1
1.1.2	Structure of Proteins . . . . .	2
1.1.3	Protein-protein interactions . . . . .	5
1.2	Force fields . . . . .	6
1.3	Coarse grained force fields . . . . .	7
1.4	Protein-protein docking . . . . .	8
1.5	Aims . . . . .	10
<b>2</b>	<b>A generic coarse grained force field</b>	<b>13</b>
2.1	Introduction . . . . .	13
2.2	Methods . . . . .	13
2.2.1	ZDOCK . . . . .	13
2.2.2	Energy minimization . . . . .	14
2.2.3	The coarse grained force field . . . . .	15
2.2.4	The test set . . . . .	20
2.2.5	Metric for success . . . . .	21
2.2.6	Different side chain parametrization . . . . .	21
2.3	Results . . . . .	23
2.3.1	Properties of the test set . . . . .	23
2.3.2	Docking with ZDOCK . . . . .	23
2.3.3	Rescoring of the ZDOCK predictions with the original BD force field . . . . .	27
2.3.4	Changing to the full Miyazawa-Jernigan matrix . . . . .	32
2.3.5	Replacing the sizes and positions of the side chains . . . . .	34
2.4	Discussion . . . . .	39
<b>3</b>	<b>OPEP for protein-protein docking</b>	<b>41</b>

<b>4</b>	<b>MC refinement of docking predictions</b>	<b>61</b>
4.1	Introduction . . . . .	61
4.2	Methods . . . . .	62
4.2.1	Data set . . . . .	62
4.2.2	Generating random structure . . . . .	62
4.2.3	Monte Carlo simulations . . . . .	63
4.2.4	Rigid body minimization . . . . .	63
4.2.5	Local minimization . . . . .	65
4.2.6	The force field . . . . .	65
4.2.7	Clustering . . . . .	65
4.3	Results . . . . .	66
4.3.1	Conformational changes . . . . .	66
4.3.2	Ranking by the energy . . . . .	75
4.3.3	Energy funnels and clustering . . . . .	78
4.4	Discussion . . . . .	88
4.5	Comparison to ZDOCK . . . . .	90
<b>5</b>	<b>Conclusion</b>	<b>95</b>

# Chapter 1

## Introduction

### 1.1 Proteins

Proteins are of utmost importance for life, they are the molecular machines in organisms that are important for movement, for information transport, and regulation of all cell functions. Proteins are polymer molecules and are constructed from amino acids which are connected in a chain. The 20 canonical amino acids are directly encoded in the DNA. This information is heritable and the sequence of triplets in the DNA defines the sequence of amino acids in a protein. The sequence of the triplets in the DNA define the sequence of the amino acids of the protein. With today's technology for the determination of the DNA sequence of an organism is feasible and relatively simple, but it is still hard to determine the structure of the proteins from this sequence. However, it is important to know the three dimensional structure of a protein, as the structure determines the function of a protein and provides insights into the possible regulation sites for medication. Experimental methods for structure determination are nuclear magnetic resonance (NMR) spectroscopy, X-Ray crystallography and cryo-electron microscopy. Hundreds of thousands of protein structures are stored in the Protein Data Bank (PDB) [1]. But there are still thousands of proteins with unknown structure, and especially the Structures of protein-protein complexes are often difficult to determine experimentally.

#### 1.1.1 Building blocks of proteins

Proteins are made from amino acids. Each amino acid has the same backbone structure consisting of  $N, C_{\alpha}, C$ , the nitrogen hydrogen and the carbon oxygen atoms. There are 20 amino acids that are directly encoded in the DNA. These amino acids have different side chains which lead to different physico-chemical behavior and also the occupied space is different for each amino acid. The natural

occurring amino acids are listed in table 1.1. In general, amino acids can be assigned to three general groups, hydrophobic amino acids, polar amino acids and charged amino acids. Hydrophobic amino acids are mostly buried in the inside of proteins. The reason is that their interaction with water is unfavoured because of their uncharged and unpolarizable side chains, which do not interact electrostatically with the water molecules, disturbing the H-Bond pattern of the water. Polar amino acids can form H-Bonds with the water and therefore these side chains are more likely found at the surface of a protein. Charged amino acids are also more likely to be located at the surface of the protein, but there amino acids can also form salt bridges with oppositely charged amino acids. Cysteine is a hydrophobic amino acid but it can additionally form sulphur bridges with other cysteines; leading to very strong because covalent interactions. Proline has a ring as side chain that also contains the nitrogen of the backbone. As a result, this nitrogen cannot form H-bonds so that the presence of proline disturbs secondary structure formation [2].

### 1.1.2 Structure of Proteins

The by the genes defined sequence of the amino acids is called primary structure, which can form secondary structures like  $\alpha$ -helices and  $\beta$ -sheets. The three dimensional structure of the whole protein is the tertiary structure and the quaternary structure contains an ensemble of protein chains. The predictions and examination of quaternary structures forming from tertiary structure is the subject of this work. In figure 1.1 the formation of a peptide bond is shown.

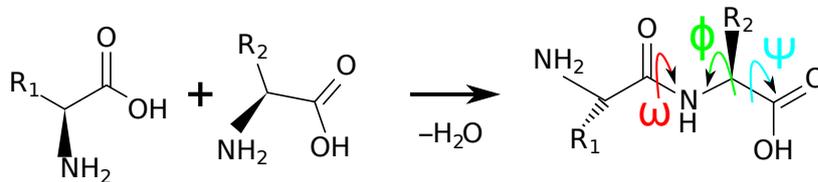


Figure 1.1: The peptide bond and the backbone dihedral angles.

Rotation around the C-N bond is (the torsional angle  $\omega$ ) is hindered, locking the peptide bond either in the cis ( $\omega = 0^\circ$ ) or trans ( $\omega = 180^\circ$ ) conformation. In either case the atoms  $C_\alpha$ , C and O from one amino acid and N, amino H and  $C_\alpha$  of the second amino acid are all co-planar. The trans conformation is the dominant conformation in proteins [3]. An exception is proline where the peptide bond is more flexible and the cis conformation is also frequently ( $\approx 5\%$ ) adapted [3].

The dihedral angles  $\phi$  and  $\psi$  are more flexible and define the secondary structure of a protein. The main secondary structure conformations are  $\alpha$ -helix

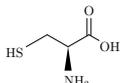
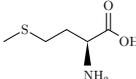
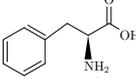
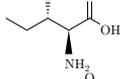
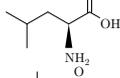
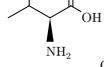
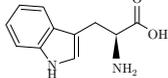
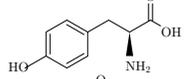
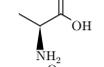
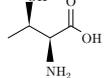
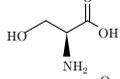
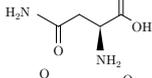
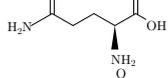
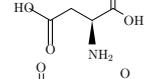
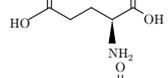
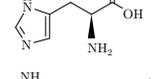
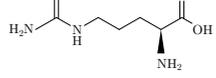
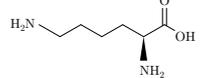
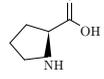
Cystein		Can form sulphur bridges with other cysteins, these are quite strong (up to 60 kcal/mol).
Methionine		hydrophobic amino acid.
Phenylalanine		hydrophobic amino acid.
Isoleucine		hydrophobic amino acid.
Leucine		hydrophobic amino acid.
Valine		hydrophobic amino acid.
Tryptophane		hydrophobic amino acid.
Tyrosine		hydrophobic amino acid.
Alanine		hydrophobic amino acid.
Glycine		glycine has no side chain.
Threonine		The OH-group makes the side chain polar.
Serine		The OH-group makes the side chain polar.
Asparagine		The amide group makes the side chain polar.
Glutamine		The amide group makes the side chain polar.
Aspartic acid		The carboxyl group makes the side chain acidic.
Glutamic acid		The carboxyl group makes the side chain acidic.
Histidine		The ring can be protonated, making the side chain basic.
Arginine		The guanidine group can be protonated, making the side chain basic.
Lysine		The amine group makes the side chain basic.
Proline		Hydrophobic amino acid. The backbone N can not form H-Bonds.

Table 1.1: The 20 canonical amino acids.

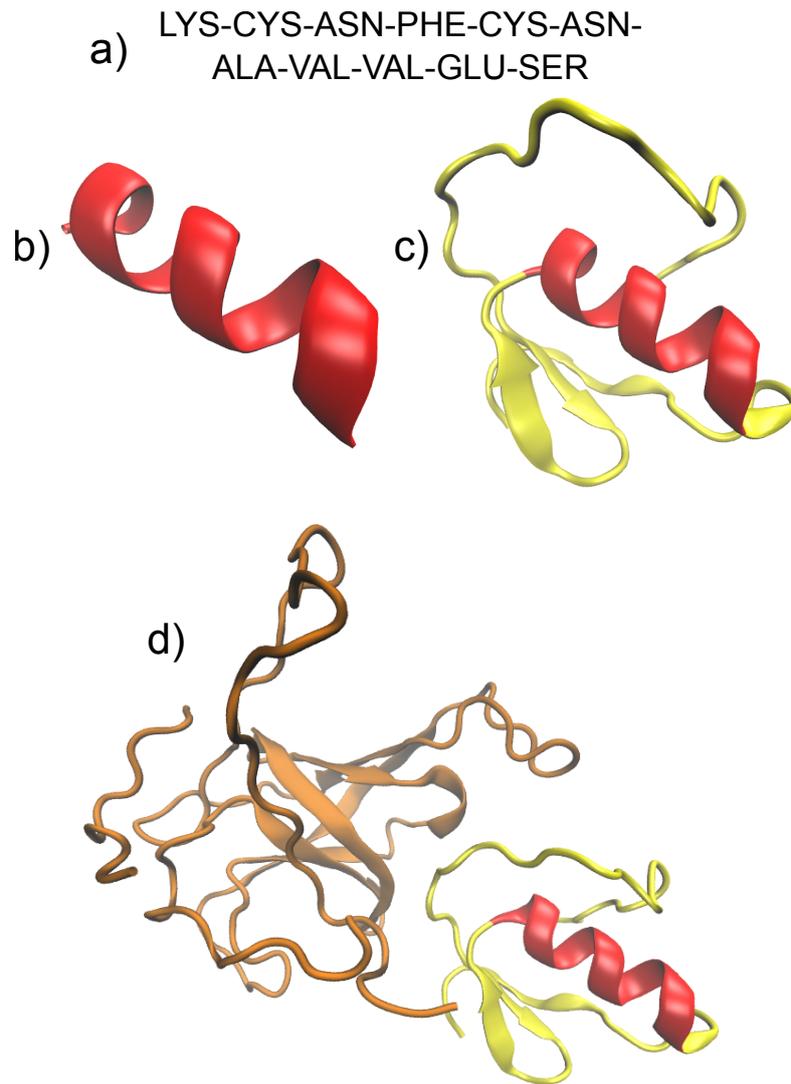


Figure 1.2: Protein structure: The amino acid sequence is the primary structure (a). This sequence forms a helix as secondary structure (b), the protein fold is called tertiary structure (c). The complex structure of proteins is called quaternary structure (d).

and  $\beta$ -sheet. These secondary structure elements form further contacts with each other, leading to the tertiary structure. These folded proteins can interact with each other and form the quaternary structure (see figure 1.2). The interactions between proteins are important for the regulation and communication between the cellular entities. However, there are also protein-protein complexes that only exist as homo-oligomers and only function in this conformation. These

interactions are normally quite strong and are not transient. Here, we are only interested in the prediction of hetero-oligomers where the binding partners are different proteins and the binding is not obligatory.

### 1.1.3 Protein-protein interactions

Protein-protein interactions can involve two or more proteins that interact with each other. A prominent example of protein-protein interaction is the regulation of the blood sugar level by insulin. Here, the protein insulin binds to a receptor which is a protein in the membrane of the cell allowing the uptake of glucose by the cell which leads to a reduction of the glucose concentration in the blood. Protein-protein interactions are typically divided into classes, namely enzyme/inhibitor, antigen/antibody and ‘other’ complexes [4], where the preceding insulin example is part of the ‘other’ complexes. Enzymes catalyze many functions in the cell. Their functions include breaking bonds, transferring functional groups or synthesizing reagents to bigger products. All these functions themselves are not protein-protein interactions, but the regulation of proteins is. Enzymes are often regulated by inhibitors, where another protein, the inhibitor, binds to the enzyme and thereby lower the speed of the reaction. The inhibitor could compete with the substrate by binding in the same pocket like the substrate, or the inhibitor could bind somewhere else on the surface of the enzyme leading to conformational changes that prohibit the binding of the substrate [5]. These processes can be reversible and irreversible. In the reversible case the speed of the reaction decreases by increasing concentration of the inhibitor but it never reaches zero, if the inhibition is irreversible the speed also decrease with increasing inhibitor concentration, but here the activity can reach zero for some inhibitors [5]. The interaction between the enzyme and the inhibitor are mostly interactions between hydrophobic side chains and the backbone does not move much during the binding process [6].

Antigen/antibody interactions are protein-protein interactions between antibodies, that are produced by the immune system, with exposed protein structures of a bacterium, virus or parasite. Antibodies may also detect sugar complexes, RNA and DNA fragments as an exposed part of a virus or bacterium, named antigen. The antibody consists of a scaffold with a light and a heavy chain, and only a part of the antibody is flexible and adopts for binding to the current antigen [7]. Because of the adoptive antibodies, the interaction properties can be very different, but in general the interfaces are not as hydrophobic like for enzyme/inhibitor complexes [6]. Once the antibody is bound to the intruder, it can be removed by the macrophage.

The category ‘other’ complexes contains all the other complexes that are neither enzyme/inhibitors nor antigen/antibody complexes. It includes, for example, a complex that is involved in removing damaged actin filaments. As

the name ‘other’ already implies, the complexes in this category do not have so much in common like complexes in the antigen/antibody or enzyme/inhibitor complex class.

## 1.2 Force fields

For simulating proteins the most accurate description would be with quantum chemical calculations. However, these calculations are computationally demanding and thus not suitable for simulating large molecular ensembles. All-atom force fields describe the interactions between the atoms with classical mechanics and they were started to be used for modeling proteins in the beginning of 1980s [8]. Their predecessors were force fields for chemical compounds and preliminary work on molecular mechanics was already done [9]. The first protein force fields were potentials like CHARMM published in 1983 [8] and Amber from 1984 [10]. The initial force fields used a united atom description, where non-polar hydrogens are neglected and instead are modeled as bigger atom together with the heavy atom they are bound to. Different parametrizations for force fields exists but generally the potentials are similar and dependent on the Cartesian positions  $\vec{r}$  of the atoms:

$$V(\vec{r}) = V_{\text{bonded}}(\vec{r}) + V_{\text{non-bonded}}(\vec{r}) \quad (1.1)$$

$$V_{\text{bonded}}(\vec{r}) = \sum_{\text{bonds}} k_b(b - b_0)^2 + \sum_{\text{angles}} k_\theta(\theta - \theta_0)^2 \quad (1.2)$$

$$+ \sum_{\text{dihedrals}} k_\psi[1 - \cos(n\psi - \psi_{n,0})] + \sum_{\text{impropers}} k_\omega(\omega - \omega^0)^2$$

$$V_{\text{non-bonded}}(\vec{r}) = \sum_{\substack{\text{nonbonded atom pairs} \\ i,j}} \frac{q_i q_j}{r_{ij}} + 4\epsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 \right]. \quad (1.3)$$

Here,  $V_{\text{bonded}}(\vec{r})$  consists of a harmonic potential for covalently bonded atoms, a harmonic potential for angles described by three consecutively bonded atoms, a dihedral potential describing rotation around bonds, and another harmonic potential for out-of-plane bending.  $V_{\text{non-bonded}}(\vec{r})$  describes electrostatic and van der Waals interactions between atoms. The van der Waals interactions are described by a Lennard-Jones-potential, for which the parameters  $\sigma_{ij}$  can be obtained from measurements of crystal structures of the elements. The electrostatic parameters are very crucial for the performance of the actual force field. The electron density partial charges are assigned to the centers of mass of the atoms. These parameters are calculated by quantum chemical calculations and the charges are slightly different from force field to force field. Parameters

for the bond lengths and bond angles can be taken from crystal structures of proteins. The dihedral angles describe the rotation around the bonds, and are typically very flexible, so that proteins can adopt a wide range of different three dimensional conformations. The dihedral parameters are often chosen in the last step during the parametrization of the force field, ensuring that the force field can reproduce the typical conformations of proteins in the molecular dynamics (MD) simulations [11]. Force fields are used together with Monte Carlo (MC) and MD simulations, or just for energy minimization. A further approximation to all-atom force fields are coarse grained force fields. They are the subject in the next section.

### 1.3 Coarse grained force fields

Coarse grained force fields unite atoms into larger beads. This leads to a computational speedup for several reasons. The first and most obvious reason is that the number of atoms is reduced. This leads to a faster calculation of the potential energy. In addition, in an MD simulation small bond vibrations are omitted, which is also frequently done for all-atom simulations by constraining the bond length. The absence of small but fast vibrations allows bigger integration steps. The third effect is the reduction of friction, which normally changes the time scale so that configuration changes happen faster [12].

The first coarse grained force field for proteins was developed by Levitt and Warshel [13, 14]. They used a two bead representation per amino acid, one bead for the side chain and one bead for the backbone at the  $C_\alpha$  position. The side chains were the only interaction partners for the non-bonded interactions and their hydrophobicity was parametrized with experimental data. At this time, 1975, computational resources were so limited that an all-atom representation was out of range, but with the simplified representation they could fold a protein of 58 residues from an extended conformation close to the native conformation using cycles of energy minimization followed by a normal mode analysis and a structure perturbation along the lowest normal mode [13]. A similar coarse grained representation is used in UNRES [15] which was introduced in 1993. Each amino acid is described by an united peptide group and one bead for the side chain. The virtual  $C_\alpha$  atom is used to calculate the side chain position and the conformation of the backbone, but does not serve as interaction site. The force field is parametrized by a potential of mean force derived from all-atom MD simulations. The backbone parameters were derived by fitting the model against the results from quantum chemical calculations of different conformations. The parameters involving the angle between peptide groups and the orientation of the side chain are fitted to distribution values calculated from protein structures

in the PDB [1]. Both coarse grained force fields mentioned thus far use implicit water and only one side chain bead. A more recently developed coarse grained model is the MARTINI force field. Here more beads are used per amino acid and it also includes a coarse grained water model. The potential was originally developed for lipids [16] and was later extended to proteins [12]. The force field uses one bead for the backbone and the number of beads for the side chains depends on the amino acid. For example, alanine and glycine are each represented by one backbone bead and have no side chain beads, cysteine has two beads in total and tryptophane is modeled with five beads. The backbone parametrization depends on the initial secondary structure and the secondary structure can not change during the simulation, but a movement of secondary structures relative to each other is possible. The non-bonded interactions are parametrized by fitting the values to experimental free energies of transferring amino acids from oil phase to water. The MARTINI water model is based on a four-to-one mapping of water molecules and later this model was further developed to a polarizable water model including two extra charged particles [17].

## 1.4 Protein-protein docking

Protein-protein docking is the process of *in silico* protein-protein complex prediction. The prediction method takes the unbound binding partners in their native conformation as input. These structures could be from computer simulations, but mostly protein crystal structures are used. The main problem is the size of the protein-protein complexes, which are too big to use all-atom force fields and simulation techniques like MD simulations for the prediction of the complex structure. Connolly provided a nice description of the complexity of this problem by considering the number of possible solutions [18]. He proposed to imagine that a protein can be approximately described as a sphere with a radius of 15 Å. A distribution of points on this surface with a spacing of 1 Å would lead to  $\approx 3000$  points on this sphere. These 3000 points have to be matched with the points of the binding partner, which leads to  $3000 \times 3000 = 9 \times 10^6$  possible solutions. Moreover, the proteins can rotate around the vector connecting their centers of mass. Connolly took  $3^\circ$  as spacing value, which would lead to a search space of  $9 \times 10^6 \times 120 \approx 10^9$  unique solutions [18]. Under the optimistic assumption that one could test 1000 predictions per second, scanning of the whole search space would take eleven days for these quite small proteins. The previous thought experiment shows that there is a need for specialized methods that can tackle this problem. The first computational study of protein-protein interaction was the study of fiber formation of sickle cell haemoglobin [19]. In this study, the crystal structure of the wild type of haemoglobin was taken and

under the assumption that the point mutation leading to sickle cell anaemia does not change the conformation, several fiber models were produced in a guided way. These models were manually examined for their physico-chemical interactions, and two different models of the fiber formation were proposed. Unfortunately, both of these models are wrong. Wodak and Janin used the coarse grained model originally developed by Levitt and Warshel (see above), to perform, for the first time, a systematic protein-protein interaction search with unbound protein structures and compared the results to the complex crystal structure [20]. In 1986, attempts were made to use pattern recognition algorithms to find possible binding poses of protein-protein complexes [18]. To this end, the surface of the binding partners is divided in different shapes and the agreement between these shapes can be tested with pattern recognition algorithms. The approach allows to further decrease the computational costs by omitting the testing of concave patches with concave patches and convex patches with convex patches.

The biggest improvement in protein-protein docking was the application of the fast convolution to this problem, first published 1992 by Katchalski-Katzir et al. [21]. The algorithm uses a grid representation of the unbound proteins and the grid cells are parametrized in a manner such that an overlap of the surfaces is awarded but an overlap of the core of one protein and the surface or core of the other protein is rejected. The algorithm is fast because the fast convolution is applicable, which has a complexity of  $\mathcal{O}(N^3 \log(N^3))$  compared to the trivial algorithm with a runtime of  $\mathcal{O}(N^6)$ . The constant  $N$  is the number of grid cells in each dimension [22]. The original algorithm just accounted only for shape but was later extended to include the physicol-chemical properties of the interaction. This approach is still today state of the art and the first step in many protein-protein docking pipelines often followed by rescoring with a finer energy function or clustering of the docking decoys. The complex candidates could also be a starting point for MD simulations. With increasing computing power also computationally more expensive methods for flexible docking became possible. In 2003, the flexible docking program Attract was published, which uses a coarse-grained representation and does a multistart energy minimization with simultaneous side chain optimization in Cartesian space [23]. In 2015, Hui et al. [24] presented a method which is based, like earlier works, on pattern recognition methods by matching concave surface with convex one on the other protein. The novelty here is that for each protein before binding different conformations are produced simulating side chain flexibility by using rotamer sampling. For every surface several models are generated and these are then matched. The results are better than for grid based methods and the calculation is also slightly faster. For studying flexibility all this methods are not suitable, but full atom models are to computational expensive to use them for protein-protein docking.

A coarse grained model which allows full flexibility would be a suitable choice. The examination of these models is the aim of this work.

## 1.5 Aims

Protein-protein interactions are involved in nearly every process in living organisms like intracellular signaling, immune response and enzyme regulation [25]. Besides the existence of several experimental methods for determining the interaction of proteins [26] these methods do not reveal the interactions on an atomic level. It is estimated that the human body alone contains around 650,000 protein-protein interactions [27]. Determining the structure of all these protein-protein complexes is an experimentally challenging task. Protein-protein complexes are often too transient for X-Ray crystallography and too big for NMR. As a result only  $\approx 1/5$  of the protein structures in the PDB are protein-protein complexes.

Protein-protein docking is the method to predict the interaction of protein-protein complexes *in silico* from the folded binding partners and thus offers the possibility to fill the gap of missing protein-protein complexes. Yet, because of the complexity of the search space, current methods for global search do not account for full flexibility of the proteins. In this work, we examine the possibility to use coarse grained force fields to predict protein-protein interactions, as they could be a good choice for simulating bigger systems without restricting the flexibility of the proteins.

We use two different coarse grained force fields, the force field developed by Berezina and Deserno (BD) [28] and the OPEP force field developed by Derreumaux and co-workers [29].

We study the suitability of these force fields by energy minimization and rescoring of the rigid body decoys produced by ZDOCK. In chapter 2, which focuses on the BD force field, we gradually change the parameters of this force field to improve the discrimination between decoys that are close to the bound conformation and mis-docked decoys produced by ZDOCK. We find that an important feature for modeling the shape of the protein-protein interface is how well the sizes and positions of the side chains are described. Our conclusion is that in OPEP a better description of the side chains is provided. Therefore, the studies in chapter 3 and 4 are based on OPEP. As with the BD force field we first energy minimize ZDOCK predictions and we develop a soft rescoring function based on OPEP to score these minimized complexes. The resulting scoring function has a softer representation of the interface energy than the original OPEP force field. The scoring function is further trained with an incremental learning approach and tested on an independent test set.

In chapter 4 we present an OPEP-based method for the creation of the initial decoys, which is based on MC runs starting around a given initial conformation. Here, we use a conformation close to the bound complex as initial conformation. Subsequently we minimize these structures in Cartesian and in the rigid body space. The limitation to a local search leads to a denser sampling allowing us to study the presence of binding funnels. We also cluster the final conformations to test for the presence of binding attractors and the possibility of the force field to detect them. Finally we examine the effect of the initial sampling and compare the results for decoys created by ZDOCK and those created by our MC approach.



# Chapter 2

## Applicability of a generic coarse grained force field to protein-protein docking

### 2.1 Introduction

In this chapter we employ the coarse grained force field developed by Bereau and Deserno (BD) [28]. This force field was originally implemented in ESPResSo [30]. For performance reasons and for having more flexibility we implemented the BD force field in FORTRAN and made modifications to the parametrizations of the force field: we changed the interaction energy of the side chain beads and in a further step their sizes and positions.

The chapter organized as followed: first we present the force field and describe methods and the data set we used in this chapter. Then we show the performance for docking for the original force field and afterwards the performance of the modifications.

### 2.2 Methods

#### 2.2.1 ZDOCK

ZDOCK is a rigid body global search protein protein docking approach [31]. For this the binding partners are mapped from all-atom representation onto a grid and each grid cube gets a value assigned, corresponding to the atom it belongs to. The interaction energy between the two protein grid maps is calculated by a fast convolution. This reduces the run time for each convolution to  $\mathcal{O}(N^3 \log(N^3))$ , instead of  $\mathcal{O}(N^6)$ , where  $N$  is the number of grid cubes in

each dimension. This convolution has to be done for each rotation of the ligand, here we use a spacing  $15^\circ$  in rotation which leads to 3600 angles for the ligand. From each convolution only the best translation is stored leading to 3600 decoys. We use, as recommended by the developers of ZDOCK, only the first 2000 decoys.

## 2.2.2 Energy minimization

Energy minimization relaxes the system to a nearby local minimum. We use a quasi Newton method to this end.

In a Newton method the object function  $F$  is approximated by a Taylor series  $TF$  of second order at the initial coordinate  $x_0$ :

$$TF_{x_0}(x) = F_{x_0} + \nabla F_{x_0}^T x + \frac{1}{2} x^T B_k x. \quad (2.1)$$

We name the approximated Hesse matrix at step  $k$   $B_k$ . On this quadratic approximation, the Newton method for finding zeros is applied on the first derivative and an update step towards the minimum has the form

$$x_{k+1} = x_k - B_k^{-1} \nabla F(x_k). \quad (2.2)$$

Thus one needs to calculate the first derivative and the Hessian matrix. In the case of molecular modeling the first derivatives are calculated by the force fields (the opposite vector of the force on the particle). Calculating the Hessian is quite costly and it is also not always possible. In quasi-Newton methods the Hessian matrix is approximated.

The Hesse matrix is iteratively updated and for the updating different schemes exist which are the main differences of the quasi-Newton approximations. The following equation has to be solved:

$$B_{k+1}(x_{k+1} - x_k) = \nabla F(x_{k+1}) - \nabla F(x_k) \quad (2.3)$$

This means that the transformation by the approximated Hesse matrix  $B_{k+1}$  should fulfill the change of the gradient of the last iteration. The matrix  $B_{k+1}$  is under-determined and the new matrix should be close to the matrix  $B_k$ . In the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [32] the update scheme of Hesse matrix has the form

$$s_k = x_{k+1} - x_k \quad (2.4)$$

$$y_k = \nabla F(x_{k+1}) - \nabla F(x_k) \quad (2.5)$$

$$B_{k+1} = B_k + \frac{y_k y_k^T}{y_k^T s_k} - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k}. \quad (2.6)$$

As we showed in equation 2.2 we need the inverted Hesse matrix to make a step. Inverting a matrix is computationally expensive and so the inverted matrix is calculated directly

$$B_{k+1}^{-1} = \left(I - \frac{s_k y_k^T}{y_k s_k}\right) B_k^{-1} \left(I - \frac{y_k s_k^T}{y_k^T s_k}\right) + \frac{s_k s_k^T}{y_k^T s_k}. \quad (2.7)$$

Saving the whole Hesse matrix is too memory demanding. For this reason we use the limited memory BFGS (LBFGS) minimizer [33]. In this algorithm the  $m$  last  $s_k$  and  $y_k$  are stored. So  $B_k^{-1} \nabla F(x_k)$  from 2.2 is created by these vectors at every iteration. The main idea is the same like in the BFGS algorithm but one only applies the update formula  $m$  times on an initial guess of the Hessian, which can be approximated as  $H_k^0 = \gamma_k I$  [32], with

$$\gamma_k = \frac{s_{k-1}^T y_{k-1}}{y_{k-1}^T y_{k-1}}. \quad (2.8)$$

Equation 2.7 can also be computed without temporary matrices. Thus for the first  $m - 1$  iteration steps the BFGS and LBFGS algorithms have the same result when the same  $B_0^{-1}$  is chosen [32].

### 2.2.3 The coarse grained force field

Protein-protein docking is computationally expensive. Thus a coarse grained model could be a suitable choice. Especially the number of unbound interactions that have to be calculated is largely reduced, because the number particles is reduced. Another advantage of coarse grained models is that they can smooth the rough energy landscape [34]. In this work a coarse grained model originally developed by Bereau and Deserno is used [28]. In this model, every amino acid is represented by four beads (figure 2.1), one bead for the side chain and three

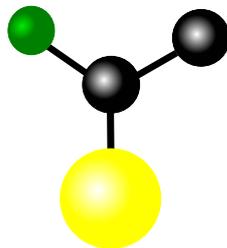


Figure 2.1: The coarse grained model by Bereau and Deserno. All heavy backbone atoms are modelled. The side chain is modelled as one big bead (yellow), which has the same size for every amino acid.

beads for the backbone (N, C<sub>α</sub>, C'). The side chain beads are centered at the C<sub>β</sub> position of the amino acids and the same van der Waals (vdW) radius of 5 Å is used for all amino acids. An exception is glycine, which has no side chain and therefor also has no side chain bead in the coarse grained model.

### Local interactions

The local interactions are modelled by harmonic and dihedral potentials. Bonds between beads are modelled with a harmonic potential, with  $r_0$  as equilibrium distance between the beads,

$$V_{\text{bond}}(r) = \frac{1}{2}k_{\text{bond}}(r - r_0)^2, \quad (2.9)$$

and also angles are described by a harmonic potential, with  $\theta_0$  as the equilibrium angle between three connected beads.

$$V_{\text{angle}}(\theta) = \frac{1}{2}k_{\text{angle}}(\theta - \theta_0)^2. \quad (2.10)$$

The force constants are  $k_{\text{bond}} = 180 \text{ kcal}/(\text{mol } \text{Å}^2)$  and  $k_{\text{angle}} = 180 \text{ kcal}/(\text{mol deg}^2)$ . These leads to rather small changes in bond lengths and angles. The main change of the conformation originates from the rotations around bonds. This is simulated by Fourier series restricted on singles modes based on dihedral angles:

$$V_{\text{dih}}(\varphi) = k_n[1 - \cos(n\varphi - \varphi_{n,0})]. \quad (2.11)$$

Only for peptide bonds before proline the sum of two modes are used to allow modeling of cis and trans conformations. Dihedral angles are defined by four consecutive beads, for example N, C<sub>α</sub>, C', N along the backbone. In this example, the rotation is around the C<sub>α</sub> - C' bond and zero degree is the conformation when the two N show in the same direction [28]. Rotation is measured via the angle between the planes defined by NC<sub>α</sub>C' and by C<sub>α</sub>C'N. The protein backbone consists of three different dihedral angles. The  $\phi$  angle, formed by the beads C'NC<sub>α</sub>C', and the  $\psi$  angle, formed by the beads NC<sub>α</sub>C'N, are rather inflexible. The third dihedral angle  $\omega$ , formed by C<sub>α</sub>C'NC<sub>α</sub> is not so flexible because its the rotation around the peptide bond. This has usually just two conformations of 0° and 180°, which correspond to the conformations cis and trans that are separated by a high energy barrier. In unfolded polypeptides 96 % of the amino acids are in the trans (180°) conformation as it is sterically more favoured [35]. The only exception are peptide bonds before proline where 20 % of the amino acids are in the cis state in unfolded polypeptides. [35]. The frequency of a peptide bond in the cis state decreases for folded proteins to 0.03 % for peptide bonds which

are not before a proline and to 5.21 % for peptide bonds before proline [3]. That is the reason why for peptide bonds before proline the sum of two modes are used. For keeping the amino acids in the favoured L-form chirality an improper dihedral angle is introduced. This dihedral is defined by  $\text{NC}_\alpha\text{C}'\text{C}_\beta$ , which keeps the side chain bead in the right tilt compared to the backbone plane.

Bereau and Deserno added a dipole interaction to the force field. The carbonyl and amide groups at the peptide bond form dipoles that interact with each other. Chen and co-workers [36] showed that the dipole interaction is important to favour  $\beta$  content over  $\alpha$  content. If one would really like to model this interaction one should introduce a dipole term. However, because it is a quite local interaction and the dipoles do not move far away from each other, the nature of this interaction only depends on the local dihedral conformation of  $\phi$  and  $\psi$ . Thus, one can also model it with the local dihedrals, leading to the potential

$$U_{\text{dip}}(\phi, \psi) = k_{\text{dip}}[(1 - \cos \phi) + (1 - \cos \psi)]. \quad (2.12)$$

The resulting potential does not cover the original influence of the dipole interaction on the dihedral angles, but ensures to stabilize the  $\beta$  conformation.

### Long range interactions

For avoiding steric overlaps, a repulsive-only Weeks-Chandler-Andersen potential is used. It has the form

$$V_{\text{steric}} = \begin{cases} 4\epsilon_{\text{steric}} \left[ \left(\frac{\sigma_{ij}}{r}\right)^{12} - \left(\frac{\sigma_{ij}}{r}\right)^6 + \frac{1}{4} \right], & r \leq r_c, \\ 0, & r > r_c, \end{cases} \quad (2.13)$$

where  $\epsilon_{\text{steric}} = 0.012$  kcal/mol,  $r_c = 2^{1/6}\sigma_{ij}$  and  $\sigma_{ij}$  is the van der Waals radius. The cutoff  $r_c$  is the distance where the potential would have its minimum if it would be a 12-6 Lennard-Jones potential. The potential 2.13 is not calculated between beads that are less than three bonds apart, because angles and dihedral potentials are sufficient to avoid overlaps. The side chain-side chain interactions are modeled as following

$$V_{\text{hp}}(r) = \begin{cases} 4\epsilon_{\text{hp}} \left[ \left(\frac{\sigma_{C\beta}}{r}\right)^{12} - \left(\frac{\sigma_{C\beta}}{r}\right)^6 \right] + (\epsilon_{\text{hp}} - \epsilon'_{ij}), & r < r_c, \\ 4\epsilon_{\text{hp}}\epsilon'_{ij} \left[ \left(\frac{\sigma_{C\beta}}{r}\right)^{12} - \left(\frac{\sigma_{C\beta}}{r}\right)^6 \right], & r_c \leq r \leq r_{\text{hp,cut}} \\ 0, & r > r_{\text{hp,cut}}. \end{cases} \quad (2.14)$$

The parameters  $\epsilon_{\text{hp}} = 2.7$  kcal/mol and  $r_{\text{hp,cut}} = 10$  Å. The formula 2.14 is not continues at  $r = r_c$  and at  $r = r_{\text{hp,cut}}$ . To fix this we change the formulation of

the repulsive part and also introduce a shift  $c_{\text{shift}}$  that smooths the potential at the cut off  $r_{\text{hp,cut}}$ . The revised formula has the form

$$V_{\text{hp}}(r) = \begin{cases} 4\epsilon_{\text{hp}} \left[ \left( \frac{\sigma_{C\beta}}{r} \right)^{12} - \left( \frac{\sigma_{C\beta}}{r} \right)^6 \right] + \epsilon'_{ij} \epsilon_{\text{hp}} (-1 + 4c_{\text{shift}}), & r < r_c, \\ 4\epsilon_{\text{hp}} \epsilon'_{ij} \left[ \left( \frac{\sigma_{C\beta}}{r} \right)^{12} - \left( \frac{\sigma_{C\beta}}{r} \right)^6 + c_{\text{shift}} \right], & r_c \leq r \leq r_{\text{hp,cut}}, \\ 0, & r > r_{\text{hp,cut}}, \end{cases} \quad (2.15)$$

where  $c_{\text{shift}}$  is given as

$$c_{\text{shift}} = -1 \left[ \frac{\sigma_{C\beta}^{12}}{r_{\text{hp,cut}}^{12}} - \frac{\sigma_{C\beta}^6}{r_{\text{hp,cut}}^6} \right] \quad (2.16)$$

The side chain interactions are parametrized based on the Miyazawa-Jernigan (MJ) matrix [37]. However, not the whole matrix is used. Instead for each residue  $i$  one parameter  $\epsilon_i$  is defined, and the geometric mean of  $\epsilon_i$  and  $\epsilon_j$  approximates the original interaction value  $\epsilon_{ij}^{\text{MJ}} \approx \epsilon'_{ij} = \sqrt{\epsilon_i \epsilon_j}$ . The  $\epsilon_i$  values for each amino acid are normalized such that all amino acids are in the range  $0, \dots, 1$ , with 0 for the interaction of the most hydrophilic residue (here lysine) and 1 for the most hydrophobic residue (leucine). This implies that all interactions with lysine are zero. The new values have a correlation of 95 % with the original MJ matrix and the interaction energies can be seen in figure 2.2.

