Forschungszentrum Karlsruhe
in der Helmholtz-Gemeinschaft

Wissenschaftliche Berichte
FZKA 7323

High-Throughput Simulation Methods for Protein-Ligand Docking

B. K. Fischer

Institut für Nanotechnologie

Juni 2007

Forschungszentrum Karlsruhe

in der Helmholtz-Gemeinschaft

Wissenschaftliche Berichte

FZKA 7323

High-Throughput Simulation Methods for Protein-Ligand Docking

Bernhard Karl Fischer

Institut für Nanotechnologie

vom Fachbereich Physik der Universität Dortmund
genehmigte Dissertation

Forschungszentrum Karlsruhe GmbH, Karlsruhe
2007

Forschungszentrum Karlsruhe GmbH
Postfach 3640, 76021 Karlsruhe

Mitglied der Hermann von Helmholtz-Gemeinschaft
Deutscher Forschungszentren (HGF)

ISSN 0947-8620

urn:nbn:de:0005-073235

HIGH-THROUGHPUT SIMULATION METHODS FOR PROTEIN-LIGAND DOCKING

Dissertation

zur Erlangung des Grades eines
Doktors der Naturwissenschaften
des Fachbereiches Physik
der Universität Dortmund

vorgelegt von

Bernhard Karl Fischer

27. April 2007

Abstract

With the work reported in this thesis, we aim to contribute to the field of computational drug discovery. With our program *FlexScreen*, we attempt to estimate the ligand affinity to a protein model by simulating the formation of protein-ligand complexes. In this approach, the affinity of a ligand to a protein is determined as the energetic difference between the energetically optimal protein-ligand conformation and the state in which protein and ligand are not interacting with each other. We show that our approach helps to identify good binding compounds in a large database of ligands.

The structure of this thesis follows the chronology of our research efforts. We first started with testing the accuracy of our docking algorithm. To calculate binding energies in a good approximation, we first determine realistic protein-ligand conformations.

In another study, we analyze the problem of protein flexibility and the shortcoming of using only one rigid protein structure for docking simulations. In large-scale database screens we compare the influence of rigid and flexible protein models with each other. We show that flexible protein models result in an increased reliability of the screen and in the identification of a higher number of good binding ligands.

Receptor-ligand interactions are calculated using many approximations. In a further study we investigate, if the accuracy of binding energies could be improved by employing parameters obtained from quantum mechanical calculations. We show that by incorporating the results of quantum mechanical calculations for the receptor, the overall accuracy of the whole simulation can be increased. This is an important result for high-throughput screening, because the time consuming quantum mechanical calculations can be performed separately in advance.

We have thus developed a high-throughput docking approach which allows us to identify good binding ligands in large databases. Including protein flexibility by allowing the side chains to alter their conformations results in a more realistic model of proteins. Applied to docking simulations of databases, our approach is less biased to the rigid, experimentally measured protein crystal structure which gives us the possibility to discover more diverse good binding ligands. In addition, the overall accuracy of our approach is enhanced further by integrating quantum mechanical calculations into our description of the proteins.

High-Throughput Simulationsmethoden für Protein-Liganden Docking

Zusammenfassung

Mit dieser Arbeit leisten wir einen Beitrag zum Forschungsfeld der rechnerunterstützten Medikamentenentwicklung. Mit Hilfe unseres Programms *FlexScreen* versuchen wir mittels energetischer Kriterien potentielle Wirkstoffe zu identifizieren. Für jeden Liganden aus einer großen Moleküldatenbank ermitteln wir anhand stochastischer Simulationsmethoden die beste Protein-Ligandenkonformation und bestimmen damit dann die Affinität des Liganden zum Protein.

In einer ersten Studie untersuchen und belegen wir zunächst die Zuverlässigkeit und die Genauigkeit unseres Ansatzes. Für ein Protein-Liganden Dockingsimulationsprogramm ist es wichtig, dass experimentell gemessene Bindungsorientierungen des Liganden zuverlässig reproduziert werden können; dies ist notwendig, um zuverlässig die Affinität eines Liganden zu bestimmen. Wir zeigen, dass dies mit unserer Methode erfüllt ist.

In einer weiteren Studie wenden wir uns dem Thema der Proteinflexibilität zu. Bindende Liganden können Konformationsänderungen der Proteinstruktur veranlassen. Besonders bei Protein-Liganden Dockingsimulationen stellen wir Nachteile fest, wenn die Proteinstruktur als unveränderliche drei-dimensionale Struktur angenommen wird. In unserem Ansatz werden die Konformationsänderungen des Proteins mittels Flexibilität in den Proteinseitenketten approximiert. Wir weisen nach, dass auf diese Weise zuverlässiger und genauer gut bindende Liganden erkannt werden können.

Das Problem, dass häufig Proteine und Liganden sich mit traditionell verwendeten Methoden nur ungenau beschreiben lassen, behandeln wir in einer dritten Studie. Hier können wir zeigen, wie durch das Extrahieren von Parametern aus vorausgehenden quantenmechanischen Berechnungen die Genauigkeit der Affinitätsbestimmung erhöht werden kann. Für Protein-Liganden Dockingsimulationen konnten wir feststellen, dass es für diese Steigerung der Genauigkeit auch ausreicht, wenn nur das Protein mit solchen extrahierten Parametern beschrieben wird. Diese Methodik hat den Vorteil, dass langwierige quantenmechanische Berechnungen nur einmal für das Protein durchgeführt werden müssen und damit die Rechenzeit für die Protein-Ligand Dockingsimulationen unveränderlich bleibt.

Mit unserer Arbeit haben wir erfolgreich "high-throughput" Protein-Liganden Docking-simulationsmethoden entwickelt. Wir zeigen die Zuverlässigkeit unseres Ansatzes und den Vorteil unseres flexiblen Proteinmodells. Des Weiteren stellen wir eine Methode vor, die eine höhere Genauigkeit zur Bestimmung der Affinität von Liganden erlaubt.

Contents

Preamble **1**

1 Introduction **5**
1.1 Drug activity . . . 5
1.1.1 Drugs acting on enzymes . . . 5
1.1.2 Drugs acting at protein receptors . . . 8
1.1.3 Bioavailability . . . 11
1.2 Brief overview of rational drug discovery . . . 12
1.2.1 Strategies without knowledge of the protein . . . 12
1.2.2 Strategies employing knowledge of the protein . . . 13
1.2.3 Recent developments . . . 15

2 Thermodynamics of Protein-Ligand Association **17**
2.1 Thermodynamic Basis . . . 17
2.2 Thermodynamic view of protein-ligand affinity . . . 19
2.3 Separability of the binding energy . . . 22
2.4 Necessary approximations of the binding free energy . . . 22
2.4.1 Thermodynamic Cycle . . . 23

3 Stochastic Optimization Methods **25**
3.1 Monte Carlo Simulation . . . 25
3.1.1 Principles of Monte Carlo simulations . . . 26
3.1.2 Algorithm of Monte Carlo simulations . . . 27
3.2 Related optimization techniques . . . 27
3.2.1 Simulated Annealing (SA) . . . 28
3.2.2 Stochastic Tunneling (STUN) . . . 28

4 Biomolecular Force field **31**
4.1 Molecular Mechanic Interactions . . . 31
4.1.1 Interaction of chemically bonded atoms . . . 32
4.1.2 Experimental parametrization . . . 34
4.1.3 Interactions of non-bonded atoms . . . 34

4.2 Treatment of the solvent . . . 38
4.2.1 Non-polar contribution . . . 38
4.2.2 Electrostatic contribution . . . 39

5 Docking Strategy **45**
5.1 Process Description . . . 46
5.2 Ligand representation . . . 47
5.2.1 Analysis of ligand flexibility . . . 48
5.3 Scoring function . . . 49
5.4 Methods for generating an initial protein-ligand conformation . . . 51

6 Binding Accuracy Evaluation **55**
6.1 Introduction . . . 55
6.2 Methods . . . 56
6.2.1 Docking Protocol . . . 57
6.3 Receptor Structures . . . 58
6.4 Results . . . 58
6.4.1 Steric Hindrance . . . 61
6.4.2 Presence of water and small ions . . . 62
6.4.3 Comparison with AutoDock . . . 63
6.5 Discussion . . . 63

7 Importance of Protein Flexibility **65**
7.1 Introduction . . . 65
7.2 Method: Flexible docking to Thymidine Kinase . . . 65
7.2.1 Preparation of the ligands and the docking site . . . 66
7.3 Results . . . 66
7.3.1 Screen using a rigid enzyme structure . . . 66
7.3.2 Identification of important side chains . . . 68
7.3.3 Flexible Protein Screen . . . 71
7.3.4 Comparison: Rigid screen vs. flexible screens . . . 72
7.4 Discussion . . . 73

8 Influence of QM Descriptors on Docking Simulations **75**
8.1 Introduction . . . 75
8.2 Ligands, Receptor Structure and Partial Charges . . . 79
8.3 Results . . . 80
8.4 Discussion . . . 86
8.5 Conclusion . . . 87

9 Summary **89**

A Analytical Calculation of Electrostatic Energies **93**

B Tables **95**

C Parameter for the Scoring Function **101**
C.1 Lennard-Jones Parameter 101
C.2 Hydrogen bond parameters 102

D Used programs and definitions **103**

Abbreviations **104**

List of Figures **108**

List of Tables **110**

Bibliography **112**

Acknowledgments **126**

Preamble

Diseases have been treated with medications for thousands of years. The effects of these drugs were usually discovered over centuries by trial and error. Many drugs used today have been discovered by such observations. However, as the cellular and molecular mechanisms behind many diseases are increasingly understood new avenues for rational drug development emerge and a systematic search for drugs began. Over time, newly developed techniques and an ever increasing knowledge led to new, but complementary, strategies for drug discovery.
First, animals were used as models for the human organism. However, because many potential drugs could not be adequately tested with animals, in-vitro experiments became more and more important. In in-vitro experiments the activity of various chemical compounds on cells or on specific proteins is measured in the laboratory. About thirty years ago computational drug research started to complement experimental techniques.

The biological system of the human body is very complex and far from being fully understood. Chemical reactions, electrostatic and chemical signals occur constantly in our organism. Complex reactions are needed, for example, to maintain the cell, to react to signals and to convert chemical energy. The central nervous system, to give just one example, controls the functionality of the whole organism. It reacts to stimuli from the environment and maintains the functionality of the organism by actively sending signals to the organs. The signal pathway is either of a pure chemical nature or uses combined electric and chemical signals.
Proteins, the building blocks of the cells, participate in such reactions and fulfill a variety of different functions in the human body. They act, for example, as catalysts and lower the activation energies for chemical reactions; proteins with this functionality are called enzymes. Proteins also work as signal handlers and respond in a particular way to specific small compounds, i.e. small molecules also called ligands that bind to a protein. In that functionality the proteins are referred to as receptors.
The major goal in drug research is to find ligands that influence the organism, unwanted cells, viruses or alien bacteria in a way that will lead to desired effects. The inhibition of enzymes or of receptors is a successful approach to influence the activity of the protein. By hindering the naturally binding compounds to dock to the protein, one can stop the metabolic cascade in which the protein is involved. For such purposes, drugs with a very high affinity, i.e. the binding energy, to a protein are necessary. These drugs should bind stronger to the protein than the natural compound that usually interacts with it.

A good drug acts specifically on the target only; i.e. in thermal equilibrium the drug is bound to the target at the active site of the protein, where the natural compound normally binds. The affinity of a compound to a target can be determined as the difference in the free energy between the bound and unbound state of a ligand and a protein. Experimentalists usually measure affinities of proteins and ligands in the solution. By comparing the concentration of bound and unbound molecules in solution, the affinity is expressed in terms of the concentration required to saturate the protein with a given probability. As an example, a good binding drug has a concentration in the nanomolar range.
Most drugs discovered today are rather small (less than 100 atoms). However, there is still a vast range of chemical possibilities to construct small compounds by connecting different chemical elements with each other. Therefore, strategies were developed to search for suitable compounds for specific targets, i.e. for finding suitable ligands that will bind to very specific protein binding sites.

The wide field of computational drug discovery developed over time from a rather abstract approach, using mainly statistical analysis, to a more and more concrete calculation of protein-ligand interaction energies. This became feasible due to a deeper understanding of the cell processes and due to the possibility to determine the three-dimensional structure of proteins and protein-ligand complexes experimentally. The three-dimensional structure of proteins can often be determined by analyzing the scattering pattern of X-rays on a protein crystal or by the method of nuclear magnetic spectroscopy. Analyzing the scattering pattern of X-rays on a protein crystal is the most common method and most accurate method to obtain the structural information of proteins. Difficulties in this method stems from problems to create crystals for a given protein.
For most solved protein structures, scientists deposit the measured three-dimensional structure of proteins and macromolecules, i.e. aggregates of large molecules, in a publicly available database on the internet: the Protein Data Bank (PDB). Presently, this database contains about 40000 entries of three dimensional information for protein structures. With the advent of structural genomics more and more people are working on determining the structural information of proteins and measuring methods improve steadily, resulting in a steady increase of the rate of new entries in the PDB database.
In addition, a large amount of three-dimensional information of small chemical compounds is publicly available. Companies that provide the profitable service of synthesizing these compounds allow scientists to use the three-dimensional information of the small molecules for computational studies. Due to the availability of the structural information of many compounds, a large library of ligands can be easily constructed allowing scientists to perform 'computational drug experiments' on a large-scale. Instead of screening many compounds in the laboratory for a high affinity to a target, computational approaches are used to identify potential drug candidates. Virtual screening provides a very quick and an inexpensive approach to pre-select promising drug candidates. Of course, the computationally identified high-affinity compounds must be experimentally verified.

With the work reported in this thesis, we aim to contribute to the field of computational drug discovery. We attempt to estimate the ligand affinity to a protein model by simulating the formation of protein-ligand complexes. In this approach, the affinity of a ligand to a protein is determined as the energetic difference between the energetically optimal protein-ligand conformation and the state in which protein and ligand are not interacting with each other. We show that our approach helps to identify good binding compounds in a large database of ligands.
The ultimate goal of this project is to develop a high-throughput drug screening approach, which can be used for large databases. Because one of our main objectives is computational efficiency, we are not free to use all possible methodologies for our approach. Many presently available computational methods can not be employed to identify potential drug candidates in large databases of chemical compounds because they are not sufficiently fast. For this reason, we can neither use quantum mechanical calculations nor molecular dynamics simulations, even though these methods very accurately model protein ligand interactions..

Proteins and ligands are no static objects in a living organism. Because of thermal activation energy, they are constantly in a state of motion. When a ligand binds to a protein, both ligand and protein can change their conformation and adapt to each other. Proteins are constructed of different amino acids which are joined together by peptide bonds. A succession of peptide bonds generates a main chain, or backbone, from which various side chains project outwards. These side chains differ from one amino acid to the next. Conformational changes of the side chains are local conformational changes. They do not alter the position of the connected amino acids. Conformational changes of the backbone, however, also alter the positions of the connected amino acids.
If we look at the structure of a protein-ligand complex as measured by crystallography, we see it as if we were looking at one single frame of a movie only. As a result, we do not get the whole picture, i.e. the whole information of the binding process. Consequently, if drug design were based solely on this particular 'frame', it is not possible to describe the whole system adequately, which prevents us from discovering all possible high affinity molecules. However, because the complexity of simulating the system increases drastically with the number of degrees of freedom of the system, conformational changes of the protein upon ligand binding are usually neglected in high-throughput screening. Conformational changes of proteins upon ligand binding can often be accounted for by changes in the side chain orientation. In our approach, we allow selected side chains to change their conformation as a first step towards treating protein flexibility. We show that with a suitable search strategy, the side chains can adapt to the ligand during the docking simulation process.

The structure of this thesis follows the chronology of our research efforts. We first started with testing the accuracy of our docking algorithm (see chapter 6). To calculate binding energies in a good approximation, it is necessary to first determine a realistic protein-ligand conformation. We show that with our approach we can reliably and reproducibly predict ex-

perimentally observed binding orientations. In this chapter, we also discuss general problems in protein-ligand docking and point out specific problems which are not yet captured with our scoring function.

In the next chapter, we analyze the problem of protein flexibility and the shortcoming of using only one rigid protein structure for docking simulations (see chapter 7). We implemented the possibility of using flexible side chains in our approach. In large-scale database screens we compared the influence of rigid and flexible protein models with each other. We show that flexible protein models result in an increased reliability of the screen and in the identification of a higher number of good binding ligands.

Receptor-ligand interactions are calculated using many approximations. In a further study we investigated if the accuracy of binding energies could be improved by employing parameters obtained from quantum mechanical calculations (see chapter 8). We show that by incorporating the results of quantum mechanical calculations for the receptor only, the overall accuracy of the whole simulation can be increased. This is an important result for high-throughput screening, because the time consuming quantum mechanical calculations can be done separately in advance. The more accurate parametrization of the receptor can then be used for in the classical screening application. The computational effort for the screening remains the same, but the accuracy of the docking results increase.

We have thus developed a high-throughput docking approach which allows us to identify good binding ligands in large databases. Including protein flexibility by allowing the side chains to alter their conformations results in a more realistic model of proteins. Applied to docking simulations of databases, our approach is less biased to the rigid, experimentally measured protein crystal structure which gives us the possibility to discover more diverse good binding ligands. In addition, the overall accuracy of our approach can be enhanced further by integrating quantum mechanical calculations into our description of the proteins.

Chapter 1

Introduction

Drugs work in many different ways in biological systems. Focusing on the human organism I will introduce the main mechanisms of drug interactions in this chapter and also briefly summarize the historic and present development of the rational design of drugs.

1.1 Drug activity

The human biological system, even the part which is already understood, is very complex. In the following I will describe some general examples of why and how drugs act. For a more complex, introducting, overview, the reader may consult the books of G.L. Patrick [128] and A. Gringautz [57].
Pharmaceutical drugs can be classified in different ways: By the protein they bind to, by their chemical structure or by their pharmacological effect. In the following, I will focus on the function of the proteins, drugs interact with. Drugs often act on enzymes, which may then perform a reaction or which are hindered at performing a reaction, or may also act on receptors, which respond to a signal or are blocked to respond.

1.1.1 Drugs acting on enzymes

Enzymes are proteins that act as biological catalysts in the human body. Many complex reactions required by our metabolism do not take place spontaneously under physiological conditions. Enzymes have the ability to decrease the activation energy and consequently allow these reactions to take place at a lower temperature, as illustrated in figure 1.1. Enzymes have an active site, to which substrates can bind. These substrates are the reactants of the catalytic reaction which is performed by the enzyme. After this reaction all products are released and the enzyme, which did not alter during or after the process, is ready to repeat the cycle. Such an enzyme activity is illustrated in figure 1.2.

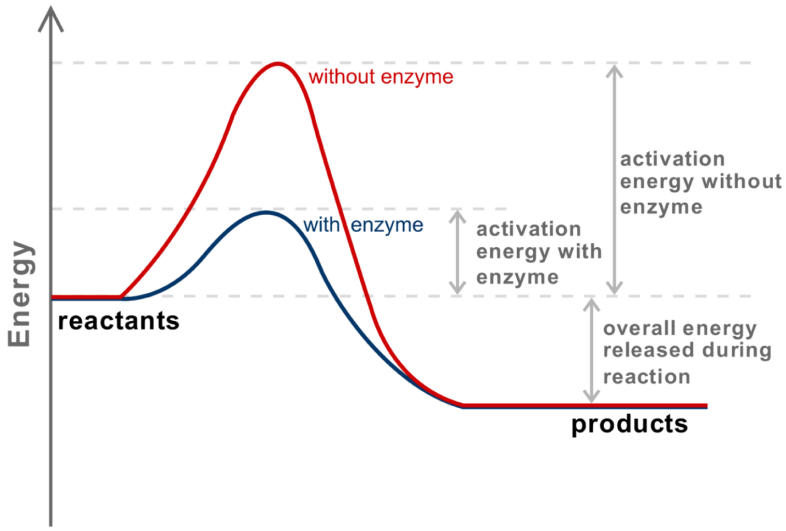

Figure 1.1: Illustration of a chemical reaction with and without enzyme. The red curve illustrates a reaction without any enzyme activity. As illustrated, if an enzyme acts catalytically at the chemical reaction (blue curve), the necessary activation energy to start the reaction is significantly lower.

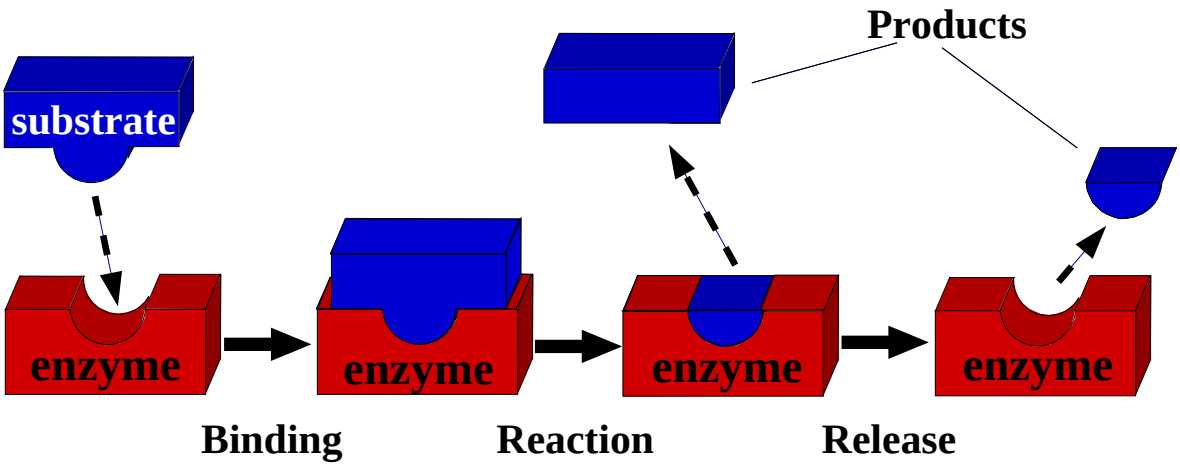

Figure 1.2: Illustration of a chemical reaction involving an enzyme. After a substrate has bound (first and second picture from the left) a chemical reaction takes place. After the reaction is completed, the products are released and the enzyme is ready for the next reaction (first and second picture from the right).

Enzymes are very important for cell metabolism. They are specifically adapted to fulfill certain tasks in the cell and thus keep the cell maintaining. Usually, one enzyme participates only in one chemical reaction. The specificity of the enzyme originates from its unique three-dimensional conformation [128].
Many enzymes, which are present in human cells, can not be found in bacteria or vice versa. The cells of bacteria and human cells seem to be quite different. This fact is also one of the starting points for eliminating, for example, dangerous bacteria in the human body. Using specific tools (drugs), the cell processes of unwanted bacteria can be disturbed and the cells can be killed by disrupting their metabolism. Penicillin, for example, acts just in that way. It blocks enzymes, which bacterial cells need to maintain the bacterial membrane [155]. If their activity is disrupted, the bacterial cell can not maintain itself and is stopped from multiplying itself.
Because of these reasons it is important in drug discovery to investigate and understand the metabolism and mechanism of these alien intruders in the human body, unwanted cells in the biological system, and then search for strategies. If specific enzymes are recognized as an integral part of a disease, they can be structurally resolved and one can search for drugs, that would stop these enzymes to work.
Figure 1.3 illustrates this idea. If a drug has a stronger affinity to bind to the enzyme than the natural substrate, the catalytic reaction can be stopped. Because the substrate does not have the enzyme as a partner no chemical reaction can take place. If this chemical reaction is important for the cell metabolism or for cell functions, the drug will lead to biological effects. The described process and the illustration is only abstract and exemplary. There are different ways to hinder an enzyme at working [15].

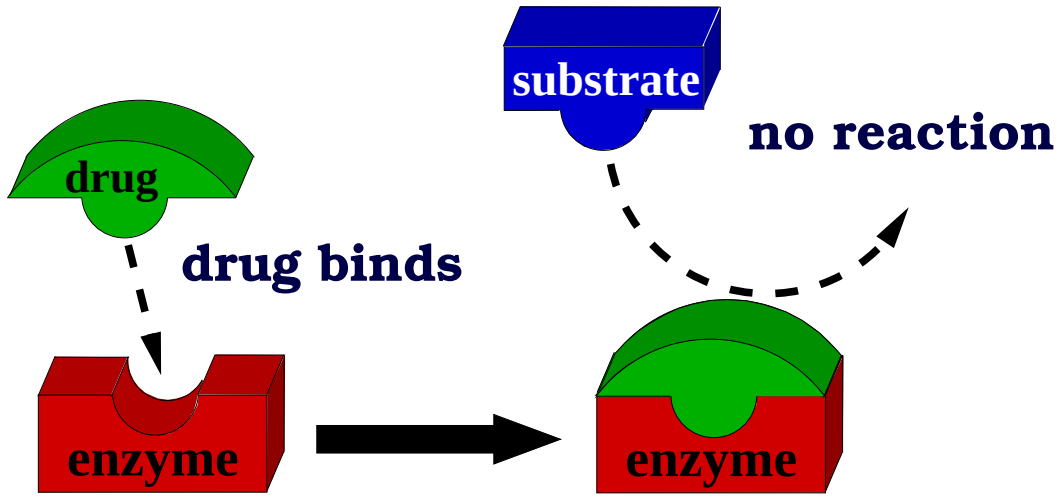

Figure 1.3: Illustration of drug activity upon an enzyme. Because of the bound drug (left picture), the natural substrate can not bind and the enzymatic process is stopped (right picture).