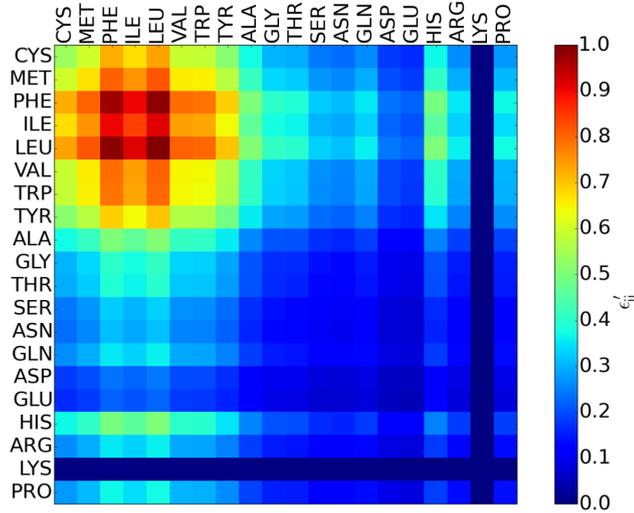


Figure 2.2: The contact matrix resulting from the coarse-grained force field by Bereau and Deserno. The bead size of the side chains is 5 Å. Higher values indicating more attractive interactions.

Hydrogen bonds are modeled between the backbone N-H and the backbone C-O. However, the hydrogen and oxygen are just virtual and not explicitly modeled (see Fig. 2.1). Therefore, the position of the surrounding beads ( $C_\alpha, N, C_\beta$  for N-H,  $C_\alpha, C', N$  for C-O) are considered and the hydrogen bond is calculated dependent on the positions of these particles. The hydrogen bond potential is angle dependent and the angles  $\theta_N$  and  $\theta_C$  are calculated. The angle  $\theta_N$  is the angle between the vector connecting N and H, and the vector connecting N and  $C'$ . The angle  $\theta_C$  is defined by the vector connecting  $C'$  and O, and the vector connecting  $C'$  and N. The resulting formula has the form

$$V_{\text{hb}}(r, \theta_N, \theta_C) = \epsilon_{\text{hb}} \left[ 5 \left( \frac{\sigma_{\text{hb}}}{r} \right)^{12} - 6 \left( \frac{\sigma_{\text{hb}}}{r} \right)^{10} \right] \quad (2.17)$$

$$\times \begin{cases} \cos^2 \theta_N \cos^2 \theta_C, & |\theta_N|, |\theta_C| < 90^\circ, \\ 0, & \text{otherwise,} \end{cases} \quad (2.18)$$

where  $\epsilon_{\text{hb}} = 3.6$  kcal/mol,  $\sigma_{\text{hb}} = 4.11$  Å and the cutoff for the interaction is 8 Å.

### Parameter tuning

The parameter tuning was done separately for local and global properties. For producing the local Ramachandran plot GLY-GLY-GLY and GLY-ALA-GLY peptides were used and the parameters for the potential in equation (2.13) determined, i.e. parameters for the size of beads and  $\epsilon_{\text{steric}}$ . Side chain effects and hydrogen bonds are cancelled out by using a three bead system, therefore larger proteins had to be considered for the determination of the other parameters. For obtaining the parameters of local properties parallel tempering simulations with eight replicas combined with the weighted histogram analysis method [38] was performed. This approach returned the free energies of the different states of the Ramachandran plot. Different parameters were tested and the one that produced the most similar Ramachandran plot to them in the publication [39] were kept.

For the determination of the global parameters the three helix bundle with PDB ID 2A3D was used as reference. The parameters that were tuned with this protein are the weighting parameter  $\epsilon_{\text{hp}}$ , for the side chain-side chain interactions, the parameter  $\epsilon_{\text{hb}}$  for hydrogen bonds and  $k_{\text{dip}}$  for the dipole interaction. They performed several replica exchange molecular dynamics (REMD) runs with eight replicas starting with structures with random dihedrals and changed different parameters. The quality of the parameters were measured by the root-mean-square deviation (RMSD) to the NMR structure. A structure with RMSD  $\approx 4$  Å could be found with the final parameter settings for 13 of the 15 REMD simulations at room temperature. The final parameters can be seen in table 2.1.

Bond length			
	$r_0$ (Å)	$k_{\text{bond}}$ (kcal/(mol·Å <sup>2</sup> ))	
NC <sub>α</sub>	1.455	180	
C <sub>α</sub> C'	1.510	180	
C'N	1.325	180	
C <sub>α</sub> C <sub>β</sub>	1.520	180	
Bond angles			
	$\theta$ (deg)	$k_{\text{angle}}$ (kcal/(mol·deg <sup>2</sup> ))	
NC <sub>α</sub> C <sub>β</sub>	108	180	
C <sub>β</sub> C <sub>α</sub> C'	113	180	
NC <sub>α</sub> C	111	180	
C <sub>α</sub> C'N	116	180	
C'NC <sub>α</sub>	122	180	
Dihedrals			
	k (kcal/mol)	$n$	$\varphi_0$ (deg)
$\phi$	-0.18	1	0
$\psi$	-0.18	1	0
$\omega$	40.2	1	180
$\omega_{\text{PRO}}$	1.8	2	0
Improper	10.2	1	±120
Bead sizes			
$\sigma_{\text{N}}$ (Å)	$\sigma_{\text{C}_\alpha}$ (Å)	$\sigma_{\text{C}'}$ (Å)	$\sigma_{\text{C}_\beta}$ (Å)
2.9	3.7	3.5	5.0
Interaction strength			
$\epsilon_{\text{steric}}$ (kcal/mol)	$\epsilon_{\text{hp}}$ (kcal/mol)	$\epsilon_{\text{hb}}$ (kcal/mol)	$k_{\text{dip}}$ (kcal/mol)
0.012	2.7	3.6	-0.42

Table 2.1: The parameters for the interactions in the BD force field.

### 2.2.4 The test set

Our test set is a subset of the ZDOCK benchmark 4.0 [40]. We used the bounded structures and selected only dimers from this test set. This leads to 47 complexes, comprising 23 enzyme/inhibitor, one antigen/antibody and 23 ‘other’ complexes. The one antigen/antibody complex 2I25 is removed from our test set, because one complex for a whole category has no statistical meaning. The remaining 46 complexes are shown in table 2.2.

enzyme/inhibitor		
1BVN	1CLV	
1D6R	1DFJ	1EAW
1EWY	1F6M	1FLE
1JIW	1JK9	1M10
1OPH	1PPE	1PXV
1TMQ	1UDI	2A9K
2OOB	2OUL	2SIC
2SNI	3SGQ	7CEI
'other' complexes		
1AK4	1B6C	
1E96	1FFW	1FQJ
1GCQ	1GHQ	1GPW
1GRN	1HE1	1IBR
1IRA	1J2J	1KAC
1LFD	1MQ8	1QA9
1SYX	1XQS	1Z0K
2HRK	3CPH	3D5S

Table 2.2: The PDB IDs of the 46 complexes, for which the BD force field was tested.

### 2.2.5 Metric for success

The results are evaluated by the ligand RMSD (LRMSD). The LRMSD is the RMSD between the ligand of the prediction and the ligand of the crystal complex after the receptor of the prediction and the receptor of the crystallized bound complex are superimposed. As a second metric we use the interface RMSD (IRMSD). The IRMSD is defined as the RMSD between  $C_\alpha$  interface atoms of the co-crystallized model and the prediction after superposition. Interface  $C_\alpha$  atoms are all atoms that are within 10 Å distance to the binding partner in the co-crystallized complex [41]. For the superposition we use the corresponding function of Biopython [42]. A hit is a prediction with an IRMSD lower than 4 Å.

### 2.2.6 Different side chain parametrization

The force field by Bereau and Deserno is coarsest for the side chain of each amino acid, which is represented by only one bead. Moreover, the position relative to the backbone and the size of all side chain beads are assumed to be identical for all amino acids. However, we found this assumption leads to an insufficient representation of the shape complementary between receptor and

ligand. Therefore, we implemented the side chain positions and sizes from the OPEP force field [43] and tested this modification. The new side chain sizes and positions are shown in table 2.3. These parameter are taken from [43].

The depth of the minimum of the Lennard-Jones potential is parametrized with values  $\epsilon'_{ij}$  that approximate the MJ matrix. This approximation is not motivated and there is also no need for it because a computer can easily handle the amount of 210 parameters, i.e., the number of unique contacts between the 20 amino acids. In the following we change these parameters back to the original full MJ matrix [37] and evaluate which impact they have on the conformation.

Res	$r_0$ (Å)	$r_{\text{CA-SC}}$ (Å)	N. CA. Sc (deg)	Sc. CA. C (deg)
ALA	4.08	1.52	116.60	111.10
CYS	4.32	1.95	108.50	117.65
ASP	4.82	2.14	110.74	119.16
GLU	5.28	2.77	110.95	120.87
PHE	5.94	2.62	110.72	124.15
GLY				
HIS	5.48	2.60	109.93	124.39
ILE	5.24	2.27	109.16	118.96
LYS	5.60	3.11	112.92	121.57
LEU	5.22	2.40	112.72	124.94
MET	5.32	2.70	113.54	121.70
ASN	4.98	2.16	109.66	121.47
PRO	4.76	1.81	68.77	133.11
GLN	5.36	2.76	111.24	122.32
ARG	6.08	3.59	112.07	119.79
SER	4.16	1.77	106.78	107.98
THR	4.72	1.90	107.21	114.11
VAL	5.14	1.44	126.10	99.66
TRP	6.54	2.86	117.40	119.81
TYR	6.02	2.79	112.54	123.56
BD force field $C_\beta$	5.0	1.50	108.0	113.0

Table 2.3: The side chain parameters of the OPEP force field that we implemented into the BD force field. The side chain center are not longer on the  $C_\beta$  position but on average positions calculated from crystal structures. In the bottom line the original parameters for the side chains in the BD force field are shown for comparison.

hydrophobic (h)	ALA, VAL , MET , LEU, ILE, PRO, TRP, PHE
polar (p)	TYR, THR, GLN, GLY, SER, CYS, ASN
basic (b)	ARG, LYS, HIS
acidic (a)	ASP, GLU

Table 2.4: Classification of the amino acids according to their physico-chemical properties.

## 2.3 Results

### 2.3.1 Properties of the test set

As mentioned in section 2.2.4 we used 46 complexes: 23 enzyme/inhibitor and 23 ‘other’ complexes. First, we evaluated the properties of the binding site of the crystal complexes. For that we calculated and visualized different properties of the test set, namely the buried surface area (BSA), the number of amino acids of the binding partner and the residue/residue contacts in the interface. For the characterization of the binding, we put each residue into one of four groups: hydrophobic (h), polar (p), basic (b) and acidic (a). The assignment of the amino acids can be seen in table 2.4. For measuring the contact we use the all-atom structures and calculate, for comparability to the coarse grained structures the center of mass of each side chain and determine the contact between side chain centers. For the definition of a contact we follow the work of Ravikant and Elber [44], where two residues are assumed to be in contact when their side chain centers are within 8 Å of each other. In figure 2.3 one can see on the left the total number of residues in the receptor and the ligand for each complex, the number of contacts in the interface, the BSA, and the ratios of contacts categorized by their physico-chemical nature. Interestingly, but not too surprisingly, there is a correlation between the number of contacts and the BSA, which is the amount of solvent accessible surface area (SASA) that gets lost during complex formation. For most complexes in the test set the amount of h/h interactions is lower than 20 %. Interactions involving hydrophobic residues (i.e. h/h, h/pc) are involved in 50 % of the interactions for most of the complexes. For some of the complexes there is no h/h interaction at all (1XQS,1QA9,1DFJ). Interestingly, in these complexes salt bridges play a larger role for receptor-ligand binding.

### 2.3.2 Docking with ZDOCK

For producing docking predictions, we restricted the search space on the receptor to residues that are closer than 15 Å distance from any residue in the ligand in the bound conformation. The docking is done with ZDOCK 3.0 and a 15°

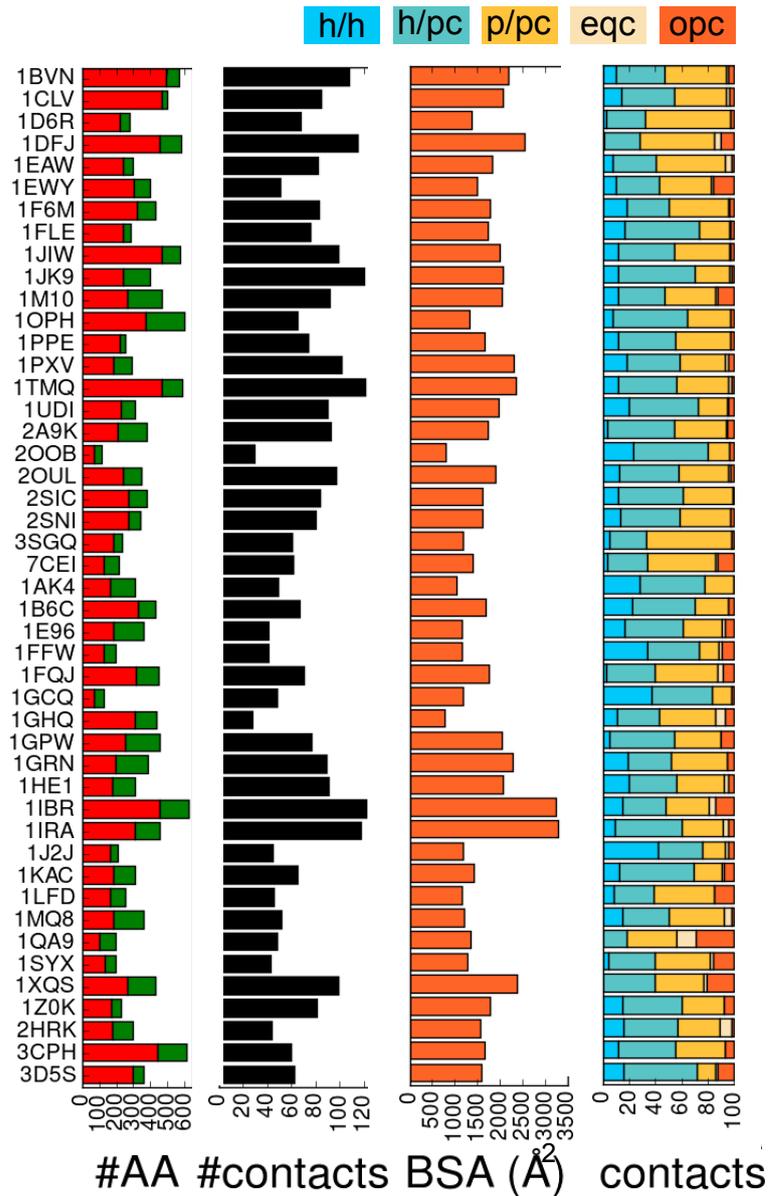


Figure 2.3: Properties of the target complexes of the test set. #AA shows the number of residues in the receptor (red) and the ligand (green). #contacts is the number of contacts in the interface. #BSA is the buried surface area due to complex formation and contacts shows the frequency of contacts between different categorized amino acids according to their category.

sampling is used [45]. For each target, we kept the first 2000 predictions and ZDOCK finds for all targets a hit (with IRMSD < 4 Å). For 32 targets a hit was

positioned as rank one, i.e., it has the highest ZDOCK score. We evaluated the top 1 predictions per target again using the amount of contacts, the BSA and the type of the contact (see figure 2.4). If the complex is marked with an asterisk, than the first ranked ZDOCK prediction is a hit. Additionally we give the number of hits found by ZDOCK. ZDOCK regularly overestimates the BSA. Figure 2.5

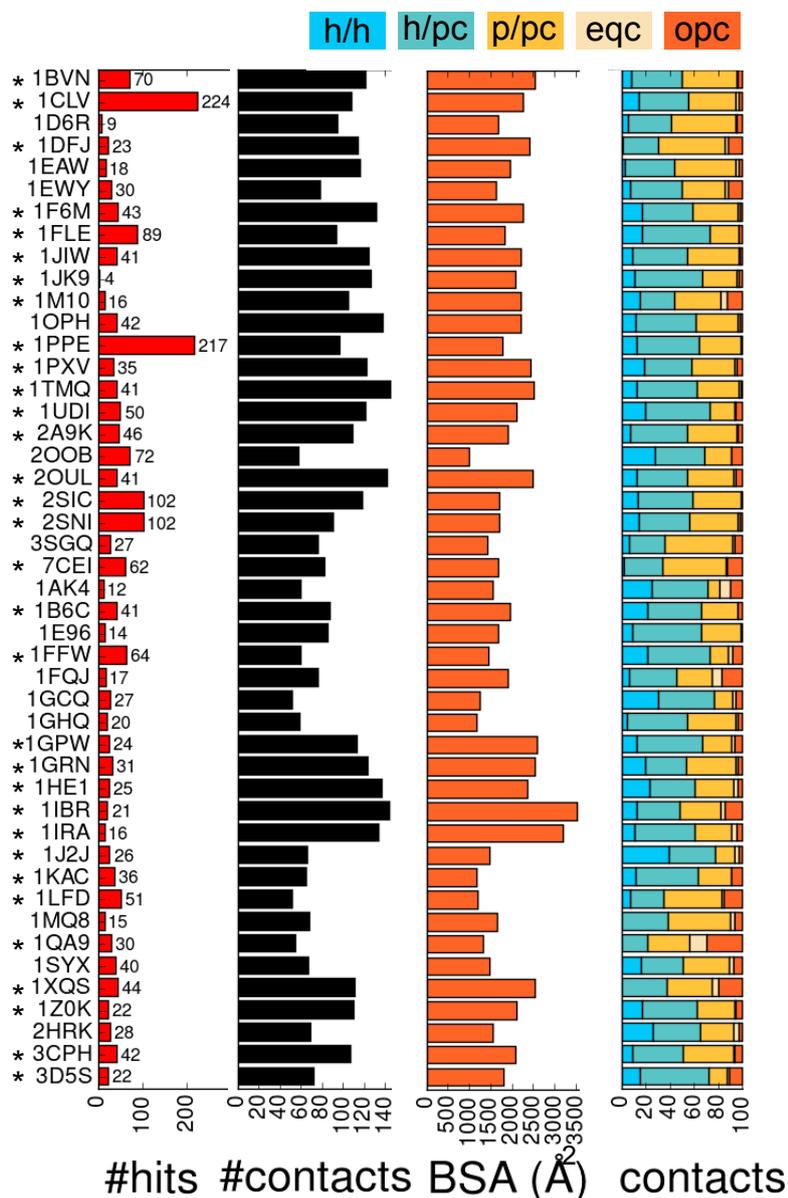


Figure 2.4: Properties of the top 1 predictions predicted by ZDOCK, #hits is the number of hits found by ZDOCK. See figure 2.3 for further explanation.

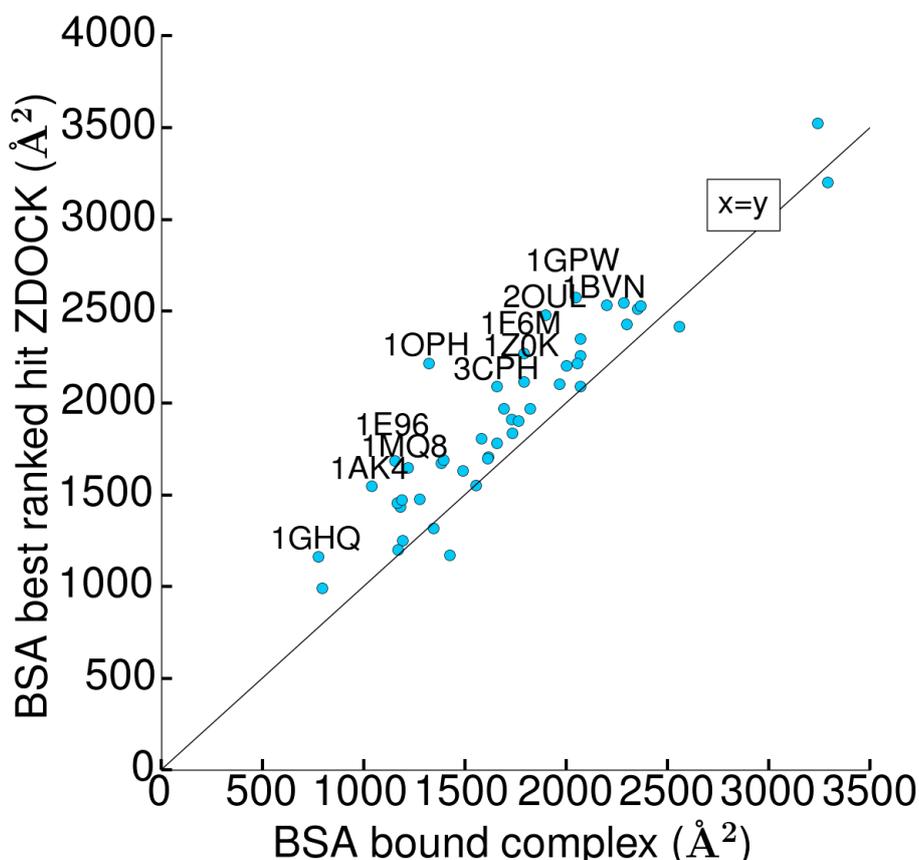


Figure 2.5: BSA of the highest scored ZDOCK prediction is plotted against the BSA of the corresponding crystal complex. Targets with a BSA difference of more than  $300 \text{ \AA}^2$  are marked by the PDB ID.

shows that the BSA is only for five targets lower in the ZDOCK prediction than in the crystal complex. We labeled the targets where the BSA differs by more than 10 %. Interestingly the overestimation is only a problem for complexes with a crystal BSA up to  $1500 \text{ \AA}^2$  (1GHQ, 1AK4, 1MQ8, 1E96, 1OPH), where the overestimation prevents the positioning of a hit as rank one. In figure 2.6 the relative distribution of the contact types of the top ZDOCK prediction is plotted against the contact distribution in the corresponding crystal structure. Targets with a deviation of more than 10 % in one of the contact types are labeled. Most of the points are close to the line  $x = y$  thus ZDOCK predicts the correct type and number of contacts. In general, ZDOCK seems to slightly overestimate contacts involving hydrophobic residues, like for 1SYX, 2HRK, 1GHQ and 1E96, where ZDOCK did not position a hit as rank one. On the other hand, for

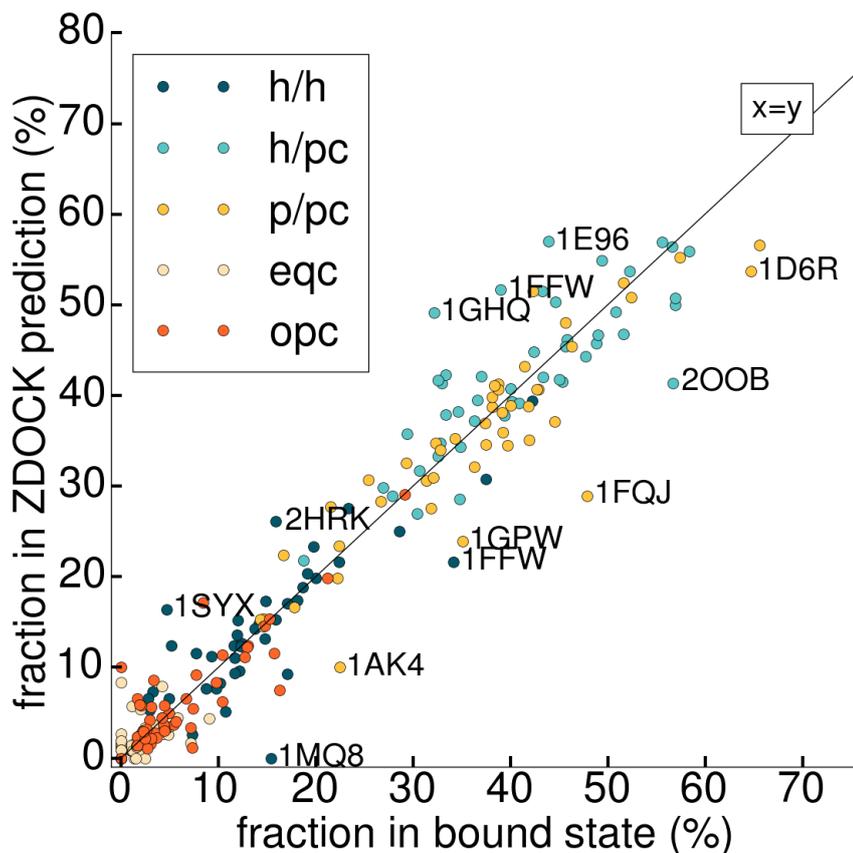


Figure 2.6: Relative distribution of contact types in the highest scored ZDOCK prediction against the relative distribution of contact types in the corresponding crystal structure. Complexes with a difference higher than 10 % are marked.

1MQ8 and 200B the hydrophobic contacts are underestimated and for these targets no hit is positioned as rank one by ZDOCK. For 1AK4, 1FQJ and 1D6R ZDOCK does not find a hit because the interactions in the interface involving hydrophilic residues are underestimated.

### 2.3.3 Rescoring of the ZDOCK predictions with the original BD force field

We rescored the ZDOCK predictions using the original BD force field and parametrization as described above, following the publication of Berau and Deserno [28]. The contact matrix for residue-residue interactions at the equilibrium distance  $r_c = 2^{1/6}\sigma_{ij}$  can be seen in figure 2.2. We take the 2000 ZDOCK

predictions plus the bound conformation for each target and first coarse-grain them and then minimize their energy using 400 steps with the LBFGS minimizer. The success in ranking can be seen in table 2.5. In column two one can see the rank using the full complex energy, while column four lists the best rank using only the interface energy. The “rank of the first hit” is the position of the best ranked hit, and, as a reminder, a hit is a prediction with an IRMSD lower than 4 Å. The column “#top5 hits” gives the number of hits that are in the top 5 after rescoring. If there are at least three hits among the top 5 predictions we assume that the scoring of this target leads to an energy funnel [46], where the conformations with lowest potential/free energy are near the native state. For enzyme/inhibitor complexes there are funnels for seven targets if the ranking is done by the complex energy and for four targets if the ranking is done by the interface energy. For ‘other’ complexes a funnel is found for one target if the ranking is done by the complex energy and for five targets if the ranking is done by the interface energy. Thus the ranking by the complex energy leads to eight funnels, while the ranking by the interface energy leads to nine funnels in our test set.

If the ranking is done by the complex energy, for nine targets a hit is ranked on position one: for six enzyme/inhibitor targets and for three ‘other’ targets. The interface energy ranks for eleven targets a hit on position one, for six enzyme/inhibitor and five ‘other’ complexes. In figure 2.7 success rate is shown

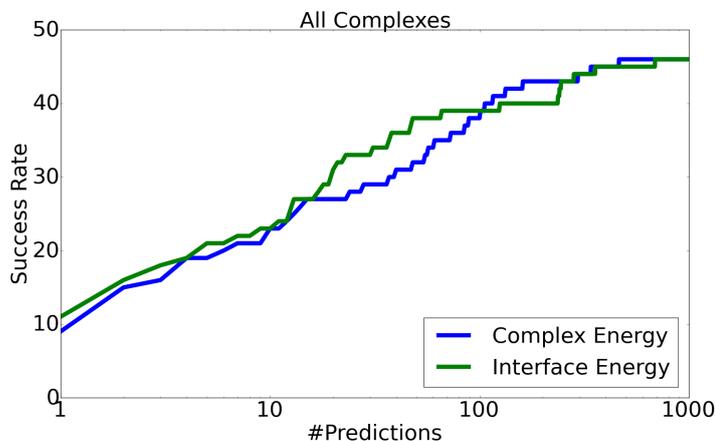


Figure 2.7: Success rate for rescoring with the original BD force field using either the interface or the complex energy.

for all complexes using either the complex energy or the interface energy for rescoring. The interface energy works slightly better if less than twelve predictions are considered, i.e. with the interface energy slightly more hits are found for the targets. This effect gets even stronger if more than twelve hits are taken and

lasts up to the top 100 predictions. Thus the interface energy is a good choice for ranking the predictions and it performs better than using the complex energy, especially if one considers a manageable subset of less than 100 predictions.

	Full Complex Energy		Interface Energy	
	Rank first hit	#top5 lower 4 Å	Rank first hit	#top5 lower 4 Å
enzyme/inhibitor				
1BVN	1	4	5	1
1CLV	4	1	1	2
1D6R	340	0	688	0
1DFJ	1	1	47	0
1EAW	116	0	2	1
1EWY	1	2	18	0
1F6M	2	3	1	1
1FLE	2	3	20	0
1JIW	28	0	38	0
1JK9	48	0	11	0
1M10	2	1	1	2
1OPH	106	0	13	0
1PPE	2	1	1	4
1PXV	61	0	13	0
1TMQ	3	2	3	2
1UDI	2	4	1	5
2A9K	10	0	3	1
2OOB	55	0	66	0
2OUL	1	3	2	3
2SIC	1	3	2	2
2SNI	1	4	1	4
3SGQ	57	0	37	0
7CEI	40	0	23	0
∅ enzyme/inhibitor	38.43	1.39	43.35	1.21
'other' complexes				
1AK4	85	0	4	1
1B6C	7	0	2	1
1E96	15	0	7	0
1FFW	10	0	1	1
1FQJ	73	0	241	0
1GCQ	37	0	9	0
1GHQ	133	0	125	0
1GPW	2	2	1	3
1GRN	4	1	21	0
1HE1	1	2	5	1
1IBR	1	1	1	4
1IRA	89	0	2	3
1J2J	4	1	17	0
1KAC	12	0	31	0
1LFD	24	0	20	0
1MQ8	161	0	357	0
1QA9	295	0	245	0
1SYX	463	0	282	0
1XQS	101	0	237	0
1Z0K	14	0	1	3
2HRK	13	0	13	0
3CPH	6	0	48	0
3D5S	1	5	1	5
∅ 'other' complexes	67.43	0.52	72.65	0.96
∅ all complexes	52.93	0.96	58.00	1.09

Table 2.5: Ranking obtained after rescoring with the original BD force field.

### Change of the RMSD

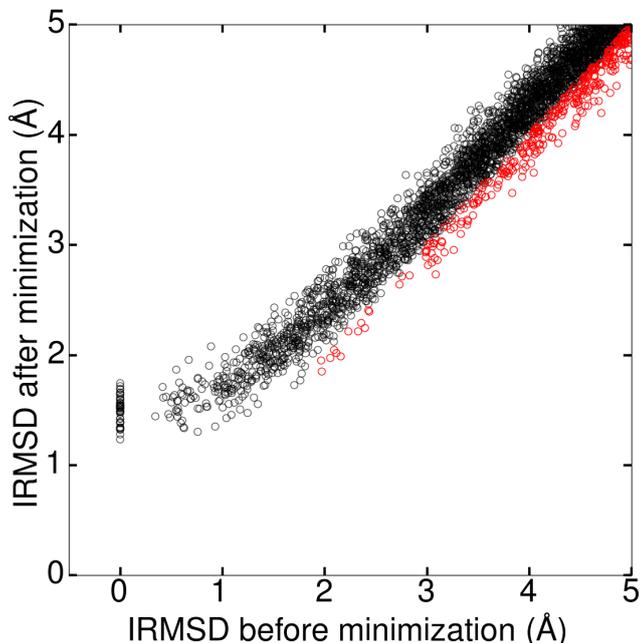


Figure 2.8: IRMSD change due to energy minimization with the original BD force field: Black dots correspond to decoys where the IRMSD increases, red indicate an IRMSD decrease.

Before we rescore the ZDOCK predictions, we minimize their energy using the BD force field. As a result, the complex structure may change, which we test by calculating the RMSD before and after minimization. We measure both the LRMSD and the IRMSD before and after the minimization. In figure 2.8 we show the change of the IRMSD for predictions that have an IRMSD up to 5 Å IRMSD before minimization. This is the range where the IRMSD is meaningful if we use the 4 Å criterion to define a hit. The IRMSD does not change much. Nonetheless, it becomes worse in most cases. For 74 % of the predictions the IRMSD increases due to minimization. In the range of up to 5 Å the IRMSD increases for even 91.2 % of the predictions. In figure 2.9 the LRMSD for predictions with an LRMSD lower than 15 Å before minimization is shown. The LRMSD increases for 57.5 % of the predictions considering all predictions. For cases with a starting LRMSD lower than 15 Å the LRMSD increases for 58.9 % of the predictions. Thus, the IRMSD increases more often than the LRMSD. This means that the whole ligand stays in the right location or close to it, while the interface area deforms during minimization. To test this hypothesis we superimposed the crystal

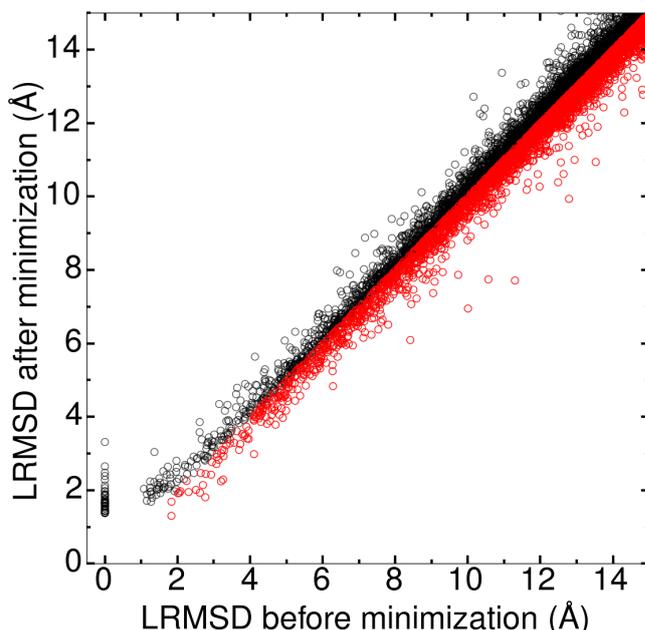


Figure 2.9: LRMSD change due to energy minimization with the original BD force field: Black dots correspond to decoys where the LRMSD increases, red dots indicate an LRMSD decrease.

structure of the receptor on the minimized receptor structure and the crystallized ligand structure on the minimized ligand structure. The result is a complex containing only the rigid movement of the ligand relative to the crystal structure as the secondary structures are the same. We calculated the IRMSD for these conformations, the plot of this IRMSD change can be seen in figure 2.10. For 56.06 % of the predictions the IRMSD increases, if all predictions are considered. In the range of up to 5 Å the IRMSD increases for only 38.55 % of the predictions, i.e. the IRMSD gets smaller or stays equal for 62.45 % of the predictions. In these cases the structures moved closer to the corresponding crystal structure due to minimization if one ignores the distortion of the secondary structure.

To check where the distortion happens we measure the RMSD change of the complex, the receptor, the ligand and the interface of each prediction. In addition, we measure the RMSD for each secondary structure element present in the crystal structure. The secondary structure is assigned with Stride [47] and we assign it to one of four groups. The assignments for  $\alpha$ -helix, 3-10 helix and  $\pi$ -helix form one group called “helix”. The assignment for extended conformation and isolated bridge form the “beta” group. The third and fourth groups are “turn” and “coil”. We calculated the mean value and standard deviation for all these

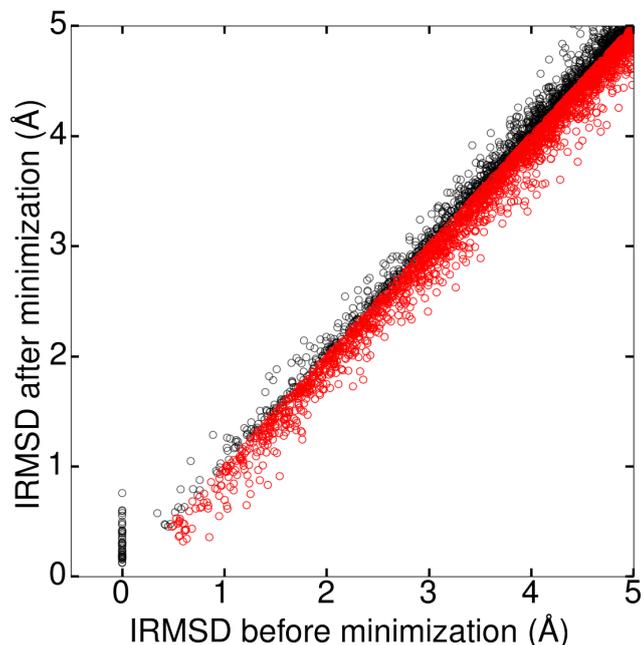


Figure 2.10: IRMSD change when the original crystal binding partners are superimposed on the minimized binding partners. Red dots show predictions where the IRMSD got lower, black dots show predictions where the IRMSD got higher due to minimization.

RMSDs. These values can be seen in table 2.6. The highest RMSD changes due to minimization can be seen for coil and turn secondary structures. This is also expected because these regions are the most flexible ones in proteins. Thereafter, the next most changing region is the interface region. The reason for this is the overlap between the receptor and ligand introduced by ZDOCK, causing a change during minimization in order to resolve the overlap between the beads. The well structured beta and helix elements do not change so much. The ligand of the complex changes a little bit more than the receptor. The reason could be that the a smaller molecule (the ligand is the smaller binding partner) is more flexible than the larger receptor.

### 2.3.4 Changing to the full Miyazawa-Jernigan matrix

The original force field by Bereau and Deserno uses an approximation of the MJ matrix, so that the whole matrix can be expressed based on 20 values. This approximation describes the matrix quite well, but is also unmotivated. Therefore, we decided to switch back to the original MJ matrix [37]. We transform the

	mean value (Å)	standard deviation (Å)
RMSD complex	1.57	0.09
RMSD receptor	1.54	0.13
RMSD ligand	1.59	0.19
RMSD interface	1.65	0.21
RMSD beta	1.32	0.21
RMSD coil	1.76	0.18
RMSD turn	1.80	0.17
RMSD helix	1.48	0.16

Table 2.6: The mean RMSD change for different parts of the complexes due to energy minimization.

matrix in this way that we multiply the energy values by  $-1$  to get  $\epsilon$  values for the Lennard-Jones-Potential (eq. 2.14). After this transformation the highest value is the LEU/LEU contact with a value of  $\epsilon = 7.37$ . We then divide the whole matrix by 7.37 in order to stay in the same  $\epsilon$  range as before. The sum over all 210 side chain-side chain interactions is 90.48, the same sum for the original parametrization of the force field was 61.95. This confirms that a lot of information got lost due to the reduction of the parameter space. Because of the higher impact of the side chain interactions by the full matrix, we have to reduce the parameter for side chain interactions from  $\epsilon_{\text{hp}} = 2.7$  to  $\epsilon_{\text{hp}} = 1.848$  kcal/mol to keep the balancing between the different contributions to the force field.

### Ranking of docking predictions

In table 2.7 the results for the reranking with the full MJ matrix are listed. If the predictions are reranked by the full complex energy, for ten predictions a hit at position one could be found. If the predictions are reranked with the interface energy, for nine predictions a hit is ranked on position one. With the original BD force field parametrization nine predictions were ranked on position one by the complex energy and eleven by the interface energy. The ranking by the complex energy ranks 25 prediction into the top 20. With the original parametrization it were 27. The ranking by the interface energy ranks 32 predictions in the top 20. With the original parametrization it were 31. The ranking by the complex energy produces eight funnels, which is the same as with the original parametrization. The reranking by the interface energy produces nine funnels, with the original parametrization it were also nine. Thus, the number of funnels is the same for both parametrizations and independent of using the complex or the interface energy for ranking.

The parametrization with the full MJ matrix performs better than the original

parametrization when the interface energy is considered for the top 20 predictions. But the interface energy ranking leads to a loss of three top 1 hits compared to the original parametrization: 1FFW is now on rank three, 2SNI is on rank two and 1CLV is on rank six. On the other hand, 1HE1 was before on position five but now has a top 1 position.

We now compare the ranking of each target by the different parametrization independent how good the ranking is. If we take the full complex energy for ranking, the original parametrization ranks 16 targets better than the full MJ matrix. The complex energy with the full MJ matrix ranks 25 targets better than the original parametrization. For five complexes the ranking does not change. If we use the interface energy for ranking the original parametrization ranks 16 targets better than that with the full MJ matrix. The full MJ matrix parametrization ranks 19 targets better than the original parametrization. For eleven targets the ranking is not changed. Thus, we loose some top 1 hits but more targets are improved in ranking than for the original parametrization. This can also be seen in the average of the ranking, which improves.

### Change of the RMSD

The change of the IRMSD during minimization is shown for  $\text{IRMSD} \leq 5 \text{ \AA}$  before minimization in figure 2.11. For most complexes the IRMSD got worse during minimization. For predictions with an IRMSD lower than  $5 \text{ \AA}$  before minimization the IRMSD increases or stays the same for 91.47 % of the decoys, which is similar to the 91.2 % obtained with the original force field.

### 2.3.5 Replacing the sizes and positions of the side chains

One of the coarsest approximations of the force field is that all side chains have the same size and that they are centered at the  $C_\beta$  position. But size and shape is crucial for protein-protein interactions [48] and is one of the main effects used by grid based docking programs. Thus we decided to change the positions and sizes of the side chain to their values in OPEP [43]. OPEP describes the sizes and positions of the side chains as a rotamer independent average calculated from a data set consisting of 2248 PDB structures.

The bead sizes are smaller than  $5 \text{ \AA}$  (see table 2.3). This means that the Lennard-Jones potential is much narrower than before. Under the assumption that every side chain contact is in a distance between  $\sigma_{ij}$  and  $10 \text{ \AA}$  we integrate for every side chain the Lennard-Jones potential for this area. First with the side chain sizes of  $5 \text{ \AA}$  and afterwards with the OPEP side chain sizes. Furthermore, we calculate the ratio for each side chain contact of the integration with  $5 \text{ \AA}$  and the new parameter. For all contacts besides glycine we obtain mean values of the

	Full Complex Energy		Interface Energy	
	Rank first hit	#top5 lower 4 Å	Rank first hit	#top5 lower 4 Å
enzyme/inhibitor				
1BVN	2	3	14	0
1CLV	2	1	6	0
1D6R	299	0	384	0
1DFJ	10	0	25	0
1EAW	377	0	2	2
1EWY	3	2	16	0
1F6M	1	3	1	3
1FLE	1	2	10	0
1JIW	24	0	23	0
1JK9	6	0	3	1
1M10	1	3	1	3
1OPH	112	0	16	0
1PPE	5	1	1	4
1PXV	24	0	7	0
1TMQ	3	2	5	1
1UDI	1	4	1	4
2A9K	2	1	3	2
2OOB	37	0	61	0
2OUL	4	2	2	2
2SIC	1	3	3	2
2SNI	1	4	2	3
3SGQ	39	0	35	0
7CEI	29	0	7	0
∅ enzyme/inhibitor	42.78	1.35	27.30	1.17
'other' complexes				
1AK4	69	0	17	0
1B6C	15	0	13	0
1E96	29	0	50	0
1FFW	8	0	3	1
1FQJ	62	0	93	0
1GCQ	61	0	11	0
1GHQ	22	0	142	0
1GPW	1	4	1	3
1GRN	66	0	3	1
1HE1	1	2	1	2
1IBR	2	1	1	3
1IRA	24	0	8	0
1J2J	1	1	15	0
1KAC	17	0	51	0
1LFD	21	0	17	0
1MQ8	173	0	465	0
1QA9	149	0	146	0
1SYX	343	0	176	0
1XQS	122	0	61	0
1Z0K	6	0	1	4
2HRK	8	0	9	0
3CPH	27	0	95	0
3D5S	1	4	1	5
∅ 'other' complexes	53.39	0.52	60.00	0.83
∅ all complexes	48.09	0.93	43.65	1.00

Table 2.7: Ranking obtained after rescoring the BD force field with the full MJ matrix.

ratio of 1.83, i.e. the energy contribution with the bigger 5 Å side chain sizes is 1.83 higher than for the new side chain sizes. To get again the same contribution

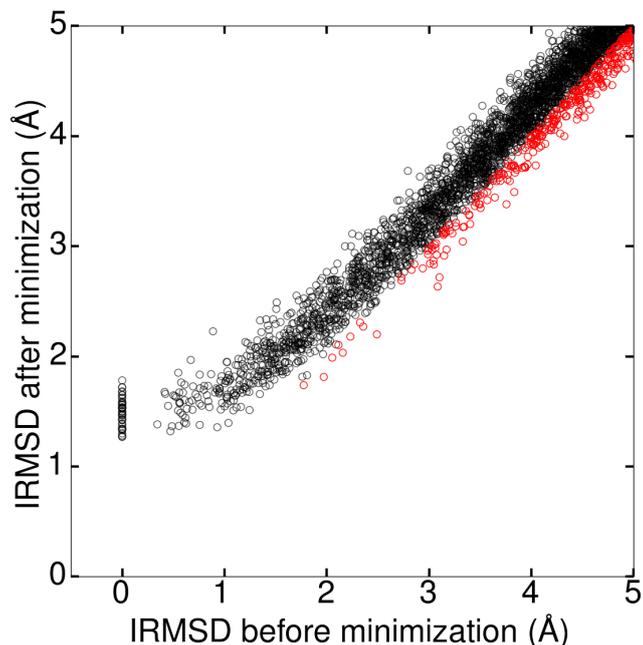


Figure 2.11: The IRMSD change due to minimization with the BD force field using the full MJ matrix: Black dots correspond to decoys where the IRMSD increases, red dots indicate an IRMSD decrease.

to the potential energy from side chains we set  $\epsilon_{hp} = 1.85 \text{ kcal/mol} \cdot 1.83 = 3.37 \text{ kcal/mol}$ .

### Ranking

In table 2.8 the results for reranking with the BD force field using the full MJ matrix and OPEP side chains are shown. If the predictions are ranked by the full complex energy, for five complexes a hit is ranked at position one. If the predictions are reranked by the interface energy, for nine complexes a hit is ranked on position one. In table 2.7, showing the ranking for equal sized side chains, the complex energy ranking resulted in more top 1 hits, while the ranking by the interface energy also gave nine top 1 hits. The complex energy ranks for 20 complexes hits in the top 20, which is less than the 25 obtained in table 2.7. The interface energy ranking ranks for 31 complexes a hit in the top 20, similar to the original BD force field. The ranking by the complex energy leads to three funnels, which are clearly fewer than the eight funnels obtained with the previous parametrizations. The ranking by the interface energy leads to nine funnels, the same number as before. The full complex energy in table 2.8 ranks for 31

	Full Complex Energy		Interface Energy	
	Rank first hit	#top5 lower 4 Å	Rank first hit	#top5 lower 4 Å
enzyme/inhibitor				
1BVN	38	0	11	0
1CLV	3	1	16	0
1D6R	66	0	358	0
1DFJ	2	1	4	1
1EAW	645	0	4	1
1EWY	43	0	29	0
1F6M	6	0	1	1
1FLE	7	0	12	0
1JIW	23	0	13	0
1JK9	8	0	4	1
1M10	30	0	1	1
1OPH	114	0	6	0
1PPE	32	0	4	2
1PXV	39	0	1	2
1TMQ	1	1	4	1
1UDI	8	0	2	2
2A9K	4	1	6	0
2OOB	49	0	23	0
2OUL	1	5	1	5
2SIC	1	3	1	4
2SNI	1	3	2	3
3SGQ	60	0	47	0
7CEI	155	0	20	0
∅ enzyme/inhibitor	58.087	0.65	24.78	1.04
'other' Complexes				
1AK4	5	1	16	0
1B6C	110	0	176	0
1E96	41	0	35	0
1FFW	13	0	21	0
1FQJ	64	0	22	0
1GCQ	27	0	8	0
1GHQ	691	0	437	0
1GPW	11	0	1	2
1GRN	60	0	6	0
1HE1	51	0	1	1
1IBR	6	0	1	4
1IRA	2	1	6	0
1J2J	22	0	16	0
1KAC	33	0	17	0
1LFD	21	0	83	0
1MQ8	421	0	167	0
1QA9	252	0	61	0
1SYX	3	1	52	0
1XQS	140	0	83	0
1Z0K	1	2	1	3
2HRK	105	0	68	0
3CPH	4	2	3	1
3D5S	17	0	4	2
∅ 'other' complexes	91.30	0.30	55.87	0.57
∅ all complexes	74.70	0.48	40.33	0.80

Table 2.8: Ranking obtained after rescoring with the BD force field with the full MJ matrix and different side chain sizes.

predictions a hit on a higher position than for the parametrization with equal side chain sizes, for three predictions it resolves the same rank and for twelve

predictions it ranks a hit on a better position. When only the interface energy is used, the ranking can improved for 20 targets, worsen for 19 and the ranking for seven targets is the same compared to parametrization with original side chain sizes and full MJ matrix parametrization. The ranking by the interface energy has the best average ranking compared to all other rescoring functions.

### Change of the RMSD

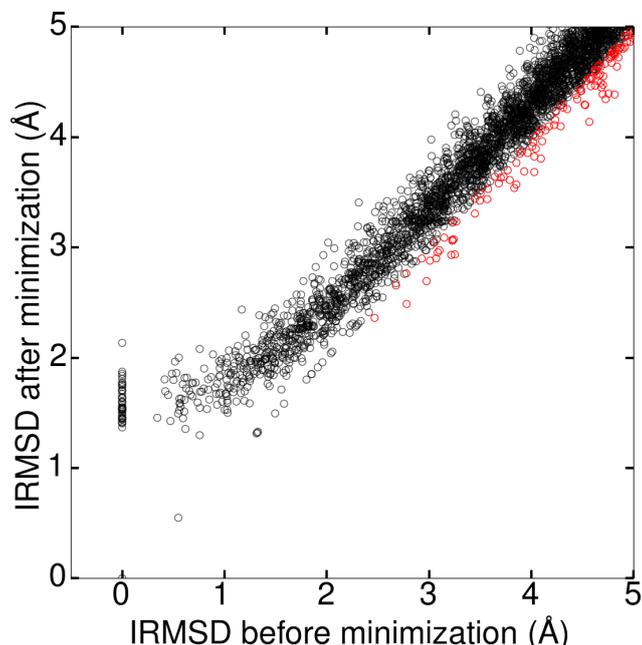


Figure 2.12: The IRMSD change due to minimization with the BD force field using the parametrization with the side chain topology of OPEP and the full MJ matrix: Black dots correspond to decoys where the IRMSD increases, red dots indicate an IRMSD decrease.

The change of the IRMSD during minimization is shown for  $\text{IRMSD} \leq 5 \text{ \AA}$  before minimization. For most complexes the IRMSD got worse during minimization. For predictions with an IRMSD lower than  $5 \text{ \AA}$  before minimization the IRMSD increases or stays the same for 96.92 % of the predictions, what is  $\approx 5 \%$  higher than for the other parametrizations of the BD force field.

## 2.4 Discussion

We tested the BD force fields for protein protein docking. The docking poses were created by ZDOCK and different parametrizations of the BD force field were introduced and their performance in rescoring the ZDOCK predictions was evaluated. The IRMSD for over 90 % of the predictions increases due to energy minimization, yet the LRMSD increases for less than 60 % of the predictions by the minimization. The reason for the more frequent increase of the IRMSD compared to the LRMSD is the overlap between the binding partners introduced by ZDOCK. If the secondary structure change of the interface is ignored the IRMSD increases for only 38.55 % of the predictions with an IRMSD up to 5 Å before minimization. This shows that an optimization of the quaternary structure by the minimization is possible.

We further showed that application of the full MJ matrix parametrizes the force field better than the approximation of the MJ matrix in the original BD force field. By introducing the full MJ matrix into the force field, the performance could be improved for ranking by the complex energy and the interface energy. In the next step we changed the side chain positions and sizes to the parameters used in OPEP. While the ranking by the complex energy worsened, the ranking by the interface energy further improved.

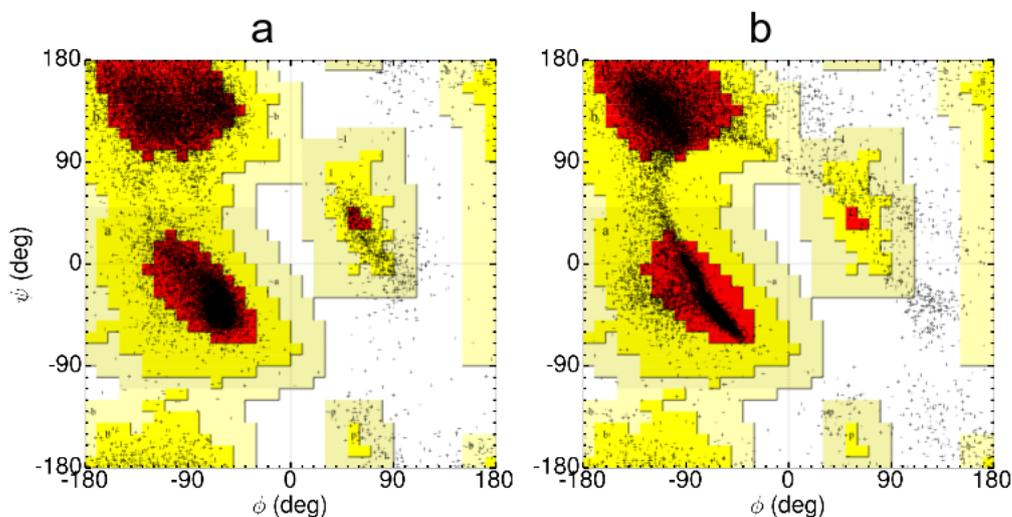


Figure 2.13: Ramachandran-plots for the 46 complexes in their bound conformation. (a) The Ramachandran-plot for the crystal structures. (b) The Ramachandran-plot after minimization with the BD force field. The colored map, showing the core, allowed and generous regions for the dihedral angle, was taken from PROCHECK [49].

In comparison to the ranking by ZDOCK, which ranks for 32 complexes a

hit on position one, the original BD force field ranks at maximum for twelve predictions a hit on position one. However, applying the side chain topology used by OPEP decreased the number of top one hits but improved the average ranking. Moreover another shortcoming of the BD force field is that the distribution of the dihedral angles is getting worse by the minimization (figure 2.13). Comparing to the expected distribution for the Ramachandran angles as introduced by Morris et al. ([50]) the percentage of dihedral pairs that are in the core region decreases from 77.16 % to 74.96 % by the minimization of the 46 bound complexes. The number of dihedral pairs that are in a disallowed region increases from 1.16 % to 3.31 %. This finding and the performance improvement by implementing OPEP topologies for the side chains led us to the decision to abandon the BD force field and to perform our further studies with the OPEP force field.

## Chapter 3

# Evaluation of the coarse-grained OPEP force field for protein-protein docking

The results in this chapter were published in:

Philipp Kynast, Philippe Derreumaux, and Birgit Strodel. Evaluation of the coarse-grained OPEP force field for protein-protein docking. *BMC biophysics*, 9:4, 2016.

Complete execution of the programming, simulations, data evaluation and wrote 90 % of the manuscript.



RESEARCH ARTICLE

Open Access



# Evaluation of the coarse-grained OPEP force field for protein-protein docking

Philipp Kynast<sup>1</sup>, Philippe Derreumaux<sup>2,3,4</sup> and Birgit Strodel<sup>1,5\*</sup>

## Abstract

**Background:** Knowing the binding site of protein–protein complexes helps understand their function and shows possible regulation sites. The ultimate goal of protein–protein docking is the prediction of the three-dimensional structure of a protein–protein complex. Docking itself only produces plausible candidate structures, which must be ranked using scoring functions to identify the structures that are most likely to occur in nature.

**Methods:** In this work, we rescore rigid body protein–protein predictions using the optimized potential for efficient structure prediction (OPEP), which is a coarse-grained force field. Using a force field based on continuous functions rather than a grid-based scoring function allows the introduction of protein flexibility during the docking procedure. First, we produce protein–protein predictions using ZDOCK, and after energy minimization via OPEP we rank them using an OPEP-based soft rescoring function. We also train the rescoring function for different complex classes and demonstrate its improved performance for an independent dataset.

**Results:** The trained rescoring function produces a better ranking than ZDOCK for more than 50 % of targets, rising to over 70 % when considering only enzyme/inhibitor complexes.

**Conclusions:** This study demonstrates for the first time that energy functions derived from the coarse-grained OPEP force field can be employed to rescore predictions for protein–protein complexes.

**Keywords:** Protein–protein docking, Coarse graining, Rescoring, Flexible docking

## Background

One of the main goals of proteomic research is to understand the biological function of proteins. Many proteins generate their function not as monomers but as part of complexes. Thus knowledge about protein–protein interactions is fundamental and allows regulation of protein structure and function. The Protein Data Bank (PDB) [1] contains more than one hundred thousand protein structures. However, structures of protein–protein complexes are often difficult to determine experimentally. These complexes are usually very big, which is a problem for elucidating structure via nuclear magnetic resonance (NMR), and the interactions are often too transient to be captured by X-ray crystallography.

Protein-protein docking is an *in silico* method for predicting the structures of protein–protein complexes. One can predict possible binding sites in a complex based on the protein structures in their unbound state. The binding partners can be single proteins or smaller protein–protein complexes. To increase computing efficiency, the proteins are usually modelled as rigid bodies at the first six-dimensional (6D) global search stage. Most of these global search methods are based on the convolution of grids, where the surface of the binding partners are parametrized such that an overlap between the surfaces of the two binding partners becomes possible. The aim of this surface description is to implicitly account for conformational changes upon binding. The convolution of the grids is accelerated by fast Fourier transformation (FFT) [2–5]. In the simplest approach, the convolution produces possible docking positions based solely on the shape of the proteins. However, more sophisticated grid maps exist which take chemical and knowledge-based properties into

\*Correspondence: b.strodel@fz-juelich.de

<sup>1</sup> Institute of Complex Systems: Structural Biochemistry (ICS-6), Forschungszentrum Jülich GmbH, 52425 Jülich, Germany

<sup>5</sup> Institute of Theoretical and Computational Chemistry, Heinrich Heine University Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf, Germany  
Full list of author information is available at the end of the article

account. For refining the initial predictions, various methods are commonly applied, for instance Monte Carlo (MC) simulations [6, 7], clustering [8, 9], or side-chain optimization using rotamer libraries [10]. As computation time is usually the limiting factor, an MC simulation should start from a conformation close to the binding site. A complete global search with this method in a reasonable computing time would be impossible.

The global search, which is performed via ZDOCK in this study [11], usually finds many similar solutions [4]. Therefore, it is common practice to cluster and rerank the docking predictions. Reranking classifies and distinguishes native or near-native solutions from non-native or wrong predictions [12, 13]. The number of predictions in a cluster can also be used for reranking [14]. The aim of both approaches is to narrow down the list of possible interaction sites, significantly decreasing computational cost and effort for further analysis of the remaining docking predictions.

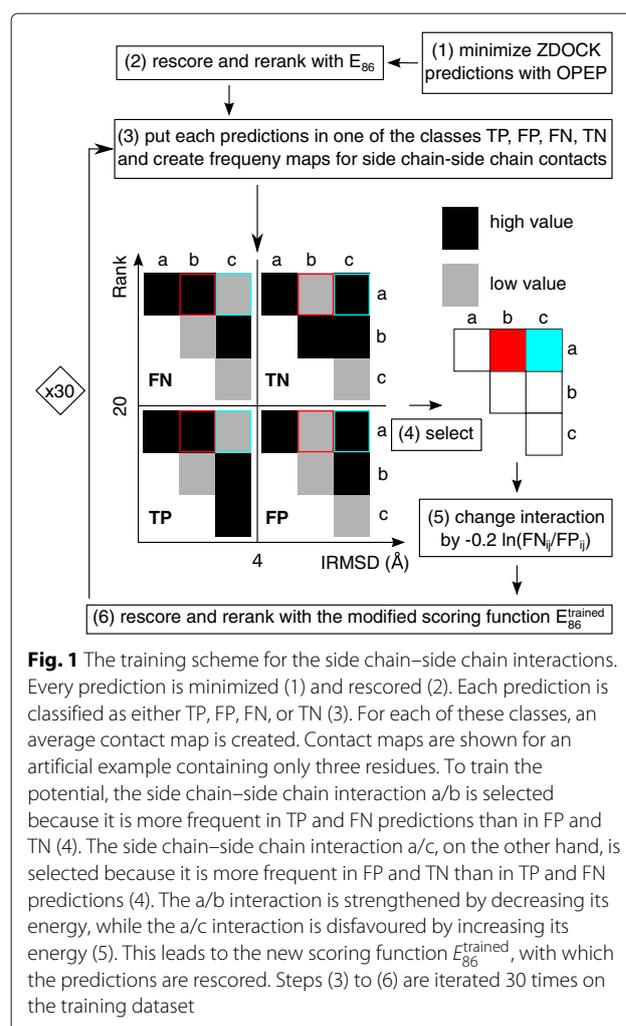
To investigate protein–protein complexes produced by ZDOCK, docking approaches that allow for more protein flexibility than ZDOCK with low time expenditure are needed. A coarse-grained force field should be a good choice here. Various coarse-grained force fields have already been developed for the treatment of protein–protein complexes, including the calculation of thermodynamic and structural properties of multi-protein complexes with relatively low binding affinities [15]. Coarse-grained models are also used for molecular dynamics (MD) simulations of protein–protein association [16, 17], where the proteins are modelled using the MARTINI force field [18, 19] or with a Go-model approach [20]. In the latter approach [17], the electrostatic and hydrophobic interactions between proteins are modelled via a Coulomb potential with a distance dependent dielectric constant and the Miyazawa-Jernigan potential [21].

In the current study, we apply the coarse-grained ‘Optimized Potential for Efficient structure Prediction’ (OPEP) [22] to the protein–protein docking problem. A coarse-grained force field is used because of the reduced number of degrees of freedom, making it computationally more efficient than an all atom potential. Moreover, it is believed that a coarse-grained model will smooth the underlying free energy landscape, facilitating exploration of the corresponding phase space [23]. OPEP has already been successfully employed with different techniques, including MD and MC simulations. It was applied to RNA/DNA/protein systems to investigate the effect of crowding, to amyloid formation, and for protein 3D structure prediction. A recent overview of OPEP and its applications can be found in [22]. This work investigates OPEP’s applicability to protein–protein complexes. To test its performance for protein–protein docking, the first step

is to investigate the discriminating power of OPEP to distinguish between correctly and wrongly docked complexes. We use global docking predictions produced by ZDOCK which we coarse grain and energy minimize using OPEP, followed by rescoring with an OPEP-based soft potential. Moreover, we enhance the performance of the rescoring function via an iterative learning procedure and test the resulting scoring function on a subset of the Dockground benchmark [24].

## Methods

We perform unbound docking, which starts from the binding partners in their native conformations. The methods applied for predicting and rescoring protein–protein complexes can be summarized via the following pipeline: For each of the 96 targets we produce 54,000 docking predictions with ZDOCK and retain the best 2000 of these complexes, as recommended by the ZDOCK developers. These predictions are energy minimized using the OPEP force field (step (1) in Fig. 1). For each prediction we



**Fig. 1** The training scheme for the side chain–side chain interactions. Every prediction is minimized (1) and rescored (2). Each prediction is classified as either TP, FP, FN, or TN (3). For each of these classes, an average contact map is created. Contact maps are shown for an artificial example containing only three residues. To train the potential, the side chain–side chain interaction a/b is selected because it is more frequent in TP and FN predictions than in FP and TN (4). The side chain–side chain interaction a/c, on the other hand, is selected because it is more frequent in FP and TN than in TP and FN predictions (4). The a/b interaction is strengthened by decreasing its energy, while the a/c interaction is disfavoured by increasing its energy (5). This leads to the new scoring function  $E_{86}^{\text{trained}}$ , with which the predictions are rescored. Steps (3) to (6) are iterated 30 times on the training dataset

perform 140 minimization steps in full Cartesian space with the limited-memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) minimizer [25], which leads to minimization times between 3.5 s for the target with PDB ID 1AY7 (185 amino acids) and 250 s for the target with PDB ID 2HMI (1413 amino acids) on a single CPU core. This amounts to an overall minimization time for the 2000 ZDOCK predictions per target of less than 24 h for 85 % of targets. Afterwards, the minimized predictions are reranked. For this, we replaced the side chain–side chain interaction potential of OPEP with a softer 8-6 Lennard-Jones-potential, while preserving the optimal distances and energies (step (2) in Fig. 1). At this stage, the OPEP potentials for salt bridges, interactions involving backbone atoms, and H-bonds are not changed. In a further step, we trained the parameters of side chain–side chain interactions, including salt bridge interactions with an iterative learning approach with the aim of further improving the performance of the OPEP-based rescoring function (steps (3)–(6) in Fig. 1). The resulting scoring function is tested on another dataset to independently prove its ability to distinguish between native and non-native complexes.

#### The dataset

We use two different benchmarks to perform unbound docking. ZDOCK benchmark 4.0 is used as training dataset, while for further evaluation we use the Dockground benchmark 2.0. We used a subset of ZDOCK benchmark 4.0 [26]. We downloaded the docking predictions for 6° angular sampling from the ZDOCK website, which were obtained using ZDOCK 3.02 [27]. Ninety-six complexes were selected, including 39 enzyme/inhibitor, 19 antigen/antibody, and 38 other types of complexes. The latter will be called ‘other complexes’ for the remainder of this paper. One condition for selecting these complexes is that ZDOCK found at least one hit in the top 2000 predictions. A hit is defined as a prediction with an interface root mean square deviation (IRMSD) from the target of lower than 4 Å. Complexes that contain small molecules like ATP and GTP, for which OPEP is not parametrized, were not considered. The 1N2C complex could not be used, because it has more than 15,000 beads after coarse graining and the fixed file format for parametrization in OPEP currently only allows for up to 9999 beads.

The second dataset is a subset of the Dockground benchmark [24]. Here we follow the same selection criteria as for the ZDOCK benchmark. Furthermore, we remove complexes present in ZDOCK benchmark 4.0 in order to generate an independent and unbiased test set. The resulting test set contains 74 targets with 18 enzyme/inhibitor, 16 antigen/antibody, and 40 other complexes. As before, to generate complex predictions we applied ZDOCK with 6° sampling, using a local ZDOCK 3.02 installation and keeping the top 2000

predictions. As in the ZDOCK dataset, the docking for the antigen/antibody complexes was restricted to the complementarity determining regions (CDRs).

#### ZDOCK

ZDOCK is an FFT-based rigid-body protein–protein docking algorithm. During the search procedure one protein is kept fixed, while the other is moved around it. The fixed protein is usually the larger of the two and is called the receptor, while the other protein is the ligand. ZDOCK generates grid-based representations from the full atom chains of receptor and ligand and after each ligand rotation the grids can be fast convoluted via FFT. The three rotational angles of the ligand are sampled with a 6° spacing, and the 3 translational degrees of freedom are sampled with a 1.2 Å spacing. For each set of rotational angles, only the best (based on ZDOCK score) translationally sampled prediction is retained [28]. This leads to 54,000 ZDOCK predictions, of which we consider the top 2000 for further refinement. To account for some flexibility in ZDOCK, a soft docking approach is used where the receptor has a 3.4 Å thick surface layer [3]. This allows for some overlap between receptor and ligand and accounts for possible movements during docking. However, it may also lead to atom clashes between receptor and ligand. The ZDOCK scoring function contains a shape-complementary term [29], a knowledge-based contact term for atoms and residues [11], and an electrostatic term [30].

#### Missing residues and atoms

Some of the complex structures considered are missing certain residues in the receptor and/or ligand. Although this is no problem for a grid-based method like ZDOCK, it must be resolved for treatment with OPEP. Missing residues lead to gaps in the backbone chain and, if untreated, they would be considered overstretched bonds. In order to resolve this problem, polypeptides with missing residues are treated as separate chains. The distance between the terminal carbon and the terminal nitrogen of the gap is kept fixed via a harmonic potential with the equilibrium distance equal to the initial gap length and a force constant of 100 kcal/(mol · Å<sup>2</sup>).

#### OPEP

As rescoring function we use the coarse-grained potential OPEP or variations of it. OPEP uses a six bead representation for every amino acid except proline and glycine. The amino nitrogen N, the C<sub>α</sub>, and the carbonyl carbon C' atoms of the backbone are each modelled by one bead. In addition, the hydrogen H of the amino-group and carbonyl oxygen O are explicitly represented. Side chains are described by only one bead, except for proline where all

heavy side chain atoms are modelled. The local energy terms in OPEP were developed based on the functional form of the Amber force field [31] and several rounds of minor adaptations to the side chain–side chain interactions have been conducted [22]. We use the latest version of OPEP, OPEPv5 [32], which for the first time includes an explicit potential for salt bridges that were parametrized with an iterative Boltzmann inversion method with parameters extracted from all atom MD simulations. A complete description of the OPEP potential can be found in the original OPEP publications [22, 31–33]. Here, we only present the nonbonded interactions, as they are used to rescore the protein–protein complexes. The nonbonded potential consists of four terms: (1) van der Waals interactions involving backbone atoms ( $E_{VDW}$ ), (2) hydrophobic and hydrophilic side chain–side chain interactions ( $E_{SS}$ ), (3) hydrogen bond (H-bond) interactions between backbone atoms ( $E_{HB}$ ), and (4) a potential for salt bridges ( $E_{SB}$ ). Interactions between side chains  $E_{SS}$  are modelled differently for attractive and repulsive interactions [34]:

$$E_{SS} = \begin{cases} \epsilon_{ij} \left[ \left( \frac{G(r_{ij}^0)}{(r_{ij}^0)^6} \right) e^{-2r_{ij}} + 0.656 \tanh(2(r_{ij} - r_{ij}^0 - 0.5)) - 1 \right], & \text{if } \epsilon_{ij} > 0 \quad (1) \\ \frac{-\epsilon_{ij}}{r_{ij}^8} & , \text{if } \epsilon_{ij} \leq 0 \quad (2) \end{cases}$$

where  $r_{ij}$  is the distance between interacting beads  $i$  and  $j$ , the equilibrium distance  $\sigma_{ij}$  is correlated with  $r_{ij}^0$  via

$$\sigma_{ij} \approx 1.0729r_{ij}^0 - 0.3992, \quad (3)$$

$\epsilon_{ij}$  is the interaction strength, and

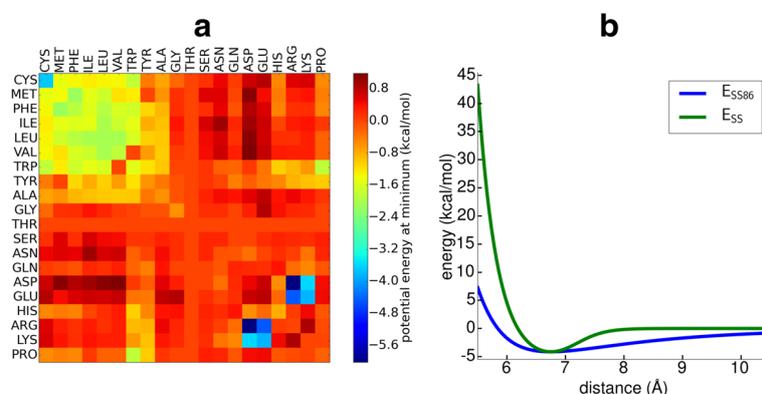
$$G(r_{ij}^0) = \left[ -0.7e^{(2(r_{ij}^0 - 0.5)/5.0)} (r_{ij}^0 - 0.5) \right]^6 \quad (4)$$

Figure 2a shows a matrix of the energies of the side chain–side chain interactions at the minimum distances  $\sigma_{ij}$ . Equation (1) replaces the common 12-6 Lennard-Jones potential in order to limit  $E_{SS}$  at longer distances.