1.1.2 Drugs acting at protein receptors

The endocrine and the nervous system are the two important control systems of the human body. Receptors are an integral part of both systems. The endocrine system directly uses chemical molecules as their messengers to cells. Of course these molecules should only deliver their message to those cells they were sent to. Due to receptors, the addressee, i.e. the recipient cell can be identified.
Likewise, receptors are also important for the nervous system. Nervous cells can not communicate directly with each other or with other cells (muscles or glands) by electric potentials. They are interconnected by chemical synapses. However, instead of communicating through electric impulses as signals are transmitted in neurons, they use neurotransmitters. An electric impulse causes many neurotransmitters to be released from the axon terminal to the synapse. Receptors situated at the outside of the cell membrane of dendrites, have the task to identify these molecules and transmit the message accordingly; they are an essential part of the biological signal transduction.
Receptors, which are proteins situated in a cell membrane, fulfill different tasks:

- Information exchange between cells
- Regulation of the ion flow in ion channels
- Regulation of protein synthesis due to binding to DNA

The tasks of various receptors are so different from each other, that they can only be described in an exemplary and abstract way. Figure 1.4 displays a possible signal pathway of hormones. When a hormone finds its corresponding addressee, then it can bind tightly to the receptor,

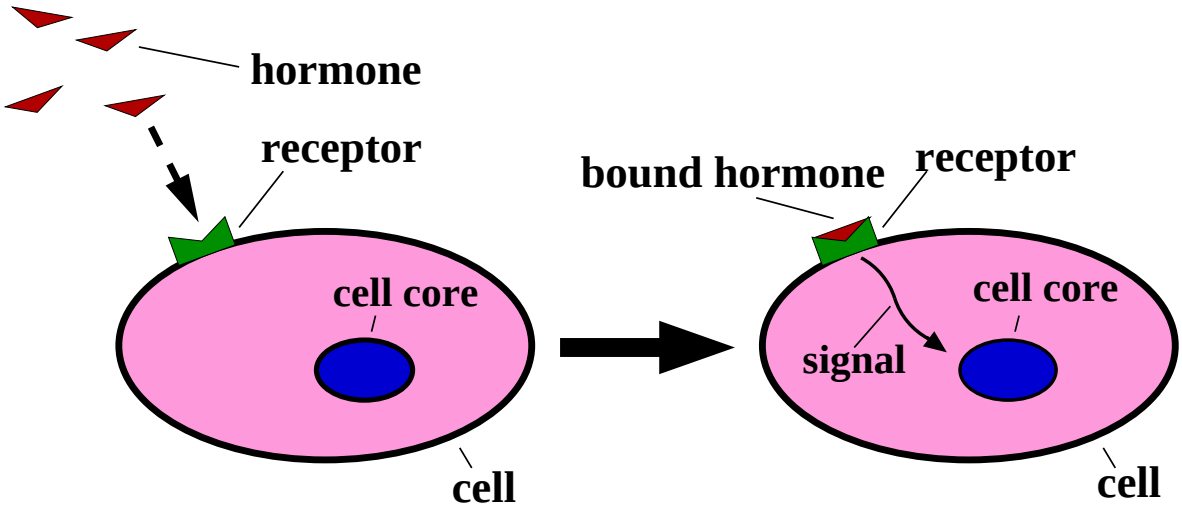

Figure 1.4: Schematic representation of the signal pathway of hormones. If a cell has the corresponding receptor to a hormone, then it can bind tightly to the receptor and a signal is transferred into the cell (here to the cell core).

as illustrated in the picture. In doing so the receptor alters its conformation slightly and a secondary messenger inside the cell is released. Such a signal could activate, for example, another receptor or enzyme, which could finally result in a new synthesis of proteins. In contrast to the interaction with enzymes (see figure 1.2), messengers acting on receptors do not change their chemical structure upon binding. Usually the binding energy is also not very strong. After some time the messenger is released from the receptor and drifts away.
Receptors are essential for inter-cell communications, because polar compounds can not cross the lipid bilayer of the membrane [15] and consequently no direct interaction within the interior of the cell is possible.
In the following, I will present two other examples to further illustrate the way signals are processed by the receptor. Figure 1.5 displays a possible receptor-induced ion-channel opening. Before a messenger molecule binds to the receptor, the relaxed receptor structure hinders ions to flow through the channel. When the receptor structure is 'activated' due to the binding of the molecule, the receptor conformation changes which leads to an un-blocking of the ion-flow. Sometimes the conformational change of the receptor can involve large parts of the structure, but often slight changes of the receptor structure, for example changes in the side-chain conformation, can have such an effect. Of course, this process could also happen in the reverse order. Due to the binding messenger, the ion-channel may get blocked.
Again, if one can design drugs that fit tighter into the receptor pocket than a respective messenger, the receptor can be successfully hindered to alter its conformation. In our example consequently the ion-channel will not open due to the blocking of the opening-mechanism. Many drugs work in such a way. Anesthetics, for example, reduce the excitability of nerves in this manner [15] so that the patient does not feel the pain anymore.

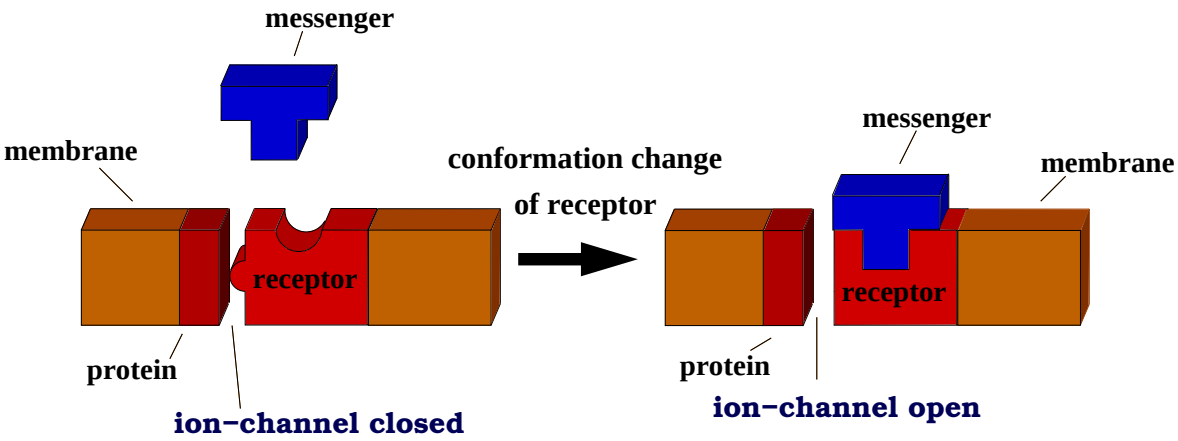

Figure 1.5: Schematic illustration of how an ion-channel is opened by an induced conformational change of a receptor. Before the messenger binds to the receptor, the receptor structure hinders ions to flow through the ion channel. After the messenger has bound, the receptor changes its conformation, which leads, in this example, to an opening of the ion-channel.

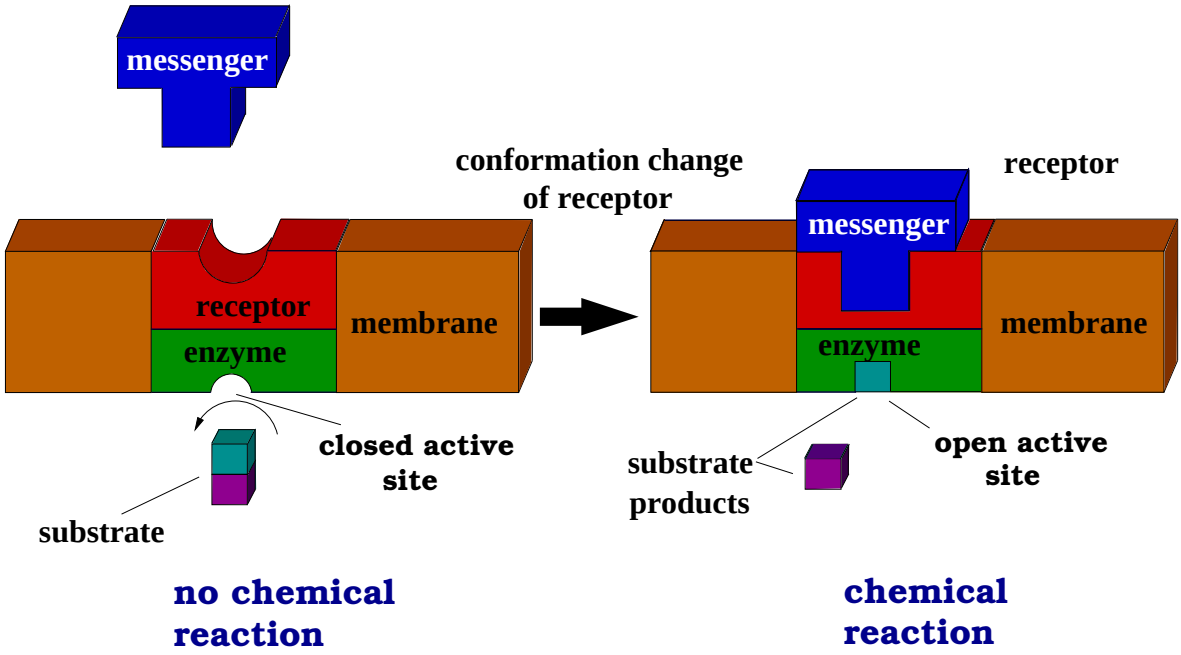

Figure 1.6: Sketch of how an induced conformational change of a receptor structure enables enzymatic reactions to take place. Before a messenger is being bound to a receptor, the active site of a neighboring enzyme in the cell interior is closed. If the messenger is bound to the receptor, the receptor changes its conformation and induces an additional conformational change of the enzyme. As a result the active site of the enzyme opens and substrates in the cell interior can start a catalytic reaction.

Figure 1.6 illustrates how an enzyme in the cell interior is activated by a signal from the cell exterior. A messenger molecule binds to a receptor and induces a conformational change of the receptor. In this example the new receptor conformation also induces directly a conformation change of a neighboring enzyme. Before the messenger was bound, the enzyme was 'switched off'; the enzyme is inactive. Consequently no bio-catalytic reaction could start. Due to the altered enzyme conformation, substrates can now bind to the active site of the enzyme and start the catalytic reaction. The resulting products may lead to further reactions.
Mechanisms, following the abstract scheme of figure 1.6, controlled by many different types of receptors are abundant in different human cells. For instance, the class of G-protein coupled receptors (GPCRs) act in a similar way. However, instead of a direct enzyme alteration, the receptor activates other proteins which then activate an enzyme for catalytic reactions. GPCRs are involved in many different stimulus-response pathways from inter-cellular communication to physiological senses.
Now drugs can be used to influence the signaling pathway of messengers. They can hinder directly or indirectly a messenger from binding to the receptor and thus disconnecting the signal-pathway.

1.1.3 Bioavailability

It is crucial that a drug binds strongly to the receptor, but no matter how well a substance binds, it is useless, if it can not be transported to the receptor. The field of pharmacokinetics describes mathematically if, in which concentration and how quick a drug gets to the target location and how and when it is removed from the human organism [15].
The most important processes in pharmacokinetics are abbreviated by the acronym LADME:

- Liberation (L): If the drug is in a solid phase, it first has to get dissolved.
- Absorption (A): To enter the blood stream the drug has to cross different barriers by passive diffusion. These barriers are usually membranes in the duodenum. Highly polar and large molecules have difficulties crossing the lipid bilayer membranes.
- Distribution (D): The bloodstream will not distribute the medicament equally in the human body. Because of different barriers in the human organism (for example the blood-brain-barrier), a drug may not reach some parts of the human organism. Consequently, the drug concentration will be higher in some regions (organs) than in others. Additionally, the molecule distribution often also depends on how good a drug flows in the blood. Molecules that are too large will have difficulties to flow quickly to the target and the chances of degeneration are higher.
- Metabolism (M): There are special enzymes in the human body that detect alien substances and convert them into products which are easy to excrete.
- Excretion (E) describes the process of how the drug leaves the human body.

Several methods have been developed that attempt to mathematically characterize the LADME processes. Through statistical evaluation of known medicaments the Lipinsky rule of five [98] characterizes already roughly which compounds have a chance to reach their target and which not.
If difficulties arise by an oral intake of a medicament, other transport ways can be investigated. Some drugs, for example, are injected locally. Presently, even research efforts progress to develop transport vessels, which can bring non-solvable drugs to the target and release them there.
In our research we are focusing on the protein-ligand interaction only. We want to help to discover compounds that bind with a high affinity to a specific target. The transport of drugs is not our main interest.

1.2 Brief overview of rational drug discovery

Drugs are rather sparsely distributed in the chemical space of possible candidates [98]. It would take too much time to blindly test every possible compound in sito. Therefore, different drug discovery strategies were developed and refined over the last 100 years based on the technology and resources of the time.
Modern drug research is often traced back to Emil Fischer. With his metaphor of 'lock' and 'keys', he gave the field of drug research the direction for the next century: Enzyme and inhibitor 'must join one another as lock and key to be able to exert a chemical effect'. [45].
Paul Ehrlich's concept, which he formulated in 1913, was also important for further developments: 'corpora non agunt nisi fixata'; which means that compounds that do not bind have no biological effect.
The strategies of rational drug design can be divided into two different classes: Methods that involve the knowledge of the three-dimensional structure of the active center of enzymes/receptors and strategies that do not require such knowledge of the protein [15].
The first class could not emerge before the 1970s. With the pioneering work of Max Ferdinand Perutz [116] and Sir John Cowdery Kendrew [80], honored both with the Nobel prize in 1962, methods were developed to determine the three-dimensional structure of macromolecules. These methods progressed with the time and became more and more to a difficult routine. In this section, I will briefly discuss the main strategies of the two classes.

1.2.1 Strategies without knowledge of the protein

In the search for well binding drugs, chemists attempted to extract useful properties of a known binding compound in order to use this knowledge to find a good binding drug. With the analogy of E. Fischer a new 'key' should be found by looking at the 'key' only. The idea is that all binding compounds to one target are similar in some way. Even though not belonging to the same 'chemical family', there is often a similarity in hydrophilic, hydrophobic, aromatic or other properties present.
These properties, descriptors, are used to search databases of ligands for similarity. Over time the descriptors became more and more complex: From one dimensional to three dimensional. Now databases are even searched with the help of neural networks [183]. Especially the pharmacophore model, a three-dimensional descriptor strategy, is widely used. This method uses the three-dimensional orientation of functional groups (hydrophobic, hydrogen binding groups and further) to describe a molecule [15] and then search for similarity in a database of compounds. One of its problems lies in sampling the conformational space of the molecules and aligning it to the pharmacophore. An additional problem is that the search based on similarity applies only well to targets which do not change much their conformation upon ligand binding.
In drug discovery, the search for new types of well binding ligand structure and the optimization of those should be differentiated. When a promising ligand has been found, the newly

found ligand structure, the lead structure, can be used as the starting point for further studies to increase the affinity or the transport properties of the ligand to the target. By varying the chemical constituents of the lead structure different new molecules can be synthesized and tested. The optimization of these compounds is not only based on the interaction of the new drug with the protein, but also on an improvement of the drug transport in the human body. Hydrophilic and hydrophobic properties of molecules influence in what concentration they reach certain regions in the human body [15]. With the method of quantitative structure-activity relationship (QSAR) the molecule is decomposed in different groups each contributing differently to a biological activity scale. The method allows to compare different well binding analogs of a new lead structure and to optimize them for transport in the biological system and therefore for the best biological activity.

1.2.2 Strategies employing knowledge of the protein

Such strategies, also called molecular docking, analyze the interaction properties of a small molecule to a protein. The approaches can be differentiated by the strategy used for sampling different protein-ligand conformations and for estimating how well a ligand can bind.
Molecular docking methods have been developed in the last thirty years. In the beginning both protein and ligand were treated as rigid entities and binding was simply based on geometric criteria; the programs searched for shape only [48]. But soon the importance of the chemistry in ligand docking [151] and also of ligand flexibility was realized and several improved strategies were developed.
I will now distinguish between two approaches: Molecular mechanics and fragment based approaches.

Molecular mechanics based approach

At the beginning both ligand and protein were considered as rigid unities. Nowadays it is common to allow for some degree of ligand flexibility during docking, i.e. some inherent adjusting to the protein environment. The protein on the other hand is usually kept rigid.
In molecular mechanics (MM) docking approaches (see chapter 4) the ligand as a whole is simulated. Through a step by step process the ligand adapts to its environment and the system energy is optimized by changing flexible bonds of the ligand and its orientation. The available methods are distinguished by the scoring function that evaluates the binding energy by a score and the search algorithm.
The scoring functions can be classified either in regression-based or interaction-based scoring functions. The former indirectly uses information of known complexes. Through statistical analysis distant-dependent potentials for different atom pairs or functional groups are constructed on the assumption that the individual contributions obey Boltzmann statistics. The frequency with which a specific geometry appears is related to the energy of that geometry. The advantage of statistical analysis scoring functions is that relationships can be evaluated, which are either not fully understood or difficult to model in a classical approach. On the

other hand docking failures, the reason why some ligands do not bind, may also be more difficult to understand.
Interaction-based scoring functions are similar to MM force fields. Both describe the interaction of the ligand with the protein by classical potentials based on physical interactions. The protein-ligand interactions are rooted in physical principles and have problems with features whose origins are not clearly understood. But even in this approach many interaction parameters are usually fitted to experimental data.
Using such scoring functions or force fields already solves one problem that pharmacophore models have to face: The ligand deformation is treated in the same way as the interaction of the ligand to the protein. Consequently, it is easier to estimate if the resulting docked geometry is realistic or not.
Because the screening programs must be very fast to be applicable, different techniques have been developed to speed up the search for the best protein-ligand conformation. Monte Carlo methods (see chapter 3) are used frequently as are genetic algorithms, which is used in the program Gold [73] for example.

Fragment based approach

Fragment based approaches use a different technique to sample the huge conformational space of a protein-ligand complex. The ligand is separated into different fragments which are separately docked to the protein and then finally assembled into the whole molecule again [160]. Different strategies are used for the reconnection of the broken bonds. One very popular approach is the incremental construction algorithm which is implemented into the program FlexX [134]. This algorithm starts with docking a base fragment to the protein and then sequentially adds the other fragments in energetic favorable directions [183]. The final protein-ligand conformation is usually evaluated by an empirical scoring function to determine the affinity of the ligand to the protein. This approach does not fully explore conformational space of the ligand. However by concentrating first on a larger or quite characteristic fragment, this strategy has proven to be successful and is, for example, intensively used on supercomputers of the Bayer CropScience Deutschland GmbH.
Interestingly, the de-novo ligand design methodology is strongly linked to this approach. Instead of screening a database of ligands, a totally new ligand is constructed by docking fragments of a database to the protein.

1.2.3 Recent developments

Over the years many rational drug discovery methods have been developed and are being used. Search strategies without the knowledge of the protein structure are still very important. For example, most membrane proteins are difficult to resolve. What is more, drugs can have side-effects, i.e. they may also bind to not designated proteins. The strategies, that compare molecules just by similarity, can help to quickly test drugs for potential unwanted effects. But unfortunately these methods are not working successfully in cases, in which no good binding compound is known, because the potential compounds can not be compared with already known ligands.
With our work we aim to contribute to the molecular docking field. Up to now, conformational changes of the protein structure are usually not considered for docking and are, if considered, still on an experimental footing. The necessity to account for a flexible protein structure is well known and published [24]. The present status regarding protein flexibility can be compared to the time docking of flexible ligands emerged. At the beginning it was computationally too expensive to allow for ligand flexibility. As a consequence, methods were developed to first generate different ligand conformations and to then dock them rigidly to the rigid protein structure [78]. Now ligand flexibility is standard in most methods but protein flexibility not. As first methods to treat ligand-flexibility one approach to solve this problem is to generate

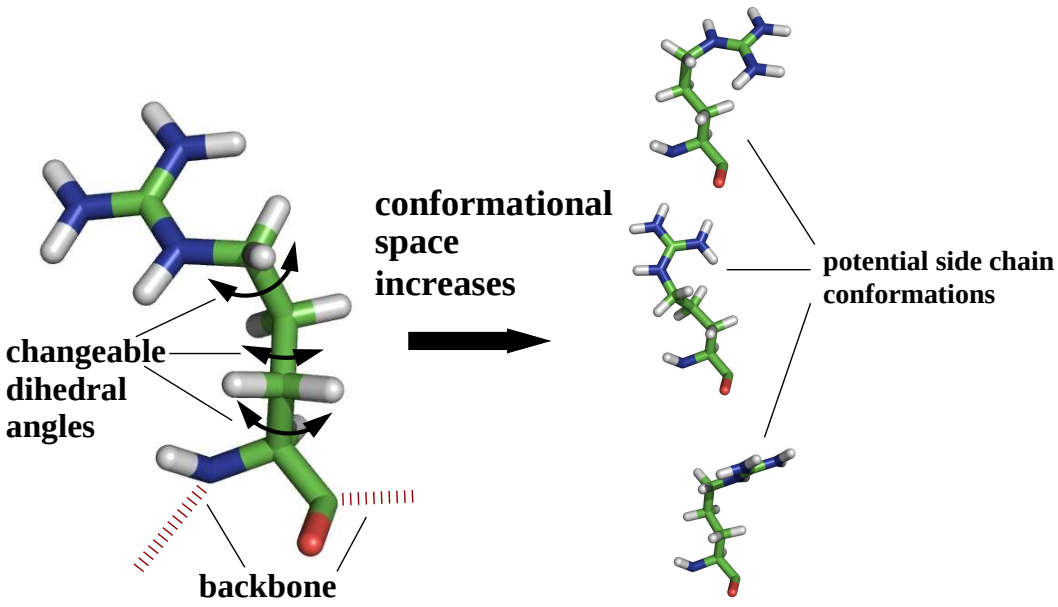

Figure 1.7: Illustration of possible side chain flexibility, as used by our program *FlexScreen*. *FlexScreen* allows, for example, the side chain arginine a maximum of three degrees of freedom: three dihedral angles are allowed to be changed. Possible dihedral angles depend on low energy interactions of the side chain with itself and the environment.

a set of different protein structures and then to dock flexible ligands to each of them.
Full protein flexibility for high-throughput screening is not yet possible at the moment, since accurate and affordable simulation strategies are lacking. In our approach, *FlexScreen*, we allow for full side chain flexibility (as illustrated in figure 1.7). We hope that using full side-chain flexibility will enable us to describe the majority of conformational changes upon ligand binding.

Chapter 2

Thermodynamics of Protein-Ligand Association

In the previous chapter different mechanism of drug activity in the human organism were characterized. In this chapter protein-ligand interactions are examined in a more theoretical manner. The affinity of a ligand to a protein is the free energy difference between a ligand bound to the protein and both being separated in solution.
Protein-ligand complexes can be structurally resolved by several methods: X-ray crystallography and nuclear magnetic resonance measurements. The affinity of a ligand to a protein is measured by stochastic observations: Either the binding processes are counted during a period of time or macroscopic properties are measured. Usually experimentalists measure the concentration difference between bound and unbound ligands in a protein-ligand solution.
Therefore, I will first relate the microscopic binding free energy with macroscopic properties and then quantify the binding free energy. Because the binding free energy is difficult to calculate for a specific protein-ligand system, necessary approximations and assumptions will be also introduced and explained.

2.1 Thermodynamic Basis

Protein-ligand docking describes the interaction between a ligand and a protein in solution under constant pressure P and constant temperature T. Systems under these environmental conditions are described with the Gibbs free energy [137] as the corresponding thermodynamic potential of the system

$$G = H - TS = E + PV - TS \tag{2.1}$$

and the Gibbs free energy change $\Delta G = \Delta E + P\Delta V - T\Delta S$, with enthalpy H, energy E, volume V, pressure P and entropy S.
The volumetric work PV in liquid systems is negligible, because the volume per molecule is small and the pressure in these liquid systems is rather low; the value PV is usually less than the thermal energy kT [41].

In the following, I will neglect the marginal difference between H and E. Consequently, Gibbs free energy and Helmholtz free energy are not differentiated and will here be referred to solely as the free energy

$$A = E - TS. \tag{2.2}$$

The thermodynamic equilibrium of a liquid system corresponds to the global minimum of the free energy. Guided by experimental observations Anfinsen [5] formulated in 1962 the hypothesis that the native structure of proteins, observed by X-ray scattering, corresponds to the global minimum of the free energy. He showed that proteins spontaneously assume their native structure.

The hypothesis of Anfinsen is well supported by experiments and has been directly applied to computational studies of protein folding. In the case of medium sized proteins (40 to 150 amino acids) several stochastic studies show that the native state of the protein corresponds indeed to the global minimum of the free energy of the system [145, 65].

In protein-ligand docking we assume the situation to be similar. We suppose that if the ligand has bound to a protein, the system is in thermodynamic equilibrium. This enables us to apply a thermodynamic methodology.[1]

In statistical physics the free energy is defined as the logarithm of the partition function Z, obtained by an integration over the propability density states of the thermodynamic system

$$F = -kt \ln Z = -\beta^{-1} \ln Z, \quad \text{with} \quad \beta = \frac{1}{kT}. \tag{2.3}$$

The propability density states $\rho(\mathbf{X})$ obeys a Boltzmann distribution $\rho(\mathbf{X}) = \frac{e^{-\beta H(\mathbf{X})}}{Z}$.

In general no defined ground state of thermal equilibrium has to exist. If the energy difference between several lowest free energy states is smaller than the thermal energy kT and the corresponding conformations to these minima differ widely from each other, then no ground state can be defined. However, by assuming one unique global minimum of the free energy, the free energy of the system can be expressed as an expansion around the ground state with the free energy A_0 at the minimum:

$$A = A_0 - T(S_{water} + S_{conformation}). \tag{2.4}$$

The free energy A is expressed as the sum of the energy at the groundstate A_0 and additionally the entropic contributions of water (S_{water}) and of the protein and the ligand ($S_{\text{conformation}}$). If essentially only two states in the protein-ligand systems (bound and unbound ligand) exist, it is a good approximation to determine the binding affinity by calculating the free energy difference between these two states.

[1] The strong postulation of assuming the bound state as the state of thermal equilibrium is not necessary. Membrane proteins, for example, have to be considered in the environment of the cell membrane, separately in solution they are in a different conformation. Because we do not assume global structurally changes of a protein upon ligand docking, we postulate that the protein conformation, we use for docking, corresponds to a low free energy minimum taking also the environment into account. We consider the bound ligand as a perturbation to the protein in its environment. Due to the docked ligand a new free energy minimum is obtained, which can also lead to structural changes of the protein, whose energetic cost is not very high.

2.2 Thermodynamic view of protein-ligand affinity

In the following, I will first relate the binding free energy of an individual ligand to macroscopic measurable properties, experimentalists can measure. Then, the binding free energy is analyzed in detail.
A very dilute solution of proteins (P) and ligands (L) is considered. If the protein and the ligand form a protein-ligand complex (PL) without the formation of any covalent bonds, such a behavior can be interpreted as a reversible chemical reaction

$$P + L \Longleftrightarrow PL. \tag{2.5}$$

If the system is in thermal equilibrium, the reaction rate into both directions is the same. An open system is considered that interchanges energy with a constant temperature bath. Consequently, in thermal equilibrium the temperature and the pressure of the system are constant. In thermal equilibrium the Gibbs free energy of the system, the corresponding thermodynamic potential of the system has its global minimum. Thus, the differential of the Gibbs free energy is $dG = -SdT + VdP + \sum_i \mu_i dN_i = \sum_i \mu_i dN_i = 0$, with μ_i being the chemical potential of particle type i, defined as $\frac{\partial G}{\partial n_i}$ (n_i being the number of particles type i). Using the Euler theorem the Gibbs free energy can be expressed as [105]

$$G = \sum_i \mu_i N_i \tag{2.6}$$

and the differential of the Gibbs free energy can be rewritten in the following form [105]:

$$dG|_{const\ T,\ p} = \sum_i \mu_i dN_i = \left(\sum_i \nu_i \mu_i \right) d\lambda = 0, \tag{2.7}$$

being $dN_i = \nu_i d\lambda$. Considering one complete reaction ($\nu_i = -1$ for the products and $\nu_i = 1$ for the reactant), the chemical potentials of reactant and products are equal at thermal equilibrium:

$$\sum_i \nu_i \mu_i = 0 \qquad \Longrightarrow \qquad \mu_{PL} = \mu_p + \mu_L. \tag{2.8}$$

With the Gibbs-Duhem relation, $SdT - Vdp + \sum_i N_i d\mu_i = -Vdp + \sum_i N_i d\mu_i = 0$ [137], one can show by an integration (in dilute solutions the osmotic pressure of one species is the same as for an ideal gas [169, 40]) that each chemical potential can be expressed in the following way

$$\mu_i = \mu_i^0 + kT \ln \frac{p_i}{p_i^0} = \mu_i^0 + kT \ln \left(V[i] \right). \tag{2.9}$$

μ_i^0 is a reference chemical potential corresponding to the reference pressure p_i^0 for one element or as usual for one Mole of element type i and $[i] = \frac{N_i}{N_{ref,i} V} = \frac{\phi_i}{V}$ is the fraction density of species i at the reference volume V with the number fraction ϕ_i in relation to the reference unit.

The binding free energy of one ligand and a protein is the free energy difference of bound and unbound state. According eq. 2.8 we express it with the reference chemical potentials

$$\Delta G_{PL-P,L} = G_{PL} - G_{P,L\,\text{sep.}} = \mu^0_{PL} - \mu^0_P - \mu^0_L. \tag{2.10}$$

Combining eq. 2.10 with eq. 2.8, the binding free energy can be expressed in the macroscopic properties of concentrations

$$\Delta G_{PL-P,L} = -kT \ln \frac{[PL]}{V[L][P]}, \tag{2.11}$$

whereas the association constant is defined by concentration densities as

$$K_{PL} = \frac{[PL]}{[P][L]} = V \frac{\phi_{PL}}{\phi_P \phi_L}. \tag{2.12}$$

This equation relates the observable macroscopic properties to the binding free energy. Now, the binding free energy will be expressed by the integral of state. In the following, K_{PL} is derived from thermodynamical arguments [101, 148].
However, instead of a solution of proteins and ligands, solely one ligand and one protein in solution are considered. Instead of comparing concentrations, thermodynamic expectation values are compared of being in the bound and in the unbound state in volume V. Thus, the macroscopic observation value 'concentration' is related to averages over a sufficiently long time of one protein and one ligand in solution.
In the dilute limit, $V \to \infty$, the thermodynamic expectation values of the bound state have the limit value of 1: $\phi_P = \phi_L \to 1$. Thus, the association constant crosses over to the limit value $K_{PL} = V\phi_{PL}$. I will now briefly summarize the determination of the bound volume density and so the association K_{PL}.

A system consisting of a ligand, a protein and water molecules, with N_L, N_P and N_S atoms respectively, has the total number of $3(N_S + N_L + N_P) - 6$ degrees of freedom. $3N_S$ solvent, $(3N_L - 6)$ internal ligand and $(3N_P - 6)$ internal protein degrees of freedom and 6 additional degrees of freedom which correspond to the relative position of the ligand to the protein.
The coordinate system shall now be defined by placing the protein (P) at the coordinate origin. Then, a particular configuration of the system is determined by the coordinates of the solvent atoms $\mathbf{r}_V$, the position and orientation of the ligand L to P with $(\mathbf{r}, \mathbf{\Omega})$ and their internal coordinates $\mathbf{q}_L$ and $\mathbf{q}_P$.
The interaction between L and P can be described by a potential of mean force (PMF) [101]

$$\omega(\mathbf{r}, \mathbf{\Omega}) = -\int_{r=\infty}^{r,\Omega} \left\langle \frac{dU(\mathbf{r_V}, \mathbf{q}_L, \mathbf{q}_P, \mathbf{r}', \mathbf{\Omega}')}{d(\mathbf{r}', \mathbf{\Omega}')} \right\rangle d\mathbf{r}' d\mathbf{\Omega}', \tag{2.13}$$

with the potential of the whole system U. The PMF $\omega(\mathbf{r}, \mathbf{\Omega})$ describes the reversible thermodynamic work to bring the ligand L from infinitely far away to the position and orientation

$(\mathbf{r}, \mathbf{\Omega})$. This potential is independent of the solvent and internal molecular coordinates, since they are already averaged over by a Boltzmann weighted integration.
$C(\mathbf{r}, \mathbf{\Omega})$ shall be defined as a function that is 1 if the ligand is bound to the protein and 0 otherwise. As the criteria for a bound state, we define that ω must be significantly lower than the negative thermal energy kT. Then the association constant can also be expressed as a thermodynamic observable [101]

$$K_{PL} = \frac{1}{8\pi^2} \int C(\mathbf{r}, \mathbf{\Omega}) e^{-\omega(\mathbf{r},\mathbf{\Omega})/kT} d\mathbf{r} d\mathbf{\Omega}. \tag{2.14}$$

Eq. 2.14 is exact but difficult to calculate. Assuming that the potential energy surface around the ground state of the system can be approximated by harmonic potentials, this assumption helps to gain further insight [101]. Now the motions of the system can be described by a multivariate Gaussian propability distribution with a covariance matrix of the coordinate fluctuations $\sigma = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle$. The potential of mean force can be rewritten as $\omega(\mathbf{r}, \mathbf{\Omega}) = \omega_{min} + \frac{1}{2} \Delta \mathbf{x}^T \bar{\mathbf{F}} \Delta \mathbf{x}$, where $\mathbf{x}$ includes the remaining coordinates $\mathbf{r}$,$\mathbf{\Omega}$, $\mathbf{q_L}$ and $\mathbf{q_P}$ and the matrix $\mathbf{F}$ contains the elements $F_{ij} = kt(\sigma^{-1})_{ij}$ [4]. Then eq. 2.14 can be expressed with the factors of the fluctuations of each coordinate [101]:

$$K_{PL} = e^{-\beta\omega_{min}} \sqrt{8\pi^3} \sigma_x \sigma_y \sigma_z \frac{\sigma_\chi^3}{\sqrt{6^3\pi}} \sqrt{\frac{|\sigma^2_{L,bound}||\sigma^2_{P,bound}|}{|\sigma^2_{L,free}||\sigma^2_{P,free}|}}. \tag{2.15}$$

In order to understand the main binding contributions in a more qualitative manner, we assume that the internal fluctuations are zero for the ligand and the protein and that the PMF ω is a square well with a constant binding energy (E) with the width σ_x, σ_y, σ_z for the x,y,z coordinates [148] and in total $\sigma^3_{coord} = \sigma_x \sigma_y \sigma_z$. Then, eq. 2.14 can be evaluated as

$$K_{LP} = \sigma^3_{coord} e^{-E/kT} \qquad \text{and} \tag{2.16}$$

according to eq. 2.11, the free energy change can be expressed by

$$\Delta A = E - kT \ln \frac{\sigma^3_{coord}}{V}. \tag{2.17}$$

The free energy change is calculated as a sum of the PMF (E) and the translational entropy of the whole ligand.
In case of the more general evaluation (eq. 2.15) ΔA also includes the contributions of the internal vibrations of the ligand and of the protein (relative to the unbound state) and of the fluctuation of the orientation of the ligand to the protein.
The important result is that it is thermodynamically sound to calculate the free energy change as a sum of a thermodynamically averaged binding energy and an entropic contribution. However, it has not yet been explained how the potential of mean force can be calculated in practice. This will be discussed in the next section.

2.3 Separability of the binding energy

In biomolecular force fields the potential energy of a system is decomposed into several energy contributions (Coulomb energy, van der Waals energy, etc. [see section 4]). Even if this is correct, with the above derivation it is not yet certain, if such a decomposition is also possible for the free energy change ΔA. That this is possible, will be shown in the following.
The separation of the potential energy is now indexed by the variable i

$$U_{System} = \sum_i U_i. \tag{2.18}$$

Consequently, the canonical ensemble averaged configuration energy $\langle U_{System} \rangle$ is

$$\langle U_{System} \rangle = \langle \sum_i U_i \rangle = \sum_i \langle U_i \rangle. \tag{2.19}$$

Also, the free energy change can be expressed by temperature derivatives of the mean binding free energy [18, 19] (the kinetic free energy contributions result a constant and are therefore neglected)

$$\Delta A = \sum_{n=0}^{\infty} \frac{(-\beta)^n}{(n+1)!} \frac{\partial^n \langle U \rangle_\beta}{\partial \beta^n} = \sum_i \left(\sum_{n=0}^{\infty} \frac{(-\beta)^n}{(n+1)!} \frac{\partial^n \langle U_i \rangle_\beta}{\partial \beta^n} \right). \tag{2.20}$$

Such a derivation is the formal justification for separating the free energy into different energy contributions like Coulomb, van der Waals interactions, entropic contributions and many other according to the separation strategy of the force field (see chapter 4).

2.4 Necessary approximations of the binding free energy

The purpose of our study is to develop a high-throughput screening tool for drug discovery. We aim to calculate the binding free energy for as many ligands as possible in an acceptable amount of time. Necessarily, such calculations can not be totally accurate.
Under these conditions, we restrict ourselves to approximate only the relative binding free energy difference of ligands. Our approach should be able to rank good or bad docking ligands for one specific target (protein) according to their binding free energy.
One problem when approximating the binding free energy change, as we have seen with eq. 2.14, is to calculate the covariance matrix of the coordinate fluctuations [101]. Such calculation and analysis is nowadays standard. For example, it is included as an additional tool into the molecular dynamics (MD) AMBER program [130]). But nevertheless of its availability, it is very time consuming and is only then a good approximation to the entropic contribution, if suitable for the system.[2] High-throughput docking methods can estimate such contributions

[2]There are also other methods to determine the binding or relative free binding energy: Thermodynamic integration [104] or free energy perturbation methods show promising results in that direction [125]. However, both methods have the disadvantage that many different system configurations have to be simulated and equilibrated. Another quicker and more general, but less accurate, approach is the linear interaction method [6], which samples only the bound and unbound state. Also this methods shows promising results, when properly parameterized [6, 20].

only through empirical parameters. Presently they are simply neglected in high-throughput docking programs [93, 114, 50, 74].
Due to the size of the protein cavity, also the size of the ligand is restricted and only ligands of a similar size can dock well. We assume that entropic contributions at a specific protein cavity are similar for different ligands. Under that assumption, the ligands can still be ranked 'accurately', according to their calculated binding affinity.
Nowadays, it is a common routine to use further docking programs for those ligands, which are selected as promising good binding ligands by a high-throughput docking tool. Then, these entropic effects can be taken into account more accurately.

2.4.1 Thermodynamic Cycle

The interaction of the solute with the solvent can have an important impact on the binding affinity of protein-ligand interactions. The best way to treat the solvent in computational calculations is to include the solvent molecules as additional entities in the protein-ligand system (as in MD simulations). Again, such very accurate methods are also computationally very time consuming.

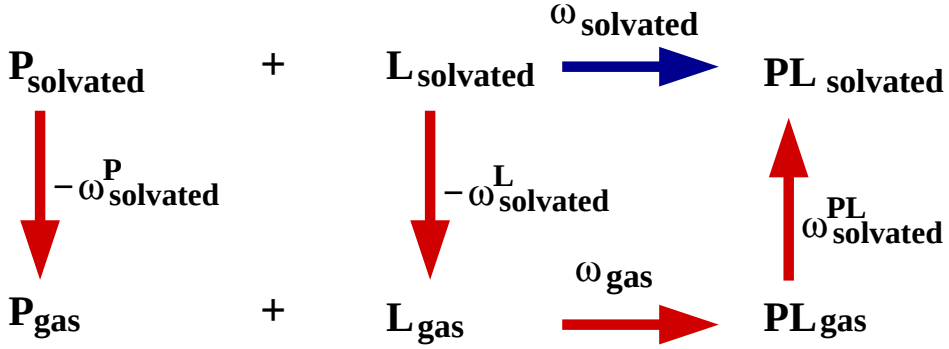

Figure 2.1: Illustration of the thermodynamic cycle. Two thermodynamic paths are shown. The blue arrows indicate the path at which the ligand and the protein are solvated and are getting bound in solution. The second pathway is sketched by the red arrows: Ligand and protein are separately de-solvated and first bind in vacuum/gas phase. Afterwards the whole complex is solvated again.

In the following, I will briefly introduce the thermodynamic cycle approach which determines the potential of mean force $\omega(\mathbf{r}, \mathbf{\Omega})$ by a different pathway (see figure 2.1).
The underlying idea of the thermodynamic cycle is that $\omega(\mathbf{r}, \mathbf{\Omega})$ is path independent. In figure 2.1 the upper line (blue arrows) describes the usual path: Ligand and protein are in the solution and are being bound with the binding potential $\omega_{solvated} = \omega_{min}$. The red arrows indicate the second path: Ligand and protein are first de-solvated, then they are bound in the gas phase (or vacuum) with the potential ω_{gas} and finally the protein-ligand complex is de-solvated again (brought from infinity into the water solution).

When each work, performed on each process, is energetically evaluated, then $\omega_{solvated}$ can be also expressed in the following way:

$$\omega_{solvated} = \omega_{min} = -\omega^{P}_{solvated} - \omega^{L}_{solvated} + \omega_{gas} + \omega^{PL}_{solvated}. \tag{2.21}$$

The thermodynamic cycle approach is a popular method, because usually it is assumed that the binding conformation of protein and ligand is the same in solution and in vacuum. Since in vacuum the docking simulations are by far more efficient than in solution, in which water molecules would have to be considered, less computational effort is needed. If the solvent is treated implicitly (see section 4.2), the solvation energies can be approximated in a computationally efficient manner.[3]

In many cases, it is justifyable to use such an approach, for example if the cavity does not have a direct connection to the bulk solvent and also if upon ligand binding almost all water molecules are removed from the cavity.

However, this is not a generally valid approximation. For example, the binding conformation in gas phase and solution are considered as being identical. But in open pockets (widely accessible to the bulk solvent) a favored binding conformation in the gas phase can be very different from the binding conformation in solution.

[3]The thermodynamic cycle would be accurate, if methods like the thermodynamic integration methods were used.

Chapter 3

Stochastic Optimization Methods

The native binding conformation of ligand and protein is usually assumed to be the global minimum of the free energy of the system: ligand, solvent and protein.
In the research field of protein-ligand docking two important objectives are:

a) To determine the groundstate of the ligand-protein system (neglecting all the protein and ligand entropic contributions)

b) To determine the free energy change upon binding at room temperature in order to compare the binding affinity of different ligands with each other

In this chapter, I will discuss the simulation methods for reaching these objectives.

3.1 Monte Carlo Simulation

Monte Carlo (MC) and other simulation methods are used to determine thermodynamic expectation values. For this purpose two approaches are often used. In kinetic simulations the equation of motion of the system are solved and thermodynamic expectation values are computed as time averages over the observables. In stochastic methods a chain of trajectories is constructed such that the frequency of each state approaches its thermodynamic propability for long simulation times. The Hamiltontian, which describes the microscopic state of a system, can be separated into kinetic and potential energy components. Since the kinetic energy contribution can be integrated to a separate constant, we do not have to consider it to describe the system. Each configuration of the state $\mathbf{q}$ is occupied according to the probability density distribution of the thermodynamic system

$$\rho(\mathbf{q}) = \frac{e^{-\beta H(\mathbf{q})}}{\int e^{-\beta H(\mathbf{q})} d\mathbf{q}} = \frac{e^{-\beta H(\mathbf{q})}}{Z}. \tag{3.1}$$

A thermodynamic expectation value $\langle X \rangle$ can be calculated by an integration over the whole configuration space and weighted with the probability density distribution of each state,

$$\langle X \rangle = \int X(\mathbf{q}) \rho(\mathbf{q}) d\mathbf{q}.. \tag{3.2}$$

3.1.1 Principles of Monte Carlo simulations

A Monte Carlo simulation is based on several ideas, which I will characterize briefly.

1. Stochastic process:
   The dynamic deterministic process as employed in molecular dynamic simulations is replaced by a stochastic process [142]. This fundamentally changes the simulation procedure. The Monte Carlo simulation can speed up the whole simulation, because the iterative change of the system conformation is independent of a realistic physical conformational change which molecular dynamic simulations would have to mimic. On the other hand, the construction of the next conformation is not as straightforward as in molecular dynamics. The transition probability from one state to the next must reflect the probability density of the two conformations. For a stochastic process, solely the ratio of the two probability density states is important

$$\frac{\rho(\mathbf{Y})}{\rho(\mathbf{X})} = \frac{e^{-\beta H(\mathbf{Y})}}{e^{-\beta H(\mathbf{X})}}. \tag{3.3}$$

2. Importance Sampling:
   A phase space of a complex system is very large. In theory, each system state has to be evaluated to achieve the thermodynamic averaging of an observable X, as in eq. 3.2. However, this is computationally impossible. This is why, for MC simulations, the simulation is focused on those regions of the phase space which are important for the accuracy of $\langle X \rangle$. Consequently, the simulation concentrates on the low energy areas of the conformation space, which contributes highest to the thermodynamic averaging [62, 142]. In other words, Monte Carlo simulations are biased towards low energy regions.

Monte Carlo simulations generate Markov chains (of first order) with special properties. For these Markov chains the probability $W(\mathbf{Y}, \mathbf{X})$ for a system state $\mathbf{X}$ to go to state $\mathbf{Y}$ is independent of all the previous conformations that have occurred before $\mathbf{X}$ [63].
As mentioned above the transition probability $W(\mathbf{Y}, \mathbf{X})$ has to obey certain rules, which I will describe briefly

1. Norm:

$$\sum_{\mathbf{Y}} W(\mathbf{Y}, \mathbf{X}) = 1 \tag{3.4}$$

   In a thermodynamic system each state has a certain probability density. Therefore the sum of the probability transition to all possible states must result into the probability one.

2. Ergodicity:

$$W(\mathbf{Y}, \mathbf{X}) > 0 \tag{3.5}$$

In principle, it must be possible to reach each state in the configuration space through a Monte Carlo simulation.

3. Detailed Balance:
   The thermodynamic system, we describe, is in thermodynamic equilibrium. This property is described by the principle of detailed balance [142]. At equilibrium the transition between two states takes place in both directions at the same frequency [88]

$$W(\mathbf{Y},\mathbf{X})\rho(\mathbf{X}) = W(\mathbf{X},\mathbf{Y})\rho(\mathbf{Y}). \tag{3.6}$$

3.1.2 Algorithm of Monte Carlo simulations

Metropolis et al. [110] introduced a simple construction procedure to calculate thermodynamic properties which obeys all the required rules. The transition probability $W(\mathbf{Y},\mathbf{X})$ is separated into two parts [63]: A selection probability $T(\mathbf{Y},\mathbf{X})$ which proposes a move to go from state $\mathbf{X}$ to state $\mathbf{Y}$ and an acceptance probability $AC(\mathbf{Y},\mathbf{X})$ expressing the probability of this move being accepted.
Now, we assume that we want to go from state $\mathbf{X}$ with energy $E(\mathbf{X})$ to state $\mathbf{Y}$ with energy $E(\mathbf{Y})$. If $E(\mathbf{Y}) < E(\mathbf{X})$, then the new conformation is accepted with probability one to become the starting conformation for the next MC step. On the other hand, if $E(\mathbf{X}) < E(\mathbf{Y})$, then the acceptance probability for this MC step is $AC(\mathbf{Y},\mathbf{X}) = \frac{e^{-\beta E(\mathbf{Y})}}{e^{-\beta E(\mathbf{X})}}$, according to eq. 3.3. These two cases can be combined to

$$AC(\mathbf{Y},\mathbf{X}) = min\{1, e^{-\beta(E(\mathbf{Y})-E(\mathbf{X}))}\}, \tag{3.7}$$

which is also called the Metropolis criterium.
Using an MC simulation to calculate the thermodynamic observable, the observable is then calculated by averaging over all visited states

$$\langle X\rangle = \frac{\sum_{i=0}^{N} X_i}{N}. \tag{3.8}$$

3.2 Related optimization techniques

MC simulations are a powerful tool to sample system conformations at low energies. However for very complicated potential energy surfaces with many hills and valleys, a MC simulation is likely to get trapped at some local minimum and computes incorrect thermodynamic properties.
Since in our approach we are solely interested in finding the global optimum of a complex energy landscape, a standard MC simulation is less suited for our purposes.
In this section, I will introduce two other strategies which are better suited to locate the global energy minimum. Both methods are implemented in our program *FlexScreen*.

3.2.1 Simulated Annealing (SA)

The SA simulation technique [81] is a very useful algorithm for global optimization problems. Annealing describes a process in which the temperature of a molten system is slowly reduced starting from a very high temperature. The system is heated to allow the system to escape a local energy minimum and by gradual cooling the system can reach its groundstate [96].
SA can be integrated into the MC simulations, for which a high temperature[1] is slowly reduced.
The challenge lies in finding an optimal cooling strategy. The temperature must be reduced so slowly that the system is equilibrated at each temperature (adiabatic process). If the temperature is cooled down to a temperature close to T=0 ($\beta \to \infty$), the acceptance criteria allows only those transitions to the next system state, which are lower in energy than the previous state (moves into the direction of the minimum)

$$\lim_{\beta\to\infty} AC(\mathbf{Y},\mathbf{X}) = \lim_{\beta\to\infty} e^{-\beta(E(\mathbf{Y})-E(\mathbf{X}))} \to 0, \qquad \text{if} \quad E(\mathbf{Y}) - E(\mathbf{X}) > 0. \tag{3.9}$$

Optimal cooling strategies can be derived [54] and estimated [66]. However, these cooling schedules take too much time for docking purposes. Therefore, we decided to use a geometric cooling strategy: For each accepted SA step n we update the temperature β_n with

$$\beta_n = \beta_0 \gamma^n. \tag{3.10}$$

β_0 is the starting temperature and parameter γ is adapted to the desired length of the Markov chain accordingly. This process is not adiabatic and does not guarantee finding the global optimum. Because of this, we use the annealing procedure only for refinements of promising protein-ligand-conformations.

3.2.2 Stochastic Tunneling (STUN)

The Stochastic Tunneling (STUN), a generalization of the optimization method SA, was developed by Wenzel et al. [177] and is well suited for optimizations of protein-ligand complexes [144]. In SA, entrapments can be only avoided if a suitable cooling strategy is employed. In STUN, the dynamical process corresponding to the formation of the protein-ligand complex explores not the original, but a transformed potential energy surface (PES), which dynamically adapts and simplifies during the simulation.
For the docking simulations we use the following transformation

$$E_{STUN} = 0.3 \ln\left(x + \sqrt{x^2+1}\right), \tag{3.11}$$

[1] In MC simulations the temperature is related to the probability density states of the configurational system space. The configurational energies of the systems are described as potentials which are parametrized at a specific temperature (in our study the body temperature). Therefore, if we speak of a high temperature, we still assume that these properties of the system are valid and are not altered and that in changing the temperature we solely change the acceptance criteria of MC simulations.