Figure 2b shows an example of the form of the potential for the Phe/Phe interaction. For proline and glycine the center of interaction is the  $C_\alpha$ -atom, while for all other side chains the interaction center is a bead representing the center of mass of the side chain [33]. The potential  $E_{SS}$  is not used for salt bridges between side chains. Instead, salt bridges are modelled with a potential,  $E_{SB}$ , derived from all atom MD simulations [32], where the distance dependent contact probability is translated to free energy profiles. These free energy profiles have one minimum for Arg/Asp and Arg/Glu pairs and two minima for Lys/Asp and Lys/Glu interactions. To describe backbone–backbone and backbone–side chain interactions, OPEP contains a van der Waals term,  $E_{VDW}$ , which is modelled via a 12-6 Lennard-Jones potential. H-Bond interactions,  $E_{HB}$ , are modelled between the backbone N-H and the backbone C'-O atoms. In addition, OPEP has special terms for stabilizing  $\alpha$ -helices and  $\beta$ -sheets. The two-body term for H-bonds between residues in the same chain has different equilibrium distances for H-bonds less than five residues apart and for H-bonds further than four residues apart. For stabilizing  $\alpha$ -helices, the intra-chain potentials also contain a 4-body H-bond term. Furthermore, 11 side chain–side chain interactions were identified to be more frequently found in  $(i, i + 3)$  and  $(i, i + 4)$  contacts in  $\alpha$ -helices. Therefore, these side chain–side chain interactions with this particular separation were made more attractive [34].

### The scoring function

Before rescoring the predictions, we perform an energy minimization using OPEPv5 to relax the complexes after their transformation from the grid presentation to the coarse-grained model. We perform 140 minimization steps, as we found this to be the best compromise between computational efficiency and optimization result. We tested the effect of fewer and more minimization steps.



**Fig. 2** The OPEP force field. **a** The potential energy  $E_{SS} + E_{SB}$  for side chain–side chain interactions is shown at the energy minimum, which is at  $\sigma_{ij}$  for  $E_{SS}$ . For Arg/Asp and Arg/Glu, the average of the two minima for  $E_{SB}$  is shown. Repulsive interactions, corresponding to energies higher than zero, are also plotted at  $\sigma_{ij}$ . **b** The OPEP potential  $E_{SS}$  is shown together with the soft function  $E_{SS86}$  for Phe/Phe

Extending the minimization beyond 140 steps does not change the outcome of the rescoring result as for  $\sim 90\%$  of the structures the energy only changes marginally at this point. Moreover, it happens especially for misdocked complexes that the energy minimum has not been reached within 140 minimization steps. However, there is no need to further optimize such misdocked decoys. Reducing the number of minimization steps below 140 bears the risk that also near-native structures have not been properly minimized yet, which would lead to a poor ranking for them. For the scoring function we found that it becomes more reliable if we introduce a softer potential, which allows for more overlap between the beads than the original OPEPv5 energy function. To obtain a softer scoring function we replace both the side chain-side chain interaction potential,  $E_{SS}$  from Eq. (1), and the 12-6 Lennard-Jones potential  $E_{VDW}$  with an 8-6 Lennard-Jones potential. This kind of soft potential is also used in the Attract force field that was developed for protein-protein docking [35]. We call the new potentials  $E_{SS86}$  and  $E_{VDW86}$ , and the formula for  $E_{SS86}$  is given as:

$$E_{SS86} = \begin{cases} \epsilon'_{ij} \left[ \left( \frac{\sigma'_{ij}}{r_{ij}} \right)^8 - \left( \frac{\sigma'_{ij}}{r_{ij}} \right)^6 \right], & \text{if } \epsilon'_{ij} > 0 \\ \frac{-\epsilon'_{ij}}{r_{ij}^8}, & \text{if } \epsilon'_{ij} \leq 0 \end{cases} \quad (5)$$

Here,  $\sigma'_{ij} = 0.866\sigma_{ij}$  and  $\epsilon'_{ij} = 9.481E_{SS}(\sigma_{ij})$ , with  $\sigma_{ij}$  given in Eq. (3). The values  $\sigma'_{ij}$  and  $\epsilon'_{ij}$  are chosen such that the minimum energies at the equilibrium distances are identical for  $E_{SS}$  and  $E_{SS86}$ . From Eq. (6), one can see that the repulsive-only potential is not modified. An example of the attractive  $E_{SS86}$  term is shown in Fig. 2b for the Phe/Phe side chain interaction. As the 8-6 potentials  $E_{SS86}$  and  $E_{VDW86}$  have broader wells than in OPEPv5, some overlap between beads is tolerated and, in addition, imperfectly fitted contacts are more strongly attractive at larger distances. The potentials for H-bonds and salt bridges were not modified, leading to our new scoring function,  $E_{86}$ , with the modified potentials  $E_{VDW86}$  and  $E_{SS86}$ :

$$E_{86} = E_{VDW86} + E_{SS86} + E_{HB} + E_{SB}, \quad (7)$$

which calculates the binding energy between receptor and ligand for scoring purposes. It should be noted that each binding partner can consist of several proteins (chains). We consider all chains from one binding partner as a single protein. Hence, we only consider non-bonded energies between the two binding partners, e.g., between receptor and ligand.

### Interface RMSD

The interface RMSD (IRMSD) is defined as the RMSD between  $C_\alpha$  interface atoms of the co-crystallized model and the prediction after superposition. Interface  $C_\alpha$  atoms

are all atoms within 10 Å distance of the binding partner in the co-crystallized complex [36]. For the superposition we use the corresponding function from Biopython [37].

### Definition of a hit

As is standard [38, 39], we define a hit as a docked conformation with an IRMSD lower than 4 Å.

### Performance evaluation

The performance is evaluated by ranking the predictions according to their (re)scoring energy in increasing order. From this list, the best ranked prediction with an IRMSD lower than 4 Å is reported. Furthermore, we calculate the success rate, which is a function of the number of predictions,  $N_{\text{pred}}$ , that we consider from the sorted prediction list. This is averaged over the number of targets,  $N_{\text{target}}$ , and is calculated according to following equation:

$$\text{success rate}(N_{\text{pred}}) = \frac{1}{N_{\text{target}}} \sum_{i=1}^{N_{\text{target}}} S_i(N_{\text{pred}}), \quad (8)$$

where  $S_i(N_{\text{pred}}) = 1$  when the subset of  $N_{\text{pred}} = 1, 2, \dots, 2,000$  predictions contains at least one hit, otherwise  $S_i(N_{\text{pred}}) = 0$ . Thus, the success rate corresponds to the probability of finding the native complex among the  $N_{\text{pred}}$  first models based on the (re)scoring energy.

### Training the scoring function

After minimization, a residue-residue contact map between receptor and ligand is produced for each prediction. A contact is present if any of the beads of two residues are closer than 8 Å. Depending on the ranking with  $E_{86}$ , one can classify the predictions for each complex into one of the four groups: true positive (TP), false negative (FN), false positive (FP), and true negative (TN). TPs have an IRMSD  $< 4$  Å and rank lower than or equal to 20, while the TN predictions have IRMSD  $\geq 4$  Å and a rank higher than 20. All other predictions are either FNs or FPs depending on whether their IRMSD is  $<$  or  $\geq 4$  Å and their ranking is  $>$  or  $\leq 20$ . We only consider the first  $N = 20$  TPs or, if  $N < 20$  hits are found, we consider only those, because ideally one wants the correct predictions within the top hits. Twenty complexes is a small enough number for further processing by computationally more expensive approaches and visual inspection. We further limit the number of FNs and FPs to  $20 - N$  for training purposes. Thus, we do not consider FN and FP predictions if  $\geq 20$  hits are found for a target, as for such targets  $E_{86}$  already produces satisfying results. For each TP, FN, FP, and TN prediction considered, we calculate the frequency map for residue-residue contacts and average them over all targets for the enzyme/inhibitor, antigen/antibody, and other complexes. Next, we select

residue-residue contacts where the frequency is higher in the maps for TP and FN than for the FP and TN maps. We assume these contacts need to be strengthened, so current FN predictions become TP without further favoring FP predictions. Therefore, we decrease the energy value  $E_{SS86}$  or  $E_{SB}$  for this contact. The other contacts, for which we modify the potential, are those where the frequency of TPs and FNs is lower than FPs and TNs. It appears these contacts are not important for the complex class in question and should thus be disfavored, with the aim of transforming a current FP prediction into a TN prediction. Therefore, we increase  $E_{SS86}$  or  $E_{SB}$  for such contacts. Figure 1 illustrates the training procedure.

The amount of change for the selected interaction between residues  $i$  and  $j$  is determined by the ratio between the corresponding  $FN_{ij}$  and  $FP_{ij}$  frequencies. A value greater than one means this interaction energy has to be decreased, while the opposite indicates this interaction energy has to be increased. We do this by changing the interaction potentials  $E_{SS86}(i, j)$  and  $E_{SB}(i, j)$  according to

$$E_X^{\text{trained}}(i, j) = E_X^{\text{old}}(i, j) - k \ln \left( \frac{FN_{ij}}{FP_{ij}} \right), \quad (9)$$

where  $E_X = E_{SB}$  or  $E_X = E_{SS86}$  depending on the residue contact  $(i, j)$ . For the parameter  $k$ , values between 0.1 and 0.6 were tested, and  $k = 0.2$  was found to be optimal. Equation (9) was iteratively applied. Thus, we had to determine when to stop the training for best parametrization and to avoid overfitting. To this end, we performed a 4-fold cross-validation on the enzyme/inhibitor training dataset, which gives us meaningful numbers for training and validation. This enzyme/inhibitor set contains 39 targets, of which 29 complexes were used for training, with the remaining 10 used for cross-validation. For these 10 targets, we measured the quality with  $\sum_{i=1}^{10} \ln(\text{rank}(\text{target}_i))$ , where  $\text{rank}()$  returns the rank of the best ranked hit. This function should decrease during training, while an increase is indicative of overfitting. We observe that overfitting becomes an issue after 30 iterations of Eq. (9). Therefore, we set the number of learning iterations to 30, yielding our new scoring function  $E_{86}^{\text{trained}}$ .

$$E_{86}^{\text{trained}} = E_{VDW86} + E_{SS86}^{\text{trained}} + E_{HB} + E_{SB}^{\text{trained}} \quad (10)$$

## Results

### Overall performance

The ranks of the first hit using ZDOCK and after rescoring are shown in Table 1. The ZDOCK column gives the results for ZDOCK 3.02. The  $E_{86}^{\text{initial}}$  column shows

the rank after rescoring using Eq. (7) before energy minimization with the OPEP potential, while the  $E_{86}$  column reports the rank after minimization. Column five reports the rank of the first hit when using all intra- and interprotein contributions of the original OPEPv5 potential [32], while column six shows the rank of the first hit when the predictions are ranked by OPEPv5 energy when only the non-bonded energies between beads from the receptor and ligand are considered. These rescoring energies are denoted by  $E_{\text{OPEP}}$  and  $E_{\text{OPEP}}^{\text{int}}$  in the following. Figure 3 represents the success rate as defined in Eq. (8) for the different complex classes. In general, ZDOCK and  $E_{86}$  perform better than  $E_{\text{OPEP}}$  and  $E_{\text{OPEP}}^{\text{int}}$  and their performance is about equal if one considers the overall performance for all complex classes (Fig. 3a). However, there are differences between the three complex classes.

### Enzyme/inhibitor

For enzyme/inhibitor complexes,  $E_{86}$  finds equal or more hits if more than four predictions are considered, i.e.,  $N_{\text{pred}} \geq 5$  (Fig. 3b). When considering more than 50 predictions,  $E_{86}$  becomes substantially better than ZDOCK. Table 1 shows that we can improve or maintain the rank using  $E_{86}$  for 25 out of 39 enzyme/inhibitor targets. For 1AVX, the rank is only slightly worse, increasing from 1 with ZDOCK to 3 with  $E_{86}$ . Comparing the performance of  $E_{86}$  to  $E_{\text{OPEP}}^{\text{int}}$ , it becomes evident that the 140 minimization steps are not always sufficient to put every side chain in the minimum of the well, because the rank with  $E_{\text{OPEP}}^{\text{int}}$  is considerably higher than for  $E_{86}$ . Thus rescoring with the softer potential is necessary. When using  $E_{\text{OPEP}}$  for ranking, the ranks of only 16 targets are kept or improved. The average rank shows that  $E_{86}$  is generally better than ZDOCK, while  $E_{\text{OPEP}}^{\text{int}}$  produces a similar ranking to ZDOCK, and  $E_{\text{OPEP}}$  performs worst.

### Antigen/antibody

For antigen/antibody complexes, rescoring with  $E_{86}$  was least successful. For  $N_{\text{pred}} \lesssim 500$ , the success rate of  $E_{86}$  is clearly smaller than for ZDOCK (Fig. 3c). Out of 19 antigen/antibody complexes,  $E_{86}$  improves the rank for only six targets and worsens it for the other 13. Using  $E_{\text{OPEP}}$  only improves the ranking of six complexes, while the rank of only one complex can be improved using  $E_{\text{OPEP}}^{\text{int}}$ . The average rank shows that ZDOCK performs considerably better than any of the OPEP-based rescoring approaches. However, it should be noted that ZDOCK is not a perfect scoring function either for antigen/antibody complexes, as revealed by comparing the average ZDOCK ranks with enzyme/inhibitor complexes.

**Table 1** Best rank for (re)scoring with ZDOCK,  $E_{86}^{\text{initial}}$ ,  $E_{86}$ ,  $E_{\text{OPEP}}$ , and  $E_{\text{OPEP}}^{\text{int}}$  for complexes from the ZDOCK benchmark 4.0.  $\emptyset$  indicates the average rank for the complex class in question

Complex	ZDOCK	$E_{86}^{\text{initial}}$	$E_{86}$	$E_{\text{OPEP}}$	$E_{\text{OPEP}}^{\text{int}}$
Enzyme/inhibitor					
1ACB	47	11	14	1	16
1AVX	1	128	3	27	31
1AY7	330	205	358	730	858
1BVN	1	39	1	98	1
1CGI	28	52	9	2	70
1CLV	1	2	1	1	1
1DFJ	1	102	1	14	4
1EAW	332	194	151	469	124
1EZU	121	919	11	946	559
1F34	176	428	2	1664	79
1FLE	1797	179	223	35	424
1GL1	49	116	8	107	34
1GXD	229	84	2	4	47
1HIA	389	88	901	1300	1392
1IJK	1296	924	355	70	5
1JIW	1350	504	553	989	851
1JTG	1	5	1	23	9
1MAH	1	80	87	9	27
1NW9	750	105	392	857	569
1N8O	11	11	14	9	12
1OC0	225	82	240	70	83
1OPH	28	150	422	903	822
2O8V	34	20	502	112	308
1OYV	3	34	1	2	381
1PPE	1	7	1	1	2
1R0R	533	896	40	151	856
1TMQ	7	228	3	295	1
1UDI	9	446	2	158	2
1YVB	11	86	47	220	67
2ABZ	689	670	772	619	1121
2B42	1	57	40	44	192
2JOT	1730	677	179	579	178
2MTA	1	122	104	13	43
2OUL	1	30	1	23	1
2SIC	1	145	1	3	5
3SGQ	309	158	16	596	260
2UUY	258	279	65	180	361
4CPA	1	1	4	1	20
7CEI	1	30	55	51	9
$\emptyset$	275.7	212.7	143.1	291.7	251.9
Antigen/antibody					
1AHW	1387	542	1087	1161	1563
1BJ1	1	20	132	89	230
1BVK	10	143	356	6	283
1DQJ [a]	1671	150	–	–	–
1E6J	8	8	39	21	10
1FSK	5	5	181	135	373
1I9R	31	39	177	23	109
1IQD	2	41	1	4	1
1JPS	1261	221	14	1635	1930
1KXQ	6	192	34	532	9
1K4C	120	201	1077	14	166
1MLC	188	1194	104	395	314
1NCA	389	331	774	1091	1340
1VFB	45	163	198	33	586
2FD6	2	82	20	16	268

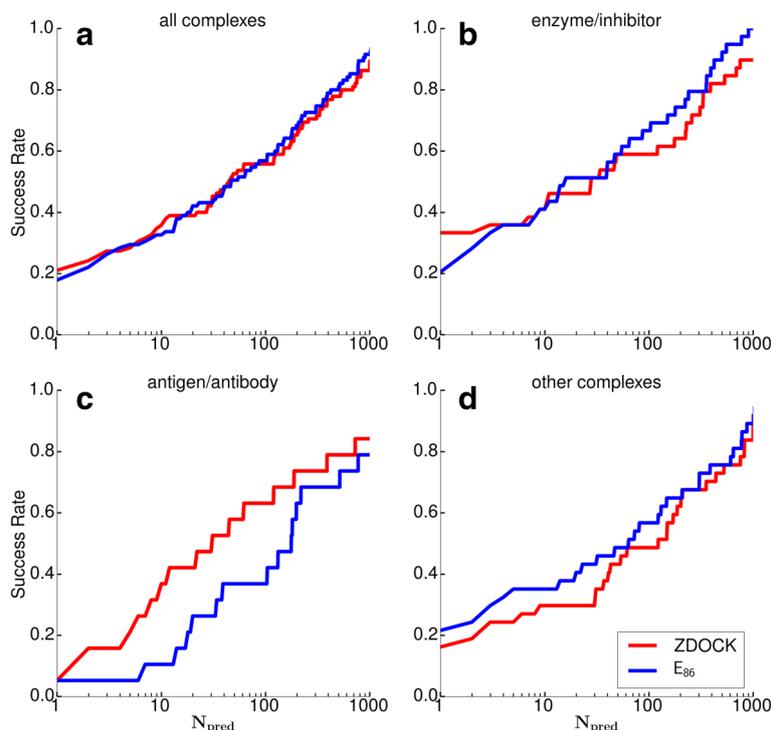
**Table 1** Best rank for (re)scoring with ZDOCK,  $E_{86}^{\text{initial}}$ ,  $E_{86}$ ,  $E_{\text{OPEP}}$ , and  $E_{\text{OPEP}}^{\text{int}}$  for complexes from the ZDOCK benchmark 4.0.  $\emptyset$  indicates the average rank for the complex class in question (continued)

2HMI	62	178	516	4	766
2I25	12	2	7	45	39
2JEL	22	85	18	49	139
2VIS	725	557	1538	625	1095
$\emptyset$	313.0	246.1	435.4	414.6	485.3
Other complexes					
1AKJ	754	47	1823	787	1887
1B6C	1	13	1	1	1
1BUH	62	25	304	10	2
1DE4	4	472	38	39	52
1F51	9	22	32	166	176
1FC2	1286	1672	1654	896	224
1FFW	41	47	47	111	433
1GLA	210	326	1013	329	1296
1GPW	3	2	1	111	997
1H9D	203	53	386	297	345
1HE1	821	1837	871	1907	1702
1I2M	354	828	3	482	262
1JK9	998	735	305	891	552
1JZD	31	141	208	3	92
1K74	1	26	1	110	717
1ML0	1	12	2	2	3
1RV6	1	188	4	1	2
1S1Q	1696	925	780	1635	932
1SYX	149	345	123	24	781
1T6B	439	224	14	130	30
1US7	150	107	772	283	341
1WDW	1	3	23	12	326
1XD3	3	25	1	1	1
1XU1	37	69	5	6	1
1Z5Y	31	1	1	1	1
1ZHI	187	43	608	156	471
2AJF	1115	318	1029	195	494
2AST	824	421	148	573	1149
2AYO	6	112	3	700	902
2CFH	1	18	1	3	3
2HLE	54	29	20	52	74
2HRK	2	76	64	30	5
2IDO	171	314	73	577	11
2NZ8	1002	281	81	793	243
2VDB	43	36	1	13	309
2Z0E	123	181	187	684	1346
3BP8	998	375	131	538	242
3D5S	524	29	1	34	246
$\emptyset$	324.6	273.1	283.1	331.1	438.2

[a] The rank is set to 2000 for calculating the average

**Other complexes**

For other complexes, the success rate is always higher for rescoring with  $E_{86}$  than scoring with ZDOCK, independent of the number of predictions considered (Fig. 3d). The  $E_{86}$  score improves or maintains the rank of 21 complexes and worsens it for the other 17; however, for 1ML0 the rank only changes from 1 to 2 and 1RV6 from 1 to 4. While  $E_{\text{OPEP}}$  improves the rank of 20 targets and worsens the rank of 18 targets, the improvements mostly occur for higher ranks, and only four predictions have



**Fig. 3** Results for the training dataset ZDOCK benchmark 4.0. The success rate is shown for (a) all complexes and separately for the three complex classes, (b) enzyme/inhibitor, (c) antigen/antibody and (d) other complexes using ZDOCK (red) and  $E_{86}$  (blue)

rank 1, compared with eight for  $E_{86}$ .  $E_{OPEP}^{int}$  can improve the rank of only 15 targets; it worsens the rank of the other 23. On average, for other complexes rescoring with  $E_{86}$  performs best,  $E_{OPEP}^{int}$  is least suited for this task, and  $E_{OPEP}$  predicts a similar ranking as ZDOCK. From the strikingly different performance of  $E_{86}$  and  $E_{OPEP}^{int}$  it seems that optimal shape complementarity implying favourable residue-residue interactions are very important for protein binding in this complex category.

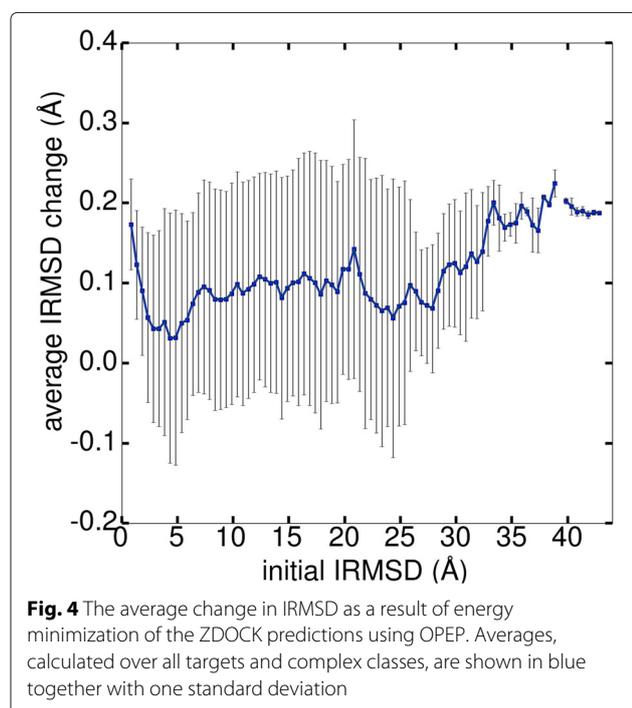
#### Structural changes upon energy minimization

We tested whether the structures of the complexes are affected as a result of energy minimization with the OPEP potential. To this end, the secondary structures of the complexes are determined before and after their energy minimization using STRIDE [40]. Since we use crystal structures of the unbound receptor and ligand as input, all 2000 ZDOCK predictions per target have the same secondary structures before minimization, while the secondary structures can change during minimization with the OPEP potential. However, we find that the changes in secondary structure are generally small (< 5%). Especially the near-native structures with IRMSD < 3 Å are least affected by energy minimization, indicating that the correct binding helps stabilize the complex structure. However, the overall changes of secondary structure are small and do not follow a pattern, which prevents

us from generalizing a dependency between IRMSD and secondary structure.

We further tested if the IRMSD is affected by minimization with OPEP and found it changes only slightly. A plot showing the average change of IRMSD as a function of the initial IRMSD as obtained from ZDOCK can be seen in Fig. 4. For most predictions, the IRMSD slightly increases due to minimization with the average IRMSD change fluctuating around 0.1 Å. For some of the complexes, the IRMSD also decreases: for 4.3 % of the predictions with IRMSD < 4 Å before minimization, which increases to 8.7 % if one considers all predictions. The preferred IRMSD increase for near-native predictions is likely to be an effect of the tight packing at the binding site, which leads to more bead clashes after transformation from the grid to the coarse-grained representation, causing the atoms or beads to reorient during minimization. Nonetheless, the structures stay close to the conformations predicted by ZDOCK, as Fig. 4 testifies. Only for severely misdocked complexes (IRMSD  $\gtrsim$  35 Å) the IRMSD change increases to around 0.2 Å.

Comparison of columns three and four of Table 1 reveals the effect of minimizing the energy before rescoring with  $E_{86}$ . Column three reports the best rank without energy minimization, which we denote as  $E_{86}^{initial}$ . For the comparison we concentrate on the complexes for which either



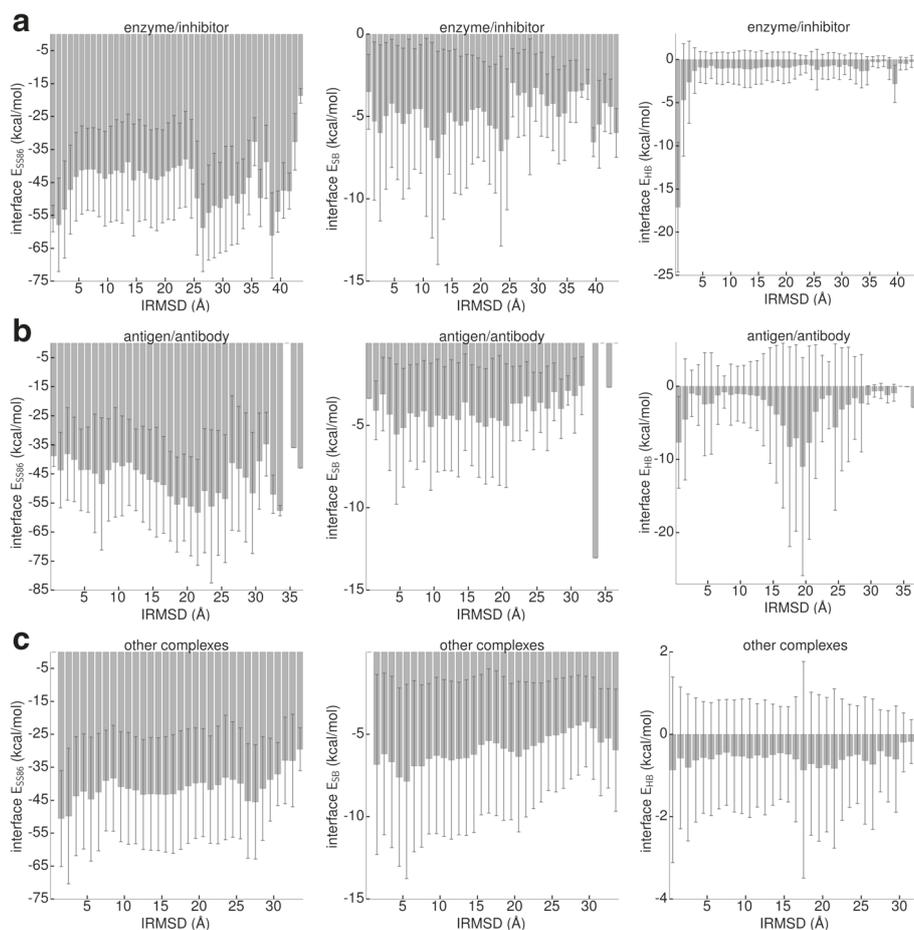
$E_{86}$  or  $E_{86}^{\text{initial}}$ , or both, predict a best rank  $\leq 10$  as in the Critical Assessment of PRedicted Interactions (CAPRI) experiment [41] one can only upload 10 predictions per target. Thus, the aim is to score the decoys closest to the native structure in the top 10. For enzyme/inhibitor complexes, energy minimization is most successful as  $E_{86}$  identifies for more than 38 % a hit in the top 10 predictions (see success rate for  $N_{\text{pred}} = 10$  in Fig. 3b). For only four of these 15 complexes (namely 1CLV, 1JTG, 1PPE and 4CPA) also  $E_{86}^{\text{initial}}$  predicts best ranks in the top 10, while it does not occur for enzyme/inhibitor complexes that  $E_{86}^{\text{initial}}$  finds a hit in the top 10, which is lost upon energy minimization. In two cases (1F34 and 1UDI) energy minimization improves the rank by more than 400 places, leading to second places in the rank list. A similar picture emerges for other complexes, for which for more than 34 % of the complexes a best rank in the top 10 is found with  $E_{86}$  (see success rate for  $N_{\text{pred}} = 10$  in Fig. 3d). With  $E_{86}^{\text{initial}}$ , on the other hand, for only three complexes a top-10 rank is achieved. For one of these three (1WDW) the rank increases from 3 to 23 upon energy minimization, while the other two are also top-10 ranked with  $E_{86}$ . Only for antigen/antibody complexes preceding energy minimization of the complexes offers no advantage over direct application of the rescoring function.  $E_{86}$  and  $E_{86}^{\text{initial}}$  find for 2 and 3, respectively, of the 19 complexes a hit in the top-10 rank list. For two complexes (1E6J and 1FSK) the top-10 rank is lost after energy minimization, while for 1IQD the best rank climbed 40

places and is ranked first with  $E_{86}$ . However, it should be noted that the average rank for  $E_{86}^{\text{initial}}$  is considerably lower than for both ZDOCK and  $E_{86}$ . Thus, energy minimization of antigen/antibody complexes is not absolutely necessary. Though apart from saving us computing time, omitting this step would also not (considerably) increase our chances of identifying the right prediction as the increase of the average rank for  $E_{86}$  originates mainly from further deterioration of the already high ranks obtained with  $E_{86}^{\text{initial}}$  (e.g., complexes 1AHW, 1K4C and 2VIS). More crucial would be a general improvement of the  $E_{86}$  scoring function for its application to antigen/antibody complexes.

#### Energy contributions to the protein-protein interactions

Figure 5 shows the different contributions to the  $E_{86}$  energy for predictions sorted by their IRMSD using a bin size of 1 Å. We show the averaged values of  $E_{\text{SS86}}$ ,  $E_{\text{SB}}$ , and  $E_{\text{HB}}$  for the three complex classes. For the enzyme/inhibitor complexes, a minimum in  $E_{\text{SS86}}$  is present for predictions up to 5 Å. However, for IRMSD values above 25 Å  $E_{\text{SS86}}$  becomes small again, in some cases even smaller than for the hits. This is more than counterbalanced by the H-bond energy, as only near-native hits have more and better oriented H-bonds, leading to  $E_{\text{HB}}$  values more than 10 kcal/mol smaller than for all other predictions. Salt bridges seem to be of minor importance for the protein binding in enzyme/inhibitor complexes, as there is no correlation between the  $E_{\text{SB}}$  values and the IRMSD, and the contribution of  $E_{\text{SB}}$  to  $E_{86}$  is generally small, with all values fluctuating around  $-5$  kcal/mol. Thus, the sum of  $E_{\text{SS86}}$  and  $E_{\text{HB}}$  is mainly responsible for distinguishing between correct and incorrect complex predictions. This partly agrees with previous findings that protease-inhibitor complexes interact predominantly through main chain-main chain interactions [42], which are represented by H-bonds in the  $E_{86}$  function.

For antigen/antibody complexes, none of the three energy contributions clearly decreases with decreasing IRMSD. Instead, both  $E_{\text{SS86}}$  and  $E_{\text{HB}}$  adopt their smallest values for IRMSD  $\approx 20$  Å, which explains why  $E_{86}$  does not perform well for this complex class. Compared to enzyme/inhibitor complexes, backbone H-bonds are less important for the native complex. This agrees with the previous observation that antigen and antibody complexes predominantly bind through side chain-side chain or side chain-main chain interactions [42], which are represented by other contributions from  $E_{86}$  but not by  $E_{\text{HB}}$ . For antigen/antibody complexes, the formation of salt bridges is also of minor importance. There is only one exception, at IRMSD  $\approx 34$  Å, where with  $E_{\text{SB}} \approx -13$  kcal/mol the smallest salt bridge energy is observed, also taking the other two complex classes into account.



**Fig. 5** The different contributions of the rescoring function  $E_{86}$  for the different complex classes, **(a)** enzyme/inhibitor, **(b)** antigen/antibody and **(c)** other complexes as a function of the IRMSD: (left)  $E_{SS86}$ , (middle)  $E_{SB}$ , (right)  $E_{HB}$ . Averages, calculated over all targets and predictions for each target belonging to one of the three complex classes, are shown together with one standard deviation

The hits for other complexes are stabilized by side chain–side chain interactions, as the lowest values for  $E_{SS86}$  are found for the complexes with  $IRMSD < 4 \text{ \AA}$ . H-bonds seem to be of minor importance for binding receptor and ligand in this complex category, as all  $E_{HB}$  values are  $> -1 \text{ kcal/mol}$ , an order of magnitude higher than those in enzyme/inhibitor and antigen/antibody complexes. On the other hand, other complexes are the only ones where salt bridges contribute to stabilizing the complexes, as for  $IRMSD > 5 \text{ \AA}$ ,  $E_{SB}$  increases. This trend only breaks for  $IRMSD \leq 5 \text{ \AA}$  as  $E_{SB}$  does not further decrease for the near-native predictions. This means that either  $E_{SS86}$  dominates these binding modes or the  $E_{86}$  potential can be further improved in this range.