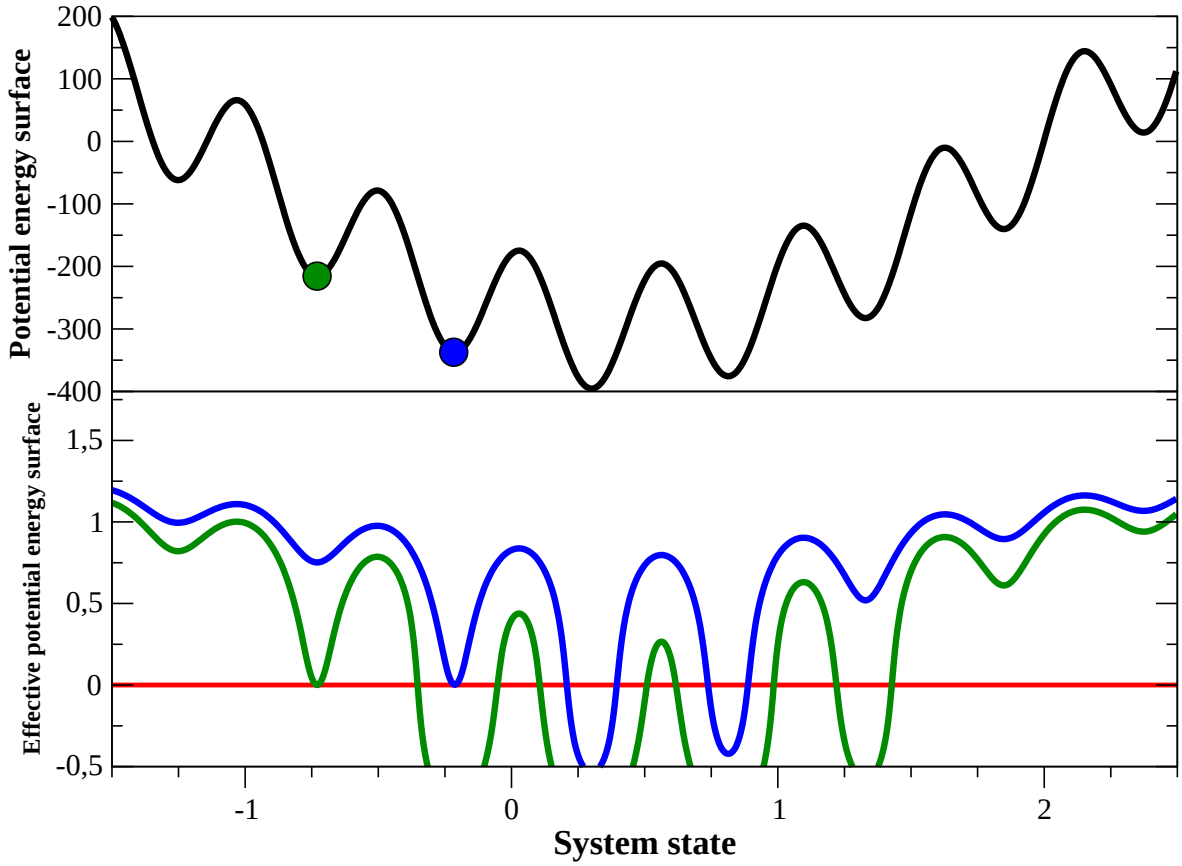

Figure 3.1: 1) The panel above illustrates a fictitious potential energy surface. This PES is transformed with STUN to an effective PES. 2) At the lower panel two effective PES are shown. The green curve shows the transformed PES under the assumption that the best energy found so far is at the green dot in the PES above and the blue curve represents the transformed PES in relation to the so far best found energy at the blue dot in the untransformed PES above.

where $x = \gamma(E - E_0)$. E is the energy of the present conformation and E_0 the best energy found so far. The parameter γ controls the steepness of the transformation.
The general idea of this approach is to flatten the potential energy surface in all regions that lie significantly above the best estimate for the minimal energy E_0. Even at low temperatures the dynamics of the system become diffusive at energies $E \gg E_0$, independent of the relative energy differences of the high-energy conformations involved (see figure 3.1). The transition of the conformations on the untransformed PES then appear to 'tunnel' through energy barriers of arbitrary height while low metastable conformations are still well resolved.
Applied to protein-ligand docking this mechanism ensures that the ligand can reorient itself through sterically forbidden regions while remaining inside the protein pocket.
In relation to SA, STUN can be understood as a scheme which adapts the temperature to the energy of the PES and the present best estimate of the energy.

By a Taylor series of the metropolis criteria (eq. 3.7) and under the assumption that the energy differences of the present to the next conformation is small
($\gamma\Delta = \gamma\left(E_{STUN}(\mathbf{X}) - E_{STUN}(\mathbf{Y})\right) \ll 1$), the STUN MC criteria can be rewritten as [177]

$$AC_{STUN}(\mathbf{Y}, \mathbf{X}) = min\{1, e^{-\beta\gamma e^{\gamma(E_0 - E(\mathbf{X}))}(E(\mathbf{Y}) - E(\mathbf{X}))}\}. \tag{3.12}$$

Eq. 3.12 shows the relation to SA: The usual MC temperature is transformed to an effective temperature $\beta_{STUN} = \beta\gamma e^{\gamma(E_0 - E(\mathbf{X}))}$ for the present best estimate of the energy.

Chapter 4

Biomolecular Force field

With recent quantum mechanical methods [51] it is possible to calculate the binding energy ω_{min} (see eq. 2.21) with high accuracy even for macromolecular systems. However although possible, it is still computationally very extensive and thus not applicable for tasks such as high-throughput protein-ligand docking or for calculations of thermodynamic properties of a system.
Molecular mechanic simulations are an alternative approach which is by far less time consuming. In molecular mechanics classical forces are used to describe molecular geometries and energies. In the underlying scheme of the Born-Oppenheimer approximation the movements of nuclei and electrons are decoupled and a molecular object is considered as consisting of points, at the position of the nuclei, with the properties of mass and charge. Chemical bonds are described by strings with specific string constants for different types of chemical bonds. These approximations can be used to perform simulations of a molecular system [142]. Unlike in quantum mechanics, the potential functions are parametrized empirically.
A general application of molecular mechanics (MM) depends on the validity of two underlying assumptions.

1. The total molecular energy of a system is separable into different energy contributions.

2. The different energy terms with their distinct parameters are transferable from the limited data they were parameterized for to all molecular systems under consideration.

4.1 Molecular Mechanic Interactions

In this section, I briefly explain the possible energy contributions of a general molecular mechanic (MM) force field. A molecular system of N atoms has 3N-6 degrees of freedom [105], assuming $N \geq 3$. In MM the total energy of a system E_{tot} is expressed as the sum of bonded E_{bond} and non-bonded interaction energies E_{nbond} [30]:

$$E_{tot} = E_{bond} + E_{nbond} \tag{4.1}$$

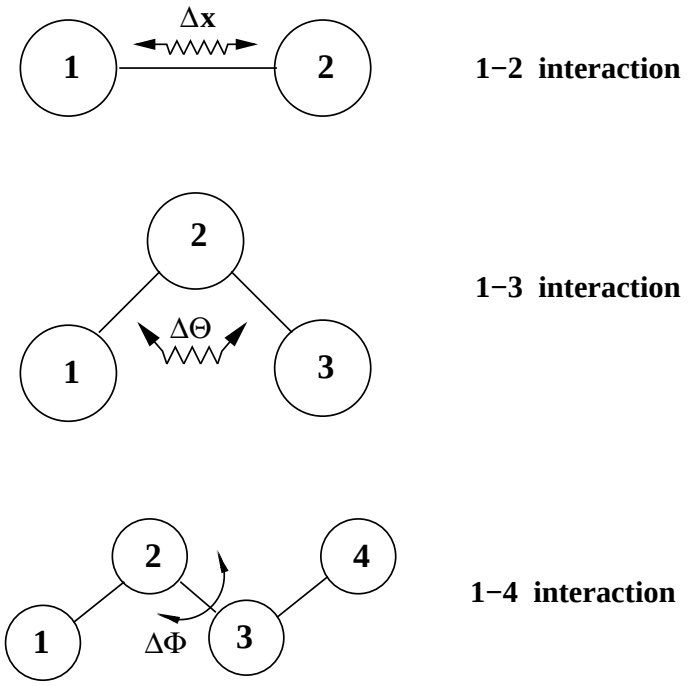

Figure 4.1: Illustration of different types of bonded interactions. 1-2 and 1-3 interactions correspond to a stretching or bending of a bond or a distance. 1-4 interactions correspond to interactions which involve a change of a dihedral angle.

4.1.1 Interaction of chemically bonded atoms

Interactions between chemically bonded atoms are described and named according to the indexed atoms which may alter their position relative to the other indexed atoms: We distinguish 1-2, 1-3 or 1-4 interactions. All other interactions shall be defined as interactions of non-bonded atoms, even though they might be part of the same molecule.
Figure 4.1 displays all the possible interactions of bonded atoms.

The 1-2 interaction or the bond stretching energy

The 1-2 (bond stretch) interactions are vibrations of the chemical bond. Force fields which do not allow chemical bonds to break often use just a simple harmonic potential to approximate the energy, which is necessary to alter the ideal bond length

$$E_{1-2} = \sum_{i,j \varepsilon S_{1-2}} D_{ij}(r_{ij} - \bar{r}_{ij})^2. \tag{4.2}$$

Here, r_{ij} is the distance of these two atoms (position of the nuclei) and $\bar{r}_{ij}$ and D_{ij} are the equilibrium distance and the 'spring constant' of the atom pair. Both parameters r_{ij} and D_{ij} have to be determined empirically or by quantum mechanical calculations.

The 1-3 interaction or the angle bending energy

The molecular bond angle depends on the overlapping molecular orbitals of bonded atoms which is defined by an ideal angle $\bar{\theta}$. $\bar{\theta}$ can be calculated with the Hartree-Fock method [47] or the related molecular orbital theory and is determined by the type of chemical bonds between the atoms and also the orientation of the valence electrons.
1-3 interactions evaluate the difference in the conformational energy if angle θ differs from $\bar{\theta}$

$$E_{1-3} = \sum_{i,j\varepsilon S_{1-3}} A_{ij}(\theta_{ij} - \bar{\theta}_{ij})^2. \tag{4.3}$$

The harmonic approximation is again the most simple assumption. If the bending energy is approximated more accurately (like the MMFF force field [61]), terms of higher order are included.

The 1-4 interaction or the torsional energy

The 1-4 interaction describes the energy effort to rotate atom 4 around the bond of atom 1 and 3. In molecular mechanics force fields, such rotations are often described in the following form

$$E_{1-4} = \sum_{i,j,k,l\varepsilon S_{1-4}} \sum_{n} \frac{R_{ijkl,n}}{2} \left(1 \pm \cos[n\Phi_{ijkl}]\right). \tag{4.4}$$

Eq. 4.4 above is very general. The potential is expressed as a Fourier series of angle Φ. In most force fields $n\ \varepsilon\ 1, 2, 3$ which means that there are no more than 3 energy minima during a full 360° rotation. $R_{ijkl,n}$ is the associated barrier height which depends on the connected chemical elements and n.
The origin of the barriers in internal rotations is often traced back to the interactions of the Pauli-repulsion and Coulomb-interaction of the nearest neighbors. The AMBER force field [130], for example, mixes the 1-4 bonded and non-bonded interactions by applying both a torsional term and a scaled down non-bonded Coulomb term.
In recent investigations [176, 131] the energy barriers are referred to as a form of hyperconjugative 'resonance stabilization' for the electrons of σ-type single bonds. These bonding electrons can delocalize and lead to a different binding pattern: a resonance structure between double and single bonds. Under this assumption, the minima in the torsional energy conformations have their origin not due to release steric repulsion, but to achieve optimal resonance stabilization.

Cross-Terms

The different interactions may also be interdependent. For example, due to a bending of a chemical bond, the bond length can change. This effect can be considered to be corrections to the harmonic approximations of the different potentials.

4.1.2 Experimental parametrization

As mentioned before, unlike in QM, in molecular mechanics the interaction potentials need to be properly parametrized. The so-called bonded interactions are usually fitted to vibrational spectra of the atoms. The spectra can be measured by infrared spectroscopy or Raman scattering [142].
Since not every atom in every molecule can be parametrized individually, atom types are introduced. Different chemical elements may possess several atom types, differing in which chemical bond the atom is participating in. The classification into different atom types depends on the hybridization, bond strength, bond angle of the chemical element. For simplicity, I will describe the mapping of atom i to its atom type by the general function $g_{a-type}(i)$.

4.1.3 Interactions of non-bonded atoms

By definition all interactions which are not part of the bonded-interactions are considered as non-bonded, even if these atoms are part of the same molecule. Interactions of non-bonded atoms are usually understood as point-like interactions.

Lennard-Jones potential

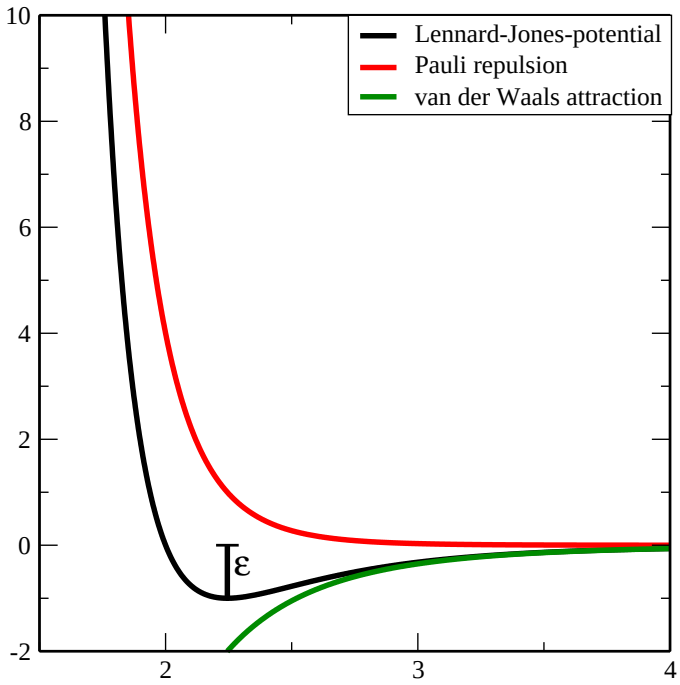

Figure 4.2: Lennard-Jones potential and its two constituents: Pauli repulsion and van der Waals attraction. The well depth ϵ is indicated additionally.

$$E_{LJ-6-12} = \sum_{i,j\varepsilon S_{LJ-6-12}} 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{4.5}$$

ε is the well depth of the potential and σ_{ij} determines the equilibrium position between the two atoms. The shape of the Lennard-Jones potential, illustrated in figure 4.2, arises from a balance of attractive and repulsive forces between two non-bonded atoms. The attraction forces are of a long-range, whereas the repulsive forces are of a shorter range.
In the 1930s London [99] showed that the attractive forces are due to fluctuations in the charge distribution of the atoms. The charge fluctuation of an atom can result in temporary dipoles which cause dipole-dipole or dipole-induced-dipole interactions. These interaction lead to the so called 'London forces' or van der Waals attractions illustrated in figure 4.2. London showed that these interaction energies are proportional to r^{-6}.
On the other hand, the short-ranged repulsive forces arise from the Pauli repulsion. The Pauli-exclusion principle states that no two fermions are allowed to occupy the same quantum state at the same position. These interaction energies are approximated with a term proportional to r^{-12} [96].
Eq. 4.5, the Lennard-Jones potential, describes these two interactions in a combined form. The Lennard-Jones potential is often used due to the mathematical simplicity, even if it is only partly correct. The repulsive part is too large in comparison to experimental results.
In order to increase accuracy some force fields use the Buckingham-potential [96] with a replaced repulsive term

$$E_{Buckingham} = \sum_{i,j\varepsilon S_{LJ-6-12}} \left[\frac{-A_{ij}}{r_{ij}^6} + B_{ij} e^{-B'_{ij} r_{ij}} \right] \tag{4.6}$$

Although this potential is more accurate than the Lennard-Jones potential, it has the disadvantage of becoming attractive for very small values of r_{ij}; but this problem can be overcome by correction terms.
The Lennard-Jones potential has its energetic minimum at $r_{ij} = 2^{\frac{1}{6}}\sigma_{ij}$. The values of ε_{ij} and σ_{ij} are atom or atom-type specific. In many force fields these parameters are defined for interactions of the same atom-type ($\varepsilon_{ij} = \varepsilon_{g_{a-type}(i)g_{a-type}(j)} = \varepsilon^{aa}$ and $\sigma_{ij} = \sigma^{aa}$ 4.1.2).
If interactions of different atom types are considered, these parameters can be calculated by

$$\varepsilon^{ab} = \sqrt{\varepsilon^{aa}\varepsilon^{aa}} \tag{4.7}$$

$$\sigma^{ab} = \sqrt{\sigma^{aa}\sigma^{bb}}. \tag{4.8}$$

Coulomb interaction

A molecule is constructed of different elements, which are covalently bonded. Because different elements attract their electrons differently strong, electronegative elements attract electrons more than elements, which are less electronegative. This results in an unequal charge distribution in the molecule, which can be represented in different ways [96]. One common way is,

to arrange the charge distribution as fractional point charges, positioned at the nuclei of the atoms. These point charges are arranged in such a way, that they reproduce to some limit the electrostatic properties of the molecule.
Many different methods are proposed to calculate the point charges. Because we want to process hundreds of thousands of ligands, we often use the point charges calculated with the force field of ESFF [149] and of Gasteiger-Marsili [53]. These methods do not employ quantum mechanical calculation for each molecule, but have different atom type parameters like the electronegativity and determine the final charge distribution iteratively.
The interaction V_{ij} between two charges q_i and q_j, which are separated by a distance of r_{ij} in a medium with the dielectric constant ϵ_r, can be calculated according to Coulomb's law

$$E_{ij}^{Coulomb} = k\frac{q_i q_j}{\epsilon_r r_{ij}} \tag{4.9}$$

The parameter k is the electrostatic constant and ϵ_r is the relative dielectric constant, describing the effect of the neighboring medium. In force fields which include explicitly water molecules, the parameter ϵ_r is constant for the whole system and depends only on the specific implementation of all the other MM potentials.
Other force fields treat the water molecules implicitly, which has the great advantage that by far less particles have to be simulated and approximate results can be achieved much quicker. But these approaches need also a more complicated description of the system (see section 4.2).
The relative dielectric constant ϵ_r was traditionally introduced to describe the polarizability of bulk materials. If water molecules are locally entrapped in the protein interior, these water molecules are difficult to treat implicitly by solely using a higher dielectric constant. These water molecules can strongly influence the dielectric constant in the protein interior.
All these influences are the reason that ϵ_r is not a constant of the system, but is rather dependent on the position and on the conformation of the solute.
Nevertheless, for densely packed proteins a relative dielectric constant of about $\epsilon_r = 1 - 4$ is often used. The bulk solvent has a dielectric constant approximately of $\epsilon_r = 80$, which is in agreement with experimental results [118].

Hydrogen bonding

A hydrogen bond is a special type of a dipole-dipole interaction. This interaction is of an intermediate range between an electron deficient hydrogen and an atom with high electronic density [90].
Such interactions play an important role for stabilizing the secondary structure, β-sheet or α-helix, of a protein and between protein and ligands. For example, in rational ligand design one approach is to modify the ligand structures to optimize possible hydrogen bond interactions. This results often in an increased affinity of the ligand to the protein [179].
Hydrogen bonds can be defined by energetic, but also by structural criteria [90]: A bond between a hydrogen (H) and an more electronegative element (X) results in a bond distance

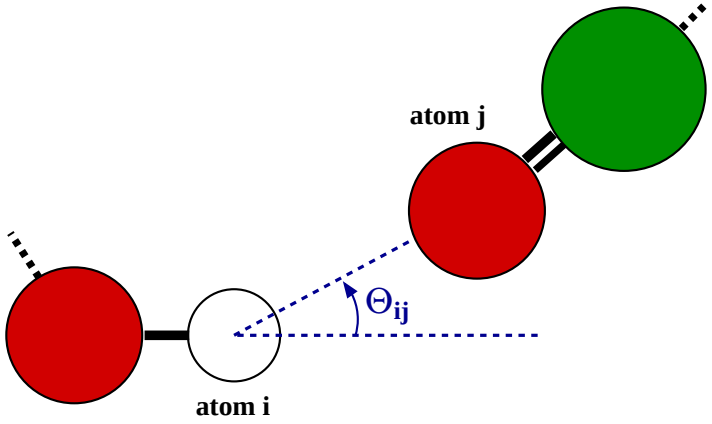

Figure 4.3: Illustration of the hydrogen bond interaction. Our scoring function uses a hydrogen bond potential, which depends on the direction of the hydrogen i to the acceptor atom j, measured by the angle θ_{ij}.

which is significantly smaller than the sum of their van der Waals radii.
Because hydrogen bonding is still difficult to model with a classical potential, it can only be quantified in a very approximate manner. In most of the MM programs [130], it is often described only by a dipole-dipole interaction, of the involved atoms.
We have therefore introduced an additional direction dependent potential which depending on the direction replaces the Lennard-Jones potential of eq. 4.5.

$$E_HB = \cos\Theta_{ij}\left(\frac{\tilde{R}_{ij}}{r_{ij}^{12}} - \frac{\tilde{A}_{ij}}{r_{ij}^{10}}\right) + \sin\Theta_{ij}\;\; 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{4.10}$$

If $\Theta_{ij} = 0$ the hydrogen potential E_HB replaces completely the LJ-potential. On the other hand if $\Theta_{ij} \geqq 90°$, then only the usual LJ-potential is used.
A similar potential is also used by the program AutoDock [114].
As mentioned before, it is very difficult to energetically evaluate different hydrogen bond geometries by classical potentials. We notice that other factors (for example the angle that evaluates how good a bond is in one plane or the dipole-dipole angle) are important. Presently quantum mechanical calculations are necessary to do such an evaluation accurately [86]. It is too difficult to treat hydrogen bonding between different chemical partners with only very simple rules for a detailed comparison.

4.2 Treatment of the solvent

As discussed before, it is difficult to approximate the solvent contribution implicitly. In docking simulations we want to approximate the potential of mean force (PMF) $\omega(\mathbf{r}, \mathbf{\Omega})$ of eq. 2.13 between protein and ligand. This potential is partly calculated by an Boltzmann weighted average over all possible solvent distributions at temperature T.
Implicit solvent models decompose this potential into a solvent independent potential $\omega(\mathbf{r}, \mathbf{\Omega})_{\text{ipt. v}}$[1] and solvent dependent PMFs $\langle U_{vv} \rangle$ and $\langle U_{uv} \rangle$:

$$\omega(\mathbf{r}, \mathbf{\Omega}) = \omega(\mathbf{r}, \mathbf{\Omega})_{\text{ipt. } v} + \langle U_{vv}(\mathbf{r}, \mathbf{\Omega}) \rangle + \langle U_{uv}(\mathbf{r}, \mathbf{\Omega}) \rangle = \omega(\mathbf{r}, \mathbf{\Omega})_{\text{ipt. } v} + \Delta B(\mathbf{r}, \mathbf{\Omega}), \tag{4.11}$$

with the indices u for the solute and v for the solvent. As for the method of thermodynamic integration [104], the reversible thermodynamic work to insert the solute into the solvent can be calculated by

$$\Delta B(\mathbf{r}, \mathbf{\Omega}) = \int_{\lambda=0}^{\lambda=1} d\lambda \left\langle \frac{\partial \left(U_{vv}(\mathbf{r}, \mathbf{\Omega}) + U_{uv}(\mathbf{r}, \mathbf{\Omega}, \lambda) \right)}{\partial \lambda} \right\rangle_{\mathbf{\Omega}, \mathbf{r} = const}, \tag{4.12}$$

with λ being the coupling parameter that describes with $\lambda = 1$ a full solute solvent interaction and with $\lambda = 0$ no such interaction at all. U_{vv} is the potential which describes the solvent-solvent interaction. In our consideration, this potential is independent of the solute and thus is the same for different solutes. Because we are only interested in binding affinities, we neglect this contribution.
$\Delta B(\mathbf{r}, \mathbf{\Omega})$ can now be further decomposed into non-polar ($\Delta B^{np}(\mathbf{r}, \mathbf{\Omega})$) and electrostatic contributions ($\Delta B^{elec}(\mathbf{r}, \mathbf{\Omega})$)[138]. Then each different contribution can be calculated in the manner of the thermodynamic integration 4.12.

4.2.1 Non-polar contribution

Analytical derivations with the scaled particle theory [135] and computer simulations with explicit solvent molecules show that the contribution of $B^{np}(\mathbf{r}, \mathbf{\Omega})$) is approximately proportional to the solvent accessible surface area [97] (SASA) of a solute which exceeds a specific length scale [7]. Instead of simulating a whole system with explicit water molecules, it is possible to quantify the non-polar contribution $B^{np}(\mathbf{r}, \mathbf{\Omega})$ of a solute by calculating the SASA[2].

[1] This potential is not the same as calculating the potential in the gas phase. Other enthalpy contributions can well be dependent on the averaged solvent; the environment of the present solvent influences other interaction energies.

[2] Actually Su et al. [158] showed by studying dimerizations of alanine dipeptide that the non-polar contribution can be better approximated by the sum of a non-polar cavity creation and the solute-solvent van der Waals interaction, which are parametrized separately. The solute-solvent dispersion energies have a medium range and it seems that they can not be neglected for an accurate calculation of $B^{np}(\mathbf{r}, \mathbf{\Omega})$). We argue that for protein-ligand docking the van der Waals interaction contribution can be neglected. Because ligands are rather small molecules, the energy inaccuracy by incorporating the dispersion energies into the SASA term should also be little.

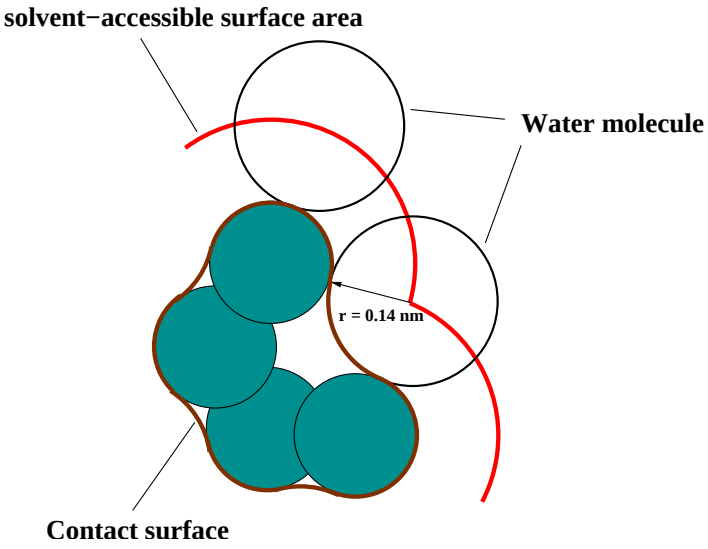

Figure 4.4: Illustration of the solvent-accessible-surface area (SASA). SASA is defined as the trace of the center of a water molecule rolling over the contact surface of the molecule.

As shown in figure 4.2.1, the solvent accessible surface area is defined as the trace of the center of a water molecule rolling over the contact surface of the molecule. The water molecules is assumed to be a sphere with a radius of 1.4 Å. Under this assumption the non-polar thermodynamic work to insert a solute into the solvent is expressed as

$$\Delta B^{np}(\mathbf{r}, \mathbf{\Omega}) = \gamma \mathrm{SASA}_{total}. \tag{4.13}$$

γ is a surface tension term with the dimension of $\frac{\text{energy}}{\text{area}}$.
The entropic effect of mixing different particles can be neglected (Flory-Huggins-theory [46]), for very dilute solutions.

4.2.2 Electrostatic contribution

The second term $\Delta B^{elec}(\mathbf{r}, \mathbf{\Omega})$ can be similarly calculated by assuming a linear response of the solvent to the charge distribution of the solute. By thermodynamic integration this results in

$$\Delta B^{elec}(\mathbf{r}, \mathbf{\Omega}) = \frac{1}{2} \left\langle U_{uv}^{elec}(\mathbf{r}, \mathbf{\Omega}) \right\rangle. \tag{4.14}$$

Continuum electrostatics offer an approach to calculate these interaction energies, without an explicit representation of the solvent particles. The averaged solvent is replaced by a featureless dielectric medium with a constant relative dielectric scalar ϵ_v and a specific ionic strength. Additionally, a constant relative dielectric scalar ϵ_u is assumed for the volume of the solute. Such approaches have been successfully applied for long and originate from the well known scientists Born [17], Kirkwood [82, 159] and Onsager [124].

It can be shown that the influence of ions in the solution to solvation energies and intramolecular electrostatic interaction, also between the ligand bound to a protein, is small and can be neglected [59]. This is why for all further (and also the previous) considerations we do not consider ions in the solution.
For the purpose of calculating the electrostatic contribution $\Delta B^{elec}(\mathbf{r}, \mathbf{\Omega})$, the Poisson equation of the system has to be solved [71]

$$\nabla \left[\epsilon(\mathbf{r})\phi(\mathbf{r})\right] = -4\pi\rho(\mathbf{r}) \tag{4.15}$$

and the electrostatic energy is computed using this potential [140]:

$$E_{elec} = \frac{1}{2}\int_{\Re^3} \rho(\mathbf{r})\phi(\mathbf{r})d\mathbf{r} = \frac{1}{8\pi}\int_{\Re^3} \frac{\vec{D}^2}{\epsilon(\mathbf{r})}d\mathbf{r} = \sum_i \frac{1}{8\pi}\int_{\Re^3} \frac{\vec{D}_i^2}{\epsilon(\mathbf{r})}d\mathbf{r} + \sum_{i<j} \frac{1}{4\pi}\int_{\Re^3} \frac{\vec{D}_i\vec{D}_j}{\epsilon(\mathbf{r})}d\mathbf{r}. \tag{4.16}$$

Here, we label the charge density $\rho(\mathbf{r})$, electrostatic potential $\phi(\mathbf{r})$ and the displacement field of the different charges i $\vec{D}_i$. Thus, $\Delta B^{elec}(\mathbf{r}, \mathbf{\Omega})$ can be considered as the correction to the case of calculating the electrostatic energies solely with a dielectric constant of ϵ_u.
Several numerical solution for solutes of arbitrary shape and charge distribution are available. Programs like APBS [8] or Delphi [121] use finite element techniques to discretize the system and solve the Poisson-Boltzmann equations iteratively. They are very accurate, but still too slow to become a general solution for high-throughput screening solutions.
Proteins are often, or can often be approximated to be, of globular shape, especially if the ligand is bound to the protein. Under this assumption eq. 4.16 can be calculated analytically [82]. We implemented and tested the electrostatic calculations based on this approach (details see appendix A). To any protein complex we selected the sphere that approximates best the electrostatic interactions of the protein cavity[3] and calculated the electrostatic contributions using this virtual sphere . As one can imagine this approach has problems for ligands, which try to find a position outside of the cavity. We did not find a satisfying solution for such cases.
If we neglect the solvent contribution for the electrostatic interactions between the bound ligand and the protein, the resulting error is for buried ligands small enough to be neglected; especially, if the resulting error is compared with inaccuracies that originate from the calculation of the partial charges by a standard force field. This is why at the moment we use a standard Coulomb interaction (as in eq. 4.1.3) with a relative high dielectric constant.
In case of protein pockets with no or only very little contact to the bulk solvent, the desolvation energy of the protein is similar for all ligands that fill the cavity. For such cases it is more important to calculate the solvation energy of the ligands than for the protein.

[3]For the optimization, we calculated for the charge distribution close to the cavity center, of a docked native ligand and several surrounding sidechains, the solvation energies with the APBS program [8]. Then, by changing the center and radius of a sphere we searched for the sphere which reproduces these results with the least square deviation.

For the ligands the analytical solution [82] that assumes a spherical or ellipsoidal shape of the ligand is not applicable. By rewriting eq. 4.16, I want to introduce briefly the Coulomb field approximation [140]

$$E_{elec} = \frac{1}{8\pi\epsilon_v}\int_{\Re^3}\vec{D}^2 d\mathbf{r} + \frac{1}{8\pi}\left(\frac{1}{\epsilon_u}-\frac{1}{\epsilon_v}\right)\int_{U^3}\vec{D}^2 d\mathbf{r}. \tag{4.17}$$

In eq. 4.17 the integral of the first addend is over the whole space ($\Re^3$), whereas the integral of the second addend is only over the volume of the solute (U^3). In the Coulomb field approximation it is assumed that the distortion field from the boundary $\vec{R}_{\text{reac}}(\mathbf{r})$, separating two regions of different dielectric constant, can be neglected: The effect of the reaction field is not taken into account.

$$\vec{D}_{i,v}(\mathbf{r}) = \frac{q_i}{\epsilon_v(\mathbf{r}-\mathbf{r_i})^2}\frac{(\mathbf{r}-\mathbf{r_i})}{|(\mathbf{r}-\mathbf{r_i})|} + \vec{R}_{\text{reac}}(\mathbf{r}) \tag{4.18}$$

This Coulomb field approximation can be applied to the first term of eq. 4.17 for the following reason, because it can be shown that, at the integration outside the solute volume $\Re^3 \setminus U^3$, $\vec{R}_{\text{reac}}(\mathbf{r})$ contributes only very close to the solute and also because at the integration over the solute volume, the distortion field is small compared to its influence in the second term, at which the integration is over the same volume (if $\epsilon_v \gg \epsilon_u$).
Assuming that the partial charges are equally distributed on spheres with the atom radii R_i and consequently avoiding the infinite energy self-contributions of the charges, E_{elec} can be written to leading orders as [140]:

$$E_{elec} = \sum_i \left\{ \frac{q_i^2}{2\epsilon_v R_i} + \frac{q_i^2}{8\pi}\left(\frac{1}{\epsilon_u}-\frac{1}{\epsilon_v}\right)\int\limits_{U^3\setminus V_i}\frac{1}{(\mathbf{r}-\mathbf{r}_i)^4}d\mathbf{r}\right\} \tag{4.19}$$

$$+\sum_{i>j}\left\{\frac{q_i q_j}{2\epsilon_v r_{ij}} + \frac{q_i q_j}{4\pi}\left(\frac{1}{\epsilon_u}-\frac{1}{\epsilon_v}\right)\int_{U^3}\vec{D}_i\vec{D}_j d\mathbf{r}\right\}, \tag{4.20}$$

with $U^3 \setminus V_i$ being the whole space in the solute excluding only the volume of atom i with radius R_i. Under this assumption the self-solvation energies per partial charge are slightly overestimated. But the advantage of the approximation is that it can be easily calculated by a numerical calculation over the solute volume.
The interaction part can not be simplified as easily. This is why we use an an empirical approximation to calculate that contribution, which is provided with the generalized Born approximation [157].

Using this approximation the interaction energy of eq. 4.20 can be rewritten [139] to

$$E_{int} = \sum_{i>j} \left\{ \frac{q_i q_j}{\epsilon_u r_{ij}} - \frac{q_i q_j}{R_{ij}^{GB}} \right\}, \tag{4.21}$$

$$\text{with} \quad R_{ij}^{GB} = \left[r_{ij}^2 + R_i^{\text{eff}} R_j^{\text{eff}} e^{\frac{-r_{ij}^2}{4R_i^{\text{eff}} R_j^{\text{eff}}}} \right]^{\frac{1}{2}}. \tag{4.22}$$

R_i^{eff}, R_j^{eff} are effective atom radii; these depend on the shape of the molecule and are defined by the generalized Born approximation [157]. As Scarsi et al. showed these effective radii can be calculated by the results of the numerical integration of the first term in eq. 4.20. Being aware of that the approximation works only well for smaller solutes, the total electrostatic solvation energy of a ligand can be expressed as [139]

$$\Delta B^{elec,\text{ligand}} = \sum_i \frac{q_i^2}{2} \left(\frac{1}{\epsilon_u} - \frac{1}{\epsilon_v} \right) \left\{ \frac{1}{4\pi} \int\limits_{U^3 \setminus V_i} \frac{1}{(\mathbf{r} - \mathbf{r}_i)^4} d\mathbf{r} - \frac{1}{R_i} \right\} - \sum_{i>j} \left(\frac{1}{\epsilon_u} - \frac{1}{\epsilon_v} \right) \frac{q_i q_j}{R_{ij}^{GB}}. \tag{4.23}$$

We implemented eq. 4.23 and we could very well reproduce the solvation energies of 400 randomly selected ligands from a larger molecule database. Our calculated solvation energies are compared with those which were calculated using the program APBS [8] (see figure 4.5). Solvation effects of the protein-complex are not considered in our studies, we present here. Nevertheless in the newest program version we treat the solvation effects of the protein-complex in the following way. We use a SASA-model that calculates the solvation free energies as a linear sum of atomic contributions [37, 138]

$$\Delta B(\mathbf{r}, \mathbf{\Omega}) = \sum_i \gamma_{g_{atype}(i)} A_{g_{atype}(i)}(\mathbf{r}, \mathbf{\Omega}), \tag{4.24}$$

with $g_{a-type}(i)$ 4.1.2 being the function that maps each atom to a defined atom-type. The SASA-model does not take charges buried in the protein-ligand complex into account. This is why this model underestimates the self-solvation energies of these charges. The model has the advantage that it is quicker than any of the other presently methods used and still has proven to be reasonably accurate [37]. It is widely used in biomolecular simulations [37, 146]. Especially for our purpose it has the advantage that no pre-calculation of the protein is necessary and ligands, which try to dock far outside of the cavity, are evaluated on a reasonable footing.

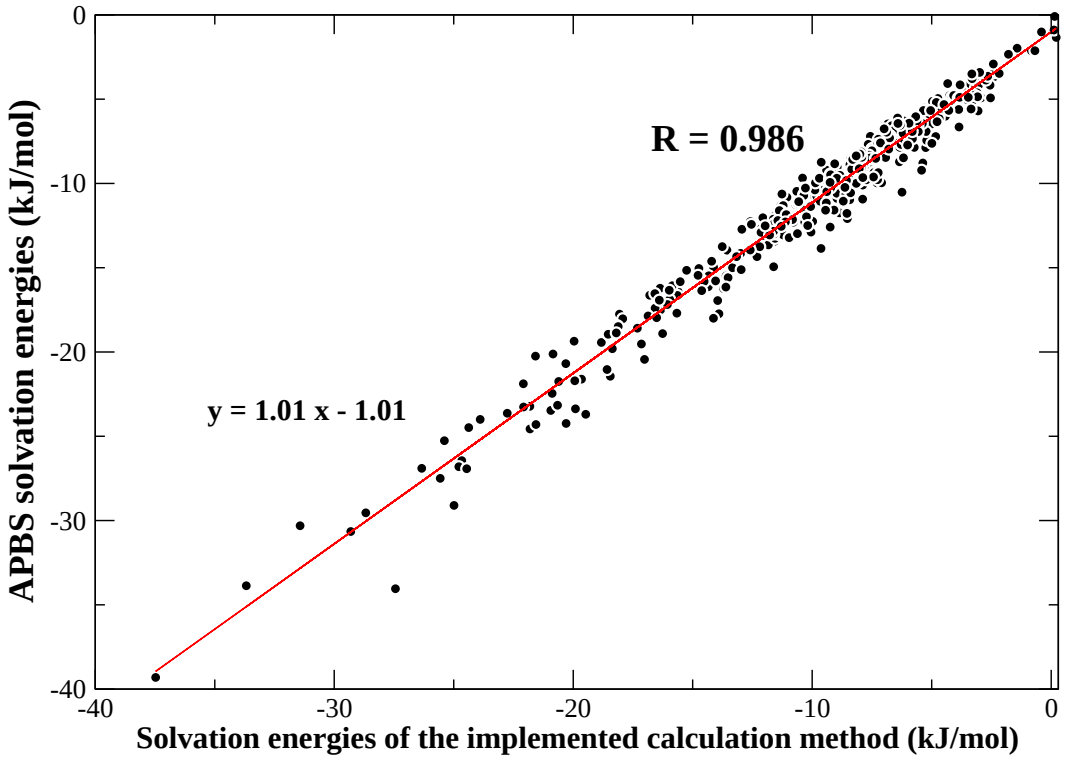

Figure 4.5: Comparison of numerical calculated solvation energies. Compared are the results for the solvation energies of 400 ligands of the implemented routine, as in eq. 4.23, with the very accurate numerical Poisson-Boltzmann solver APBS [8]. Due to optimizing the correlation the atomic radii are decreased by 0.02nm. The correlation of the solvation energies is with $R = 0.986$ very high. Our calculated values also do not have to be rescaled by a factor, since they match already very well the numerical values.

Chapter 5

Docking Strategy

In this chapter, I introduce and describe the most important steps of the simulation process (see figure 5.1). It is described how the ligand and the protein are represented in a simulation and how their changeable dihedral angles are identified. At the end, I want to introduce one of our new techniques which improves the simulation accuracy: Reasonable protein-ligand conformations are pre-generated before the usual docking simulation starts.

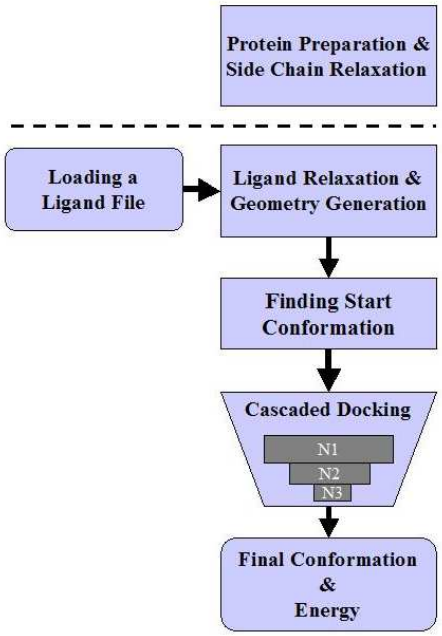

Figure 5.1: Abstract process flow of a complete docking simulation for one ligand using our approach

5.1 Process Description

As Figure 5.1 indicates, the whole simulation process can be divided into two parts: Into a ligand independent and a ligand dependent part. Already before any ligand structure is investigated, the protein is analyzed by an automated procedure. All the parts of the protein which do not alter during the simulation are pre-processed to ensure a quick evaluation of the binding energies. For this purpose, position dependent Coulomb potentials and position dependent atom neighbor lists are pre-calculated and stored on grids.
We allow some side chains, but not the backbone, to change their conformation. Side chain flexibility can be assigned manually or by an automated procedure (see section 5.2). Treating side chains flexible during the simulation can be interpreted as the first approximation of a completely flexible and thus as a realistic representation of the protein. The necessity for flexible-protein docking has been widely discussed and recognized [24]. By including side chain flexibility, many important protein deformations upon binding can be modeled.
Because the program optimizes the ligand as well as side chain conformations during a simulation, the final binding energy E_{Bind} depends on the energetic cost of the conformational deformation of the ligand $\Delta E_{L,\text{deform}}$ and the side chains $\Delta E_{P,\text{deform}}$:

$$E_{Bind} = E_{LP} - E_L - E_P = E_{LP,inter} + \Delta E_{L,\text{deform}} + \Delta E_{P,\text{deform}}. \tag{5.1}$$

As a consequence, the relaxed conformations of ligand and side chains have to be determined before the actual docking simulation can start. These conformations serve as an energetic reference with which all other conformations are compared and with which $\Delta E_{L,\text{deform}}$ and $\Delta E_{P,\text{deform}}$ are calculated.
After the protein preparation, one or several ligands can be selected for a docking simulation. The docking simulations are performed in the active site of the protein cavity. This position has to be specified before the simulation can start. As a result, only side chains close to the cavity and usually only the side chains which hinder ligands to bind or which improve the ligand side chain interaction by a change of the side chain conformation should be selected as flexible side chains.
The specified position of the protein cavity is also the starting point for our docking simulation. For this purpose, an acceptable protein-ligand conformation initial of the simulation has to be found. This initial protein-ligand conformation should be a 'realistic' conformation: i.e. neighboring atoms should not overlap with each other[1]. Section 5.4 discusses the techniques of finding initial protein-ligand starting conformation in more detail.
After the starting conformation has been found, the general docking simulation can start. With the STUN-MC simulation we aim to find the global optimum of the PES. Different docking strategies were compared but the cascaded docking procedure proved to be very effective as well as superior to a sequential procedure [107]. The cascaded docking procedure

[1] In our experience, it is very difficult to obtain meaningful docking results with a fixed number of computational steps if the docking simulation is started at conformations which are sterically forbidden. Due to Pauli repulsion these conformations have an enormous binding energy in our force field.

emerged as a compromise between the reduction of the statistical noise and the necessary computational effort. As indicated in figure 5.1, the process is divided into three simulation parts (stages). Starting with N_1 independent docking simulations, their number is reduced with each following stage: $N_1 > N_2 > N_3$. In the cascaded docking procedure only the energetically best simulations of the previous stage are continued. As the number of simulations decreases, the computational effort spent on each stage increases. As a result of the cascaded docking procedure, the statistical noise and also docking failures are reduced while limiting the computational effort to an affordable amount [107]. Usually, we choose the N_i independent simulation lengths for each stage in such a way that each stage uses approximately the same computational effort in total.
During one STUN-MC step, either the conformation of the ligand or of one side chain is altered. A conformational change of the ligand comprises, for example, a random translation and rotation around the center of mass and additionally a change of one dihedral angle, if the ligand is flexible.
At the end of stage 3, N_3 results are returned. If $N_3 > 1$, then the binding energies and the different ligand conformations can be compared with each other. Such a comparison helps to determine if the binding energies are reliable [44] and if the binding mode was unique.

5.2 Ligand representation

The ligands as well as the protein are represented classically (see chapter 4). In most protein-ligand complexes, the ligand and the protein do not form any covalent bonds. Because of this, an approach like ours that does not use quantum mechanical calculations can be applied successfully.

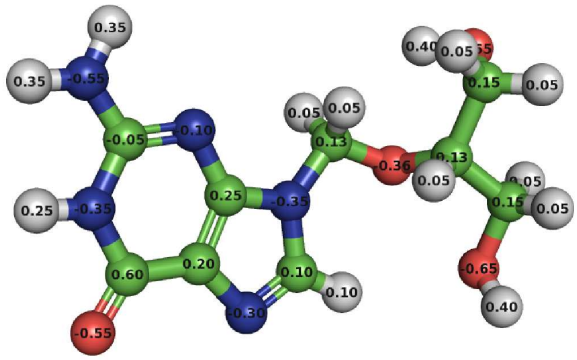

Figure 5.2: Illustration of our ligand representation. As shown in the picture, we distinguish different atom types (represented by different colors) and different covalent bonds (single or double bonds). The molecular charge density of the molecule is approximated by assigning to each nucleus a partial charge. The partial charges of the molecule are labeled in the picture as an example.

A full description of a ligand contains:

1. Partial charges for each atom
2. Van der Waals parameter for the atom type
3. Bonding topology
4. Bond type: σ, π or σ-π mixed bond

As we will see later (see the study in chapter 8), the better the description of the charge density for the ligand and the protein the more reliable are the calculated electrostatic interactions of the complex, and, as a consequence, also the docking accuracy.

5.2.1 Analysis of ligand flexibility

Unlike Molecular Dynamics (MD) simulations, our simulations do not allow bonds to be stretched and to be squeezed. This is unnecessary, because we are not interested in the dynamics of the system but rather in the binding energy and in the resulting receptor-ligand complex. Consequently, we fix the bond distances and bond angles (described as 1-3 interactions in section 4.1.1) and allow changes of dihedral angles for the ligand and for the side chain, (described as 1-4 interactions in section 4.1.1).
Through a statistical analysis of accurate small X-ray structures, Engh et al. [38] derived parameters to refine less accurate X-ray structures. With the accurate structures they created a parameter set for different atom types and showed that the measured bond lengths and bond angles highly correlate with those calculated by the parameter set. The standard deviations for the bond length and the bond angles are with 0.06 Å and 5 degree respectively very small [38]. Consequently, in our program the ligand and the protein structure have to be refined according to such parameters only once in the beginning. Afterwards, i.e. during the docking simulation, it is not necessary to alter bond lengths or bond angles in order to refine the structure. The deviations are so small that fixing the bond lengths and bond angles is justified.
All bonds of the ligand are automatically investigated for possible flexibility. If a bond is recognized as a σ-bond, then an energetic free rotation around that axis is allowed without any torsional potential; restrictions arise only indirectly due to non-bonded interactions with other ligand atoms (definition of non-bonded interactions see section 4.1.3). On the other hand, if two atoms are connected by a π-bond, we do not allow any rotation. QM calculations show that the participating electrons are de-localized between these bonded atoms and that in relation to the dihedral angle often two low energy minima exist, which are separated by high energies. The conformation corresponding to the two minima are referred to as the cis- and the trans-orientation [96]. If such a bond is treated classically, a detailed analysis of the rotational energies is necessary. Bonds which are partly π- and σ-bonds are even more complicated to describe with rotational potential energies.

Because of this, we presently do not allow rotations for these kinds of bonds.[2]
To identify π-bonds we use the following rule:
First, we identify all carbons and nitrogen atoms which have 3 bonded neighbors as sp2-hybrids. We consider all bonds which interconnect these sp2-hybrids as π-bonds and do not allow any rotation around these bonds. These rules also apply for the side chains. But in addition to the automatic identification of π-bonds, it is also possible to identify them manually.

5.3 Scoring function

Many different scoring functions have been proposed in recent years [87, 174] but no clear consensus has emerged to date on the superiority of force-field-based or knowledge-based approaches. In our investigations, we employ a force field based scoring function originating from physical principles:

$$S = \sum_{\mathrm{i}\,\epsilon\,\mathrm{Ligand}} \sum_{\mathrm{j}\,\epsilon\,\mathrm{Protein}} \left(E_{ij}^{Coul} + E_{ij}^{LJ} + E_{ij}^{HB} \right) + E_{Lig.}^{\mathrm{deform.}} + E_{Prot.}^{\mathrm{deform.}} \tag{5.2}$$

with :

$$E_{ij}^{Coul} = k \frac{q_i q_j}{\varepsilon_r \, r_{ij}} \tag{5.3}$$

$$E_{ij}^{LJ} = 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{5.4}$$

$$E_{ij}^{HB} = \cos\Theta_{ij} \left(\frac{\tilde{R}_{ij}}{r_{ij}^{12}} - \frac{\tilde{A}_{ij}}{r_{ij}^{10}} \right) + \sin\Theta_{ij} \; 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right], \tag{5.5}$$

derived from the AutoDock scheme [114]. The scoring function of eq. 5.2 considers interaction energies as well as the deformational energies of protein and ligand. The interaction energies between the protein and the ligand contains a term for the Coulomb interaction (see eq. 5.3 and also section 4.1.3), a Lennard-Jones potential which combines the empirical Pauli repulsion with the van der Waals attraction (see eq. 5.4 and also section 4.1.3) and a angular dependent term which considers the interactions due to hydrogen bonds (see eq. 5.5 and also section 4.1.3). As explained in section 5.1, the actual conformation of the protein and the ligand are compared with their respective relaxed conformation; i.e. the conformation of the protein or the ligand with the least internal energy. The internal energy itself is again calculated with the Coulomb interaction energies of eq. 5.3 and the Lennard-Jones potential

[2] Another approach is to statistically evaluate the angular population for different bond types and to use the frequency a specific dihedral angle appears to construct rotational potential energies. Programs like, for example, Gold [73] use rotamer libraries of allowed and forbidden angles. In our approach we have not yet included such libraries. We believe that, for example, the cis-trans conformational change should not be done during the docking simulation, but before a docking run. Two different ligand conformations should be generated (according to cis-trans) and used for separate docking simulations. Since a conformational change from cis to trans is a large conformational change, it is unlikely that such a conformational change is accepted by the STUN-MC criteria (see section 3.12).

of eq. 5.4. We use the Lennard-Jones parameter σ_{ij} of OPLSAA [75] as our established parameter set for most of our docking studies instead of the smaller AutoDock radii [114]; both parameter sets as well as the parameters $\tilde{R}_{ij}$ and $\tilde{A}_{ij}$ of the hydrogen bond term are listed in appendix C. We experimented with a number of other Lennard-Jones parameterizations and found that binding modes depend to a large extend on the size of the radii. Using smaller radii as, for example, in AutoDock increases the likelihood that the ligand will find a good binding mode in validation studies, but will also increase the number of false positives in screening applications.

We compute the partial charges (q_i) for both receptor and ligand with the ESFF force field [149] with the program InsightII [70] (at pH 7.4). This provides an adequate assignment of the charges for both the protein and a wide variety of ligands: ESFF succeeded to automatically assign consistent charges for over 180,000 ligands of the NCI Open database (210,000 ligands).

In previous studies, we found that a purely Coulomb-based representation of hydrogen bonding significantly decreases the accuracy of the docking results, and we incorporated the hydrogen bond parameters $\tilde{R}_{ij}$, $\tilde{A}_{ij}$ from AutoDock [114] as one established empirical model.

When specifically mentioned in the presented docking studies, we calculate the de-solvation energies for the ligands with a term for the non-polar de-solvation energy (see section 4.2.1) and for the electrostatic contribution to the de-solvation energy (see eq. 4.23) (atomic radii are reduced by 0.02nm). The non-polar de-solvation energy is calculated with

$$E_{\text{non-polar}} = -30 \frac{cal}{mol\ \text{Å}^2}\ \text{SASA}_{molecule}. \tag{5.6}$$

5.4 Methods for generating an initial protein-ligand conformation

In our approach several independent docking simulations are performed for each ligand. Each simulation starts from a different initial protein-ligand conformation. As discussed in section 5.1, it is advantageous to use such an initial conformation which corresponds to an energy below a defined limit. This energy is calculated by our scoring function. As a consequence for the energetic limit of the conformation, initial protein-ligand conformations with minor atom penetrations are allowed, whereas those with a larger overlap between atoms are forbidden. Several strategies to find these allowed protein-ligand conformations are implemented and used:

1. Using the relaxed ligand conformation and placing it by random translation and rotation in the cavity until an allowed protein-ligand conformation has been found.

2. Same as point 1, but with three different ligand conformations: Conformations resulting after a small stochastic optimization to find the conformation of lowest energetic energy (relaxed), of largest end to end distance (straightened), and of smallest radius of gyration (globular).