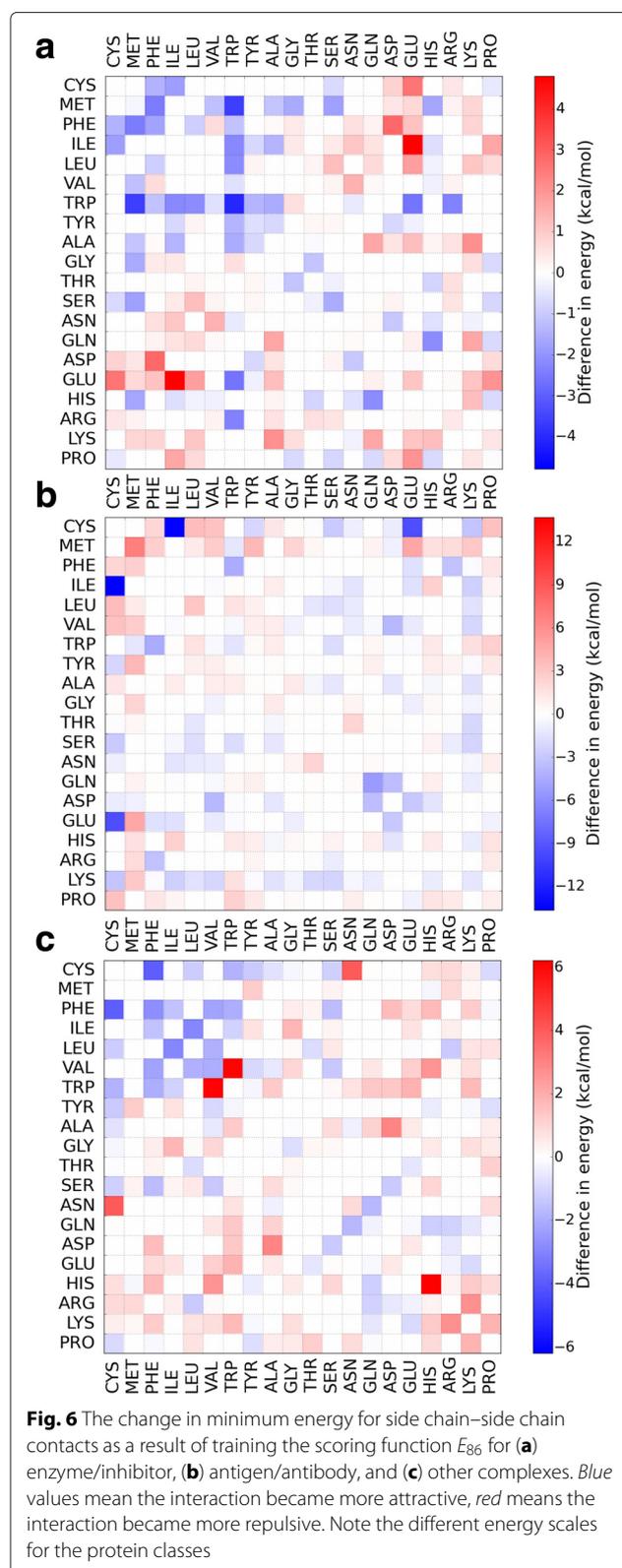
### Improving the rescoring function

Next we tested if the performance of  $E_{86}$  can be enhanced by training it according to Eq. (9), yielding the new rescoring function  $E_{86}^{\text{trained}}$  defined in Eq. (10). As the energy analysis revealed that complex formation in the three

categories is driven by different interactions, we decided to optimize  $E_{86}$  separately for enzyme/inhibitor, antigen/antibody, and other complexes. The resulting  $E_{86}^{\text{trained}}$  leads to new energies at the optimal distances between the side chains at the binding sites, which can be presented as a matrix. Subtracting the new energy matrix from the original potential energy matrix shown in Fig. 2a gives a matrix for each complex category that represents the change in interaction energies. These matrices are shown in Fig. 6.

### Enzyme/inhibitor

With few exceptions, the change in interaction energy follows the hydrophobicity of the amino acids. This confirms the findings from Fig. 5 that the enzyme/inhibitor complexes are stabilized by the interactions modeled by  $E_{SS86}$ . The amino acids Phe to Ala are the hydrophobic amino acids, and interactions between them got stronger, except for Phe/Val and Phe/Ala. Most interactions involving polar amino acids do not change much,



while some of the interactions involving charged residues become more repulsive. Previous studies also found that enzyme-inhibitor complexes contribute more hydrophobic interactions at the expense of polar contributions [42]. The most pronounced changes occur for Trp/Trp, Met/Trp, Glu/Ile, and Asp/Phe. The increased stability for Trp/Trp agrees with the Ravikant-Elber matrix [43], which was derived as the most likely interaction from a statistical analysis of protein-protein complexes. The Met/Trp interaction is also favoured by the Ravikant-Elber matrix. Both interactions were already attractive in the  $E_{86}$  rescoring function, but become even stronger as a result of the optimization procedure. Glu/Ile and Asp/Phe, on the other hand, were repulsive in  $E_{86}$  and become more so. Glu/Ile is also slightly repulsive in the Ravikant-Elber matrix, while Asp/Phe is slightly attractive. However, the Ravikant-Elber matrix includes protein-protein interactions independent of complex class, while our current finding only applies to enzyme/inhibitor complexes.

#### Antigen/antibody

Figure 6b shows the change in interaction energies for antigen/antibody complexes. Surprisingly, two interactions involving cysteine, namely Cys/Ile and Cys/Glu, considerably increase in strength. This probably results from the presence of a Cys residue just before the start of most of the CDRs [44], which is thus in contact with the antigen. The interaction between Met residues becomes the most repulsive. Before training it was slightly attractive. This change in energy is difficult to rationalize; many of the other changes are correlated to the frequency of residues at the antigen-antibody interface. At the paratope of the antibody, the residues that contribute most to the binding are Tyr, Trp, Asp, Glu, Asn, Ser, Thr, and Gly, while at the antigen epitope these are Arg, Lys, Asp, Tyr, Glu, Asn, Ser, Thr, and Gly [42]. Many of the interactions involving these residues become more attractive, while the remaining interactions do not change much in strength. This shows that our training scheme can strengthen interactions which have been previously shown to play an important role in antigen-antibody binding [42, 45].

#### Other complexes

The difference map for the other complexes can be seen in Fig. 6c. As with enzyme/inhibitor complexes, most hydrophobic interactions become more attractive. The exception is Trp/Val, for which the interaction became more repulsive. Previously, this interaction was only slightly repulsive. Almost all of the polar/hydrophobic interactions become more repulsive. Interestingly, the His/His interaction becomes considerably more repulsive, which corresponds well with repulsion of equal

charges when His is positively charged. Before training, this interaction was attractive. The repulsion between the equally charged residues Glu/Asp and Arg/Lys also increased, but these were already repulsive before optimization. The salt bridges, with the exception of Lys/Asp, got stronger. Overall, this shows that electrostatics interactions play a more important role here than in the enzyme/inhibitor complexes. It also confirms the trend from Fig. 5, which revealed a general decrease (apart from a few exceptions) for  $E_{SS86}$  and  $E_{SB}$  with decreasing IRMSD.

### Test on a new dataset

To test the optimized rescoring function, we use the protein–protein complexes from the Dockground test set [24], removing all complexes which are also in the ZDOCK Benchmark 4.0 and were already used for training. The remaining 74 complexes are listed in Table 2. As before, we perform unbound docking with ZDOCK producing 2000 predictions for each target. However, ZDOCK is not able to produce a hit for all targets in the top 2000 predictions. In particular, ZDOCK is not very successful for other complexes, generating hits for only 19 out of 40 of these complexes. However, it is successful for 15 out of 18 enzyme/inhibitor complexes and 12 out of 16 antigen/antibody complexes. For complexes for which ZDOCK produced one or more hits, the 2000 predictions are rescored using  $E_{86}$  and  $E_{86}^{\text{trained}}$ .

### Enzyme/inhibitor complexes

Both OPEP-based rescoring functions can significantly improve the average ranking compared to ZDOCK, and  $E_{86}^{\text{trained}}$  performs better than  $E_{86}$ . Compared to ZDOCK,  $E_{86}^{\text{trained}}$  can improve or maintain the rank for 11 targets and worsens the rank for 4 targets. However, for 1T6G the ranking decreased by only three places, from two to five. The standard soft rescoring function  $E_{86}$  can improve or maintain the ranking for 10 and worsens the ranking for 5 complexes. Figure 7a shows the success rate for the enzyme/inhibitor complexes. Both OPEP-based rescoring functions produce at least one hit in the top 1000 for all targets (i.e., the success rate is one for  $N_{\text{pred}} = 1000$ ), which is not the case with ZDOCK. The performance of  $E_{86}$  is weak for  $N_{\text{pred}} < 10$ , but when considering more than 10 predictions the results improve, and  $E_{86}$  performs then better than ZDOCK and similar to  $E_{86}^{\text{trained}}$ . This means the selectivity of the  $E_{86}$  function near the native complex structure is not high enough; this is improved by training the scoring function, yielding  $E_{86}^{\text{trained}}$ . For  $N_{\text{pred}} < 10$ , the performance of  $E_{86}^{\text{trained}}$  is equal or better to ZDOCK. This finding shows that training the OPEP-based scoring function was successful for the enzyme/inhibitor complex class.

### Antigen/antibody complexes

ZDOCK finds hits in the top 2000 predictions for 12 out of 16 targets.  $E_{86}^{\text{trained}}$  can improve or maintain the rank for five complexes. For 1G9M and 1S2, the rank only decreases from 1 to 3 and from 1 to 2, respectively.  $E_{86}$  performs less well, only improving the ranks of three targets and worsening them for the other nine targets. Figure 7b shows  $E_{86}$  has as many top-1 hits as ZDOCK has, but its success rate dwindles when more predictions are taken into account.  $E_{86}^{\text{trained}}$ , on the other hand, performs best when between 2 and 12 predictions are considered, yet for  $N_{\text{pred}} > 12$  ZDOCK is still most successful for antigen/antibody complexes. Nonetheless, training  $E_{86}$  was worthwhile, as for  $N_{\text{pred}} > 1$  the trained potential always performs better than or equal to  $E_{86}$ , improving the average rank by more than 120 places (see Table 2).

### Other complexes

For the other complexes,  $E_{86}^{\text{trained}}$  can (considerably) improve the average ranking compared to ZDOCK and  $E_{86}$ . Both  $E_{86}^{\text{trained}}$  and  $E_{86}$  improve the ranks of nine targets and worsen them for the other 10. However, with  $E_{86}$  the ranking of these 10 targets is considerably increased, leading to an average rank more than 120 places higher than the ZDOCK average. Figure 7c shows that  $E_{86}^{\text{trained}}$  performs slightly better than ZDOCK for  $N_{\text{pred}} > 20$ . However, the selectivity of  $E_{86}^{\text{trained}}$  should be further improved for near-native predictions, i.e., its performance should be increased for the top 20 predictions. However, this may prove difficult, as the other complexes are a collection of protein–protein complexes from different classes. Thus, the protein–protein binding may be driven by different interactions for the different complexes, making it difficult to fully accommodate all peculiarities within a scoring function.

### Medium and high accuracy predictions

In the CAPRI evaluation [41], where the predictions are made blindly (i.e., without any knowledge of the correct answer), the predicted models are classified into four categories: incorrect, acceptable, medium, and high accuracy. To this end, the combination of three parameters is used, namely the fraction of native residue–residue contacts ( $f_{\text{nat}}$ ), the RMSD of the ligand molecules in the predicted versus the target complexes (LRMSD), and the IRMSD. A detailed description of these parameters and the corresponding thresholds used in classifying predictions can be found in previous CAPRI reports [41, 46]. In this work, only the IRMSD is used to assess the quality of the predictions. Application of  $f_{\text{nat}}$  requires an atomistic representation of the predicted complexes as it is defined based on contacts between any atoms of interacting residues. Therefore, a transformation from the coarse-grained OPEP to an atomistic representation would be

**Table 2** Best rank for (re)scoring with ZDOCK,  $E_{86}$ , and  $E_{86}^{\text{trained}}$  for complexes from the Dockground 2.0 benchmark

Complex	ZDOCK	$E_{86}$	$E_{86}^{\text{trained}}$
Enzyme/inhibitor			
1ARO	-	-	-
1AVW	2 (56/-)	12 (76/-)	46 (48/-)
1BTH	366 (-/-)	218 (-/-)	106 (-/-)
1CHO	3 (3/86)	5 (5/59)	1 (1/21)
1GPQ	1271 (-/-)	438 (-/-)	109 (-/-)
1ID5	72 (-/-)	11 (-/-)	4 (-/-)
1KU6	10 (62/-)	19 (74/-)	14 (103/-)
1OFH	-	-	-
1PPF	12 (12/36)	12 (12/109)	1 (1/13)
1T6G	2 (2/579)	22 (22/311)	5 (6/121)
1TX6	539 (610/-)	28 (28/-)	87 (1090/-)
1UGH	1 (1/1098)	1 (1/104)	1 (1/44)
1XX9	279 (-/-)	17 (-/-)	8 (-/-)
2BKR	4 (4/-)	18 (24/-)	33 (33/-)
2D26	-	-	-
2FI4	335 (1287/1287)	69 (69/182)	48 (48/86)
2KAI	269 (737/-)	287 (287/-)	75 (75/-)
3SIC	1 (6/24)	1 (1/1)	1 (1/1)
∅	211.3	77.2	35.9
Antigen/antibody			
1A2Y	5 (-/-)	21 (-/-)	3 (-/-)
1G6V	1344 (-/-)	1485 (-/-)	705 (-/-)
1G9M	1 (5/-)	4 (38/-)	3 (18/-)
2BNQ	-	-	-
1BZQ	13 (13/13)	22 (22/61)	7 (43/274)
1FBI	609 (-/-)	1113 (-/-)	1174 (-/-)
1FNS	729 (-/-)	1055 (-/-)	1906 (-/-)
1H0D	159 (-/-)	17 (-/-)	1 (-/-)
1JTP	13 (-/-)	1 (-/-)	2 (-/-)
1MQ8	16 (98/-)	1479 (1548/-)	565 (807/-)
1NBY	-	-	-
1NCB	-	-	-
1NSN	562 (-/-)	695 (-/-)	949 (-/-)
1PKQ	-	-	-
1SQ2	1 (16/-)	1 (6/-)	2 (8/-)
1Z3G	6 (6/-)	1273 (1273/-)	378 (378/-)
∅	288.2	597.2	474.6
Other complexes			
1BUI	343 (-/-)	1332 (-/-)	573 (-/-)
1F6A	-	-	-
1FM9	1 (14/52)	2 (3/26)	2 (24/87)
1G20	11 (132/-)	15 (178/-)	12 (34/-)
1G4A	-	-	-
1G4U	-	-	-
1GHQ	-	-	-
1GLB	1021 (-/-)	1356 (-/-)	1352 (-/-)
1HXY	-	-	-
1JWM	-	-	-

**Table 2** Best rank for (re)scoring with ZDOCK,  $E_{86}$ , and  $E_{86}^{\text{trained}}$  for complexes from the Dockground 2.0 benchmark (continued)

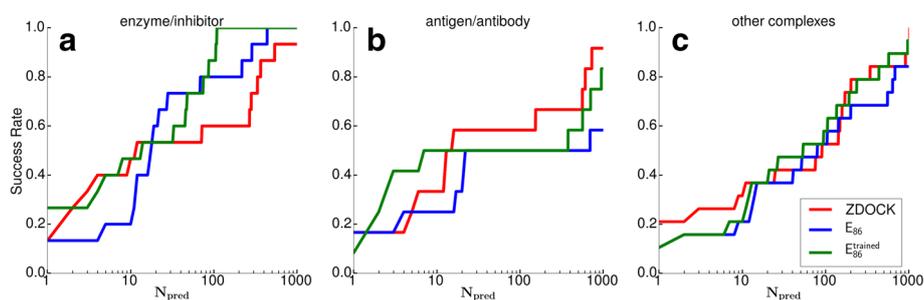
1K90	-	-	-
1K93	-	-	-
1L9B	-	-	-
1MA9	1 (4/-)	1 (3/-)	1 (1/-)
1NBF	91 (-/-)	105 (-/-)	106 (-/-)
1NVU	1020 (-/-)	555 (-/-)	192 (-/-)
1OMW	-	-	-
1OOK	171 (-/-)	639 (-/-)	237 (-/-)
1P7Q	3 (433/-)	13 (139/-)	21 (215/-)
1R4M	9 (-/-)	201 (-/-)	27 (-/-)
1RQQ	-	-	-
1S6V	1 (-/-)	9 (-/-)	7 (-/-)
1SQ0	-	-	-
1U0N	-	-	-
1U7F	149 (936/-)	1508 (1875/-)	969 (1740/-)
1UEX	25 (872/-)	14 (578/-)	11 (457/-)
1V7P	76 (-/-)	52 (-/-)	13 (-/-)
1WLI	-	-	-
1YI5	-	-	-
1ZY8	202 (325/435)	143 (527/1097)	54 (495/940)
2A42	-	-	-
2ATQ	-	-	-
2B45	-	-	-
2CKH	1 (1/21)	1 (1/178)	1 (1/68)
2G45	912 (-/-)	682 (-/-)	438 (-/-)
2GD4	-	-	-
2GOO	-	-	-
2GY7	-	-	-
3FAP	143 (-/-)	80 (-/-)	95 (-/-)
3PRO	158 (350/-)	41 (154/-)	136 (412/-)
∅	228.3	355.2	223.5

[a] The rank is set to 2,000 for calculating the average

∅ indicates the average rank for the complex class in question. Targets without a hit in the top 2,000 are indicated by '-'. Values in brackets show the best rank for predictions with an IRMSD < 2 Å and < 1 Å, respectively. If such predictions are not found, no value is being reported

first required for the calculation of  $f_{\text{nat}}$ . This would probably entail an optimization of the side chain positions so that the correct residue-residue contacts can form. While desirable, this is, however, would be beyond the scope of current study, which focuses on testing of OPEP as rescoring function for protein-protein docking. Therefore, only the IRMSD is used to classify the accuracy of predictions as high if  $\text{IRMSD} \leq 1 \text{ \AA}$ , medium if  $\text{IRMSD} \leq 2 \text{ \AA}$ , and acceptable if  $\text{IRMSD} < 4 \text{ \AA}$  [41, 46]. As we want to know whether  $E_{86}^{\text{trained}}$  finds more predictions of medium and high accuracy than  $E_{86}$ , we determined the best ranks using these thresholds for the predictions obtained for the Dockground 2.0 test set. The results are listed in Table 2, together with the ones discussed above for threshold  $\text{IRMSD} < 4 \text{ \AA}$ .

Table 2 reveals that one problem of our current approach is that ZDOCK does not produce many decoys



**Fig. 7** The success rate for the different complex classes, (a) enzyme/inhibitor, (b) antigen/antibody, and (c) other complexes for the Dockground test set when using ZDOCK (red),  $E_{86}$  (blue), and  $E_{86}^{\text{trained}}$  (green) as (re)scoring function

of medium or high accuracy in the top 2000 predictions. This is particularly the case for antigen/antibody and other complexes. For only one antigen/antibody complex (1BZQ) decoys of high accuracy are predicted by ZDOCK, while for 5 of the 16 complexes predictions of medium accuracy are produced. For other complexes ZDOCK performs even worse as for only 9 (3) of the 40 complexes predictions of medium (high) accuracy are found. The ZDOCK results are somewhat better for the 18 enzyme/inhibitor complexes with decoys of medium accuracy being found for 10 complexes and of high accuracy for 6 complexes. As in the current study only energy minimization is used to optimize the geometry of the decoys, which has only minor effects on the docking pose (IRMSD changes of around 0.1 Å only, see Fig. 4),  $E_{86}$  and  $E_{86}^{\text{trained}}$  cannot find more decoys of high or medium quality as being produced by ZDOCK. More structural refinement of the ZDOCK predictions, for instance by using MC simulations as done by RosettaDock [6, 47], would be necessary for their further improvement. Comparison of the ZDOCK,  $E_{86}$  and  $E_{86}^{\text{trained}}$  scoring of decoys of medium accuracy and with top-10 ranks shows that  $E_{86}^{\text{trained}}$  performs best for enzyme/inhibitor complexes. In this category,  $E_{86}^{\text{trained}}$  ranks the docking models for four complexes first and for a fifth complex on sixth place. Also ZDOCK has for five complexes such models ranked in the top 10, however, for none on the first place.  $E_{86}$  predicts for only three complexes top-10 ranks, however, for two of them they are on the first place. For antigen/antibody complexes, ZDOCK finds for two complexes models in the top-10 rank list, while  $E_{86}$  and  $E_{86}^{\text{trained}}$  for only one complex. For other complexes,  $E_{86}$  is slightly better than ZDOCK and  $E_{86}^{\text{trained}}$  as it has for three complexes decoys in the top-10 list, while the other two scoring functions achieve this for only two complexes.

In summary,  $E_{86}$  and  $E_{86}^{\text{trained}}$  rank docking models of medium accuracy on average better than ZDOCK (apart from antigen/antibody complexes). For complex 3SIC both OPEP-based rescoring functions even rank a high-accuracy decoy first, which ZDOCK fails to achieve for

any complex; it does not even place any decoy of high accuracy in the top 10. However, mainly due to limited refinement of the docking models obtained from ZDOCK, both  $E_{86}$  and  $E_{86}^{\text{trained}}$  do not find quantitatively more docking models of medium and high quality. This is further supported by the fact that for seven complexes (1BTH, 1XX9, 2BKR, 1G20, 1NBF, 1OOK, 3PRO), for which no decoy of high (sometimes not even of medium) accuracy and also no top-10 hit are found after rescoring, the native (i.e., target) complex is ranked first or second by  $E_{86}$  and/or  $E_{86}^{\text{trained}}$  (data not shown). In these cases rescoring with  $E_{86}$  and  $E_{86}^{\text{trained}}$  would have worked if the correct decoys had been generated. It should be noted, however, that in many other cases the native complex has a much higher rank than the other decoys, also for complexes for which top-10 predictions of medium or even high accuracy have been found. A similar observation was made by Baker and co-workers [6] when the performance of RosettaDock was for the first time tested. There, the problem was solved by performing 50 rounds of side-chain repacking and minimization. We assume that also after the transformation of the PDB structures to the coarse grained representation, energy minimization is often not sufficient for an optimal positioning of the side-chain beads. In our future work we will test whether side-chain refinement will improve the scoring of the native complexes.

## Discussion and conclusion

In this work we examined the applicability of the coarse-grained OPEP force field [22] for refining and rescoring rigid body protein–protein docking predictions. We use ZDOCK [11] to produce protein complex predictions, which also serves as quality control. The predictions from the ZDOCK benchmark 4.0 are transformed to the coarse-grained model and their energy minimized using the original OPEP potential, which is followed by rescoring with a softer energy function, denoted  $E_{86}$ , based on the interprotein OPEP interactions. This approach produces a better rank for the best prediction than

ZDOCK for 54 % of targets. However, the results differ significantly across the three complex classes: There is an improvement for 65 % of the enzyme/inhibitor complexes, for 55 % of other complexes, but for only 32 % of the antigen/antibody complexes. Furthermore, the average rank with  $E_{86}$  is for antigen/antibody complexes considerably higher than that obtained with ZDOCK. To improve these results, we developed a training scheme for the OPEP-based rescoring function based on false positive and false negative predictions. The resulting trained rescoring function called  $E_{86}^{\text{trained}}$ , which was applied to the targets from the test dataset taken from the Dockground benchmark [24], produces a lower best rank compared to the ZDOCK results for 54 % of the targets, while the untrained OPEP-based scoring function can only improve the rank of 48 % of targets. The trained scoring function performs particularly well for enzyme/inhibitor complexes, where the best rank of 73 % of targets can be improved. These figures are 47 % for other complexes, and 42 % for antigen/antibody complexes.

#### **Performance analysis for different complex classes**

Training the OPEP-based rescoring function revealed that the complexes from different classes are stabilized by different protein–protein interactions. For enzyme/inhibitor and other complexes, interactions between hydrophobic residues are of general importance, and for enzyme/inhibitor complexes backbone–backbone hydrogen bonds are also important. For antigen/antibody complexes we found that training strengthens the interactions between residues, which have been previously shown to be prevalent at the paratope of the antibody and the epitope of the antigen [42, 45]. The different performance and training potentials reflect the different protein–protein binding in enzyme/inhibitor and antigen/antibody complexes. Antibodies can recognize a wide spectrum of antigens, including proteins, polysaccharides, nucleic acids, and even lipids, while enzyme–ligand binding has developed in an evolutionary sense to enable specific binding of a ligand to its target enzyme. This diverse binding by antibodies is accommodated by the complementarity determining regions composed of six loops that are modified in shape and chemical nature to match the corresponding features of the antigen epitope. Furthermore, the paratopes are mainly discontinuous, and binding is usually mediated by only 4–13 residues. In contrast, the enzyme inhibitors are typically small proteins that form tight, substrate-like interactions with the enzyme, which is reflected in a much stronger binding energetics. The binding constants for enzyme/inhibitor complexes are in the femtomolar range, which is about six orders of magnitude smaller than the nanomolar binding constants between antigen and antibody [42]. Thus, it is not surprising that the more static and strong

enzyme–inhibitor binding is more easily predicted than the protein–protein interface in antigen/antibody complexes, where already one missing or one wrong inter-residue contact in a decoy can have a profound impact on the performance of the scoring result. Our results suggest that the collective complex class called ‘other complexes’ lies between the two ends of the spectrum bounded by enzyme/inhibitor and antigen/antibody complexes.

#### **Comparison to other rescoring approaches**

In summary, this study demonstrates for the first time that energy functions derived from the coarse-grained OPEP force field can be employed to rescore predictions for protein–protein complexes. This expands the applicability of OPEP to new problems. While the performance of OPEP is already very good for enzyme/inhibitor complexes and better than ZDOCK, for the other complexes and especially for antigen/antibody complexes, ZDOCK is still better suited. The comparison to RDOCK results [13] shows that rescoring with an all-atom force field works somewhat better than rescoring with  $E_{86}$  and  $E_{86}^{\text{trained}}$ . In RDOCK, the ZDOCK predictions are subjected to a three-stage energy minimization scheme using the CHARMM force field [48] and amounting to 130 minimization steps, followed by the rescoring based on the CHARMM electrostatic and desolvation energies. This elaborate approach improves the success rate for  $N_{\text{pred}} = 10$  (i.e., the success rate for finding a near-native hit within the first 10 predictions, as expected in the CAPRI experiment [41]) from 38 to 45 % for decoys obtained from ZDOCK(PDE), which is similar to the ZDOCK 3.02 version used in this work. In our study, the success rate for  $N_{\text{pred}} = 10$  decreases by 1–2 % after rescoring with  $E_{86}$  and  $E_{86}^{\text{trained}}$  (see Figs. 3 and 7). In case of  $E_{86}$  it is due to the poor performance of this scoring function for antigen/antibody complexes, while  $E_{86}^{\text{trained}}$  does not perform well for  $N_{\text{pred}} < 12$  for other complexes. It should be noted that also RDOCK is considerably less successful for antigen/antibody complexes compared to enzyme/inhibitor complexes, supporting our conclusion that the rescoring of the former is more challenging than that of the latter.

Comparison to other coarse-grained force fields shows that OPEP is better suited as scoring function for protein–protein docking than these. In addition to OPEP we also tested the coarse-grained force field developed by Bereau and Deserno (BD) [49] on a decoy set produced by ZDOCK consisting of 23 enzyme/inhibitor and 23 other complexes. The BD force field increased the rank of 31 complexes and decreased it for only four complexes, which is considerably worse than what we obtain with OPEP. Reasons for the failure of the BD force field when applied to protein–protein docking are that the side-chain beads have all the same size and that electrostatic

interactions between charged residues are not modelled, features that are present in OPEP. Moreover, in a study performed similarly to ours, the UNRES coarse-grained force field was tested as rescoring function [50]. The number of hits that were retained in the top-10 predictions generated by FTDock [51] decreased by more than 50 %, while with our approach the success rate decreases by only 1–2 % at  $N_{\text{pred}} = 10$ . This shows that while OPEP is still not a perfect scoring function for protein-protein docking, it is clearly better suited than other coarse-grained force fields.

### Outlook

In our future work we strive to further improve the performance of the OPEP rescoring functions. Here, special attention will be devoted to antigen/antibody complexes, where improvement is most needed. In addition, we will not only rescore the decoys generated by ZDOCK but also refine them by performing Monte Carlo simulations with OPEP. One advantage of OPEP is that it is a physics-motivated force field defined based on continuous functions and is therefore ideally suited for flexible docking. Our aim is to produce a reliable refinement and rescoring protocol based on OPEP that only needs docking decoys generated by ZDOCK or another global search algorithm as additional input. For the participation in the CAPRI experiment, however, a final transformation from the coarse-grained to the atomistic level for the top-10 decoys will become necessary as only atomistic decoys can be submitted.

### Availability of data and materials

All decoy structures as obtained from ZDOCK and after their transformation to the OPEP coarse-grained level are available upon request from the authors.

### Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

PK, PD, and BS designed the study. PK assembled input data, wrote code, ran the simulations, and analysed output data. PK and BS wrote the manuscript. All authors have read and approved the final version of the manuscript.

### Acknowledgements

We thank Holger Gohlke for fruitful discussions. P.K. gratefully acknowledges financial support from the International NRW Research School BioStruct, granted by the Ministry of Innovation, Science and Research of the State North Rhine-Westphalia, the Heinrich Heine University of Düsseldorf, and the Entrepreneur Foundation at the Heinrich Heine University of Düsseldorf.

### Author details

<sup>1</sup>Institute of Complex Systems: Structural Biochemistry (ICS-6), Forschungszentrum Jülich GmbH, 52425 Jülich, Germany. <sup>2</sup>Laboratoire de Biochimie Théorique, UPR 9080 CNRS, Institut de Biologie Physico-Chimique, 75005 Paris, France. <sup>3</sup>Institut Universitaire de France, 103 Boulevard Saint-Michel, 75005 Paris, France. <sup>4</sup>University Paris Diderot, Sorbonne Paris Cité, 75205 Paris, France. <sup>5</sup>Institute of Theoretical and Computational Chemistry, Heinrich Heine University Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf, Germany.

Received: 17 December 2015 Accepted: 21 March 2016

Published online: 21 April 2016

### References

- Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The protein data bank. *Eur J Biochem.* 1977;80(2):319–24.
- Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc Natl Acad Sci.* 1992;89(6):2195–199.
- Chen R, Weng Z. Docking unbound proteins using shape complementarity, desolvation, and electrostatics. *Proteins Struct Funct Bioinforma.* 2002;47(3):281–94.
- Garzon JI, López-Blanco JR, Pons C, Kovacs J, Abagyan R, Fernandez-Recio J, Chacon P. Frodock: a new approach for fast rotational protein-protein docking. *Bioinformatics.* 2009;25(19):2544–551.
- Kozakov D, Brenke R, Comeau SR, Vajda S. Piper: An fft-based protein docking program with pairwise potentials. *Proteins Struct Funct Bioinforma.* 2006;65(2):392–406.
- Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, Baker D, et al. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J Mol Biol.* 2003;331(1):281–300.
- Fernández-Recio J, Totrov M, Abagyan R. ICM-DISCO docking by global energy optimization with fully flexible side-chains. *Proteins Struct Funct Bioinforma.* 2003;52(1):113–7.
- Comeau SR, Gatchell DW, Vajda S, Camacho CJ. ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics.* 2004;20(1):45–50.
- Tong W, Weng Z. Clustering protein-protein docking predictions. *Conf Proc IEEE Eng Med Biol Soc.* 2004;4:2999–3002.
- Jackson RM, Gabb HA, Sternberg MJE. Rapid refinement of protein interfaces incorporating solvation: application to the docking problem 1. *J Mol Biol.* 1998;276(1):265–85.
- Vreven T, Hwang H, Weng Z. Integrating atom-based and residue-based scoring functions for protein-protein docking. *Protein Sci.* 2011;20(9):1576–1586. ZDOCK.
- Pierce B, Weng Z. Zrank: reranking protein docking predictions with an optimized energy function. *Proteins Struct Funct Bioinforma.* 2007;67(4):1078–1086.
- Li L, Chen R, Weng Z. Rdock: Refinement of rigid-body protein docking predictions. *Proteins Struct Funct Bioinforma.* 2003;53(3):693–707.
- Kozakov D, Clodfelter KH, Vajda S, Camacho CJ. Optimal clustering for detecting near-native conformations in protein docking. *Biophys J.* 2005;89(2):867–75.
- Kim YC, Hummer G. Coarse-grained models for simulations of multi-protein complexes: Application to ubiquitin binding. *J Mol Biol.* 2008;375(5):1416.
- May A, Pool R, van Dijk E, Bijlard J, Abeln S, Heringa J, Feenstra KA. Coarse-grained versus atomistic simulations: realistic interaction free energies for real proteins. *Bioinformatics.* 2014;30:326–34.
- Ravikumar KM, Huang W, Yang S. Coarse-grained simulations of protein-protein association: An energy landscape perspective. *Biophys J.* 2012;103(4):837–45.
- Marrink SJ, Risselada HJ, Yefimov S, Tieleman DP, de Vries AH. The MARTINI force field: coarse grained model for biomolecular simulations. *J Phys Chem B.* 2007;111(27):7812–824.
- Monticelli L, Kandasamy SK, Periole X, Larson RG, Tieleman DP, Marrink SJ. The MARTINI coarse-grained force field: Extension to proteins. *J Chem Theory Comput.* 2008;4(5):819–34.
- Taketomi H, Ueda Y, Gö N. Studies on protein folding, unfolding and fluctuations by computer simulation. *Int J Pept Protein Res.* 1975;7(6):445–59.
- Miyazawa S, Jernigan RL, et al. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol.* 1996;256(3):623–44.
- Sterpone F, Melchionna S, Tuffery P, Pasquali S, Mousseau N, Cragnolini T, Chebaro Y, St-Pierre JF, Kalimeri M, Barducci A, et al. The opep protein model: from single molecules, amyloid formation, crowding and hydrodynamics to dna/rna systems. *Chem Soc Rev.* 2014;43(13):4871–93.

23. Zhou J, Thorpe IF, Izvekov S, Voth GA. Coarse-grained peptide modeling using a systematic multiscale approach. *Biophys J*. 2007;92(12):4289–303.
24. Liu S, Gao Y, Vakser IA. Dockground protein–protein docking decoy set. *Bioinformatics*. 2008;24(22):2634–635.
25. Nocedal J. Updating quasi-newton matrices with limited storage. *Math Comput*. 1980;35(151):773–82.
26. Hwang H, Vreven T, Janin J, Weng Z. Protein–protein docking benchmark version 4.0. *Proteins Struct Funct Bioinforma*. 2010;78(15):3111–114.
27. Pierce BG, Hourai Y, Weng Z. Accelerating protein docking in zdock using an advanced 3d convolution library. *PLoS ONE*. 2011;6(9):24657.
28. Hwang H, Vreven T, Pierce BG, Hung JH, Weng Z. Performance of zdock and zrank in capri rounds 13–19. *Proteins Struct Funct Bioinforma*. 2010;78(15):3104–110.
29. Chen R, Weng Z. A novel shape complementarity scoring function for protein–protein docking. *Proteins Struct Funct Bioinforma*. 2003;51(3):397–408.
30. Chen R, Li L, Weng Z. Zdock: an initial-stage protein-docking algorithm. *Proteins Struct Funct Bioinforma*. 2003;52(1):80–7.
31. Derreumaux P. From polypeptide sequences to structures using monte carlo simulations and an optimized potential. *J Chem Phys*. 1999;111(5):2301–310.
32. Sterpone F, Nguyen PH, Kalimeri M, Derreumaux P. Importance of the ion-pair interactions in the opep coarse-grained force field: parametrization and validation. *J Chem Theory Comput*. 2013;9(10):4574–584.
33. Maupetit J, Tuffery P, Derreumaux P. A coarse-grained protein force field for folding and structure prediction. *Proteins Struct Funct Bioinforma*. 2007;69(2):394–408.
34. Chebaro Y, Pasquali S, Derreumaux P. The coarse-grained opep force field for non-amyloid and amyloid proteins. *J Phys Chem B*. 2012;116(30):8741–752.
35. Zacharias M. Protein–protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Sci*. 2003;12(6):1271–1282.
36. Méndez R, Leplae R, De Maria L, Wodak SJ. Assessment of blind predictions of protein–protein interactions: current status of docking methods. *Proteins Struct Funct Bioinforma*. 2003;52(1):51–67.
37. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 2009;25(11):1422–1423.
38. Viswanath S, Ravikant D, Elber R. Improving ranking of models for protein complexes with side chain modeling and atomic potentials. *Proteins Struct Funct Bioinforma*. 2013;81(4):592–606.
39. Tobi D. Designing coarse grained-and atom based-potentials for protein–protein docking. *BMC Struct Biol*. 2010;10(1):40.
40. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. *Proteins Struct Funct Bioinforma*. 1995;23(4):566–79.
41. Janin J. Assessing predictions of protein–protein interaction: The capri experiment. *Protein Sci*. 2005;14(2):278–83.
42. Jackson RM. Comparison of protein–protein interactions in serine protease-inhibitor and antibody-antigen complexes: Implications for the protein docking problem. *Protein Sci*. 1999;8(03):603–13.
43. Ravikant D, Elber R. Energy design for protein–protein interactions. *J Chem Phys*. 2011;135:065102.
44. Martin AC. Protein sequence and structure analysis of antibody variable domains. In: *Antibody engineering*. Berlin Heidelberg: Springer; 2001. p. 422–439.
45. Robin G, Sato Y, Desplancq D, Rochel N, Weiss E, Martineau P. Restricted diversity of antigen binding residues of antibodies revealed by computational alanine scanning of 227 antibody–antigen complexes. *J Mol Biol*. 2014;426(22):3729–743.
46. Méndez R, Leplae R, De Maria L, Wodak SJ. Assessment of blind predictions of protein–protein interactions: Current status of docking methods. *Proteins Struct Funct Bioinforma*. 2003;52(1):51–67.
47. Pierce B, Weng Z. A combination of rescoring and refinement significantly improves protein docking performance. *Proteins Struct Funct Bioinforma*. 2008;72(1):270–9.
48. Brooks BR, Brucoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem*. 1983;4(2):187–217.
49. Bereau T, Deserno M. Generic coarse-grained model for protein folding and aggregation. *J Chem Phys*. 2009;130(23):235106.
50. Solernou A, Fernández-Recio J. Refinement of rigid-body protein–protein docking using backbone and side-chain minimization with a coarse-grained model. *Open Access Bioinforma*. 2010;2:19–27.
51. Gabb HA, Jackson RM, Sternberg MJE. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J Mol Biol*. 1997;272(1):106–20.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)





# Chapter 4

## Monte Carlo refinement of docking predictions

### 4.1 Introduction

In the previous chapters we used ZDOCK to produce starting structures for further refinement and ranking. The ZDOCK algorithm and the OPEP force field have different energy functions and different representations of the proteins and thus prefer different conformations. This can lead to high energies even for structures that are close to the binding site because of the ruggedness of the energy landscape [51]. In addition, the coarse grained ZDOCK predictions contain overlaps which adds stress on the conformations. In this chapter we will produce our decoys using MC simulations and the OPEP force field. This will reduce the dependence on the decoy sampling algorithms and assures a continuity between sampling and rescoring. This approach, however, has the disadvantage that a global search would be computationally too expensive. Therefore, we limit the search to conformations close to the binding site and perform 1000 independent MC runs with subsequent energy minimization. This confined search has the additional advantage that one has a dense sampling of conformations close to the binding site and can examine the presence of an expected binding funnel. This kind of perturbation studies were done for the first time by Gray and co-workers for Rosetta Dock [46]. Instead of the full-atom side-chain optimization at the end of an MC run like in RosettaDock, we do a minimization with the OPEP force field. We further investigate the presence of attractors, regions in the phase space where many similar structures cluster. The binding site has the greatest breadth in the free energy landscape and should accumulate the highest number of complexes [52]. For this reason we cluster the final conformations and examine them by the number of cluster members [53].