The downside of these methods is that they all assume a spacious cavity. If the cavity is narrow, then it is left to chance, if one of the ligand conformations (relaxed, straightened or globular) can be fitted into the cavity by mere translations and rotations. Figure 5.3 displays a two dimensional representation of such a difficult case. The ligand fits tightly into the protein cavity, which is represented by dashes. If the ligand conformation is very different from the one which is displayed, it is not possible to find an allowed conformation by mere translations and rotations of the whole molecule.

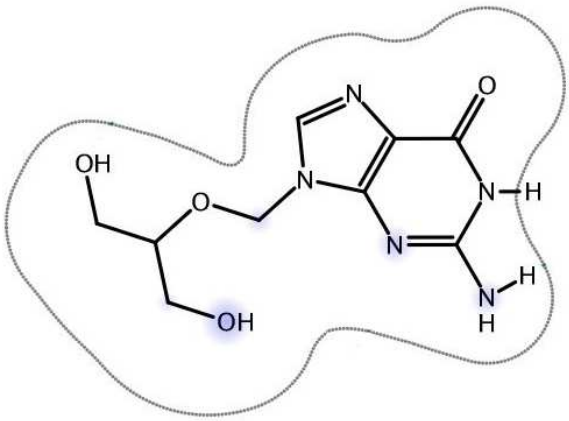

Figure 5.3: Ligand and cavity representation in 2 dimensions. The border line of the cavity is plotted by dashes. As illustrated, the ligand fits tightly into the cavity.

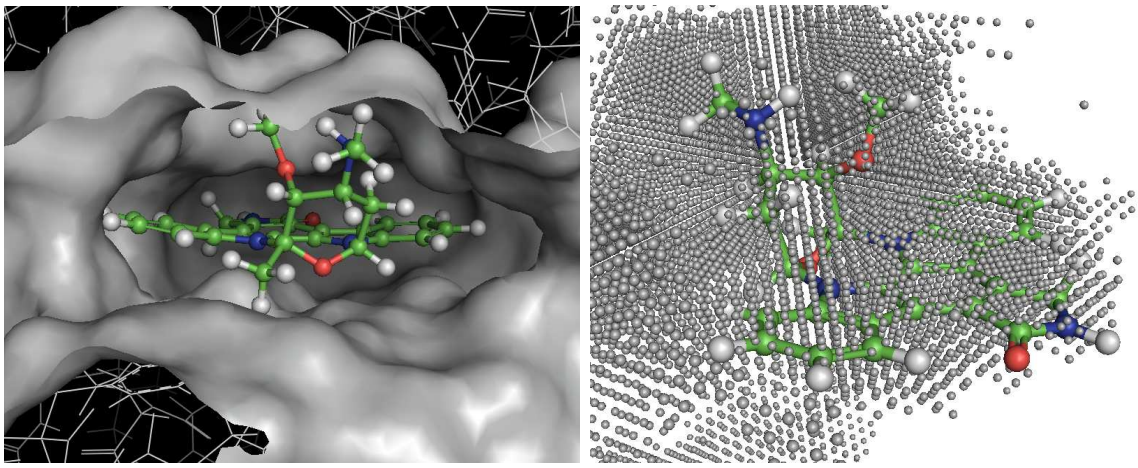

Figure 5.4: Receptor-ligand complex (pdb: 1STC); on the left picture the ligand is shown in the cavity of the receptor (grey); on the right picture the receptor structure is replaced by attractors, which constitute an attractor field (the view angle is changed by about 180°)

This is one of the reasons why we decided to implement a pre-docking method using very small MC simulations, even though MC simulations and optimizations are difficult and time consuming, when many local energy minima at the PES are between very high energies. To avoid these problems a different protein representation is used to find allowed protein-ligand conformations.
The purpose of the pre-optimization routine is to find ligand conformations in allowed regions. The ligand should have already adapted to the protein structure, before the usual docking routine starts. Since our aim is to dock many ligands in a short period of time, it is necessary that the pre-optimization routine is very quick.
Therefore, in a first step the empty space (discretized by a grid) of the protein is calculated. All grid points which are further away to any protein atom than the radius of the protein atom and of an additional atom which corresponds to a potential ligand atom are considered as points of the empty space. Since we are not interested in protein-ligand conformations that do not interact with each other, we restrict the empty space. If the grid points which correspond to the empty space are further away than 5 Å, then these points are neglected as being too far away from the protein.
The two different representations are shown in figure 5.4. The left picture displays a closeup of the cavity and the docked ligand staurosporine (PDB: 1STC). The empty space the ligand can accommodate in is replaced by grid points, attractors, which are illustrated in the right picture of figure 5.4.

After the empty space is calculated, in a second step, an attraction field is calculated that evaluates each atom of a ligand conformation in the cavity according to the distance to the next attractor, the grid point of the empty space. Thus, in the attraction field a protein-ligand conformation is calculated by

$$E_{\text{attraction}} = \sum_{i \ \epsilon \ \text{atoms}} B \ \min(\Delta r_{i,\text{attractor}}), \tag{5.7}$$

with the constant energy factor B and the distance to the closest attractor $\min(\Delta r_{i,\text{attractor}})$. Refining a conformation on this attractor-PES usually takes less than one CPU second. This enables us to refine and compare a high amount of different ligand conformations with each other.
Hydrogen bonding is important for the stabilization of a protein-ligand conformation. Possible hydrogen bond donors or attractors of the ligand should be saturated as much as possible. Under this assumption we added another criterion to identify among the allowed conformations further good or bad protein-ligand conformations. Each possible ligand hydrogen bond attractor or donor should be in the usual hydrogen bond distance (2.8 Å [86]) to the participating partner. Therefore, the least distance from a ligand donor or attractor to a protein atom partner is calculated and is energetically evaluated in a similar way as in eq. 5.7, if this least distance exceeds the 2.8 Å hydrogen bond distance (not the smallest distance is evaluated, but the difference in distance to the usual hydrogen bond distance). Both energy contributions are added together to constitute the attractor-H-PES.
All together a total Pre-docking simulation consists of:

1. Generation of at least 500 energetically allowed different ligand conformations (rotations around as many bonds as possible)

2. Search for the two best ligand orientations of each of the 500 conformations in the attractor-H-PES field (the ligand is rigidly rotated around a defined cavity center by step sizes of 30 degrees)

3. Conformation refinement of the energetically best N_1 conformations in the attractor-H-PES (For this purpose, we use the SA optimization technique, to ensure a quick optimization on the not very complicated attractor-H-PES)

4. If necessary, the conformation is briefly refined in the atomistic protein structure according to our scoring function.

After these 4 steps, the protein-ligand conformation is ready for a STUN-MC docking simulation. In total, the whole conformation generation and optimization takes less than 20 seconds on a modern CPU. The advantage of this method is that a by far larger ligand conformational space can be sampled which makes the docking simulation independent of conformations loaded from files.

It is obvious that the pre-optimization is superior to the previous methods if the cavity is very narrow and has a non-isospheric shape. In addition, however we analyzed the binding accuracy of wider cavities with the attractor-H-PES optimization. To test the behavior of our routine, we docked 10000 ligands of a randomly generated database to a homology model of the methyl transferase receptor and compared the calculated binding energies with and without pre-optimization. Methyl transferase methylates the deoxyribonucleic acid. These proteins are involved in the carcinogenesis in humans and presently it is investigated if the methylation process can be influenced by inhibiting the catalytic site of the enzyme with small compounds [102].

Binding Energy (kJ/mol)	without pre-optim.	with pre-optim.
(-230.0, -190.0]	1.0	1.0
(-190.0, -150.0]	21.0	27.0
(-150.0, -110.0]	387.0	410.0
(-110.0, -70.0]	2389.0	2466.0
(-70.0, -30.0]	4097.0	3803.0
(-30.0, 10.0]	1162.0	1173.0
(10.0, 50.0]	456.0	508.0
(50.0, 90]	253.0	306.0

Table 5.1: Comparison of the docking results with and without the pre-optimization method. It is being compared how many ligands were found at the different binding energy intervals (up to 90 kJ/mol).

The results, listed in table 5.1, show that by using the pre-optimization method, more ligands found an energetically top-ranking position than without. This behavior remains the same for structures ranking better than the binding energy of -70kJ/mol. The optimization method proved to be useful for the top $\frac{1}{3}$ ranking ligands of the database.
At the interval from -70 to -30 kJ/mol, not using the pre-optimization method seems to work better. Most of the 10000 ligands have a binding energy in this energy interval. These ligands fit into the receptor but they do not have a very stable binding mode.
At the moment the pre-optimization routine can generate very similar protein-ligand conformations as starting points for the STUN-MC docking simulation. In case of good binding ligands, the pre-optimization proves to work successfully in selecting good initial conformations. But in spacious cavities, as for the protein methyl transferase, it is disadvantageous, if the initial protein-ligand conformations are too similar. Including diversity rules, which force a higher diversity for the initial conformations, may further improve the docking accuracy.

Chapter 6

Binding Accuracy Evaluation

The reliability to determine the native binding mode is an important test for both the scoring function and the search method in protein-ligand docking. Only if the scoring function describes the protein-ligand interactions well, the native protein-ligand conformation can be identified by the best score. On the other hand, a suitable method must be implemented to find this particular protein-ligand conformation reproducibly.
In this chapter, we investigate, how well *FlexScreen* can predict experimental binding modes with a simple interaction-based scoring function, solvation effects are excluded. For this test, we perform docking simulations with eighty-three protein-ligand complexes of the high-resolution subset of the ASTEX/CCDC protein-ligand data set.

6.1 Introduction

As the number of therapeutic targets with available structural information increases, virtual screening of chemical databases to targets of known three-dimensional structure is developing into an increasingly reliable method for in-silico drug development [2, 36, 85, 160, 173]. Both better scoring functions [13, 154, 87] and novel docking strategies [1, 10, 21, 56, 93, 103, 143] contribute to this trend, although no completely satisfying approach has been established yet[13, 174]. The complexity of the physical principles governing protein-ligand interactions makes the formulation of simple, numerically tractable representations a daunting task [147]. It is presently controversial whether force field-based scoring functions based on biochemical/biophysical models are capable to adequately represent the complex interactions that stabilize ligand-protein complexes or whether knowledge-based scoring functions [14, 33, 55, 64, 67, 166, 79, 112, 115, 165] offer more promising results. Techniques, which parameterize protein-ligand interactions without explicit reference to the underlying physical interactions, promise to capture even those contributions to the interactions that are not easily accounted for by numerically tractable physical parameterizations. Such effects include the formation of hydrogen bonds [16], polarizability and the complex influence of the solvent [77]. The virtue of a simple implicit representation of these complex phenomena in

terms of potentials of mean force spurred the development of many regression-based scoring functions. Increasing computational power has permitted these techniques to adopt atomistic representations [50] of the protein-ligand complex. Partially because of their higher cost, force field-based scoring function have received much less attention [13, 174, 79]. The feasibility of direct atomistic simulations of the docking process through molecular dynamics [9, 127, 76] offers an increased ability to parameterize and validate force field-based scoring functions on individual protein-ligand complexes. Force field-based scoring functions, which explicitly refer to established modes of interaction, have been argued to offer better transferability and extendability. In addition, the key-lock-principle, crucial for protein selectivity and problems of induced fit [28, 117], is obviously well represented in a framework that treats the protein and all ligands on the same footing. In this study we have investigated an atomistic docking approach, using a very simple force-field-based scoring function, which is very similar to that used by AutoDock [114], with respect to its accuracy for binding mode prediction. The accuracy by which experimentally resolved ligand-protein complexes are reproduced, serves as one important measure to assess the reliability of in-silico screening approaches. Regression-based methods have demonstrated an impressive progress over recent years [150, 79]. This improvement stems undoubtedly from the improved performance of both the docking methods and the underlying scoring functions.

6.2 Methods

There are two major ingredients to an all-atom in-silico screening method: (1) a scoring function that approximates the binding energy (ideally the affinity) of the protein-ligand complex for different possible conformations of the complex and (2) an efficient optimization method that is able to locate the binding mode of a given ligand to the protein as the global optimum of the scoring function. In a database screen, all ligands are assigned a score, which ideally approximates the affinity of the protein-ligand complex on the basis of the predicted binding mode. Promising lead candidates are then identified by ranking the ligands in the database according to their score. Methods that fail to predict the correct binding modes of known protein-ligand complexes can give fortuitous results in large-database screens. The prediction of the correct binding modes thus emerges as a necessary, but not as a sufficient criterion for in-silico docking protocols. Limitations in the available computational resources and the large number of possible ligands enforce severe approximations in the representation of proteins and ligands. Further difficulties arise from the complexity of the interactions governing protein-ligand interactions. In this study we employ a comparatively simple, interaction-based scoring function and use a fully automated docking protocol, including receptor and ligand preparation.

The number of possible protein-ligand hydrogen bonds is constrained [171, 172] for each hydrogen bond donor or acceptor in ligand and protein to ensure only the evaluation of physically possible interactions. Omission of such constraints leads to unphysical binding conformations

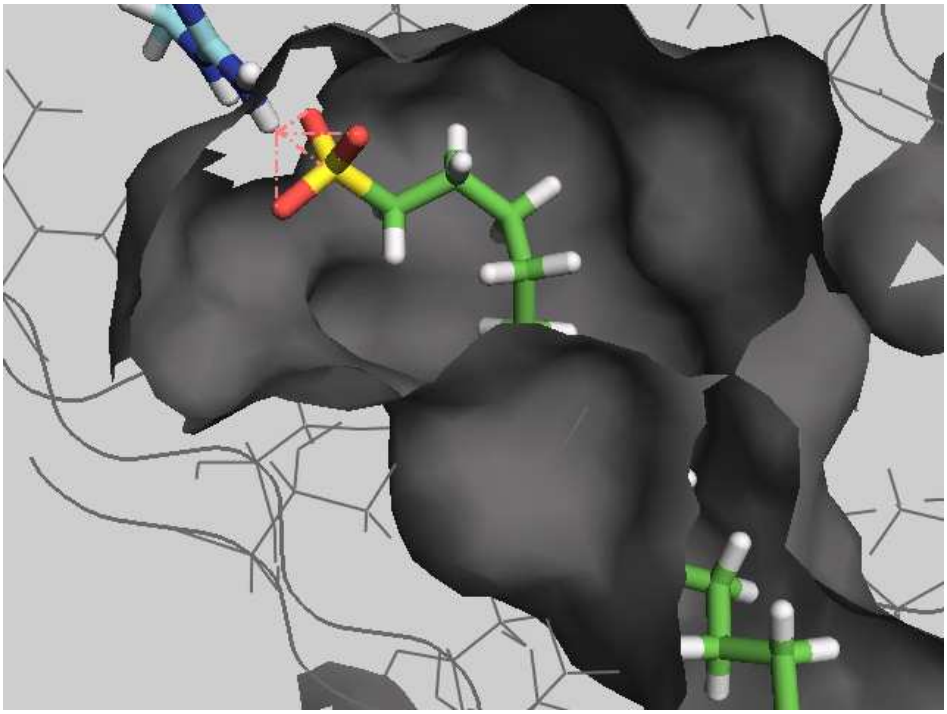

Figure 6.1: Resulting optimal protein-ligand conformation of 1LIC, when the scoring function does not account for the number of hydrogen bonds formed by each atom. The SO_3-head group of the ligand forms 3 hydrogen bonds to a single H-bond donor, leading to un-physically high stabilization energies for a conformation which differs from the X-ray conformation by 4.27 Å. If the hydrogen bond counting is enabled, as described in the methods section, the energetically lowest binding mode improved to 1.08 Å RMSD. In this figure, as well as the following illustrations of protein-ligand complexes, the protein pocket is represented as a gray cavity and atoms are color coded, highlighting specific groups of interest.

as illustrated in figure 6.1. We note that crystal water and co-ligands, such as conserved ions, are easily implemented in the model by treating them on the same footing as rigid protein atoms.

6.2.1 Docking Protocol

The docking simulations were performed using a cascaded docking approach (see section 5.1): The total number of simulation steps is divided into several partitions of similar computational effort. To obtain good statistics we chose a protocol that results in ten final conformations for each protein-ligand complex: In the first partition of the simulation a large number (1000) of short simulations (5000 steps per run) is performed on the ligand. The 50 trajectories with lowest binding energies are then continued for additional 30000 simulation steps each. Of the resulting conformations, again only the lowest ten are selected and further optimized using

another 75000 steps. Each step comprises a random rotation of the ligand, a translation of its center of mass and rotations of each of its flexible bonds, the new conformation is accepted or rejected according to the Metropolis criterion based on the effective energy of the scoring function (see section 3.2.2). This protocol generates ten nearly independent conformations for each protein-ligand complex with a total of 725000 energy evaluations. Depending on the number of ligand atoms (and hence interactions) one such simulation requires 2-3 minutes of cpu-time (on Intel Xeon 2.4 GHz). The cascaded docking strategy balances diversity and computational effort. It invests the largest computational effort into the most promising candidates at the end of each partition, while unsuccessful simulations are terminated early. The difference in the scoring function for the final conformations can be used as a posterior error estimate and serves as an indication whether the predicted binding mode is unique.

6.3 Receptor Structures

The protein-ligand complexes were taken from the Astex/CCDC validation set [123]. We concentrated on the subset of highest quality, containing 92 structures with a resolution of better than 2.0 Å. The protein conformations in the data set (including hydrogen atoms and to some extent partial charges) were frequently insufficient to implement our all-atom scoring function. In many cases, only the neighborhood of the docking site was explicitly included in the data set. For reasons of consistency we therefore prepared the entire set of protein structures on the basis of the original entries of the PDB database using InsightII [70]. This procedure has also been our standard for the screening applications reported previously [107, 106, 108]. Starting with the de-protonated protein structure, hydrogen atoms were attached using the InsightII Builder module, choosing protonation states on the basis of a physiological pH value of 7.4. Partial charges were then assigned with the universal all-atom force field ESFF [149]. In contrast to other studies [73, 50] no pre-optimization of the initial protein or ligand structures were performed, as these tend to adapt the protein pocket to the desired binding mode and hence bias the investigation. Unless explicitly noted below, all the water molecules present in the PDB structures were removed.

6.4 Results

Nine of the 92 ligand-protein complexes were omitted from the investigation because either these ligands were covalently bound to the protein (1aec, 1b59, 1tpp, 1vgc and 4est), or the ligand coordinated with a metal center (2h4n, 1lna, 1xie) or because there were no direct interactions between ligand and protein (3cla). Since the interactions to treat these cases are not implemented in the scoring function used in this investigation, results obtained for these complexes would be fortuitous. The table B.1 (see appendix) demonstrate the overall accuracy obtained for the binding modi of the high-quality Astex data set using the *FlexScreen* flexible-ligand screening approach described above. We list the (median) RMS deviations of

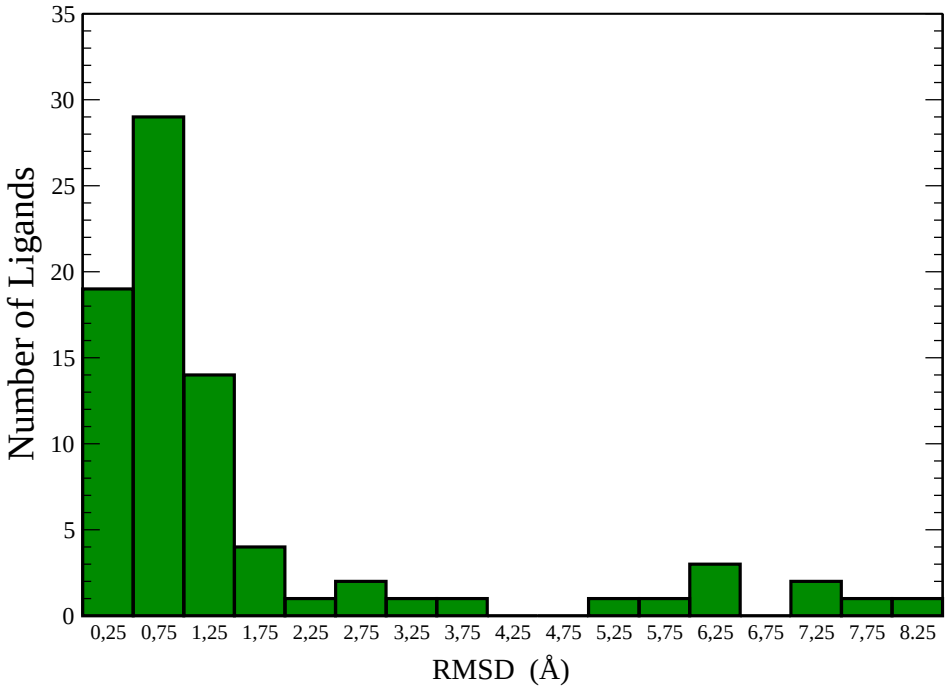

Figure 6.2: Histogram of the RMS deviations for all investigated protein ligand complexes. Only 8 simulations resulted in RMS deviations larger than 5 Å.

the ten final conformations with respect to the X-ray structures, the (root mean) fluctuations of the binding energies and the RMSD obtained with Glide [50], Gold [73] and FlexX [134] (as cited in [50]), where available. The data show that 68 of the 83 complexes (81.9%) are predicted within a median root mean square deviation (RMSD) of less than 2 Å. The distribution of RMS deviations for all runs (10 per ligand) and all ligands is shown in figure 6.2. This compares to 57.1%, 76.0% and 71.4% for FlexX, Gold and Glide respectively. Applying a threshold of 1 Å, *FlexScreen* predicts 50 complexes within a RMSD of 1 Å (60.2%). This compares to 23.2%, 44.0% and 51.8% for FlexX, Gold and Glide respectively. Averaged over all complexes the median RMS deviation to the experimental binding mode was an impressive 0.83 Å, and the unique native binding mode was found in 61 of the 68 cases, indicating the reliability of the docking protocol. An example for an ideal docking result is shown in the left side of figure 6.3, which illustrates the overlay of the 10 docked conformations (thin) in comparison with the experimental conformation (thick). In such cases where not all simulations found the same binding mode (see the right side of figure 6.3 for an example), the one binding mode to be regarded as the natural conformation was selected according to the lowest computed binding energy (this is the case in 7 complexes).

There are eleven protein ligand complexes where the *FlexScreen* protocol fails to locate the correct binding mode, one example is illustrated in figure 6.4, where a very open binding pocket leads to many badly docked conformations. A correct prediction of the binding mode

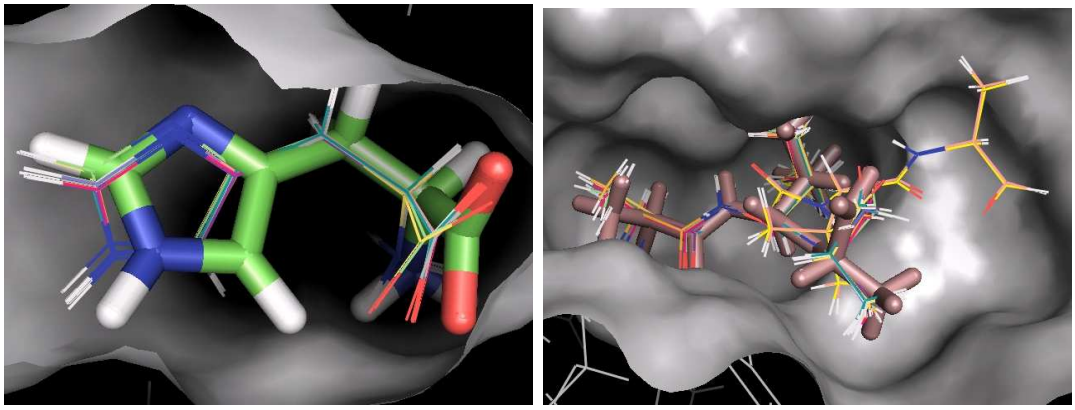

Figure 6.3: Illustration of the computed and experimental binding mode of 1hls and 1hsi. In the first case all of the ten independent simulations correctly identified the experimental binding pose, in the second example two of ten simulations found different binding modes, which are higher in energy.

of this complex requires the treatment of additional entropic contributions and possibly the inclusion of de-solvation effects in the scoring function These data demonstrate that experi-

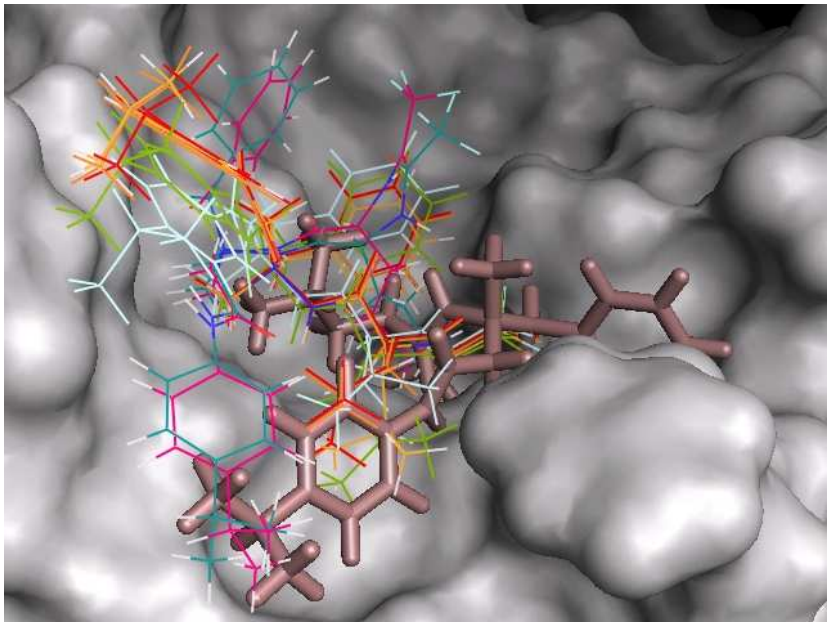

Figure 6.4: Predicted and experimental binding modes of 1bma: the scoring function fails to identify the correct binding mode in a comparatively open binding pocket. This difficulty can be traced back to the scoring function, because the correct binding mode is also sampled, but misplaced conformations repeatedly score with better binding energy.

mental binding modes are very well reproduced for the majority of ligand-protein complexes in the data set. However, there are a few notable and characteristic difficulties that are also likely to be encountered in practical screening applications. In the following, we will discuss these difficulties in detail to illustrate how an atomistic interaction-based scoring approach can lead to rational improvement.