## 4.2 Methods

We apply a multistep approach, where all steps are done using the OPEP force field or variations of it called  $E_{86}$  or  $E_{86}^{\text{soft}}$ . The approach has following steps:

1. Coarse graining of the unbound binding partners.
2. Take the initial conformation and produce 1000 random start structures.
3. Do 500 MC steps for each random start structure ( $E_{86}^{\text{soft}}$ ).
4. Minimize the energy of the final conformation of each MC run.
5. Rescore the final conformation ( $E_{86}$ ).

In step 3 the OPEP-based energy function  $E_{86}$  with an additional parameter for softening the repulsive van der Waals forces,  $E_{86}^{\text{soft}}$  is used. Rigid minimization and full flexible minimization is done, whereby for the rigid minimization  $E_{86}^{\text{soft}}$  and for the full flexible minimization the unmodified OPEP force field  $E_{\text{OPEP}}$  is used [54]. For the final rescoring in step 5 the  $E_{86}$  force field with no additional scaling of the repulsive forces is used.

### 4.2.1 Data set

The Dockground 2.0 data set is used which was already employed in chapter 3. We only use a subset of the complexes, i.e., those for which ZDOCK also found a hit. This is done for reducing the amount of complexes and to have the possibility to compare to the results in chapter 3. This leads to a data set containing 15 enzyme/inhibitor, 12 antigen/antibody and 19 ‘other’ complexes.

### 4.2.2 Generating random structure

The random structures (see step 2) are produced as following: The unbound binding partners are aligned on the bound conformation, resulting in a structure close to the bound state but with the unbound conformations of the binding partners. We take this conformation and follow a protocol from Gray and co-workers to create random structures from this conformation [46]. This is done by first randomly moving the ligand away from the receptor and then bringing the binding partners in contact again.

The random numbers are drawn from a normal distribution with a mean value of 0 and varying standard deviations. The ligand is randomly moved along the vector connecting the centers of mass of the receptor and the ligand with a standard deviation of 3 Å and with a 3 Å standard deviation along the two

perpendicular directions of this vector. Subsequently the ligand is rotated around the vector connecting the centers of mass of the binding partners with a standard deviation of  $8^\circ$ . Finally the ligand is rotated with a standard deviation of  $8^\circ$  in a random direction. It is obvious that this approach can lead to overlaps or to structures where the binding partners are not in contact anymore. To bring the structures in a weak contact we first separate the binding partners along the vector connecting their centers of mass (we chose  $100 \text{ \AA}$ ) and search for the closest pair of beads of the two binding partners that also have each a SASA greater than  $30 \text{ \AA}^2$ . Then we move the ligand along the center of mass by the distance of the two selected beads. This approach produces structures good enough for starting MC simulations. For each complex we produce 1000 decoys.

### 4.2.3 Monte Carlo simulations

MC simulations are done with rigid body movements of the ligand starting from the random structures. For each of the 1000 starting structures, we perform 500 MC steps. The translation step is drawn for each direction from a normal distribution with a mean of  $0.1 \text{ \AA}$  and a standard deviation of  $1 \text{ \AA}$ . The direction of translation is determined by the sign of a real number drawn from a uniform distribution between  $-0.5$  and  $0.5$ . The ligand is rotated around the vector connecting the centers of mass of the ligand and the receptor. In addition, it is rotated in a random direction. For both rotations the values are picked from a normal distribution with a mean value of  $1^\circ$  and a standard distribution of  $1^\circ$  and the sign is chosen from a random distribution. After each step the new energy  $E_{\text{new}}$  is compared with the previous energy  $E_{\text{old}}$ . If the value  $\Delta E = E_{\text{new}} - E_{\text{old}}$  is lower than zero, the new complex geometry is accepted. If  $\Delta E \geq 0$  the Metropolis criterion is applied and the conformation is accepted if

$$q < \exp((-\Delta E)/\beta), \quad (4.1)$$

where  $q$  is a random number drawn from a uniform distribution with values between 0 and 1, and  $\beta = 1$  was chosen, corresponding to a temperature of  $\approx 504 \text{ K}$ . The step size and temperature are not changed during the simulations.

### 4.2.4 Rigid body minimization

For rigid body minimization we implemented the methods described by Mirzaei et al. [55]. A rigid body has six dimensions of freedom, three translational degrees  $t = (x_1, x_2, x_3)$  and three rotational degrees  $r = (\alpha, \beta, \gamma)$ , where the rotation is around the three principal axes of rotation of the rigid body. Assuming that the rigid body  $X$  is centered at the origin, each particle of the body with the

coordinate  $X_n$  can then be moved by  $R^r(X_n) + t$ , where  $R^r$  is a rotation matrix for the angles  $r$  and  $t$  is the translation. The problem with the use of the Euler angles  $r = (\alpha, \beta, \gamma)$  is that this parametrization of the rotation contains singularities and an uniform sampling of these angles would not result in a uniform sampling of the rotations. This behavior prevents the use of traditional minimization methods that need an Euclidean space. In the following we show a representation of the rotation that allows us to redefine the problem as an optimization in Euclidean space.

The rotation of a rigid body can be approximated in Euclidean space by using a tangent line representation. Subsequently the change in the tangent line is mapped back with an exponential map. In the paper of Mirzaei et al. [55] they gave the example of a circle which can be approximated by a tangential line on a point on the circle. This line is in Euclidean space and the point  $\theta_N$  on the line can be mapped from the tangent space on the circle by  $\theta_N \mapsto \exp(i\theta_N) = \cos \theta_N + i \sin \theta_N$ . For applying the rotation, the map has to be applied to the original angle  $\theta_O$  by subsequently using the mapping  $\exp(i\theta_N) \exp(i\theta_O)$ . When the complex number  $\exp(i\theta_N)$  is represented in matrix form,

$$\exp(i\theta_N) = \begin{pmatrix} \cos \theta_N & -\sin \theta_N \\ \sin \theta_N & \cos \theta_N \end{pmatrix}, \quad (4.2)$$

it is obvious that the exponential mapping of the tangent space is the well known rotation matrix for the two-dimensional case.

The exponential mapping is also possible in three-dimensional space. Here the tangent space has the form

$$[\theta] = \begin{Bmatrix} 0 & -\theta_3 & \theta_2 \\ \theta_3 & 0 & -\theta_1 \\ -\theta_2 & \theta_1 & 0 \end{Bmatrix}, \quad (4.3)$$

defined by the vector  $\theta = (\theta_1, \theta_2, \theta_3)$ . The exponential mapping of the tangent space to the space of rotation groups can be done with the Rodrigues formula

$$\exp([\theta]) = I + \frac{\sin(\|\theta\|)}{\|\theta\|} [\theta] + \frac{1 - \cos(\|\theta\|)}{\|\theta\|^2} [\theta]^2, \quad (4.4)$$

where  $I$  is the identity matrix. The only needed input for formula (4.4) is the vector  $\theta$ . This vector defines a rotation around the vector  $\theta$  with an angle of  $\|\theta\|$  and  $\exp([\theta])$  is the rotation matrix  $R^\theta$ , its application to a vector in three dimensional Cartesian space will do the rotation. For using this with a minimization method like LBFGS [33], the gradient of the free parameters  $\theta_1, \theta_2, \theta_3$  is needed, thus the partial derivatives for  $\theta_i$  have to be calculated:

$$\frac{\partial \exp([\theta])}{\partial \theta_i} = \frac{\partial}{\partial \theta_i} \left( \frac{\sin(\|\theta\|)}{\|\theta\|} [\theta] + \frac{1 - \cos(\|\theta\|)}{\|\theta\|^2} [\theta]^2 \right). \quad (4.5)$$

The translation is just the sum of the gradients of all atoms  $N$

$$\nabla t = \sum_{n \in N} \nabla X_n. \quad (4.6)$$

### 4.2.5 Local minimization

After the 500 MC steps a short rigid body minimization of eight minimization steps is performed, followed by 80 minimization step with full flexibility. This minimization procedure is repeated five times for each decoy. Thus, every decoy is fully flexible minimized for  $5 \times 80 = 400$  steps and brought in contact with the binding partner by the five rigid body minimizations prior to each full minimization. Different amounts of cycles were tested and five cycles were found to be optimal, because the energy does not decrease further with more cycles. The rigid body minimization steps in between are done to keep the proteins in contact with each other and to optimize their interface energy.

### 4.2.6 The force field

We use the OPEPv5 force field [54] in the standard parametrization for every full flexible minimization. For the rigid body optimization during the MC run and also for the rigid body minimization we use the  $E_{86}^{\text{soft}}$  potential, which uses the  $E_{86}$  potential as described in chapter 3, but with the additional change that we scale the van der Waals and the side chain - side chain interactions with a factor of 0.001, when the distance  $d$  between the interacting beads  $i$  and  $j$  fulfills  $d < \sigma_{ij} = (\sigma_i + \sigma_j)/2$ , where  $\sigma_i$  and  $\sigma_j$  are the diameters of the two beads. The final conformation after minimization is rescored with the standard  $E_{86}$  potential without additional scaling.

### 4.2.7 Clustering

For clustering we use a greedy algorithm and follow the assumption that the energy minimized conformations with the lowest energy of each complex should be close to the correct binding pose. Moreover, close to the binding site should be more similar conformations than for other conformations with similar basin depth in the free energy landscape [52]. The cluster algorithm works as following:

1. Select the structure with the lowest energy.
2. Find structures that have an LORMSD  $< \text{RMSD}_{\text{cut}}$  and assign them to the cluster.
3. Remove the structures found in steps 1 and 2 from the list of structures.

4. Repeat until all structures are a member of a cluster.

The LORMSD is the ligand only RMSD between the ligand of two structures after super imposing the receptor conformations of the two structures. The LORMSD is calculated with respect to the selected low energy structure in step 1, which is considered as the center of the cluster in question.

## 4.3 Results

We are interested in how much different the random start conformations that are produced are and how they change as a result of the MC run and subsequent minimization. Furthermore, the ability to distinguish between hits and non-hits by the  $E_{86}$  energy is examined. Next, we investigate if there are binding funnels, i.e., whether it is possible to gradually improve the energy and thus to move closer to the native conformation. And another question to be answered is whether there are more structures accumulated around the crystal conformation.

### 4.3.1 Conformational changes

In figure 4.1 the change of the distribution of the IRMSD after steps 2, 3 and 4 can be seen. The algorithm for the production of the start structures creates an almost normal distributed set with most structures having an IRMSD  $\approx 6\text{\AA}$ . After the MC run the IRMSD distribution gets more rugged between 2 and 10  $\text{\AA}$ . An important finding is that the energy minimization leads to an accumulation of structures around IRMSD = 2-3  $\text{\AA}$ . In table 4.1 the number of structures after each step of the algorithm can be seen. The number of hits increases after the MC run, but also the number of structures with a large IRMSD. The number of hits becomes slightly reduced again by the minimization. Start complexes that are already a hit do have a probability of 0.56 to stay a hit during the MC runs. The minimization destroys hits in 9 % of the cases, but this effect is partially compensated by the fact that for 3 % of the structures a medium decoy is transformed to a hit. In figure 4.2 the movement of the ligand, which is quantified by the LORMSD, as a result of the MC runs is shown. For most complexes the ligand is moved by more than 10  $\text{\AA}$  (panel (a)). In panel (b) the average movement of the ligand as a function of the IRMSD of the start structures is shown. The ligand moves less if the initial structure is closer to the bound conformation. The increase of the movement correlates with the IRMSD up to 10  $\text{\AA}$  and after this reaches a plateau for values up to IRMSD = 25  $\text{\AA}$ , i.e. the attractor for the binding pose can reach up to IRMSD = 10  $\text{\AA}$ .

In figure 4.3 the change of the IRMSD by the minimization procedure is shown. The red colored circles are complexes where the IRMSD improves or stays

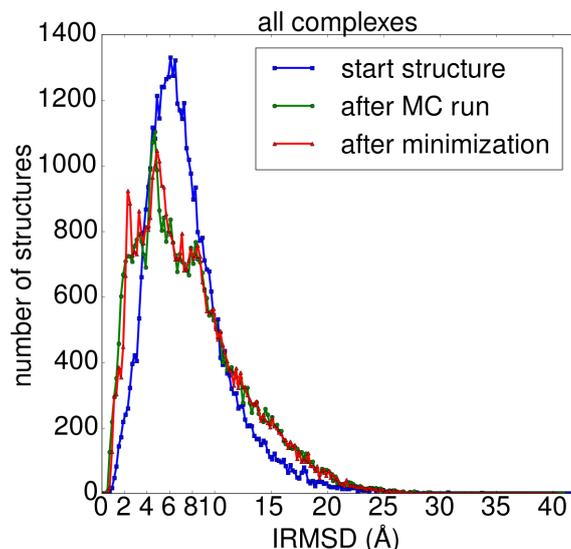


Figure 4.1: IRMSD distribution at the different stages of the algorithm for all complexes. The bin size is 0.2 Å.

		(b) After 500 MC steps		
		IRMSD < 4 Å (10497)	4 Å ≤ IRMSD < 10 Å (23163)	IRMSD ≥ 10 Å (12340)
(a) Start struc.	IRMSD < 4 Å (5748)	0.56	0.38	0.06
	4 Å ≤ IRMSD < 10 Å (31313)	0.22	0.58	0.20
	IRMSD ≥ 10 Å (8903)	0.03	0.31	0.66
(c) After min.	IRMSD < 4 Å (10236)	0.91	0.03	0.00
	4 Å ≤ IRMSD < 10 Å (23436)	0.08	0.94	0.06
	IRMSD ≥ 10 Å (12328)	0.00	0.03	0.94

Table 4.1: Number of structures for different IRMSD ranges. These numbers are given in brackets for each step of the algorithm and they include all complexes. In addition, transition probabilities between the different IRMSD ranges are given. Read it from (a) to (b) and from (b) to (c).

the same during minimization. The IRMSD decreases for 52 % of the structures and increases for 48 % of the structures. For 32 out of the 46,000 structures the IRMSD does not change. The biggest increase of the IRMSD is from 4.06 Å to 29.27 Å; this is a conformation from 1G9M where the ligand becomes tilted relative to the receptor by the minimization. For the complex 1G9M also the

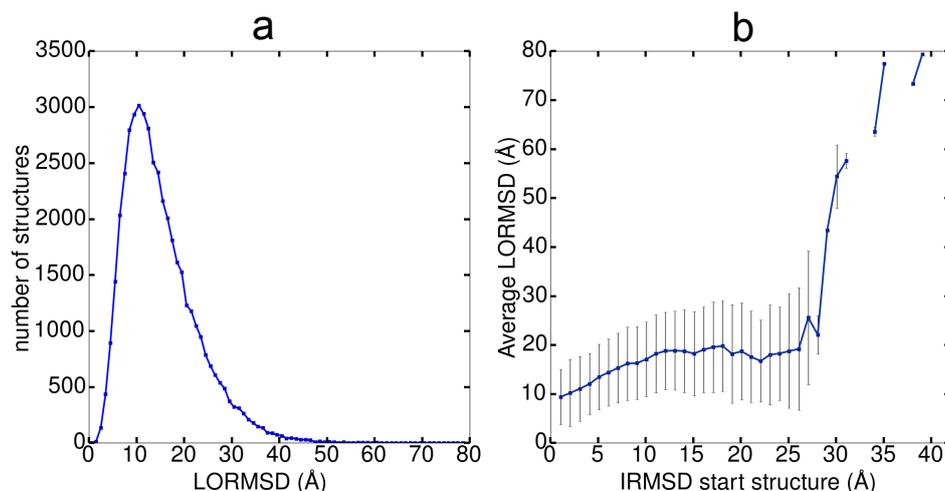


Figure 4.2: The movement of the ligand by the MC run. Panel (a) shows the distribution of the LORMSD values, while panel (b) shows the LORMSD as a function of the IRMSD of the start structure. One standard deviation is shown.

biggest IRMSD change happens, which produces a hit. It is for a structure which has an IRMSD of 11.45 Å after the MC run, that reduces to 2.66 Å after the energy minimization.

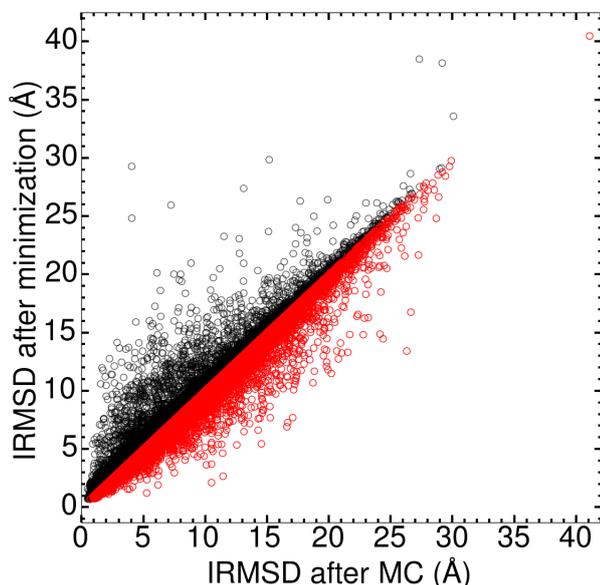


Figure 4.3: The change of the IRMSD as a result of energy minimization of decoys produced by the MC run. Red dots indicate a decrease, black dots an increase of the IRMSD after minimization.

We also measure the influence of the rigid body minimization compared to the full flexible minimization. To this end, we calculate three RMSD values.

First, the LORMSD between the structures after the MC run and after the minimization is measured:  $\text{LORMSD}_{\text{MC}/\text{Min}}$ . Next, we calculate  $\text{LORMSD}_{\text{rigid}}$ , which measures the rigid transformation of the ligand by the minimization. For this, we first replace the ligand in the minimized complex structure by the superimposed unbound structure of the ligand, and then calculate the LORMSD between this structure and that after the MC run. Here, one has to keep in mind that the ligand is still in the unbound crystal conformation after the MC run, thus the  $\text{LORMSD}_{\text{rigid}}$  measures the rigid transformation of the ligand by the minimization. Finally, we calculate  $\text{RMSD}_{\text{Min}/\text{Cryst,lig}}$ , which is the RMSD after superimposing the minimized ligand structure on the crystal ligand structure, which measures the internal structural changes of the ligand.

In figure 4.4 the distribution of the three different RMSD values is shown.  $\text{LORMSD}_{\text{MC}/\text{Min}}$  has a peak around 1.4 Å. When ignoring secondary structure changes and only considering the rigid movement, the resulting  $\text{LORMSD}_{\text{rigid}}$  has a peak around 1.3 Å.  $\text{RMSD}_{\text{Min}/\text{Cryst,lig}}$  measuring the secondary structure changes has a peak around 0.8 Å. Thus, the rigid movement is for most complexes the dominant factor for the movement of the ligand by the minimization procedure.

Another question is if the receptor or the ligand adapt conformations, which become closer to their crystal bound conformations during the docking optimization procedure. For measuring this, the RMSD of the superimposed receptor after the minimization is measured with respect to the bound receptor conformation, we call this value  $\text{RMSD}_{\text{crys,rec}}$ . The same is done for the ligand, we call this value  $\text{RMSD}_{\text{crys,lig}}$ . Moreover, to take into account the (dis)similarity between bound and unbound structures, we normalize the values

$$\text{RMSD}_{\text{rel}}^{\text{rec}} = \frac{\text{RMSD}_{\text{crys,min}}^{\text{rec}}}{\text{RMSD}^{\text{rec}}}, \quad (4.7)$$

$$\text{RMSD}_{\text{rel}}^{\text{lig}} = \frac{\text{RMSD}_{\text{crys,min}}^{\text{lig}}}{\text{RMSD}^{\text{lig}}}, \quad (4.8)$$

where  $\text{RMSD}^{\text{rec}}$  and  $\text{RMSD}^{\text{lig}}$  are the RMSDs between the bound and unbound crystal structures of the receptor and ligand, respectively. The histograms for these relative RMSD values are shown in figure 4.5. In both cases a peak close to unity is present, indicating that no adaption of the bound conformation appear. In the following the results are separately discussed for the three complex classes.

### Enzyme/inhibitor

The IRMSD distribution of the starting structures can be seen in figure 4.6. There is a peak around  $\text{IRMSD} \approx 6\text{-}7$  Å and the values range from 2.68 up to 17.63 Å. After the 500 MC steps, the distribution becomes even broader and

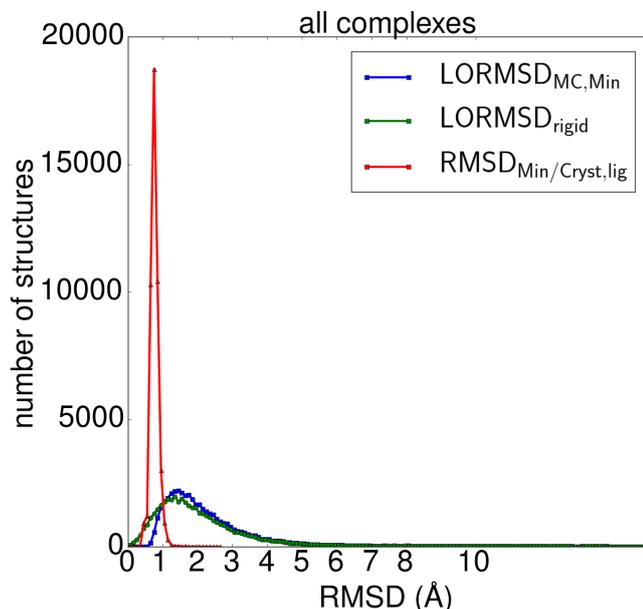


Figure 4.4: Distribution of the different RMSD values measuring the rigid movements and secondary structure changes resulting from the minimization of the complexes obtained from the MC runs. Bin size 0.1 Å.

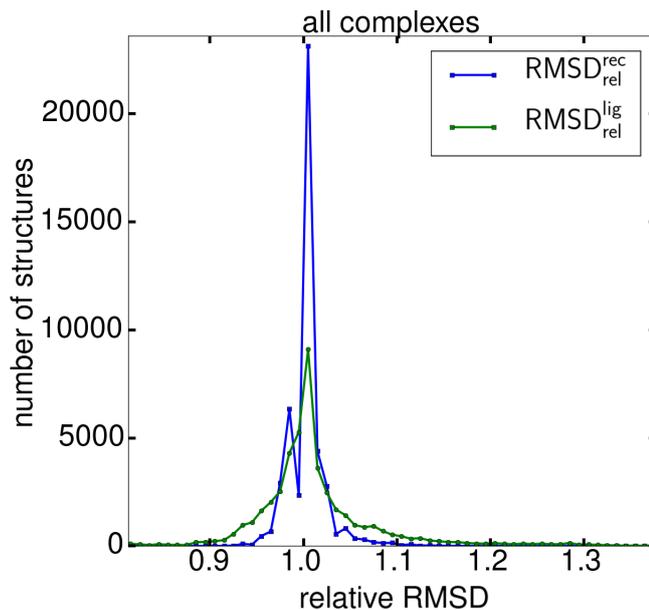


Figure 4.5: Distributions of  $\text{RMSD}_{\text{rel}}^{\text{rec}}$  and  $\text{RMSD}_{\text{rel}}^{\text{lig}}$ . Bin size 0.1.

several peaks occur in the distribution. Hits are being produced, but also more decoys further away from the bound crystal structure. The algorithm for the starting structures produced 544 structures with an IRMSD below 4 Å, after the MC run 3062 structures are below 4 Å (see table 4.2). The following minimization results in 3064 structures with  $\text{IRMSD} < 4$  Å. For starting structures far away

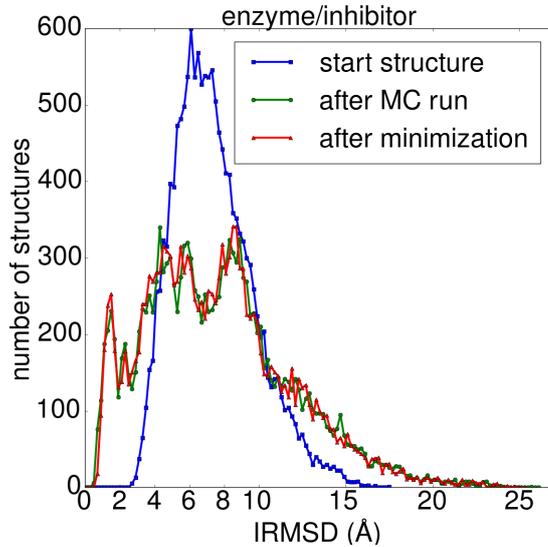


Figure 4.6: Distribution of the IRMSD values at the different stages of the algorithm for enzyme/inhibitor complexes. Bin size 0.2 Å.

		(b) After 500 MC steps		
		IRMSD < 4 Å (3062)	4 Å ≤ IRMSD < 10 Å (8147)	IRMSD ≥ 10 Å (12215)
(a) Start struc.	IRMSD < 4 Å (544)	0.55	0.37	0.08
	4 Å ≤ IRMSD < 10 Å (12452)	0.22	0.58	0.20
	IRMSD ≥ 10 Å (2004)	0.01	0.34	0.65
(c) After min.	IRMSD < 4 Å (3064)	0.92	0.03	0.00
	4 Å ≤ IRMSD < 10 Å (8222)	0.09	0.94	0.09
	IRMSD ≥ 10 Å (3714)	0.00	0.03	0.93

Table 4.2: Number of structures for different IRMSD ranges. These numbers are given in brackets for each step of the algorithm and they include only enzyme/inhibitor complexes. In addition, transition probabilities between the different IRMSD ranges are given. Read it from (a) to (b) and from (b) to (c).

from the bound crystal structure (IRMSD ≥ 10 Å) the probability to be a hit after 500 MC steps is 1 %. For around 1/3 of these conformations the IRMSD improves, but the majority (65 %) stays further than 10 Å away from the crystal structure. Medium range structures (4 - 10 Å) mainly stay in this IRMSD range, but it is slightly more probable that they become more similar to the crystal

structure than that they move further away during the MC run. Structures with an initial IRMSD lower than 4 Å are likely to remain a hit, but also 45 % of these structures move away from the crystal structure. After minimization 92 % of the hits also stay hits and another 3 % of the medium range structures become hits by the minimization, which somewhat compensates for the loss of 9 % hits due to minimization.

The quite small number of initial structures that have an IRMSD  $< 4$  Å are an indicator that the enzyme/inhibitor interfaces are not so flat and that often pockets are present, where the ligand can move into during the subsequent MC optimization procedure. The number of hits increased six-fold after the MC stage and the subsequent minimization kept the number of hits more or less fixed.

### Antigen/antibody

The algorithm for the production of random starting structures creates decoys in a range from 0.99 Å to 25.96 Å of IRMSD with a peak around 4.4 Å (figure 4.7). 3116 structures have an IRMSD lower than 4 Å, which is relatively more than for enzyme/inhibitor complexes and shows that the interfaces of antigen/antibody complexes are flatter, so that near-native binding poses can be more easily adopted. After 500 MC steps, three IRMSD peaks around 2, 3 and 5 Å appear and the range increases from 0.48 Å to 41.10 Å. The number of hits increased from 3116 to 3586 structures due to the MC run. The subsequent minimization further accumulates structures around IRMSD = 2 Å, while the peak around 5 Å decreases a bit and gets slightly broader. The number of hits decreased to 3379 by the minimization. In table 4.3 the transition probabilities between the

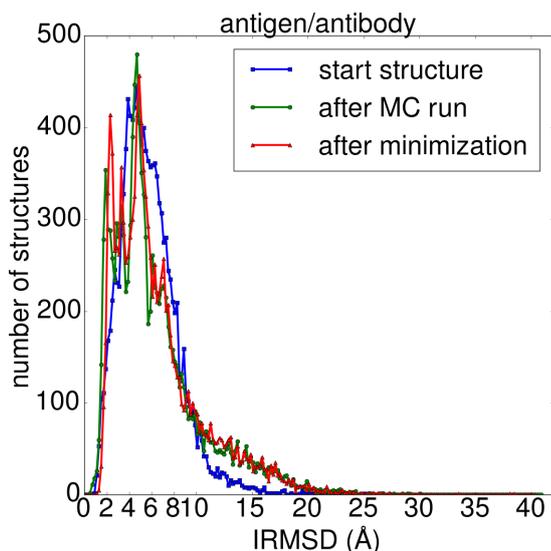


Figure 4.7: Distribution of the IRMSD values at the different stages of the algorithm for antigen/antibody complexes. Bin size 0.2 Å.

		(b) After 500 MC steps		
		IRMSD < 4 Å (3586)	4 Å ≤ IRMSD < 10 Å (6538)	IRMSD ≥ 10 Å (1876)
(a) Start struc.	IRMSD < 4 Å (3166)	0.49	0.45	0.06
	4 Å ≤ IRMSD < 10 Å (8201)	0.25	0.59	0.16
	IRMSD ≥ 10 Å (683)	0.05	0.39	0.57
(c) After min.	IRMSD < 4 Å (3379)	0.89	0.03	0.00
	4 Å ≤ IRMSD < 10 Å (6615)	0.11	0.93	0.07
	IRMSD ≥ 10 Å (2006)	0.01	0.04	0.93

Table 4.3: Number of structures for different IRMSD ranges. These numbers are given in brackets for each step of the algorithm and they include only antigen/antibody complexes. In addition, transition probabilities between the different IRMSD ranges are given. Read it from (a) to (b) and from (b) to (c).

different IRMSD ranges are shown. The probability that a hit stays a hit after the MC run is 0.49. For medium structures the probability that they become a hit is 0.25, which is the reason why the number of hits is increased by the MC run. The minimization keeps only 89 % of the decoys as hits when they were hits before, but there are still more hits than after the creation of start structures.

### ‘Other’ complexes

The distribution of the IRMSD for ‘other’ complexes at the different stages of the procedure can be seen in figure 4.8. For the start structures the IRMSD distribution follows roughly a Gaussian distribution. After the MC run, there are peaks around 2, 4 and 10 Å. The peak around 2 Å of IRMSD gets shifted to the left due to minimization and also contains more structures. The previous double peak around 10 Å becomes one broader peak. In table 4.4 the numbers of structures in the different IRMSD ranges at every stage and their exchange probabilities are shown. The start structure algorithm generates 1927 hits. After the MC run, 3451 hits are produced. After the minimization, 3391 structures are hits. For the start structures that are hits, 65 % of them stay in this IRMSD range after the MC run. The minimization removes 6 % of the hits and 2 % of the medium range structures become hits, i.e. there is less exchange between hits and medium range structures than for enzyme/inhibitor and antigen/antibody

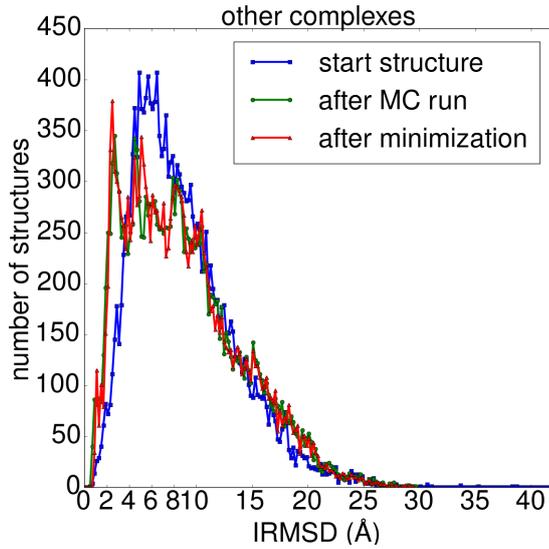


Figure 4.8: Distribution of the IRMSD values at the different stages of the algorithm for ‘other’ complexes. Bin size 0.2 Å.

		(b) After 500 MC steps		
		IRMSD < 4 Å (3451)	4 Å ≤ IRMSD < 10 Å (8001)	IRMSD ≥ 10 Å (6548)
(a) Start struc.	IRMSD < 4 Å (1927)	0.65	0.28	0.07
	4 Å ≤ IRMSD < 10 Å (9937)	0.20	0.57	0.23
	IRMSD ≥ 10 Å (6136)	0.03	0.29	0.67
(c) After min.	IRMSD < 4 Å (3391)	0.94	0.02	0.00
	4 Å ≤ IRMSD < 10 Å (8106)	0.06	0.95	0.04
	IRMSD ≥ 10 Å (6503)	0.00	0.03	0.96

Table 4.4: Number of structures for different IRMSD ranges. These numbers are given in brackets for each step of the algorithm and they include only ‘other’ complexes. In addition, transition probabilities between the different IRMSD ranges are given. Read it from (a) to (b) and from (b) to (c).

complexes.

### 4.3.2 Ranking by the energy

In this section we study if the energy can discriminate between hits and mis-docked structures. We test three different possibilities for doing the ranking: after the MC run ( $\text{rank}_{\text{MC}}$ ) using the energy  $E_{86}^{\text{soft}}$ , using the interface energy  $E_{86}$  after minimization ( $\text{rank}_{\text{min}}$ ), and taking the OPEPv5 energy of the whole complex after the minimization ( $\text{rank}_{\text{OPEP}}$ ). In the tables 4.5, 4.6 and 4.7 the best ranks for hits for the different complex classes are shown. For the enzyme/inhibitor complexes the ranking by  $\text{rank}_{\text{MC}}$  leads to the best average ranking of the hits. For antigen/antibody complexes,  $\text{rank}_{\text{OPEP}}$  leads to the best ranking, and for ‘other’ complexes,  $\text{rank}_{\text{min}}$  is the best energy function for ranking. For all complexes the average rank with  $\text{rank}_{\text{MC}}$  is 41.64, with  $\text{rank}_{\text{min}}$  the average ranking is 28.30, and with  $\text{rank}_{\text{OPEP}}$  it is 36.04.

#### Enzyme/inhibitor

complex	not minimized		minimized			
	$\text{rank}_{\text{MC}}$	IRMSD (Å)	$\text{rank}_{\text{min}}$	IRMSD (Å)	$\text{rank}_{\text{OPEP}}$	IRMSD (Å)
1AVW	2	3.85	6	3.43	11	1.90
1BTH	1	2.26	4	3.63	4	3.78
1CHO	18	3.58	4	1.24	3	2.50
1GPQ	66	3.78	65	3.77	2	3.70
1ID5	6	2.15	4	3.28	107	2.42
1KU6	5	3.38	11	3.15	1	3.59
1PPF	4	1.66	1	1.33	1	1.46
1T6G	54	3.36	29	1.69	28	3.84
1TX6	9	2.60	18	1.46	48	1.84
1UGH	1	3.17	1	2.08	2	3.82
1XX9	3	3.33	7	3.35	480	3.34
2BKR	30	2.78	94	3.93	11	2.68
2FI4	3	0.74	27	2.27	3	3.49
2KAI	30	1.10	14	1.31	67	3.37
3SIC	1	0.84	1	0.70	6	3.40
∅	15.53	2.57	19.07	2.44	51.60	3.01
rank = 1	3		3		2	
rank ≤ 10	10		8		8	

Table 4.5: Enzyme/inhibitor complexes. The position of the best ranked hit after the MC run ( $\text{rank}_{\text{MC}}$ ) and after the minimization ( $\text{rank}_{\text{min}}$ ) considering only the interface energy. For  $\text{rank}_{\text{OPEP}}$  the OPEPv5 energy after minimization is used for ranking. The IRMSD for the best ranked hits is also provided. The row ‘∅’ shows the average rank of the best ranked hits and the corresponding average IRMSD for each ranking. In the row ‘rank = 1’ and ‘rank ≤ 10’ the number of complexes with a hit with a rank = 1 or rank ≤ 10 are shown.

In table 4.5 we show the position of the best ranked hits after the MC run and after energy minimization. Both  $\text{rank}_{\text{MC}}$  and  $\text{rank}_{\text{min}}$  rank three hits on position one. For two complexes the rank does not change, but for seven complexes the

rank is better for  $\text{rank}_{\text{MC}}$  and gets worse for  $\text{rank}_{\text{min}}$ . Also the average ranking for  $\text{rank}_{\text{min}}$  increases and the number of top 10 hits decreases from ten for  $\text{rank}_{\text{MC}}$  to eight for both  $\text{rank}_{\text{min}}$  and  $\text{rank}_{\text{OPEP}}$ . The main reason for this increase is the high ranking for 2BKR. Here the best ranked structure has an IRMSD of 4.43 Å, but this is slightly higher than the threshold for a hit.

The average of  $\text{rank}_{\text{OPEP}}$  is considerably higher than those of the other two ranking schemes. The highest rank is found for 1XX9, which is also the main reason for the high average rank. The complex 1XX9 has a much lower rank if only the interface energy is considered. The complex energy of 1XX9 has an anti-correlation with the IRMSD (see section 4.3.3).

During the minimization of 1XX9 the secondary structure of the hits changes more than for the mis-docked structures. This means that the right binding site enforces a conformational change, which however causes a higher energy than for the mis-docked conformations. Conversely, the complex 1GPQ has a much better ranking by the complex energy than for  $\text{rank}_{\text{MC}}$  and  $\text{rank}_{\text{min}}$ . This is because there are two hits that have a  $\approx 10$  kcal/mol lower OPEPv5 energy than all other decoys for this complex.

### Antigen/antibody

complex	not minimized		minimized			
	$\text{rank}_{\text{MC}}$	IRMSD (Å)	$\text{rank}_{\text{min}}$	IRMSD (Å)	$\text{rank}_{\text{OPEP}}$	IRMSD (Å)
1A2Y	136	3.77	4	3.54	29	3.68
1G6V	1	3.12	4	3.18	2	3.41
1G9M	1	2.73	1	2.60	1	2.08
1BZQ	2	2.07	3	2.89	3	3.05
1FBI	29	3.45	11	3.18	1	3.88
1FNS	521	3.90	237	3.96	41	3.78
1H0D	2	3.85	2	3.88	5	2.92
1JTP	-	-	-	-	-	-
1MQ8	242	3.58	205	3.92	14	2.79
1NSN	82	3.84	59	3.99	56	3.94
1SQ2	1	3.65	1	3.37	12	3.46
1Z3G	24	2.75	2	2.59	1	3.34
∅	94.64	3.34	48.09	3.37	15.00	3.30
rank = 1	3		2		3	
rank ≤ 10	5		7		6	

Table 4.6: Antigen/antibody complexes. The position of the best ranked hit after the MC run ( $\text{rank}_{\text{MC}}$ ) and after the minimization ( $\text{rank}_{\text{min}}$ ) considering only the interface energy. For  $\text{rank}_{\text{OPEP}}$  the OPEPv5 energy after minimization is used for ranking. The IRMSD for the best ranked hits is also provided. The row ‘∅’ shows the average rank of the best ranked hits and the corresponding average IRMSD for each ranking. In the row ‘rank = 1’ and ‘rank ≤ 10’ the number of complexes with a hit with a rank = 1 or rank ≤ 10 are shown.

In table 4.6 the ranks of the best hits by the different ranking approaches are

shown for antigen/antibody complexes. The average ranks show that  $\text{rank}_{\text{OPEP}}$  performs best, followed by  $\text{rank}_{\text{min}}$ . The approaches with  $\text{rank}_{\text{OPEP}}$  and  $\text{rank}_{\text{MC}}$  predict for three complexes a hit on position one. With  $\text{rank}_{\text{min}}$  only for two complexes a hit on position one is found, but the average ranking is considerably lower than for  $\text{rank}_{\text{MC}}$  and only for two complexes the rank number increases due to minimization. Moreover, with  $\text{rank}_{\text{min}}$  for seven complexes a top 10 hit is produced while this number is five and six for  $\text{rank}_{\text{MC}}$  and  $\text{rank}_{\text{OPEP}}$ , respectively.

For the complex 1JTP no hit was found. For this complex the IRMSD of all decoys is higher than 4 Å because of the conformational changes of the secondary structure during binding. One loop of the antibody changes its conformation during complex formation, which prevents more native complex conformations being formed.

The results for complexes 1FNS and 1MQ8 are significantly better for  $\text{rank}_{\text{OPEP}}$  than for  $\text{rank}_{\text{min}}$ , due to an energy increase for the decoys around 5 Å when OPEPv5 is used (see section 4.3.3)

#### ‘Other’ complexes

In table 4.7 the rank of the hits for ‘other’ complexes are shown. For 17 of the 18 complexes a hit is found. For 1NVU no hit was produced during the MC run and also the minimization could not create a hit for this complex. The lowest average rank is found for  $\text{rank}_{\text{min}}$  followed by  $\text{rank}_{\text{MC}}$  and  $\text{rank}_{\text{OPEP}}$ . The ranking approaches  $\text{rank}_{\text{min}}$  and  $\text{rank}_{\text{OPEP}}$  identify for six complexes a hit on position one, while with  $\text{rank}_{\text{MC}}$  for only four complexes a hit on position one is found. The minimization improves the ranking by the interface energy, as for seven complexes the ranking gets better and for five it stays the same. However, the number of complexes with a hit ranked in the top 10 increased only from eleven to twelve after minimization.

For 1P7Q and 1V7P the rank is significantly better for  $\text{rank}_{\text{OPEP}}$  than for  $\text{rank}_{\text{min}}$ . The receptor of 1P7Q does not much change its conformation during minimization. To find out what the main difference in the secondary structure is, we calculated the average ligand structures for all hits with an energy lower than -1382 kcal/mol, and an average ligand structure for all hits with an average energy higher than -1382 kcal/mol. This energy is chosen because it lies in between two distinguishable energy levels in the IRMSD/energy plot (compare with section 4.3.3). Furthermore, we calculate the root mean square fluctuation (RMSF) between the two resulting average ligand structures. In figure 4.9 the hit with the lowest  $\text{rank}_{\text{OPEP}}$  is shown. Here, ligand residues with an RMSF higher than 0.1 Å are highlighted. The maximum movement between the two average structures is 0.8 Å, which happens at the  $\beta$ -strand and loop that connects the two subunits of the ligand. This rather small movement of this small section

complex	not minimized		minimized			
	rank <sub>MC</sub>	IRMSD (Å)	rank <sub>min</sub>	IRMSD (Å)	rank <sub>OPEP</sub>	IRMSD (Å)
1BUI	48	3.06	2	3.61	1	3.95
1FM9	1	3.19	1	3.53	4	1.39
1G20	3	3.84	5	2.82	5	3.41
1GLB	237	3.51	91	3.82	197	3.98
1MA9	3	3.94	1	2.16	4	3.41
1NBF	4	3.54	7	3.72	154	3.86
1NVU	-	-	-	-	-	-
1OOK	17	1.89	17	3.17	1	3.31
1P7Q	76	2.60	118	2.67	1	2.96
1R4M	3	3.80	3	3.91	151	3.75
1S6V	1	3.14	1	3.59	1	2.37
1U7F	12	3.88	3	3.34	2	2.98
1UEX	3	2.79	1	2.50	4	3.85
1V7P	74	3.92	128	3.71	11	3.60
1ZY8	3	3.44	27	2.13	1	2.12
2CKH	1	1.83	1	1.92	4	3.06
2G45	69	3.79	16	3.75	20	3.90
3FAP	1	3.00	7	3.01	1	2.38
3PRO	2	1.25	1	1.30	84	2.13
∅	31.00	3.14	23.89	3.04	35.94	3.13
rank = 1	4		6		6	
rank ≤ 10	11		12		12	

Table 4.7: ‘Other’ complexes. The position of the best ranked hit after the MC run (rank<sub>MC</sub>) and after the minimization (rank<sub>min</sub>) considering only the interface energy. For rank<sub>OPEP</sub> the OPEPv5 energy after minimization is used for ranking. The IRMSD for the best ranked hits is also provided. The row ‘∅’ shows the average rank of the best ranked hits and the corresponding average IRMSD for each ranking. In the row ‘rank = 1’ and ‘rank ≤ 10’ the number of complexes with a hit with a rank = 1 or rank ≤ 10 are shown.

leads to a significant change of the energy OPEPv5, while it does not effect the interface energy. For the complex 1V7P decoys with an IRMSD  $\approx 5$  Å have a lower interface energy than decoys with IRMSD  $< 4$  Å. This effect is also present by the scoring with  $E_{\text{OPEP}}$ , but the energy difference is lower, therefore the hits are ranked better.

### 4.3.3 Energy funnels and clustering

It is common practice to show the energy of protein-protein docking decoys as a function of the distance of these decoys to the bound conformation [56]. Here we use as distance measurement the IRMSD. In case of good docking, i.e., good sampling of complex structures and good scoring, so-called binding funnels can be seen. There is no universal definition how an ideal funnel should look like. Here we use a definition inspired by Gray and co-workers [46] and define a funnel by the number of hits in the top 5: three hits in the top 5 is assumed to indicate the presence of a funnel.

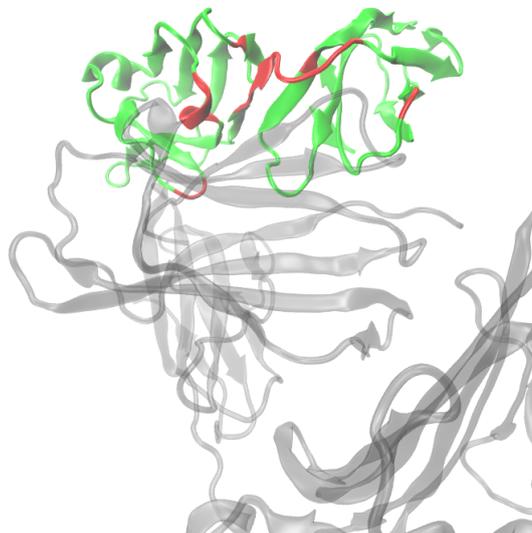


Figure 4.9: The structure with the lowest  $\text{rank}_{\text{OPEP}}$  of 1P7Q. The transparent gray structure is the receptor and the green structure is the ligand. The red sections highlight the difference in the movement for high energy hits and low energy hits in the range of 0.1 Å up to the maximum difference of 0.8 Å. The red sections are part of the interface of the protein complex.

In addition, we cluster the structures based on their LORMSDs from each other, i.e, the RMSD of the ligands after super imposing the receptors. The cluster size can give an approximation of the free energy and it is assumed that the binding site has a lower free energy than other conformations [53, 57]. The clustering of 1000 complexes by the LORMSD is a computational expensive task, because the LORMSD between all structures has to be calculated, which would be 500,500 LORMSD calculations. We can accelerate the clustering by assuming that the secondary structure does not change much during minimization and that the main difference in the LORMSD is a result of the rigid body MC run and the rigid body motion during minimization. In subsection 4.3.1 we showed that this assumption holds. This allows us to calculate the centroid for each ligand and take the distance between the centroids as a filter for clustering before we calculate the LORMSD. The LORMSD is only calculated for structures, whose centroids are separated by a distance lower than  $\text{RMSD}_{\text{cut}} + 0.5$ , we add a small value of 0.5 Å to the cutoff distance to take some internal motion into account. As a cutoff we use  $\text{RMSD}_{\text{cut}} = 4$  Å, which is the reported size for binding funnels

of antigen/antibody complexes [56], the complex group with the smallest funnels.

### Enzyme/inhibitor

In figure 4.10 one can see the IRMSD/ $E_{86}$ -plots for enzyme/inhibitor complexes. For all complexes one can observe a general trend of decreasing energy with decreasing distance to the bound protein-protein complex. But really good funnels with low energies for only the near native decoys are only present for 1PPF, 1UGH and 3SIC (compare to table 4.8). For all the other complexes, decoys with low energy exist that are further than 4 Å IRMSD away from the crystal structure. However, for the complexes 1ID5, 1BTH and 1CHO not perfect yet still acceptable funnels can be seen. For the enzyme/inhibitor complexes,

Scoring complex	#(top 5 $\wedge$ hit)	interface energy $E_{86}$		complex energy $E_{OPEP}$ #(top 5 $\wedge$ hit)
		#hits	#hits / #decoys in biggest cluster	
1AVW	0	277	134/134	0
1BTH	2	400	99/99	2
1CHO	1	296	70/70	1
1GPQ	0	59	17/43	2
1ID5	2	73	2/23	0
1KU6	0	301	4/79	5
1PPF	5	272	118/118	5
1T6G	0	97	65/68	0
1TX6	0	123	32/32	0
1UGH	5	256	57/57	2
1XX9	0	39	0/28	0
2BKR	0	136	53/54	0
2FI4	0	333	90/90	1
2KAI	0	58	0/18	0
3SIC	5	344	85/85	0
$\emptyset$	1.33	204.27	50.07/66.53	1.20

Table 4.8: Enzyme/inhibitor complexes: The second column show the number of hits among the five lowest energy decoys for  $E_{86}$ . Column three shows the total number of hits, while column four shows the number of hits and the number of decoys in the biggest cluster. In column five the number of hits in the top 5 for  $E_{OPEP}$  is shown.

none of the low energy but mis-docked decoys are part of a big cluster, which means that these complexes are not part of a well defined funnel. In table 4.8 one can see that for nine out of the 15 complexes a well defined attractor with IRMSD values lower than 4 Å exist, i.e., the biggest cluster contains mainly hits. Only for 2KAI and 1XX9 none of the complexes in the biggest cluster are a hit, but for 2KAI the second biggest and for 1XX9 the fourth biggest cluster are clusters that contain primarily hits.

For complexes that have the best ranking for the full energy in table 4.5, we show the IRMSD/ $E_{OPEP}$ -plots in figure 4.11. For 1CHO and 2FI4 a separation

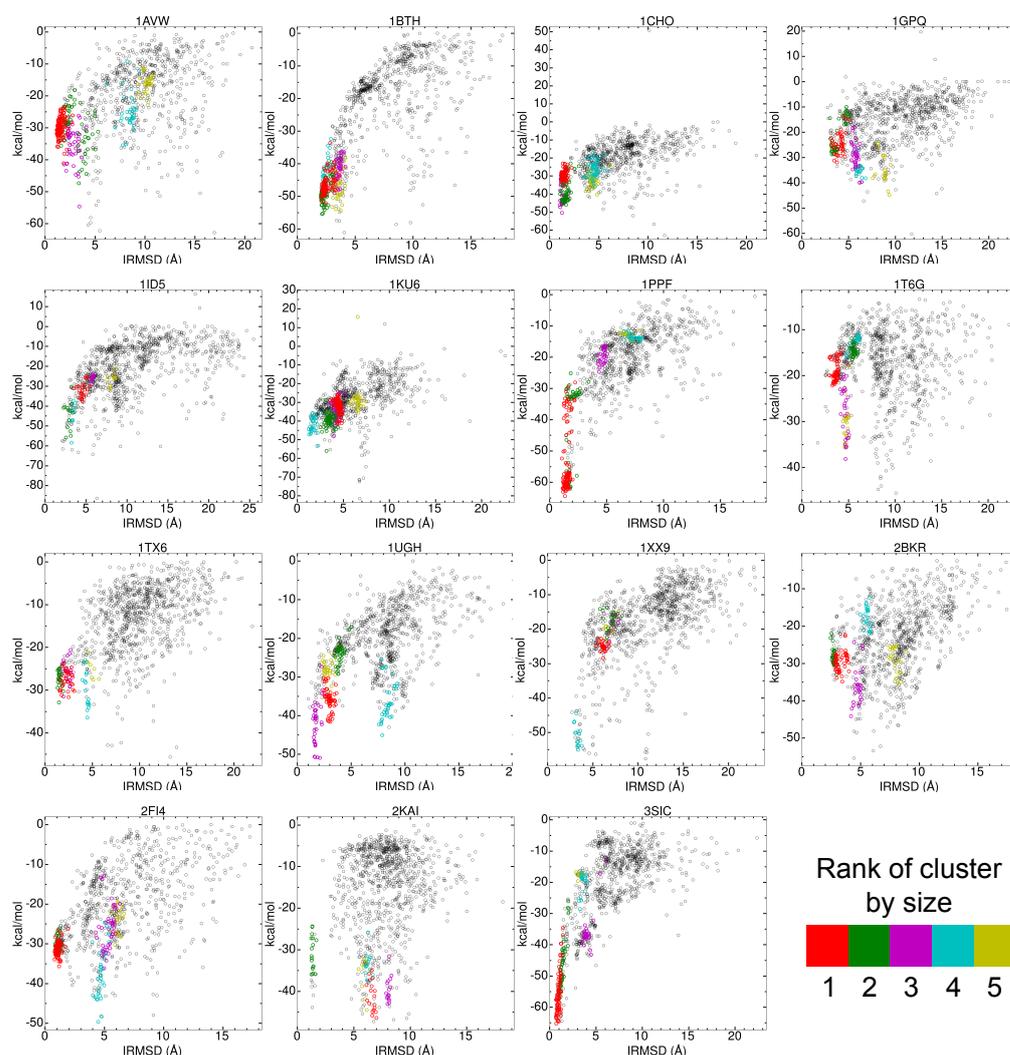


Figure 4.10: Binding funnels for enzyme/inhibitor complexes after minimization. The energy  $E_{86}$  is plotted as a function of the IRMSD. The members of the five biggest clusters are colored.

around 2.5 Å between low IRMSD and the other structures is visible. For 1CHO the better ranking with  $E_{\text{OPEP}}$  compared to  $E_{86}$  results from the fact that one single decoy with a slightly lower energy at  $\text{IRMSD} \approx 2.5$  Å exists. But there is no systematically lower  $E_{\text{OPEP}}$  energy observable for hits, and the same is true for the other enzyme/inhibitor complexes.

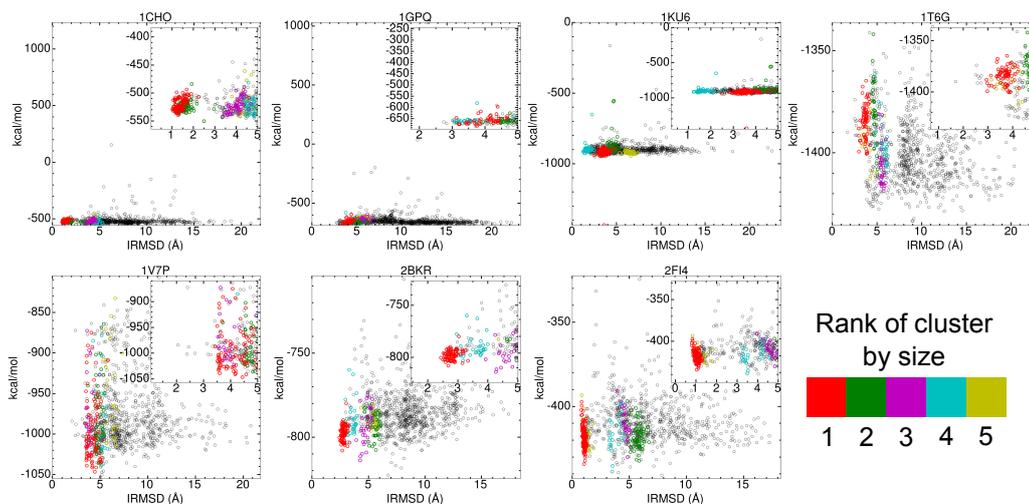


Figure 4.11: Enzyme/inhibitor complexes: The  $E_{\text{OPEP}}$  energy as a function of the IRMSD is shown for complexes where  $E_{\text{OPEP}}$  performs best for ranking. The members of the five biggest clusters are colored.

Scoring complex	#(top 5 $\wedge$ hit)	interface energy $E_{86}$		complex energy $E_{\text{OPEP}}$ #(top 5 $\wedge$ hit)
		#hits	#hits / #decoys in biggest cluster	
1A2Y	1	288	0/76	0
1G6V	2	196	0/32	1
1G9M	5	541	99/99	4
1BZQ	2	487	123/123	1
1FBI	0	419	123/123	2
1FNS	0	18	0/82	0
1H0D	1	365	9/50	1
1JTP	0	0	0/69	0
1MQ8	0	59	0/29	0
1NSN	0	64	7/52	0
1SQ2	5	322	53/53	0
1Z3G	4	620	130/130	2
$\emptyset$	1.67	281.58	45.33/76.50	0.44

Table 4.9: Antigen/antibody complexes: The second column show the number of hits among the five lowest energy decoys for  $E_{86}$ . Column three shows the total number of hits, while column four shows the number of hits and the number of decoys in the biggest cluster. In column five the number of hits in the top 5 for  $E_{\text{OPEP}}$  is shown.

### Antigen/antibody

In figure 4.12 the IRMSD/ $E_{86}$ -plots for antigen/antibody complexes can be seen. Funnels are produced for 1G9M, 1SQ2 and 1Z3G, but also the results for 1BZQ, 1G6V and 1A2Y are acceptable (compare with table 4.9). For five complexes the biggest cluster only consists of hits, while for 1H0D and 1NSN only a few

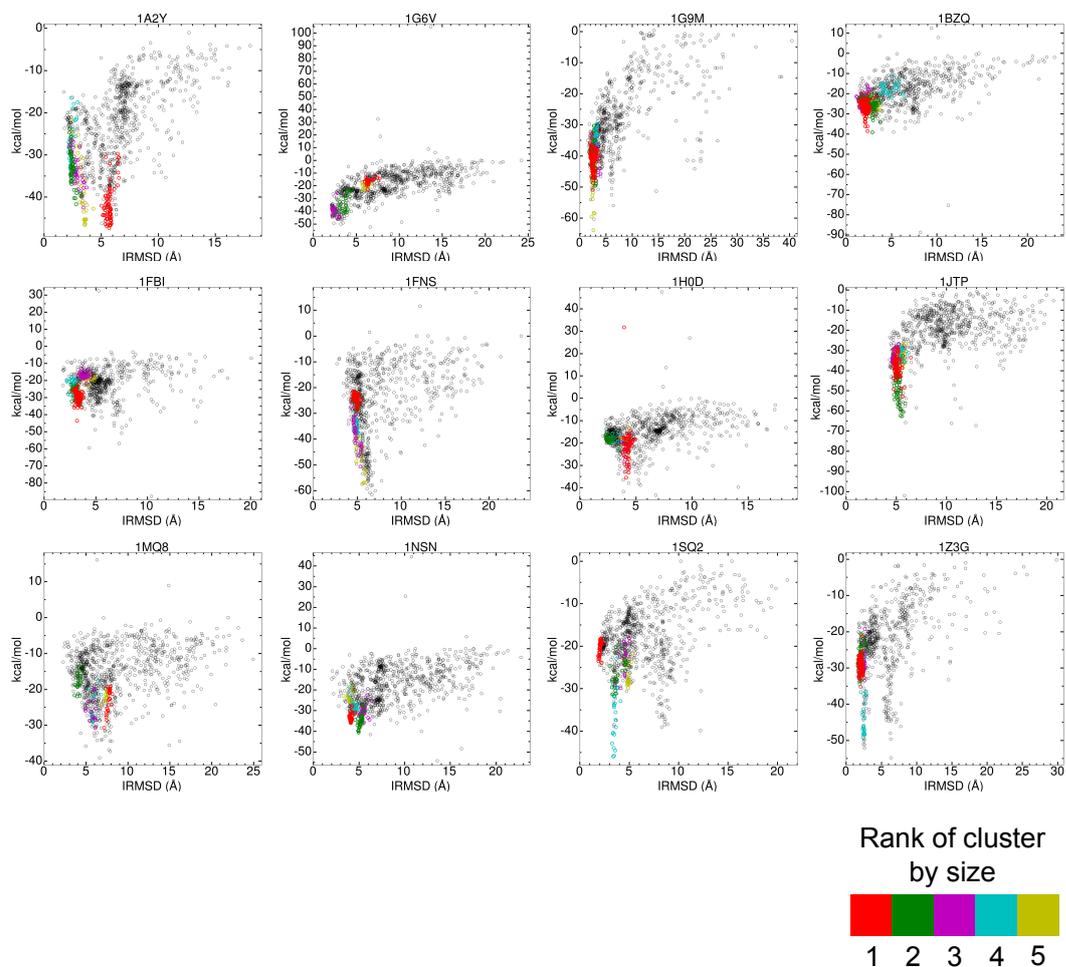


Figure 4.12: Binding funnels for antigen/antibody complexes after minimization. The energy  $E_{86}$  is plotted as a function of the IRMSD. The members of the five biggest clusters are colored.

hits are in the biggest cluster. For 1FNS and 1JTP funnels exist but the

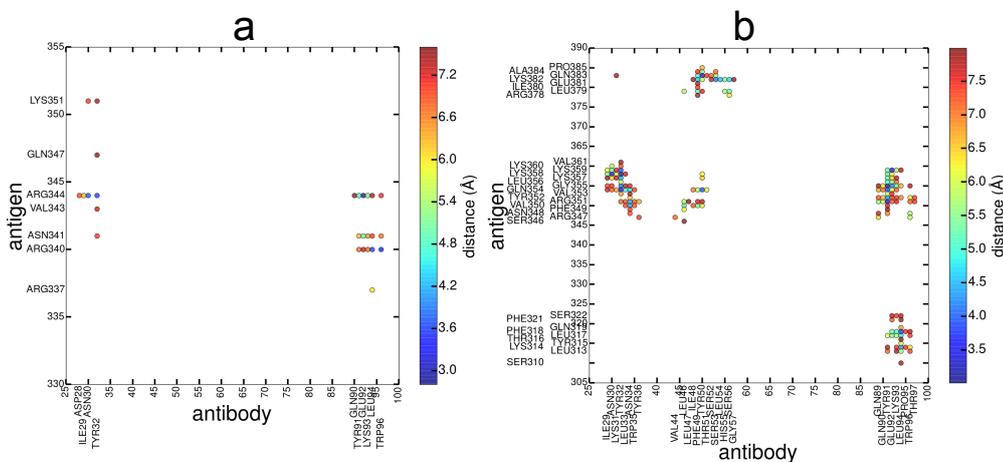


Figure 4.13: Contact maps for the all-atom representation of the bound conformation (a) and for the lowest energy decoy from the fifth cluster (b) of 1FNS.

energy minima are further than  $\text{IRMSD} = 4 \text{ \AA}$  from the crystal structure. The protein-protein interface of the native complex of 1FNS has a very low BSA of  $294 \text{ \AA}^2$ ; in comparison the lowest energy decoy of 1FNS, which is from the fifth cluster and has  $\text{IRMSD} = 6.54 \text{ \AA}$ , has a BSA of  $956 \text{ \AA}^2$ . The interface of the native conformation has 26 residue contacts using a cutoff of  $8 \text{ \AA}$  including three salt bridges (figure 4.13). Two arginins are stretched out on the antigen surface, leading to contacts with the antibody in the crystal structure. The lowest energy conformation on the other hand, has 164 contacts including four salt bridges (figure 4.13b). This large number of contacts originates from the fact that the antibody docks via three patches to the antigen, which increases the contact area. However, the native salt bridge ASP28/ARG344 gets lost as ARG344, which now has a contact to GLU92, is not anymore in the interface. ARG344 is an important crystal antigen residue because it forms ten contacts with the antibody in the native complex structure. The native ARG340/GLU92 contact also gets lost. The most remarkable contacts are the twelve contacts between the equally charged lysins in the lowest energy structure. Instead of stabilizing salt bridges like in the crystal structures hydrophobic LEU/LEU and LEU/TYR contacts are favored, leading to low energies.

The antigen/antibody complexes are the protein complex class which has the best results in ranking using  $E_{\text{OPEP}}$ . For this reason we show in figure 4.14 the funnel plots for the complex energy. As already expected from table 4.9, 1G9M has a prominent funnel with the energy minimum at around  $2 \text{ \AA}$ . The low energy structures are stabilized by a loop in the interface, which is also captured by the interface energy (compare to figure 4.9). The complex 1G6V has a weak

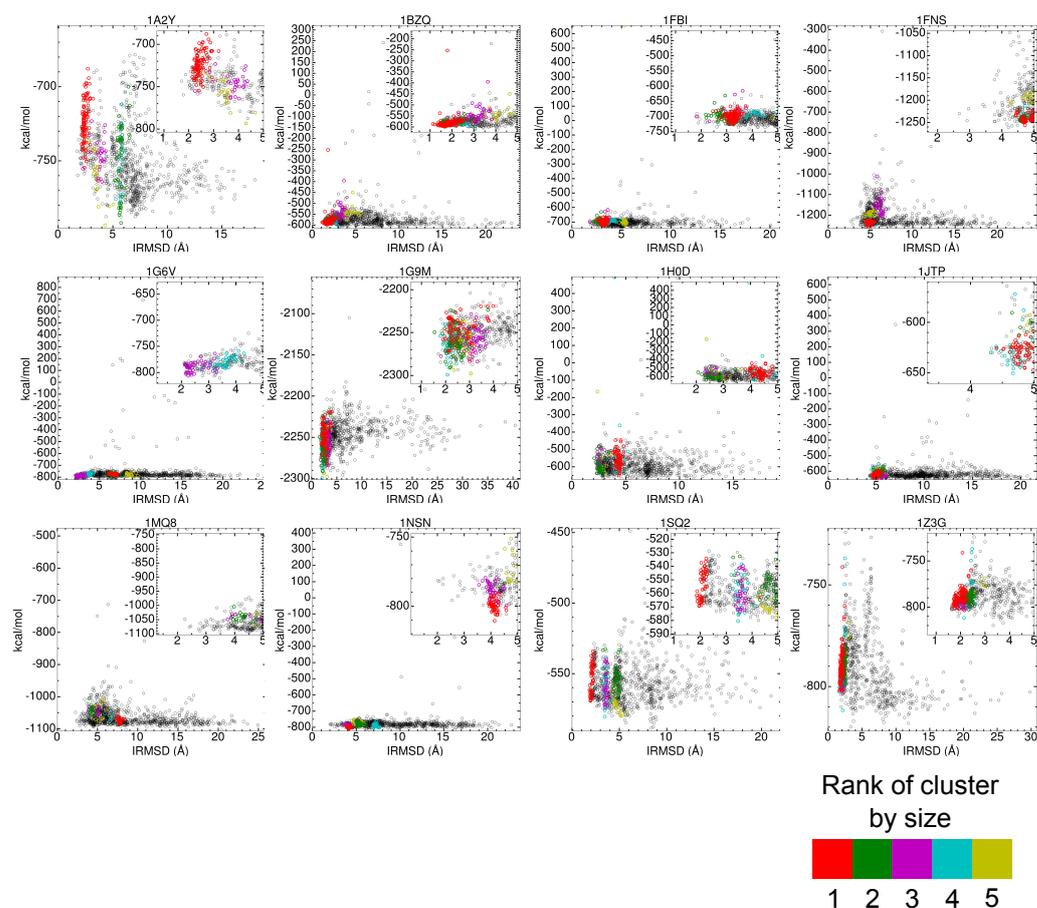


Figure 4.14: Antigen/antibody complexes: The  $E_{\text{OPEP}}$  energy as a function of the IRMSD is shown for complexes where  $E_{\text{OPEP}}$  performs best for ranking. The members of the five biggest clusters are colored.

funnel for complexes with an IRMSD  $< 4$  Å. However, the energy for complexes with an IRMSD higher than 4.5 Å decreases again, yet there is an energy barrier observable at around 4 Å. For the complex 1BZQ the energy values increase constantly between 1 and 3 Å in IRMSD, but around 3 Å the energy drops again. The complex 1FBI has an energy minimum at around 2.5 Å in IRMSD, but the main trend is that the energy increase with decreasing IRMSD. The complex 1H0D has a minimum for complexes at IRMSD  $\approx 3$  Å and the energy values also increase for higher IRMSD values, but the energy differences are not so high. The complex 1Z3G shows increasing energy for increasing IRMSD values between 1.5 and 3 Å but the energy values drop again for higher IRMSD values. In summary, only for 1G9M an energy funnel exists, which would be suitable for a systematic search for the native conformation when using  $E_{\text{OPEP}}$ .

**‘Other’ complexes**

For 11 out of the 19 complexes the biggest cluster only contains hits (see also table 4.10). For another three complexes there is at least one hit in the biggest cluster, while for the five complexes 1GLB, 1NBF, 1NVU, 2CKH and 2G45 the biggest cluster contains no structures with an IRMSD  $< 4$  Å. However, for 2CKH the third biggest cluster forms a nice funnel for IRMSD  $\lesssim 2$  Å.

Scoring complex	#(top 5 $\wedge$ hit)	interface energy $E_{86}$		complex energy $E_{OPEP}$ #(top 5 $\wedge$ hit)
		#hits	#hits / #decoys in biggest cluster	
1BUI	2	92	40/55	2
1FM9	5	491	165/165	1
1G20	1	253	58/58	1
1GLB	0	14	0/8	0
1MA9	5	185	38/38	2
1NBF	0	9	0/18	0
1NVU	0	0	0/21	0
1OOK	0	398	50/50	5
1P7Q	0	233	1/75	5
1R4M	2	35	22/32	0
1S6V	5	514	185/185	5
1U7F	3	96	0/16	1
1UEX	4	260	54/54	1
1V7P	0	86	1/66	0
1ZY8	0	196	89/89	0
2CKH	5	313	0/113	1
2G45	0	33	3/38	0
3FAP	0	470	204/204	5
3PRO	4	111	43/43	0
$\varnothing$	1.89	199.42	50.16/69.89	1.53

Table 4.10: ‘Other’ complexes: The second column show the number of hits among the five lowest energy decoys for  $E_{86}$ . Column three shows the total number of hits, while column four shows the number of hits and the number of decoys in the biggest cluster. In column five the number of hits in the top 5 for  $E_{OPEP}$  is shown.

In figure 4.15 the IRMSD/ $E_{86}$ -plots for ‘other’ complexes can be seen. For the six complexes 1FM9, 1MA9, 1S6V, 1UEX, 2CKH and 3PRO a funnel is present. For the three complexes 1BUI, 1R4M and 3FAP at least an attractor for structures with an IRMSD  $< 4$  Å can be seen, while there is no well defined attractor for the complexes 1GLB, 1NBF, 1NVU, 1U7F and 1UEX. For 1P7Q, 1V7P and 2G45 funnels exist, yet their minimum is for IRMSD values larger than 4 Å.