6.4.1 Steric Hindrance

If there are clashes between the ligand and the protein in the experimental X-ray structure, this binding mode is unattainable in the docking approach. In this study, 3 complexes (1tni, 1tnl and 1eta) showed such clashes. In 1tni and 1tnl, hydrogen pairs between protein (GLY216) and ligand exist in the experimental conformation, which have a distance of 1.50 or 1.79 Å. Such distances are too short for most accurate all-atom force fields which assume an equilibrium distance of at least 2 Å. We have used the full vdW radii of the OPLSAA potential for the docking simulations, which generates clashing experimental conformations for such protein-ligand complexes. We applied this approach here nevertheless for consistency, because it reduces the number of false positives in screening applications. In the past, two methods were suggested to remove these clashes for binding mode predictions:
In the investigation of Glide [50], such problems were largely eliminated by annealing away initial clashes. Even the positions of heavy atoms of the protein were optimized in the presence of the ligand under the condition that the RMS deviation of the complex remained below 0.3Å. This procedure improves the docking accuracy for the particular protein-ligand complex under investigation, but it cannot be generalized to screening applications, where the ligand binding mode is unknown. Annealing only the side chains of the cavity before docking also helped to eliminate the clashes in our simulations.

As an alternative approach to the problem of clashing conformations, we repeated the docking simulations with the smaller radii used by AutoDock. Then, all of the ligands enumerated above docked with median RMS deviations below 2 Å. We believe that the use of artificially small radii is preferable to the use of annealed structures, because it can also be applied, when no ligand-protein structure is available. We also note that many clashes occurring in rigid protein screens can be avoided at moderate computational cost by using a flexible protein screening tool, such as *FlexScreen*. To illustrate this point we have performed flexible protein screens for 1xid, where only one of the ten trajectories reached a conformation within 2 Å to the experimental binding mode. Even though this conformation had the lowest energy, the median RMS deviation was 3.22 Å and the energies of the metastable competitors to the experimental docking mode were very close to the optimum. Using a flexible protein screen (4 protein degrees of freedom for the dihedral rotations of the amino acids: GLU_181 (2), ASP_287 (1) and LYS_289 (1)), nine of ten trajectories reached the experimental binding mode, which reduced the median RMS deviation to 0.9 Å for this system. In the flexible protein screen the energy difference between the metastable and the near-experimental con-

formations also increased substantially, indicating that the true experimental minimum was not accessible in the search space of the rigid protein. The principal advantage of this method is that the same protein degrees of freedom can be made flexible without bias during database screens, thereby treating all database ligands on the same footing.

6.4.2 Presence of water and small ions

The treatment of conserved water molecules and small ions is an important outstanding problem in atomistic docking models [95](see figure 6.5). Removal of these molecules increases the size of the cavity, leading to potentially incorrect results, while their explicit treatment would increase the size of the search space drastically and place high demands on the accuracy of the scoring function. A dramatic illustration of the influence of crystal water is 3cla, where the ligand is almost entirely embedded in water. Various water molecules are located in the neighborhood of the ligand to stabilize the binding mode. Although some scoring functions (e.g. Glide [50], AutoDock [126]) implicitly account for solvation effects, this approach remains fruitless here, because the water molecules serve as bridges that mediate the ligand-target interaction. We note that none of the docking programs was able to predict a protein-ligand conformation within a RMSD < 2.0 Å for this complex. The determination of the correct binding mode is only possible with an explicit placement of water. Using *FlexScreen*, this modification led to an improvement of the RMS deviation from 5.9 Å to 0.93 Å. There are also examples (1atl, 1jap, 1xie, 1slt, 1lna) where the ligand appears to form no hydrogen bridges with water, but where conserved water molecules occupy and effectively block remote parts

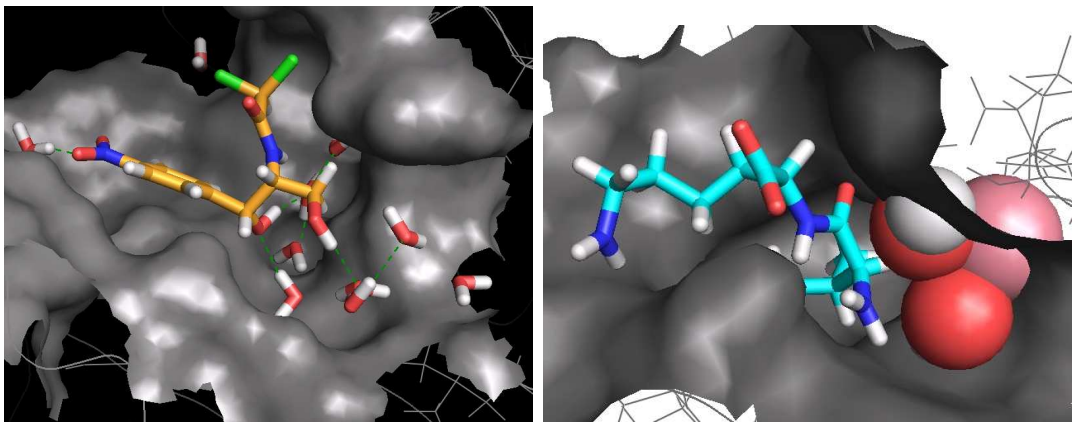

Figure 6.5: Ligand-protein complexes that are stabilized by crystal water. In the case of 3cla (left) the ligand is stabilized by a network of conserved water molecules that mediate indirect interactions between ligand and protein. For 1lna (right), water molecules bound to a calcium ion occupy parts of the protein pocket and confine the ligand to its experimental position. Even though there are no direct stabilizing interactions between ligand and these water molecules, their omission leads to the prediction of a wrong binding mode.

of the cavity. If these are removed, the ligand may find binding modes which were initially excluded and are now competing with the correct binding mode. These complexes provide suitable examples for the development of solvation models, which must take into account the affinity contribution of water molecules interacting with vacant parts of the target.

6.4.3 Comparison with AutoDock

To put these results into perspective, we must ask whether the good performance results primarily from the superiority of the scoring function or from the accuracy of the docking method. This issue can be resolved by a comparison of the *FlexScreen* results with those of AutoDock, in particular for closed protein pockets where solvent interactions have little differential effect on the docking position. Unfortunately, there is no AutoDock data available for the entire CCDC/ASTEX data set. In table B.2 (see appendix) we summarize data from a recent comparative investigation that reported AutoDock data for 11 complexes of this set. In addition we succeeded to generate data for another 14 complexes in independent investigations (2 complexes are same). This data is also included in table B.2. Comparing these results, we see that *FlexScreen* / AutoDock predict the correct binding pose to within 1 Å RMSD in 13 (52%) / 4 (20%) of the cases respectively. Reducing the criteria to 2 Å, we find success rates of 18 (72%) / 14 (56%) respectively. Comparing the data on a case-by-case basis, we see that *FlexScreen* yields more accurate binding poses in 16 (64%) of the cases.

6.5 Discussion

We have investigated the accuracy of the predicted ligand-protein conformation for 83 complexes of the high resolution ASTEX/CCDC data set, for which crystal structures with an experimental accuracy of better than 2 Å are available. For each system we obtained 10 conformations for the protein-ligand complex using *FlexScreen*, which performs atomistic docking simulations based on the stochastic tunneling method with a simple, interaction-based scoring function. The median RMS deviation between the predicted and the experimental structure is 0.83 Å. Each of the ten simulations per ligand performed here requires between 2-3 minutes of CPU time on standard off-the-shelf hardware.
We have compared the accuracy and success rate, defined as the fraction of ligands that were docked in a position with a median RMS <2.0 Å, of *FlexScreen* with other docking methods (Glide, Gold and FlexX), which have been tested for the same data set [50]. In over 80% of the cases, *FlexScreen* found a binding mode with a RMS deviation of less than 2.0 Å , compared with 57%, 71% and 76% for FlexX, Glide and Gold respectively. In a case-by-case comparison for each of the complexes, *FlexScreen* achieved a lower median RMS than the values reported for FlexX, Gold and Glide in 79%, 72% and 46% of the cases, respectively. In contrast to the investigation of Glide, no annealing of the complex prior to docking was performed. *FlexScreen* succeeded to find binding modi in some of the sterically difficult cases. *FlexScreen* approximates the computationally more involved MD based docking schemes [181]

and achieves similar accuracy for many of the targets investigated here. Where affinity differences of a few kJ/mol matter, the exploitation of relevant protein degrees of freedom may be of importance to obtain optimal hydrogen bonding. These observations correlate well with results for other MC/MD based docking methods [168], where MC/MD based docking methods outperformed GA based search strategies for large search spaces and complex docking problems. While GA based techniques are very efficient to explore conformational spaces with uncorrelated degrees of freedom, they may have difficulties to search spaces in which intramolecular rearrangement must correlate with the center-of-mass motion to find the optimal docking mode. We also note that with force field-based scoring functions, widening of the pocket with soft LJ potentials may reduce selectivity [167]. In our approach, intermediate clashes are avoided by the STUN transformation, so that the full potential may be used.

Overall, these results demonstrate that a cascaded all-atom docking protocol with a simple force field-based scoring functions can yield very accurate results that match or exceed those of recently developed knowledge-based models. Our comparison with AutoDock suggests that the reliability of the screening protocol is crucial to obtain these results. As expected for many of the closed binding pockets of the present test set, the scoring function as such appears to be less relevant, as long as minimal accuracy requirements are met.
We stress that results obtained for binding modes are not necessarily transferable to screening applications where the rank order of different ligands in the correct binding pose ultimately determines success. The scoring function has obvious limitations for lipophilic interactions and the ranking of affinities with large differential solvation effects.

Much work remains to be done, regarding both the parameterization of scoring function and docking methodology, to make force field-based methods universally applicable. The treatment of metal-centers in classical potentials will require knowledge-based components in the scoring function. The treatment of crystal water mediating a hydrogen bond network of the ligand, as well as that of other mobile ions, remains a significant challenge in screening applications where the existence and position of such co-ligands cannot be controlled on a case-by-case basis. Our approach delivered good results only when these molecules were properly accounted for. Since the search space of the docking program increases rapidly in the presence of such molecules, thermodynamics-based search methods, such as simulated annealing, genetic algorithms or stochastic tunneling, may be the best option to tackle such complex problems.

Chapter 7

Importance of Protein Flexibility

In this chapter, we investigate both benefits and present limitations of the treatment of target flexibility for high-throughput *in-silico* database screenings. Among the benefits are an improved diversity of binding modes, which allows to identify a wider class of drug candidates. The limitations are related to a diminishing docking accuracy and an increased number of false positives.

7.1 Introduction

The key-lock principle, primarily focused only on geometric criteria [45], is the starting point for rational drug design: if either one ligand of the enzymatic process is known, or the X-ray crystallographic structure of a binding site of the protein has been determined, then a blueprint of a potential drug candidate, a pharmacophore model, can be constructed and molecules designed that share a certain similarity with that blueprint. This strategy has been applied during the last two decades in many successful drug design projects [31].
Difficulties arise, if the conformation of the protein structure depends on the type of ligand, that binds to the protein. As one can imagine, different possible protein conformations permit a far more diverse set of ligands to dock well to the protein.

7.2 Method: Flexible docking to Thymidine Kinase

The following docking simulations were performed using the thymidine kinase (TK) enzyme as an example of a flexible protein. This enzyme has since long been in focus of pharmaceutical research because of its role in reproduction of the herpes simplex virus [34]. It has since then emerged as a useful benchmark system in rational drug design, because not just one, but ten active inhibitors are known and the X-ray structures of their binding modes have been identified [13]. When these ten inhibitors are mixed with a database of randomly selected compounds, the screening tool should be able to identify these as being good binding ligands, i.e. it should assign a high rank to these benchmark ligands. The present enzyme structure is

of particular interest, since the measured target conformations of the various complexes are significantly different.

7.2.1 Preparation of the ligands and the docking site

First, 10000 compounds were randomly chosen from the open NCI database[111]. Since, at this stage, no partial charges were assigned to these compounds, we used the Insight II package [70] with ESFF force field[149] and an automated script to evaluate partial charges for each ligand atom at ph 7.4.
For our analysis, the ligand-free X-ray TK enzyme structure (1e2h)[170]) was taken from the PDB database, partial charges were assigned using Insight II. The ligand-free structure was chosen in order to avoid any conformational bias, created by the ligand closely interacting with flexible side chains inside the binding pocket.

7.3 Results

7.3.1 Screen using a rigid enzyme structure

The database screen was carried out using the cascaded docking method (see sec. 5.1): we start with a population of 100 different conformations, for which we do short docking simulations with 7500 steps. The energetic best five conformations are selected for further

Inhibitor	rigid	6 flex	6 flex+SO_4
acv	221	55	38
ahiu	1454	1315	794
dhbt	2	2	1
dt	308	172	45
hmtt	3117	2934	2076
hpt	8	13	7
idu	612	97	23
mct	nd	4180	4054
pcv	437	71	25
gcv	187	14	4
Score	5225	6576	7063

Table 7.1: Comparisons of different database screens. For each screen we compare the rank of the TK inhibitors in a screen with 10000 randomly chosen ligands of the NCI database. Finally, we evaluate each screen with a score that quantifies how well the active compounds are identified as high affinity ligands. The top row designates the docking model that was used in the screen (nd = not docked)

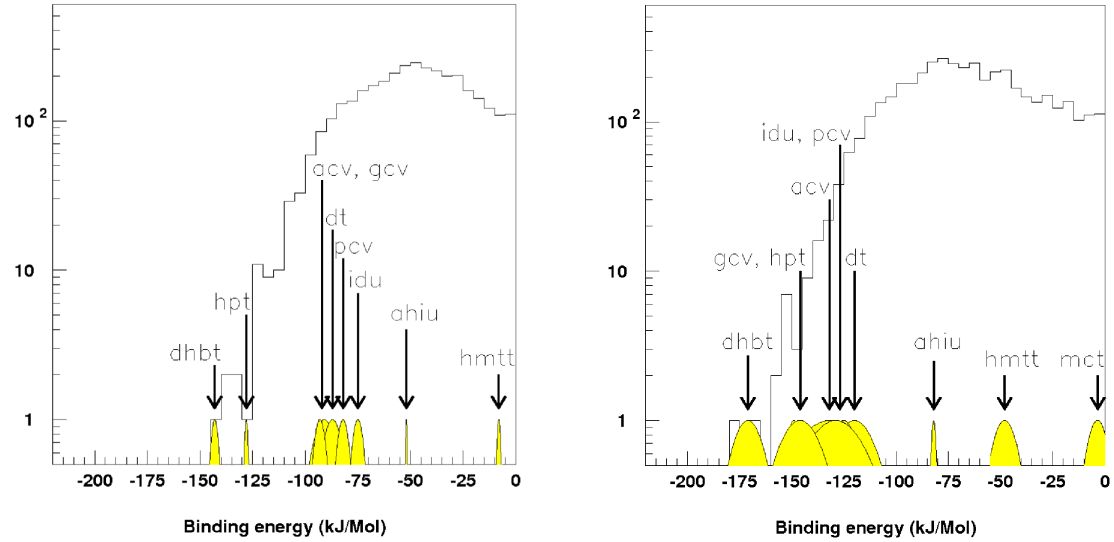

Figure 7.1: Histogram of the resulting ligand binding energies after a screen to the TK enzyme [pdb code: 1E2H] (positions of the known inhibitors are highlighted). Left: Screen to the rigid ligand-free enzyme structure. Right: Screen using a flexible target.

simulations on the next stage with 30000 simulation steps each. Finally the two best energetic conformations are again refined with 75000 simulation steps. In this screen, 3291 database ligands attained a stable conformation with negative binding affinity within the enzyme pocket. The resulting ranks for the ten inhibitors during this screen are summarized in table 7.1 (second column). The ligand dhbt and hpt were ranked with a very high affinity. The ligand hmtt and mct docked badly. Hmtt has barely reached a negative binding energy, whereas mct has never been bound. The majority of the benchmark ligands were energetically more or less close to each other but did not score especially well, as one may notice in the left panel of figure 7.1.

Repeating the docking simulations for these ligands did not substantially improve their rank in the database, eliminating the possibility of statistical fluctuations of the docking algorithm as the source for this difficulty. This enrichment rate is comparable to the results of other scoring functions that were previously investigated for this system, but the overall performance is quite disappointing [13].

In a previous study [44] we compared different database screens to different rigid enzyme structures. In these cases a high specificity of the enzyme to its complexed ligands could be observed. Only those ligands scored well which were similar in their structure to the ligand, the enzyme formed the complex with. This 'memory effect' is a straight consequence of the key-lock principle. Such a high degree of selectivity is spurious, however, since the natural receptor or enzyme, the 'lock', contains a certain degree of flexibility to accommodate a variety of ligand structures 'keys'. The rigidity of the model is thus causing a lack of diversity of

the screen, and consequently the majority of potentially good drug candidates are rejected in such a simulation. The introduction of target degrees of freedom delivers an important tool to recover the diversity of the screening method.

7.3.2 Identification of important side chains

It is immediately clear that a relaxation of all side chains which are located near the binding site is not feasible with todays computational resources, a consequence of the combinatorial explosion of the conformational space. Instead, those side chains which would play an important role in the binding modes have to be identified and partially released.
For this particular pre-study, the X-ray structure of TK in complex with the ligand hmtt [182] was used. The side chain GLN125 was quickly identified as an important hot spot for the binding motif, forming two hydrogen bonds with hmtt. When comparing different complexed crystal structures of the same enzyme with each other, GLN125 turned out to be highly flexible: Its conformation was significantly changed with the ligand it was interacting with. To simulate such a system, 3 chemical bonds of GLN125 (among others) were made flexible to allow the ligands to find their individual binding motifs. The figures 7.2 show two final conformations of a docking simulation for the two ligands dt (left) and gcv (right). Similar to the measured X-ray structures, in our docking study the side chain GLN125 has changed its conformation as well. Dt, utilizing the same binding mode as hmtt, did not modify the side chain orientation significantly compared to the X-ray side chain orientation. On the other hand, when gcv was docked into our flexible enzyme structure, the side chain was moved to

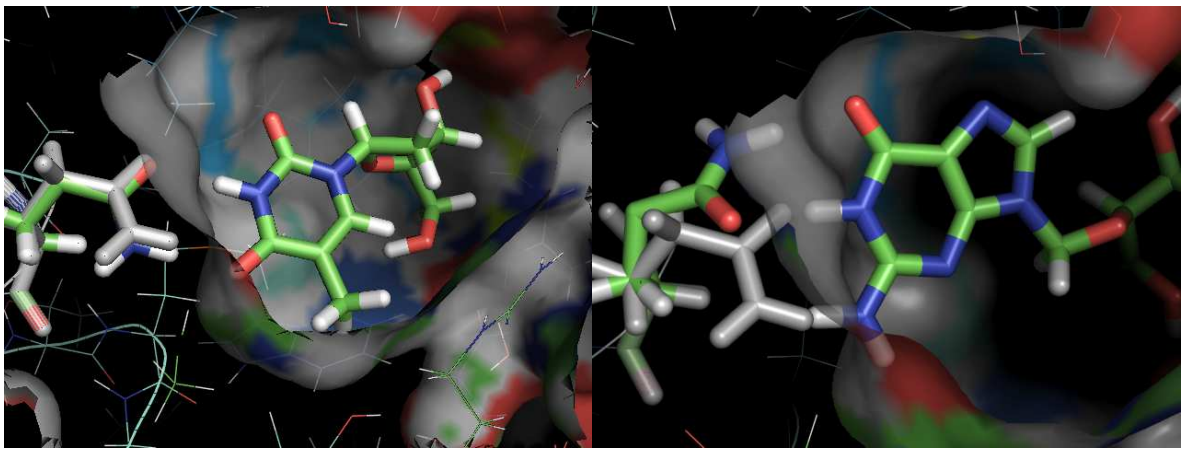

Figure 7.2: The X-ray structure (grey) along with the simulated conformation of GLN125 and the docked ligand. Left: dt (deoxythymidine) docked into 1e2n. The side chain movement is insignificant, since the binding pattern of this side chain matches to this ligand. Right: gcv (ganciclovir) is docked to 1e2n. Since the original crystal binding motif of GLN125 did not allow the ligand to form its individual interaction pattern, a new side chain conformation was energetically more favorable

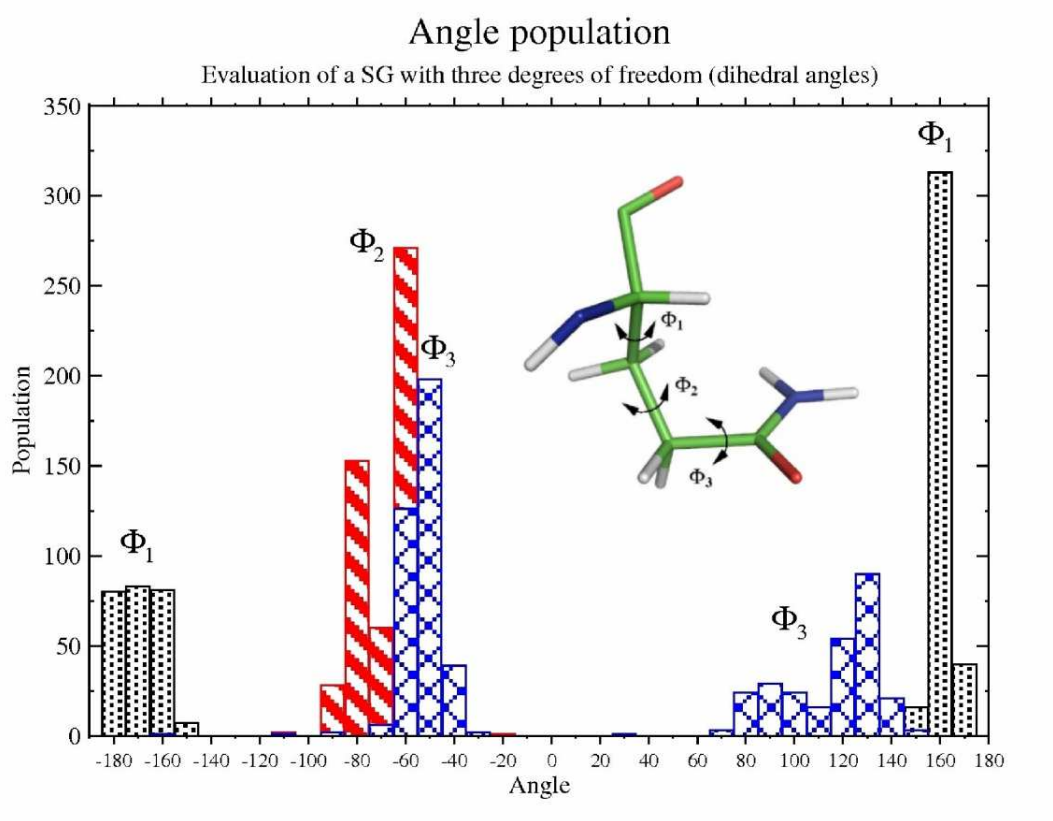

Figure 7.3: Histogram of the changeable three dihedral angles of GLN125. For 640 final conformations the number of occurrences for each dihedral is plotted

form the two important hydrogen bonds with the ligand, as can be seen in the right figure. In the following, we analyze the dihedral angle population of the side chain GLN125 for the final enzyme-ligand conformations after a docking run of the 10 known active substances [13]. For each of the active compounds 32 independent simulations were performed, and those two conformations with lowest energies were selected for the analysis. The conformation of side chain GLN125 can be represented by the three dihedral angles Φ_1, Φ_2 and Φ_3. For each dihedral angle the number of occurrences at the different angles is plotted in the histogram of figure 7.3. The histogram shows more than just three peaks for the different dihedrals: Φ_1 and Φ_2 form two separate peaks, but the dihedral distribution of Φ_3 separates into two heaps. In order to allow the two ligands, as illustrated in figure 7.2, their individual binding modes, this side chain had to flip around by 180 degree, as also observed in the crystal structures. In a similar manner, other side chains whose flexibility would contribute to increase the diversity of the database screen could be identified. For the following screen, 4 side chains with a total of 6 degrees of freedom were introduced into the structure 1e2h, namely dihedral rotations of the amino acids GLN125(3), TYR101(1), ARG222(1), and HIS58(1) (the numbers in brackets denote the respective numbers of flexible bonds).