In the lowest energy conformation of the biggest cluster of 1P7Q the ligand lies with the correct site in the correct pocket but it is slightly tilted. For 1V7P the lowest energy complex of the biggest cluster is close to the crystal binding conformation but the ligand is slightly translated to a pocket on the receptor so that the number of contacts is increased. In the lowest energy conformation of

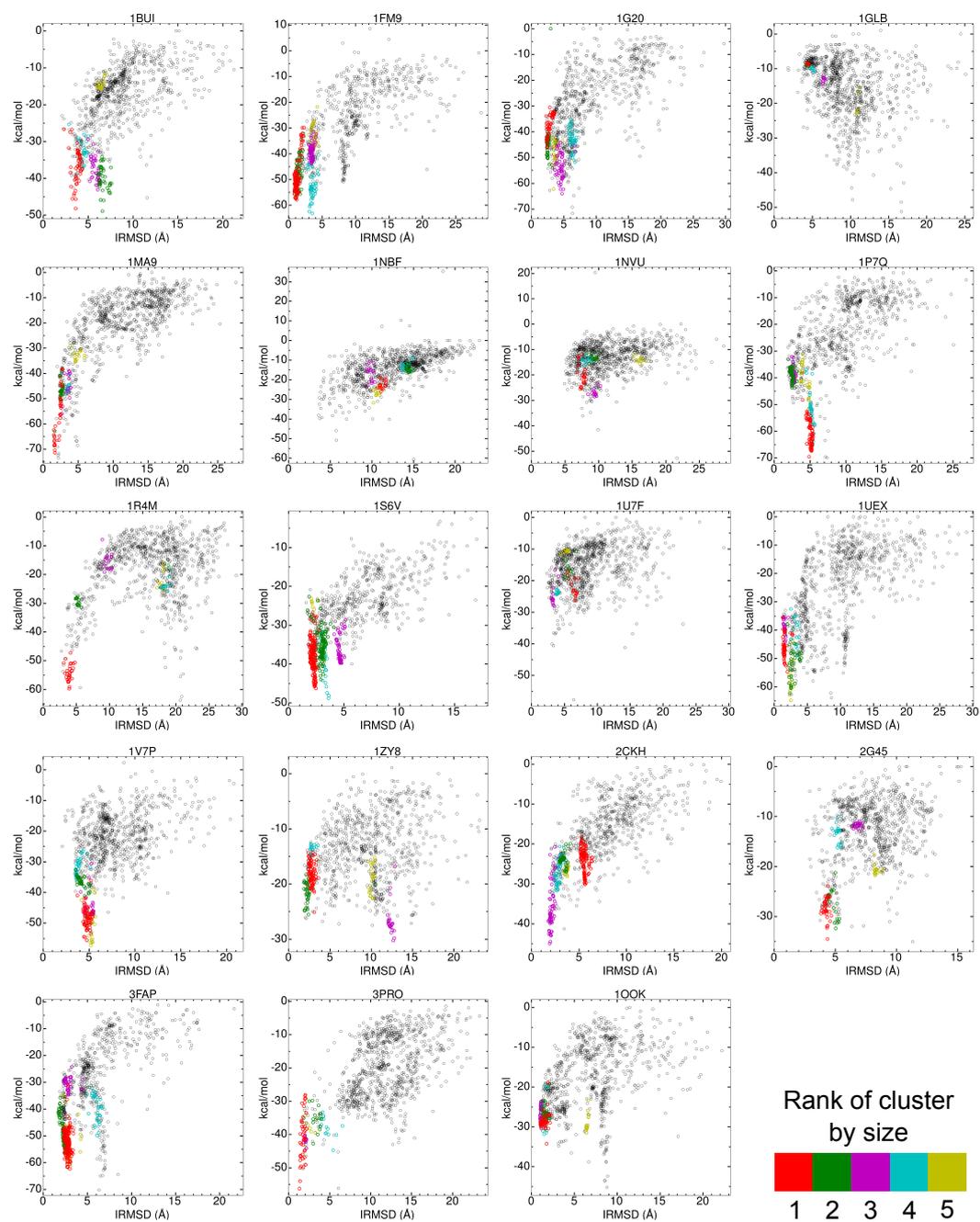


Figure 4.15: Binding funnels for ‘other’ complexes after minimization. The energy  $E_{86}$  is plotted as a function of the IRMSD. The members of the five biggest clusters are colored.

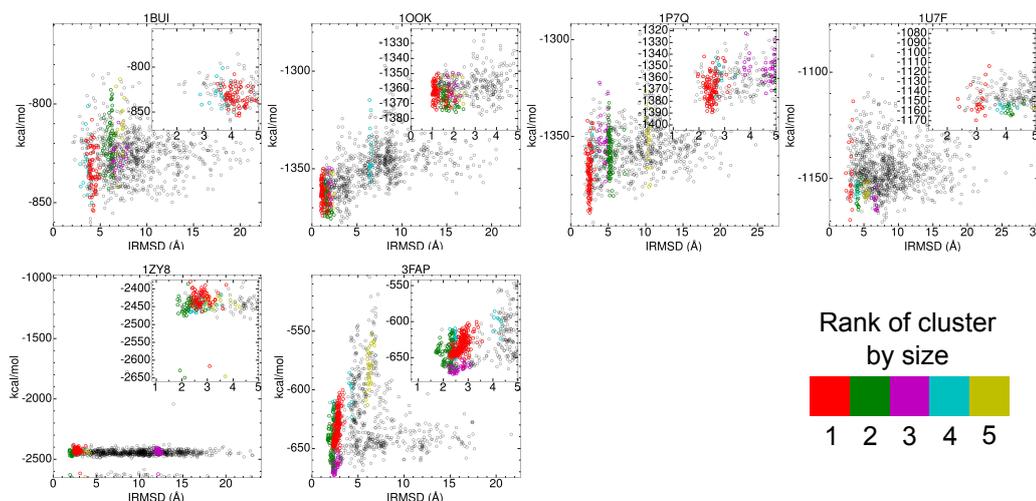


Figure 4.16: ‘Other’ complexes: The  $E_{OPEP}$  energy as a function of the IRMSD is shown for complexes where  $E_{OPEP}$  performs best for ranking. The members of the five biggest clusters are coloured.

the biggest cluster of 2G45 the ligand binds to the receptor next to the correctly bound conformation, yet this ligand conformation has no overlap with the crystal conformation. The complex 1GLB has such a frayed energy landscape because both binding partners are almost spherical and have no distinct pockets, so there are no distinct attractors which also leads to small cluster sizes.

For the complexes that have a better ranking for the full  $E_{OPEP}$  energy in table 4.7 compared to the ranking by  $E_{86}$  we show the IRMSD/ $E_{OPEP}$ -plots in figure 4.16. For the complexes 100K, 1P7Q, and 3FAP the energy decreases with decreasing IRMSD. 100K has a ligand with a large  $\beta$ -sheet connected by loops and in the binding pocket of the ligand there is also a loop. Slight differences in the arrangements of these loops results in different complex energies. The ligand of the complex 3FAP consists of four  $\alpha$ -helices that are connected by loops, the arrangement of these loops is the main difference in the conformation for decoys closer to the crystal structure and results in a lower energy.

## 4.4 Discussion

We presented a local perturbation study of protein-protein complexes using the OPEP force field. It was found that the force field can identify binding conformations and that the subsequent energy minimization can further improve the discrimination between wrongly and correctly docked conformations. The minimization leads to an improvement of the IRMSD for 52 % of the com-

plexes, whereby the rigid body minimization steps have a bigger influence on the change of the IRMSD than the full flexible minimization. Especially for the enzyme/inhibitor complexes the OPEP force field works well and for most of the complexes a binding funnel is observable. The existence of enzyme/inhibitor funnels is even more remarkable when one considers that only few hits are generated by the initial starting structure algorithm and that only the subsequent MC simulations produce more hits. The reason for this are the prominent binding pockets for enzyme/inhibitor complexes, which can be easily detected by the coarse grained force field because detailed atomistic interactions are less important. For antigen/antibody complexes the performance is worse compared to the enzyme/inhibitor complexes. An explanation for this could be the much flatter interfaces compared to the enzyme/inhibitor interfaces, which reduces the importance of the backbone shape of the interface for the interaction but increases the importance of the side chain shape, physico-chemical properties and orientation. It was shown that aromatic and polar side chains on the paratope interacting with the backbone atoms of the epitope are important for the selectivity of antibodies [58], yet in OPEP are not precisely modeled. Moreover, OPEP has one averaged side chain conformation for each amino acid, but for a proper description the modeling of different rotamer states would be necessary. For ‘other’ complexes, the average performance in ranking by  $\text{rank}_{\min}$  is slightly worse compared to the enzyme/inhibitor complexes, but the relative amount of complexes with a hit in the top 10 is the highest (0.63, antigen/antibody complexes 0.58, enzyme/inhibitor complexes 0.53) among the three protein classes. For seven out of the 19 complexes a funnel could be found, which is significantly higher than for enzyme/inhibitor and antigen/antibody complexes (3/15, 3/12). Only the clustering is not as successful as it is for enzyme/inhibitor complexes. For only 14 of the 19 complexes a hit is in the biggest cluster, compared to 13/15 for enzyme/inhibitor complexes and 7/12 for antigen/antibody complexes. Finally, it should be noted that the rescoring of the ZDOCK predictions performed much worse for ‘other’ complexes compared to enzyme/inhibitor complexes, which is not the case for structures that are sampled with our MC search approach.

It looks promising to use the algorithm without energy minimization for enzyme/inhibitor and for ‘other’ complexes, because the improvement is mostly insignificant by the minimization. However, for antigen/antibody complexes the minimization is necessary. The use of  $\text{rank}_{\text{OPEP}}$  performs very well for antigen/antibody complexes but for the two ‘other’ complex classes it is the worst ranking function. The good performance for the antigen/antibody complexes mainly derives from the fact that the best ranks of the complexes 1FNS and 1MQ8 are smaller, which reduces the average rank. However, the number of hits in the top 10 gets reduced by  $\text{rank}_{\text{OPEP}}$ . Thus, also for antigen/antibody

complexes  $E_{\text{OPEP}}$  is not an optimal rescoring function.

## 4.5 Comparison to the rescoring of ZDOCK predictions

We now compare the results of the reranking of the ZDOCK predictions with  $E_{86}$  in chapter 3 with the  $\text{rank}_{\text{min}}$  results in this chapter. A difference is that in chapter 3 we performed global docking, while in this chapter we limited the search space during the generation of start conformations and subsequent MC run. Therefore, we can not directly compare the ranks, but we can compare the relative success of each complex compared to the ‘other’ complexes. We sort the complexes in each class by their ranking of the first hit and compare these resulting orders from chapter 3 and 4. If the relative order changes dramatically between the two sampling methods, it is worth to further investigate these complexes. We will further inspect the complexes that have a difference in the relative ranking of four positions. The results shown in table 4.11 that for the most complexes the relative order does not change dramatically. Yet besides this general trend there are complexes with a big change in relative order. In the enzyme/inhibitor class, 1BTH performs much better with  $\text{rank}_{\text{min}}$  than with ZDOCK sampling and subsequent rescoring. The  $\text{IRMSD}/E_{86}$  plot for this complex (figure 4.10) shows a nicely decreasing energy with descending IRMSD. The reason for the bad performance in chapter 3 is that ZDOCK finds only 37 hits in the top 2000 for this complex. Also for 1PPF the relative ranking improves by using our MC based sampling and the complex shows a nice funnel in the  $\text{IRMSD}/E_{86}$  plot. Here ZDOCK finds many hits (239 in the top 2000); the difference in ranking by the two sampling function probably originates from the different energy landscapes produced by ZDOCK and by OPEP. Thus the two energy functions prefer slightly different conformations. Conformations that are preferred by  $E_{86}$  are more probably generated by our MC run with its  $E_{86}^{\text{soft}}$  function. The performance for complex 2BKR is worse with the MC approach used in this chapter. A reason could be the two arginines, a glutamine and a lysine on the interface surface of the ligand. The side chains of these residues are in a stretched conformation and fit into pockets on the receptor surface. With the coarse grained representation of OPEP this kind of detail can not be sufficiently modeled and scored. Here the grid representation of the side chains by ZDOCK has an advantage for this complex. However the MC procedure samples 134 hits and maybe rescoring of the conformations using an all-atom representation could deliver the right side chain packing and energy. Another complex where ZDOCK is not convincing is 2KAI, ZDOCK delivers only 25 hits in the top 2000. Here,

due to the narrow interface of the complex ZDOCK can not find a good fit by using the discrete rotational sampling. This also explains the accumulation of structures around  $\text{IRMSD} \approx 2 \text{ \AA}$  in the  $\text{IRMSD}/E_{86}$  plot (figure 4.10) and the gap in the sampling up to  $\text{IRMSD} \approx 4 \text{ \AA}$ .

Next we compare the performance of the two sampling and ranking methods for antigen/antibody complexes. One should keep in mind that the ZDOCK search is, as recommended by the developers, restricted to the CDR regions of the antibody. Nonetheless, for the complex 1G6V  $\text{rank}_{\text{min}}$  is much better than for the rescoring of the ZDOCK predictions. ZDOCK only finds two hits in the top 2000 for 1G6V. The better performance of  $\text{rank}_{\text{min}}$  is because of the better sampling, as the energy function itself works quite well for the complex as confirmed by the  $\text{IRMSD}/E_{86}$ -plot (compare figure 4.12), which shows a nice funnel for this complex. For 1JTP the right conformation is found by the MC run with subsequent minimization, but the loop on the antigen enters not deeply enough in the pocket of the antibody. ZDOCK is softer and thus the ligand can penetrate more deeply into the pocket. 1Z3G performs very well using the local MC search approach. ZDOCK also finds 117 hits for this complex and ranks a hit on position 6, but the rescoring with  $E_{86}$  worsens the result, as the two energy function favor different conformation for this complex.

The performance of the complex 1BUI, which belongs to ‘other’ complexes, is much better with our MC search approach compared to the rescoring of ZDOCK predictions. The reason is the poor sampling by ZDOCK which only finds eight hits in the top 2000 for this complex. For the complex 1NVU our algorithm does not sample any hit during any stage, the reason for this is the big conformational change of the receptor during binding. ZDOCK also find only one hit for this complex but this is then quite highly reranked by  $E_{86}$ . For 1P7Q near native solutions are found by our MC algorithm and the minimization further increases the number of hits, but conformations of  $\text{IRMSD} \approx 5 \text{ \AA}$  have a lower energy than hits. Conformations with  $\text{IRMSD} \approx 5 \text{ \AA}$  are not sampled by ZDOCK and this could be the reason for the better performance of  $E_{86}$  used for ZDOCK predictions. Regarding the complex 1R4M the scoring function based on OPEP prefers a conformation with an  $\text{IRMSD}$  slightly lower than  $4 \text{ \AA}$ . ZDOCK finds much better solutions in terms of  $\text{IRMSD}$  but these are more highly ranked by  $E_{86}$ : ZDOCK ranks a hit on position 9 and the rescoring delivers a rank of 201. Thus, the reason for the better performance of  $\text{rank}_{\text{min}}$  for 1R4M is that the MC run samples worse solutions, that are still hits, than ZDOCK, but these structures get a better  $E_{86}$  score. The rescoring by  $E_{86}$  of the ZDOCK predictions for the complexes 1UEX and 3PRO already resulted in an improvement of the ranking. With our own MC search strategy we have an even better sampling and scoring, leading to a better ranking.

In summary, the sampling with the MC approach has the advantage of the continuity as the same energy function are used for the sampling and the rescoring. Different energy models, on the other hand may cause different conformational preferences, which can lead to a bad rescoring performance. This continuity during our MC approach improves the results for 1Z3G and 1BTH. For other complexes like 1PPF, 1G6V and 1BUI, ZDOCK produces too few hits, which complicates the rescoring. Here, our MC search finds more hits. For the two complexes 1UEX and 3PRO, rescoring with  $E_{86}$  already improves the ZDOCK predictions, but sampling with the OPEP-based function and subsequent rescoring further improves the scoring. The search in the unrestricted Cartesian space is the reason for the success for 2KAI. In this complex the interface pocket is very narrow and ZDOCK does not find an optimal fit by the discrete rotational sampling. For 1R4M the results become only apparently better, as the MC run prefers solutions that are worse than the ZDOCK hits, but still fulfill the criteria for a hit. The results become worse for 2BKR because of the missing representation of side chains in OPEP. For 1JTP the force field  $E_{86}^{\text{soft}}$  is still not soft enough compared to ZDOCK and for 1P7Q ZDOCK does not predict the mis-docked conformations for which  $E_{86}$  wrongly calculates low energies.

If one compares the average ranks for the three complex classes, the order of the performance is independent from the sampling. The enzyme/inhibitor complexes are still the complex class with the best average rank followed by the ‘other’ complexes and antigen/antibody complexes. But now the difference between the enzyme/inhibitor complexes and the ‘other’ complexes is marginal. ZDOCK performs considerably better for enzyme/inhibitor complexes and has generally worse rankings for the ‘other’ complexes. We do not have this problem with our more generic energy function that was not specifically derived for protein-protein docking and therefore it does not bias towards a specific protein class. The antigen/antibody complexes are still the class with the worst results, but we could show that OPEP is not completely inapplicable to this complex class. This contrasts to the poor performance in chapter 3. It can deliver good result for some complexes and further investigation is necessary. Another future step will be that the ZDOCK predictions will be further optimized with our MC approach before rescoring with the OPEP-based energy function  $E_{86}$  that was introduced and extensively tested in this work.

complex	rank <sub>min</sub>	rank ZDOCK + $E_{86}$
enzyme/inhibitor		
1AVW	7	5
1BTH	<b>4</b>	13
1CHO	4	3
1GPQ	14	15
1ID5	4	4
1KU6	9	9
1PPF	<b>1</b>	5
1T6G	13	10
1TX6	11	11
1UGH	1	1
1XX9	8	7
2BKR	15	<b>8</b>
2FI4	12	12
2KAI	<b>10</b>	14
3SIC	1	1
antigen/antibody		
1A2Y	6	5
1G6V	<b>6</b>	12
1G9M	1	3
1BZQ	5	6
1FBI	8	9
1FNS	11	8
1HOD	3	4
1JTP	12	<b>1</b>
1MQ8	10	11
1NSN	9	7
1SQ2	1	1
1Z3G	<b>3</b>	10
'other' complexes		
1BUI	<b>7</b>	17
1FM9	1	3
1G20	10	7
1GLB	16	18
1MA9	1	1
1NBF	11	11
1NVU	19	<b>14</b>
1OOK	14	15
1P7Q	17	<b>5</b>
1R4M	<b>8</b>	13
1S6V	1	4
1U7F	8	19
1UEX	<b>1</b>	6
1V7P	18	<b>9</b>
1ZY8	15	12
2CKH	1	1
2G45	13	16
3FAP	11	10
3PRO	<b>1</b>	8

Table 4.11: The relative order of performance for the different protein classes. The complexes are sorted in each class according to their ranking of a hit. For example, in the enzyme/inhibitor class the complexes 1PPF, 1UGH and 3SIC have their best hits on position 1 with rank<sub>min</sub>. Therefore, they are jointly on position 1 in this table. 2BKR, on the other hand, has the best hit on position 94, which is the worst result for rank<sub>min</sub> in the enzyme/inhibitor class, thus its position is 15. The column ‘rank ZDOCK +  $E_{86}$ ’ corresponds to the results in table 2 in chapter 3. Positions are marked in bold if they differ at least 4 places between the ranking schemes rank<sub>min</sub> and ‘rank ZDOCK +  $E_{86}$ ’.



# Chapter 5

## Conclusion

Protein-protein docking is an important computational method allowing us to gain an understanding of the interface-protein interactions on atomic level. Experimental methods for the structural elucidation of protein-protein structures often fail, because these complexes are often too big for NMR and too transient for X-Ray crystallography. Additionally, wet lab work is time consuming and often more expensive than *in silico* experiments. The known structures of the unbound proteins is a good starting point for the prediction of protein-protein interactions. However, these predictions are computationally expensive and several simplifications have to be done, like treating both binding partners as rigid bodies [21] or keeping the backbone rigid [46, 23]. All-atom models are computationally too expensive and their representation of proteins leads to a rough energy landscape. Coarse graining can smooth out the energy landscape and can reduce the computational costs. The computational speed up provides the possibility to perform simulations without any restrictions. In this work we tested two different coarse grained models for protein-protein docking. First, the coarse grained model developed by Berau and Deserno was used to rescore decoys produced by ZDOCK. We found that the allowed overlap in the grid representation leads to a distortion of the interface during the minimization procedure. Furthermore, we revealed that the interface energy is a better discriminator between near native and mis-docked conformations than the complex energy. We could show that the parametrization of the side chains is crucial for protein-protein docking and that the initial force field parametrization does not sufficiently describe the protein-protein interaction. To this end, we replaced the original side chain parameters with the parameters of the OPEP force field and afterwards decided to continue the further studies with the OPEP force field.

The reranking of ZDOCK decoys by a softer potential based on OPEP led to a performance comparable to ZDOCK for enzyme/inhibitor complexes and ‘other’ complexes. We decided to train the rescoring function to further improve the dis-

crimination between near-native and mis-docked decoys. We trained the rescoring function with an iterative learning procedure and tested the new trained rescoring function on an independent test set and found that it performs better than the original rescoring function for all protein-protein classes. Nonetheless, the performance is still better for enzyme/inhibitor complexes than for antigen/antibody and ‘other’ complexes. The good performance for enzyme/inhibitor complexes can be explained by the good performance of ZDOCK for these complexes, but also by the dominant backbone interaction of enzyme/inhibitor complexes [6]. For antigen/antibody complexes the results could be improved by the trained scoring function but the side chains are more important for this complex type. The OPEP representation of the side chain beads is generally not sufficient for the modeling of these interactions [6]. The weak performance of the OPEP based functions for the ‘other’ complexes is likely because of the weak sampling by ZDOCK since results for ‘other’ complexes improved when the sampling was performed with OPEP.

The generation of docking decoys using a local rigid-body Monte Carlo search close to the bound conformation with an OPEP based force field was presented in chapter 4. The MC search produced near native conformations which were further energy minimized in Cartesian and in rigid body space. We could show that the rigid movement has a bigger impact on the RMSD change of the ligand than the secondary structure change. Opposite to the results in chapter 2 and 3 the interface RMSD was improved for the majority of the decoys by the minimization. The clustering of the decoys and ranking the resulting cluster by size allowed to identify near native solutions and we could proof that our approach is able to identify docking attractors and near-native complexes. Especially for ‘other’ complexes we could see an improvement of the ranking of near native complexes compared to the results in chapter 3. The reason is that the OPEP force field and ZDOCK are different representations of proteins and thus may favor different conformations. ZDOCK may find a near native solution with a high score but OPEP prefers near native solutions with a slightly different conformation. Our conclusion is that sampling and scoring work better when they are done with similar energy functions.

In summary, we showed that the OPEP force field is in general a good choice for enzyme/inhibitor complexes and ‘other’ complexes. However, we saw that the representation of the side chain by one bead is not detailed enough for antigen/antibody complexes. An all-atom or at least a more precise representation of the side chain would be necessary for these complexes. In the future it would be desirable to have an algorithm that allow back mapping of the coarse-grained OPEP structures to all-atom models, which would enable us to continue the modeling, such as MD simulations with all atom models. Back-mapping to all-

atom models is also necessary if one wishes to compete in the CAPRI experiment. Another interesting test would be to perform the MC simulations on the ZDOCK predictions without subsequent minimization and then selecting structures only by clustering and/or energy. The rigid MC sampling step performed quite well for enzyme/inhibitor and 'other' complexes and maybe one can omit the costly minimization, or restrict it to the fast rigid body minimization. In addition, one could introduce more flexibility in the MC sampling by performing MC moves leading to small changes of the dihedral angles of the protein backbone.



# Bibliography

- [1] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, "The protein data bank," *European Journal of Biochemistry*, vol. 80, no. 2, pp. 319–324, 1977.
- [2] D. Barlow and J. Thornton, "Helix geometry in proteins," *Journal of molecular biology*, vol. 201, no. 3, pp. 601–619, 1988.
- [3] A. Jabs, M. S. Weiss, and R. Hilgenfeld, "Non-proline cis peptide bonds in proteins," *Journal of molecular biology*, vol. 286, no. 1, pp. 291–304, 1999.
- [4] H. Hwang, T. Vreven, J. Janin, and Z. Weng, "Protein–protein docking benchmark version 4.0," *Proteins: Structure, Function, and Bioinformatics*, vol. 78, no. 15, pp. 3111–3114, 2010.
- [5] J. M. Berg, J. L. Tymoczko, and L. Stryer, "Biochemistry. 5th," 2002.
- [6] R. M. Jackson, "Comparison of protein–protein interactions in serine protease-inhibitor and antibody-antigen complexes: Implications for the protein docking problem," *Protein Science*, vol. 8, no. 03, pp. 603–613, 1999.
- [7] C. A. Janeway, P. Travers, M. Walport, M. J. Shlomchik, *et al.*, *Immunobiology: the immune system in health and disease*, vol. 2. Garland New York, 2001.
- [8] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, "Charmm: A program for macromolecular energy, minimization, and dynamics calculations," *Journal of computational chemistry*, vol. 4, no. 2, pp. 187–217, 1983.
- [9] S. Lifson and A. Warshel, "Consistent force field for calculations of conformations, vibrational spectra, and enthalpies of cycloalkane and n-alkane molecules," *The Journal of Chemical Physics*, vol. 49, no. 11, pp. 5116–5129, 1968.

- [10] S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta, and P. Weiner, "A new force field for molecular mechanical simulation of nucleic acids and proteins," *Journal of the American Chemical Society*, vol. 106, no. 3, pp. 765–784, 1984.
- [11] J. W. Ponder, D. A. Case, *et al.*, "Force fields for protein simulations," *Advances in protein chemistry*, vol. 66, pp. 27–86, 2003.
- [12] L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S.-J. Marrink, "The martini coarse-grained force field: extension to proteins," *Journal of chemical theory and computation*, vol. 4, no. 5, pp. 819–834, 2008.
- [13] M. Levitt and A. Warshel, "Computer simulation of protein folding," *Nature*, vol. 253, no. 5494, pp. 694–698, 1975.
- [14] M. Levitt, "A simplified representation of protein conformations for rapid simulation of protein folding," *Journal of molecular biology*, vol. 104, no. 1, pp. 59–107, 1976.
- [15] A. Liwo, S. Oldziej, M. R. Pincus, R. J. Wawak, S. Rackovsky, and H. A. Scheraga, "A united-residue force field for off-lattice protein-structure simulations. i. functional forms and parameters of long-range side-chain interaction potentials from protein crystal data," *Journal of computational chemistry*, vol. 18, no. 7, pp. 849–873, 1997.
- [16] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. De Vries, "The martini force field: coarse grained model for biomolecular simulations," *The Journal of Physical Chemistry B*, vol. 111, no. 27, pp. 7812–7824, 2007.
- [17] S. O. Yesylevskyy, L. V. Schäfer, D. Sengupta, and S. J. Marrink, "Polarizable water model for the coarse-grained martini force field," *PLoS Comput Biol*, vol. 6, no. 6, p. e1000810, 2010.
- [18] M. L. Connolly, "Shape complementarity at the hemoglobin  $\alpha 1\beta 1$  subunit interface," *Biopolymers*, vol. 25, no. 7, pp. 1229–1247, 1986.
- [19] C. Levinthal, S. J. Wodak, P. Kahn, and A. K. Dadivanian, "Hemoglobin interaction in sickle cell fibers. i: Theoretical approaches to the molecular contacts," *Proceedings of the National Academy of Sciences*, vol. 72, no. 4, pp. 1330–1334, 1975.
- [20] S. J. Wodak and J. Janin, "Computer analysis of protein-protein interaction," *Journal of molecular biology*, vol. 124, no. 2, pp. 323–342, 1978.

- [21] E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A. A. Friesem, C. Aflalo, and I. A. Vakser, “Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques,” *Proceedings of the National Academy of Sciences*, vol. 89, no. 6, pp. 2195–2199, 1992.
- [22] R. Chen and Z. Weng, “Docking unbound proteins using shape complementarity, desolvation, and electrostatics,” *Proteins: Structure, Function, and Bioinformatics*, vol. 47, no. 3, pp. 281–294, 2002.
- [23] M. Zacharias, “Protein–protein docking with a reduced protein model accounting for side-chain flexibility,” *Protein Science*, vol. 12, no. 6, pp. 1271–1282, 2003.
- [24] L. Hui, L. Feng, Y. Jianli, and L. Xiu-Ling, “Side-chain flexibility in protein docking,” in *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 IEEE Conference on*, pp. 1–8, IEEE, 2015.
- [25] I. M. Nooren and J. M. Thornton, “Diversity of protein–protein interactions,” *The EMBO journal*, vol. 22, no. 14, pp. 3486–3492, 2003.
- [26] V. S. Rao, K. Srinivas, G. Sujini, and G. Kumar, “Protein–protein interaction detection: methods and analysis,” *International journal of proteomics*, vol. 2014, 2014.
- [27] M. P. Stumpf, T. Thorne, E. de Silva, R. Stewart, H. J. An, M. Lappe, and C. Wiuf, “Estimating the size of the human interactome,” *Proceedings of the National Academy of Sciences*, vol. 105, no. 19, pp. 6959–6964, 2008.
- [28] T. Bereau and M. Deserno, “Generic coarse-grained model for protein folding and aggregation,” *The Journal of chemical physics*, vol. 130, p. 235106, 2009.
- [29] P. Derreumaux, “From polypeptide sequences to structures using monte carlo simulations and an optimized potential,” *The Journal of chemical physics*, vol. 111, no. 5, pp. 2301–2310, 1999.
- [30] H.-J. Limbach, A. Arnold, B. A. Mann, and C. Holm, “Espressoan extensible simulation package for research on soft matter systems,” *Computer Physics Communications*, vol. 174, no. 9, pp. 704–727, 2006.
- [31] T. Vreven, H. Hwang, and Z. Weng, “Integrating atom-based and residue-based scoring functions for protein–protein docking,” *Protein Science*, vol. 20, no. 9, pp. 1576–1586, 2011. ZDOCK.
- [32] J. Nocedal and S. Wright, *Numerical optimization*. Springer Science & Business Media, 2006.

- [33] J. Nocedal, “Updating quasi-newton matrices with limited storage,” *Mathematics of computation*, vol. 35, no. 151, pp. 773–782, 1980.
- [34] D. U. Ferreira, J. A. Hegler, E. A. Komives, and P. G. Wolynes, “Localizing frustration in native proteins and protein assemblies,” *Proceedings of the National Academy of Sciences*, vol. 104, no. 50, pp. 19819–19824, 2007.
- [35] J. A. Tuszynski and M. Kurzynski, *Introduction to molecular biophysics*. CRC press, 2003.
- [36] N.-Y. Chen, Z.-Y. Su, and C.-Y. Mou, “Effective potentials for folding proteins,” *Physical review letters*, vol. 96, no. 7, p. 078103, 2006.
- [37] S. Miyazawa, R. Jernigan, *et al.*, “Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading,” *Journal of molecular biology*, vol. 256, no. 3, pp. 623–644, 1996.
- [38] J. M. Rosenbergl, “The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method,” *Journal of computational chemistry*, vol. 13, no. 8, pp. 1011–1021, 1992.
- [39] B. K. Ho, A. Thomas, and R. Brasseur, “Revisiting the ramachandran plot: Hard-sphere repulsion, electrostatics, and h-bonding in the  $\alpha$ -helix,” *Protein Science*, vol. 12, no. 11, pp. 2508–2522, 2003.
- [40] H. Hwang, T. Vreven, J. Janin, and Z. Weng, “Protein–protein docking benchmark version 4.0,” *Proteins: Structure, Function, and Bioinformatics*, vol. 78, no. 15, pp. 3111–3114, 2010.
- [41] R. Méndez, R. Leplae, L. De Maria, and S. J. Wodak, “Assessment of blind predictions of protein–protein interactions: current status of docking methods,” *Proteins: Structure, Function, and Bioinformatics*, vol. 52, no. 1, pp. 51–67, 2003.
- [42] P. J. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, *et al.*, “Biopython: freely available python tools for computational molecular biology and bioinformatics,” *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, 2009.
- [43] J. Maupetit, P. Tuffery, and P. Derreumaux, “A coarse-grained protein force field for folding and structure prediction,” *Proteins: Structure, Function, and Bioinformatics*, vol. 69, no. 2, pp. 394–408, 2007.

- [44] D. Ravikant and R. Elber, “Energy design for protein-protein interactions,” *The Journal of chemical physics*, vol. 135, p. 065102, 2011.
- [45] J. Mintseris, B. Pierce, K. Wiehe, R. Anderson, R. Chen, and Z. Weng, “Integrating statistical pair potentials into protein complex prediction,” *Proteins: Structure, Function, and Bioinformatics*, vol. 69, no. 3, pp. 511–520, 2007.
- [46] J. Gray, S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C. Rohl, D. Baker, *et al.*, “Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations,” *Journal of molecular biology*, vol. 331, no. 1, pp. 281–300, 2003.
- [47] D. Frishman and P. Argos, “Knowledge-based protein secondary structure assignment,” *Proteins: Structure, Function, and Bioinformatics*, vol. 23, no. 4, pp. 566–579, 1995.
- [48] R. Chen and Z. Weng, “A novel shape complementarity scoring function for protein-protein docking,” *Proteins: Structure, Function, and Bioinformatics*, vol. 51, no. 3, pp. 397–408, 2003.
- [49] R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton, “Procheck: a program to check the stereochemical quality of protein structures,” *Journal of applied crystallography*, vol. 26, no. 2, pp. 283–291, 1993.
- [50] A. L. Morris, M. W. MacArthur, E. G. Hutchinson, and J. M. Thornton, “Stereochemical quality of protein structure coordinates,” *Proteins: Structure, Function, and Bioinformatics*, vol. 12, no. 4, pp. 345–364, 1992.
- [51] A. M. Ruvinsky and I. A. Vakser, “The ruggedness of protein-protein energy landscape and the cutoff for  $1/rn$  potentials,” *Bioinformatics*, vol. 25, no. 9, pp. 1132–1136, 2009.
- [52] D. Kozakov, O. Schueler-Furman, and S. Vajda, “Discrimination of near-native structures in protein-protein docking by testing the stability of local minima,” *Proteins: Structure, Function, and Bioinformatics*, vol. 72, no. 3, pp. 993–1004, 2008.
- [53] D. Kozakov, K. H. Clodfelter, S. Vajda, and C. J. Camacho, “Optimal clustering for detecting near-native conformations in protein docking,” *Biophysical journal*, vol. 89, no. 2, pp. 867–875, 2005.
- [54] F. Sterpone, S. Melchionna, P. Tuffery, S. Pasquali, N. Mousseau, T. Cragolini, Y. Chebaro, J.-F. St-Pierre, M. Kalimeri, A. Barducci, *et al.*, “The

- opep protein model: from single molecules, amyloid formation, crowding and hydrodynamics to dna/rna systems,” *Chemical Society reviews*, 2014.
- [55] H. Mirzaei, D. Beglov, I. C. Paschalidis, S. Vajda, P. Vakili, and D. Kozakov, “Rigid body energy minimization on manifolds for molecular docking,” *Journal of chemical theory and computation*, vol. 8, no. 11, pp. 4374–4380, 2012.
- [56] A. M. Ruvinsky and I. A. Vakser, “Chasing funnels on protein-protein energy landscapes at different resolutions,” *Biophysical journal*, vol. 95, no. 5, pp. 2150–2159, 2008.
- [57] D. Kozakov, R. Brenke, S. R. Comeau, and S. Vajda, “Piper: An fft-based protein docking program with pairwise potentials,” *Proteins: Structure, Function, and Bioinformatics*, vol. 65, no. 2, pp. 392–406, 2006.
- [58] H.-P. Peng, K. H. Lee, J.-W. Jian, and A.-S. Yang, “Origins of specificity and affinity in antibody–protein interactions,” *Proceedings of the National Academy of Sciences*, vol. 111, no. 26, pp. E2656–E2665, 2014.