Binding mode analysis

In the following we want to mention problems, that are caused by crystal water, which are also present in the cavity of TK. Analyzing the binding modes of the ten active compounds in the cross-docking simulation (averaged over 32 conformations each), six out of the 10 compounds find a binding mode with less than 2 Å RMSD (root mean square deviation of non hydrogen atoms) to its complexed structure. These results are listed in table 7.2. Three of the four

Inhibitor	acv	ahiu	dhbt	dt	gcv	hmtt	hpt	idu	mct	pcv
RMSD/Å	3.55	3.29	1.28	0.5	0.88	1.02	2.76	0.79	3.37	1.37

Table 7.2: Averaged cross docking results for the ten active compounds. The RMSD is measured to the specific ligand in its complexed X-ray structure which was aligned to 1e2n.

wrongly docking ligands (ahiu, hpt and mct) have in relation to the side chain GLN125 a symmetric binding pattern: the interaction between these ligand and GLN125 (the main side chain for hydrogen bond interactions) is not altered by a 180°-rotation of the ligand around its longest axes of inertia; these two binding conformations can also be observed in our docking results. Because the energetic difference between the two orientations is very little, it differs only by few kJ/mol, the different resulting conformations are not a problem of our search

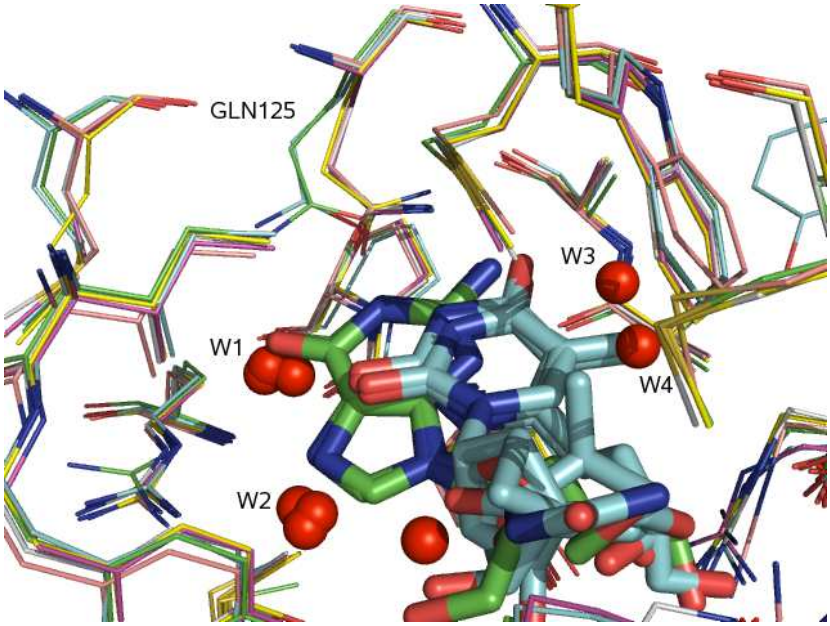

Figure 7.4: Overlay of several different aligned TK crystal structures (inhibitors and water molecules are highlighted). Ligands: two different binding orientation (carbon atoms green or blue), Water molecules: are labeled by numbers. The water molecules are an important criteria for the binding orientation of the inhibitor.

strategy, but a result due to two energetic close competing conformations.
By investigating an alignment of several TK X-ray-structures, the origin for the preference for one conformation is revealed: crystal water molecules. Their presence decides which conformation of the ligand is more favorable.
Figure 7.4 shows the aligned X-ray enzyme structures which illustrates the two important binding orientation of the ligands. The figure also displays different clusters of crystal water molecules (labeled W1, W2, W3 and W4) in the overlay of different enzyme structures.
Modeling the water molecules is very difficult. These water molecules can mediate a hydrogen bond network from the ligand to the enzyme, can be pushed aside from the ligand to become part of the bulk solvent: enthalpic and also entropic energy contributions have to be evaluated. Simulating explicit water molecules during a docking simulation is a challenging problem, since the complexity of the system increases dramatically and also since the entropic change, if the water molecule is freed or not, has to be quantified [1].

7.3.3 Flexible Protein Screen

The results of this screen are summarized in figure 7.1 (right panel), the scores of the individual inhibitors are listed in the column labeled '6 flex' in the table 7.1. Now all ten ligands achieved a negative binding energy. As expected, the number of database compounds that achieved a negative and higher binding energy increased as well, because a flexible conformation of the enzyme reduced the bias of the screen against a specific binding pattern. Since 4251 compounds had now got a negative binding energy (compared with 3291 ligands of the rigid-model run), the diversity of the docking tool had increased by roughly 30%. At the same time, the specificity had decreased because the 'lock' now allowed for a broader class of 'keys' to fit into.
It was also observed that the accuracy of the flexible model screen was lower than that of the screen using a rigid enzyme model (with the same number of function evaluations) because the number of degrees of freedom has increased. In the figure 7.1 the docking energy error is proportional to the width of the cone of the corresponding ligand.
In all investigated X-ray enzyme pockets an additional co-factor (SO_4) is present. Its position is almost invariant in the observed X-ray structures. The SO_4 co-factor is strongly stabilized by a dense hydrogen-bond network with the enzyme. On the other hand, all of our ligands have more than 20 atoms and accommodate worse to the local hydrogen-bond network than this co-factor. Consequently the energetic cost to push the co-factor aside should be very high.
Because we are interested in ligands with a high affinity, we restrict our docking simulations also only to those possible binding conformations, which may lead to a high affinity.
In a second flexible screen, summarized in figure 7.5, we additionally added the co-factor to

[1] At the current version of *FlexScreen* we allow some movable water molecules. At the end of each stage the water molecule conformations are optimized and if favorable, these interactions are evaluated for the binding energy; if not, these interactions are neglected.

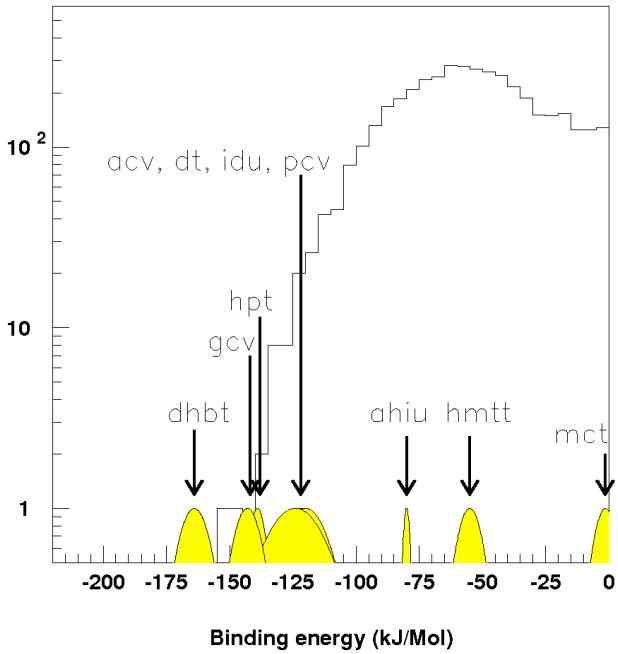

Figure 7.5: Histogram of the binding energies of the docked ligands after a screen to the TK enzyme [pdb code: 1E2H] (positions of the known inhibitors are highlighted). 5 side chains of the enzyme were treated flexible. Additionally also the co-factor SO_4 was included for the screen.

our previous flexible enzyme structure. Because its position appeared to be almost invariant in the observed X-ray structures, it was kept fixed during the simulation. Another benefit of the inclusion of the co-factor is, that the enzyme pocket is now completely closed. In previous screens some enzyme-ligand conformations were partly in contact with the bulk solvent. Such conformations are not possible here.
As a result, 4058 of 10000 ligands now docked to the target, a little less than before due to the sterical restrictions. Comparing the figure 7.5 with the flexible screen of figure 7.1 shows that the binding energies of the ten active compounds with the additional co-factor are not very different from the previous flexible screen. As in the crystal structure their binding conformations are nearly independent from the SO_4-group.

7.3.4 Comparison: Rigid screen vs. flexible screens

To quantitatively compare different screens against the same ligand database, it is sensible to assign an overall score to each screen which rates its performance [89]. We computed such a 'score' for the entire screen from the ranks of the docked known inhibitors among the $N = 1000$ best ligands (uppermost 10%). This score is computed as the sum of $N - P$ where P is the rank of the known inhibitor and shown in the bottom row of table 7.1. An inhibitor

ranking in the top of the screen contributes a score of 1000 to the sum, a badly ranked inhibitor comparatively little. Because the best N ligands are evaluated, screens which dock many known inhibitors with moderate rank may have comparable scores with screens which perform perfectly for one inhibitor, but fail for all others.
With a score of 5225, the rigid enzyme screen displayed the poorest performance among all screens, because the individual binding patterns of the inhibitors are not supported by a rigid enzyme structure.
In comparison the flexible model screens performed much better. With a score of 6576 and 7063 for the screen with the additional co-factor, these results indicate that the increase in diversity of the technique had out-balanced the decrease in specificity, leading to an overall better docking performance of the screening tool.

7.4 Discussion

The necessity to account for the dynamic behavior of a protein has been recognized and discussed [25, 24] for a long time. A single fixed protein structure is often not an adequate model for the dynamical ensemble of structures assumed by the protein. The low energy conformations of a protein may comprise several different side chain orientations which differ by less than 1 kT. Additionally the protein may slightly change its low energy conformation upon ligand binding [25].
In previous studies, Merlitz et al. [109], tested the impact of rigid protein structures on the result of docking simulations. In these studies the binding affinity of the native and the native-like ligands to the fixed protein structure was overestimated. In comparison with other good binding ligands, the ligand of the native protein structure docked significantly better than dissimilar compounds that also have a high experimental affinity.
One possibility to tackle this problem is to perform several docking simulations to different rigid protein structures and then unite all the different results in a consensus score for each ligand. For this purpose, strategies were developed to identify a minimal set of flexible side chains, that allow a generation of a sufficient large structural ensemble of the protein for later rigid docking simulations [3].
In our study we present a different approach. Before docking we identify possible side chains which may change their conformation during the docking simulation. We compared different protein structures with each other and could therefore identify flexible side chains. In this study, we use the ligand-free structure of the enzyme thymidine kinase for the docking simulations. Employing ligand-free structures of a protein with flexibility at the active site is often disadvantageous for docking studies. On the one hand the starting structure is not biased to any known native ligand, but on the other hand the active site of the protein can be narrower than in the complexed structure, because the protein can also accommodate to ligands by minor backbone movements.
Our results demonstrate the shortcomings of a rigid protein structure for virtual screening.

Similarly to consensus scoring, our results suggest that the inclusion of side chain flexibility for the enzyme leads to a less biased score of high-affinity ligands, because important different binding modes are available for a higher variety of ligands. A fixed protein structure has a high specificity for those ligands which are similar to the ligand the protein originally formed a complex with. Including at least partial flexibility for the enzyme allows a by far wider class of ligands to dock well.
As a consequence, differences in the enrichment ratio for different scoring functions [13] may depend more on the suitability of the enzyme or receptor conformation and environment than on the quality of the scoring function.

By using an additional co-factor in the enzyme structure, we investigated, how restrictions change the docking results. The SO_4-group occupies a favorable position in the enzyme structure and several polar residues are blocked for direct ligand-interactions. Since the co-factor is strongly stabilized at its position, it is energetically unfavorable to remove it.
In the docking studies without the co-factor, the affinity of the ligands which have direct interactions with those residues which are blocked if the co-factor is present should be re-evaluated. The affinity of these ligands must be lower, because of the energetic cost to push the co-factor aside. Several approaches are possible for this problem. Since we are solely interested in high affinity ligands (the top 10% of the database), we decided against an energetic penalty for replacing the co-factor. As a consequence of the restricted cavity and also of the restricted amount of different conformations, the accuracy of our docking screen increases. In the screen with the additional SO_4, we notice the highest score. From this we learn, that it is an advantage to use as much information as is available on a system and incorporate it into the docking simulations. When analysing protein structures it should be also investigated if co-factors are present and which of them are likely or unlikely to be removed. Exploiting this information may result in docking screens with a higher accuracy.

Our results demonstrate the importance of crystal water molecules for binding orientation and affinity. As in a previous study (see section 6.4.2), not all water molecules are displaced upon ligand binding. Some water molecules are stabilized in the hydrogen bond network and their dynamic displacement would need a detailed analysis of enthalpic and entropic energy contributions. These contributions would have to be considered for the affinity of a ligand.
Such an analysis is still very difficult presently. In our study, we observe the influence of these strongly bound water molecules on two competing ligand binding orientations. Comparing several enzyme crystal structures, we observe five different probable positions of crystal water molecules. These water molecules seem to determine which of the competing binding orientations is favored.
Very recently, we implemented the treatment of displaceable water molecules (translation, rotation and a total removing) during docking. This method enables us to repeat the study by including the displaceable water molecules. In a prospective study it will be very interesting to investigate the influence of the new method on the docking results.

Chapter 8

Influence of QM Descriptors on Docking Simulations

In this chapter, we investigate, if it is possible to incorporate quantum mechanical calculations into our characterization of proteins and ligands and if it is advantageous for the docking accuracy.
In our program *FlexScreen* proteins and ligands are described with the following properties: Atom types, partial charges, positions, bonding topology and bond types. Especially, the partial charges resulting from the molecular charge distribution can often not be determined accurately. It depends on the connected atoms, bond types and also the neighboring atoms. In this study, we investigate a method to improve the partial charges by ab initio quantum mechanical calculations. We present a methodology with which the accuracy of the docking results increases and still can be used for high-throughput screening.

8.1 Introduction

As our study with the Astex/CCDC study shows, recent docking methods permit the prediction of the binding mode to high-accuracy with modest computational cost. The accuracy of the scoring function for the affinity estimate thus emerges as the most important determinant for the success of this approach [13, 154, 87]. It is presently controversial whether simple scoring functions based on biochemical/biophysical models, henceforth called interaction based scoring functions (IBSF), are capable to adequately represent the complex interactions that stabilize ligand-protein complexes or whether knowledge based potentials [166, 33, 64, 14, 112, 115, 67, 55, 165, 79] offer more promising results. While many methods perform well in the prediction of binding modes, the correct estimation of the affinity remains a significant outstanding challenge. Recent developments of linear-scaling quantum chemistry methods now permit the quantum calculations on large molecules, such as protein fragments [52, 49, 84, 83, 119, 120, 129]. Large scale applications of semi empirical quantum chemical methods for protein characterization [133, 129, 132, 122, 163] have recently shown

good correlation between observed and calculated affinities for a large class of compounds. In this study we investigate two receptors, in which ligand binding is mediated not only by direct ligand-receptor interaction, but also by indirect stabilization of water molecules.
We compare purely classical force field-based scoring functions with quantum-derived scoring functions in the description of these systems: Estrogen is a steroid hormone that plays an import role in the regulation of tissue growth, differentiation and homeostasis. Estrogens also play an important role in bone maintenance, in the central nervous system and in the cardiovascular system where estrogens have certain cardioprotective effects [91, 162, 39, 69] . Estrogens diffuse in and out of cells but are retained with high affinity and specificity in target cells by an intranuclear binding protein, termed the estrogen receptor (ER). Once bound by estrogens, the ER undergoes a conformational change allowing the receptor to interact with chromatin and to modulate transcription of target genes [72, 11, 161].
All nuclear receptors function as ligand-activated transcriptional factors and possess a common domain structure, comprising a conserved DNA-binding domain, variable hinge-regions and conserved ligand-binding domains. The understanding of ER receptor function and the identification of further potential ligands to the ER thus remains an important goal. A number of compounds other than estrogen have been identified as potential ligands of the estrogen receptor, some of which might induce hormone-like effects on humans and animals.
The retinoic acid receptor (RAR) is also a member of the nuclear receptor superfamily [58]. Its ligand-binding domain (LBD, molecular mass 30 kDa) contains the ligand-dependent activation function [113]. RAR (in complex with the retinoid X receptor (RXR)) binds to their target DNA sequences and activate transcription in the presence of retinoic acids, the biologically active metabolites of vitamin A. Retinoids are involved in the regulation of cell growth, differentiation and apoptosis, processes that play an important role in embryonal development and postnatal life [26] and that are the basis for the use of retinoids in cancer prevention and treatment [100, 68, 94]. The RAR family is composed of three genes leading to the α, β and γ isotypes that correspond to distinct pharmacological targets [26], the γ subtype is studied here.
For the estrogen receptor [13, 153, 27, 29, 184, 35], the determination of binding modes and affinities is further complicated by the indirect stabilization of most ligands through a hydrogen bond network involving a conserved water molecule, binding key residues (GLU353, ARG394, LEU387 side chain and backbone) on the one side of the binding pocket in competition with electrostatic interactions with a histidine residue (HIS524), at the opposite corner of the biding pocket (see figure 8.3). A balanced description of these competing interactions thus emerges as an important prerequisite for affinity predictions. High dimensional QSAR studies demonstrate a high correlation between experimental affinities and theoretical estimates based on a multitude of possible binding modes [164], but require significant pre-existing data for parameterization. Standard docking methods [13] identify known binding modes well, but fail to correlate the estimated affinity with the experiment.
For both proteins a large environment of the ligand binding site was recently characterized using the fragment molecular-orbital method (FMO)[51]. FMO permits quantum-chemical

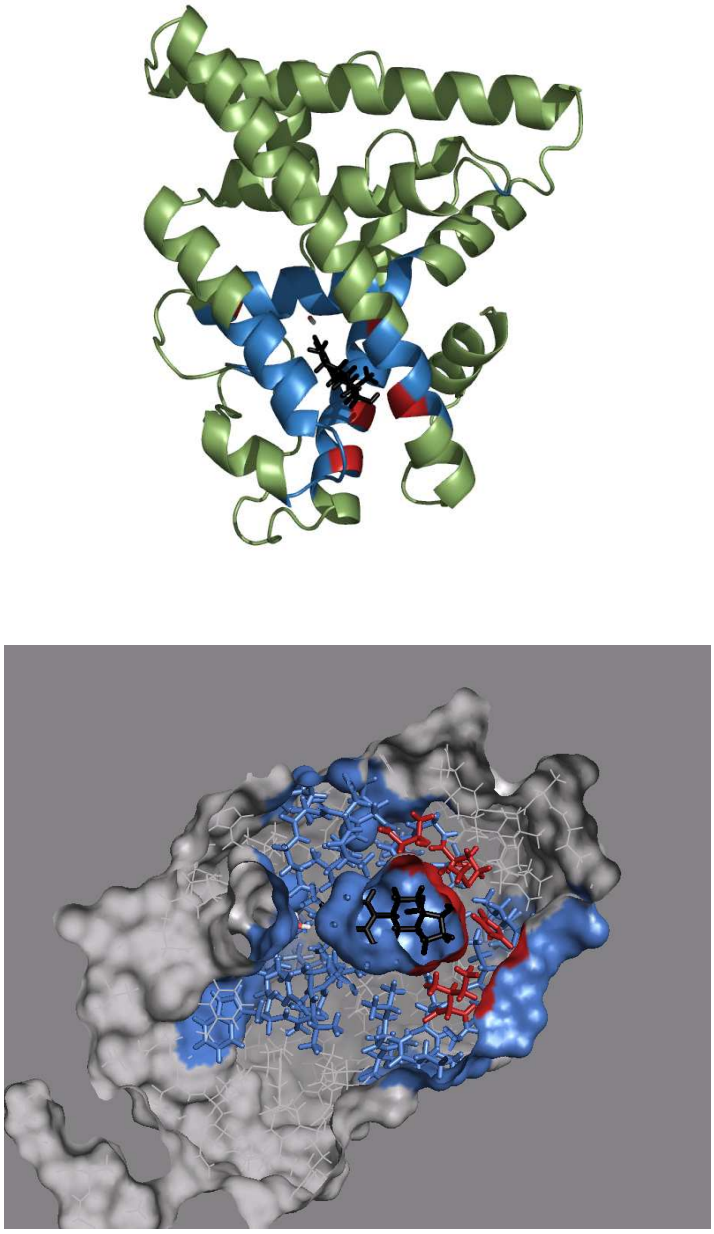

Figure 8.1: Illustration of the ligand binding site of ERα. The top panel illustrates the overall binding site in the ERα (ligand EST shown in black), the lower panel the vicinity of the binding pocket with the residues that were treated in the quantum-chemical calculation in blue and red. The residues that were flexible in the docking simulations are shown in red. The surface around the ligand (EST) illustrates the interaction hot spots with the receptor, the conserved water molecule is visible in a ball-and-stick representation of the left side of the ligand.

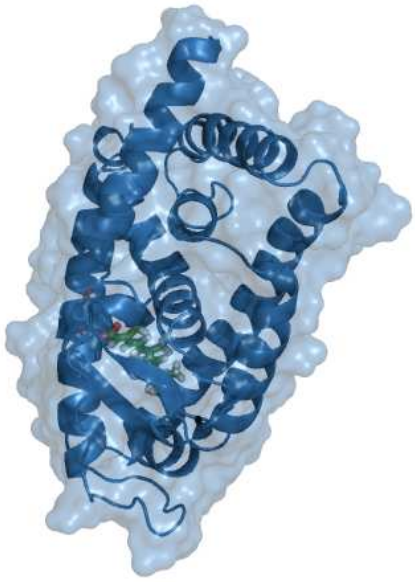

Figure 8.2: Illustration of the binding pocket of RAR in complex with AT-RA.

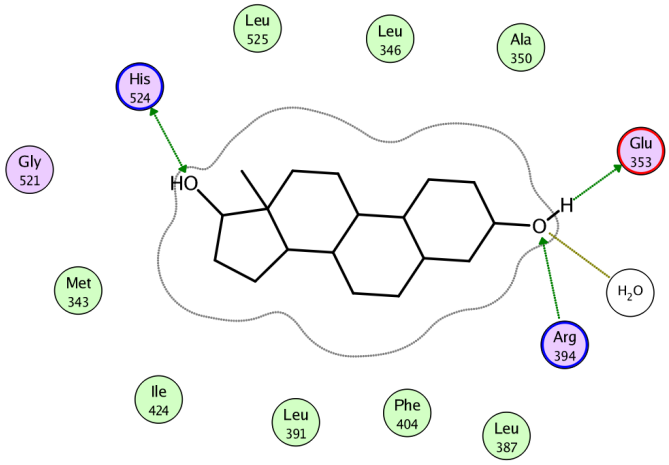

Figure 8.3: Illustration of the ligand-receptor interaction binding site of ERα. The ligand 17β-Estradiol is stabilized by a hydrogen bond network with a conserved water and the residues: GLU353, ARG394 and LEU387 .

calculations of large bio-molecules using a variety of established quantum-chemical techniques beyond the semi-empirical level. Here we derive classical electrostatic models from the partial charges of the quantum calculation and used them in our docking simulations [84, 83, 119, 120]. We docked a set of quantum mechanically characterized ligands into the ERα and RAR receptor structure using both force field based functions (FSF) and the quantum based scoring function (QSF). For ERα we find a significant increase in the correlation between computed and measured affinities for these compounds, when the QSF model is used. We demonstrate, by directly comparing QSF and FSF results, that this improvement is rooted in the quality of the quantum-based electrostatic model. For the RAR receptor we find a very good correlation between experimental and computed binding affinities using both purely classical and quantum based receptor descriptors. Our results therefore demonstrate a viable route for the development of receptor specific interaction based scoring functions that may improve the accuracy of affinity predictions for in-silico drug discovery.

8.2 Ligands, Receptor Structure and Partial Charges

The ERα (pdb code 1ERE) was previously characterized at the HF/STO-3G level using the fragment molecular orbital (FMO) method, both in a ligand free conformation and in complex with the ligands:17β-Estradiol (EST), Diethylstilbestrol (DES), Raloxifene (RAL), 4-hydroxytamoxifen (OHT), Genistein (GEN), Tamoxifen (TAM), 4-hydroxyclomifene (OHC), Clomifene (CLO), 17-Estradiol (ESTA), Bisphenol A (BISA) and Bisphenol F (BISF)[51]. Because quantum calculations are computationally very expensive, even with the FMO-technique [84], the binding energies for the ligands were calculated with respect to the most important fifty-amino acid subset of complete protein receptor structure (50 amino acids). This subset (model 2 in [51], see figure 8.1) includes the most important residues for ER-ligand. The hydrogens in the receptor structure are optimized by CHARMm force field calculations [22], in addition the hydrogen bond network of an even smaller model, consisting of a ligand, a water molecule and the residues GLU353, ARG394, LEU387, HIS524 of the receptor structure (1ERE) were optimized at the HF/6-31G(d) level. Binding energies were calculated as the difference of the FMO energy of the receptor free-structure and the ligand in isolation to the FMO energy of the complex[51].
For the docking study with RARγ we performed docking simulations to the rigid receptor structure of RARγ in complex with the all-trans retinoic acid ligand (AT-RA) (pdb code: 2LBD [136]). With three other ligands we performed docking simulations to this receptor structure: 9-cis retinoic acid (9C-RA), CD564 and BMS181156. The modeling of the ligands and the receptor followed the same procedure as for the estrogen receptor.
For the quantum-based models of the present investigation, we extracted Mulliken partial charges for the receptor and ligand atoms. These Mulliken charges are calculated from the charge distribution of the FMO calculation. For the classical based models partial charges were assigned using the ESFF-force field [149] in InsightII [70] (at pH 7.4), which provides

an adequate assignment of the charges for both the protein and a wide variety of ligands. ESFF succeeded to automatically assign consistent charges for over 180,000 ligands of the NCI Open database (210,000 ligands), and we used the same assignment procedure here for consistency.
In contrast to the FMO study [51] we use a single flexible receptor model for all docking simulations. The rigid part of the receptor was taken from the ERα in complex with EST. Based on a comparison of all receptor models in the FMO investigation the six side chains (MET343, HIS524, LEU525, ILE424, THR347, LEU354) were made flexible (see figure 8.1). Since the computational cost rises with the number of flexible sidechains we have not implemented an automated selection scheme for the flexible side chains.
The docking simulations use a cascaded approach (see section 5.1): the total number of simulation steps is divided into several partitions of similar computational effort. In the first partition 100 simulations with 7500 computational steps, in the second partition 5 simulations with 30000 computational steps and in the third partition 2 simulations with 75000 computational steps are performed. In partition 2 and 3 only the best energetic trajectories of the former partition are continued to simulate. To avoid in partition 2 a simulation of nearly identical conformations, we divide the final 100 conformations in stage 1 into three different clusters: cluster 1 includes the best-scoring conformation and all others with a similar binding pose (RMSD less than 0.8 Å), cluster 2 includes the best-scoring ligand outside cluster 1 and all conformations with a RMSD of 0.8 Å and cluster three contains all other conformations. We start stage 2 with the two top-scorers of clusters 1 and 2 and the top-scorer of cluster 3. For partition 1 we use 3 different starting conformations which are randomly selected for each of the 100 simulations: a relaxed conformation, a conformation with largest atom-to-atom distance and a conformation with smallest radius of gyration (see section 5.4).

8.3 Results

For ERα docking simulations were performed with 11 different ligands and the binding energies were compared to available experimental affinities. We employed different electrostatic models for ligands and receptor, respectively. In the force field-based scoring function (FSF) partial charges were assigned using a classical force field and in the quantum-based scoring function (QSF) partial charges were derived from the FMO model described (see section 8.2). For each ligand we perform 10 independent simulations using the same docking protocol (section 8.2) and calculate the median docking energy and the standard deviation. Table B.3 (see appendix) summarizes the estimated binding energies for all these calculations. Commensurate with previous studies *FlexScreen* (see chapter 6) reproduces the binding modes of the quantum chemical calculations to within less than 1 Å root-mean-square deviation (RMSD) in either model. Figure 8.4 shows typical examples for the calculated receptor-ligand configurations in the QSF model: the left panel of figure 8.4 shows the binding mode of Diethylstilbestrol (DES) in complex with ERα. With a RMSD of just 0.63Å, we

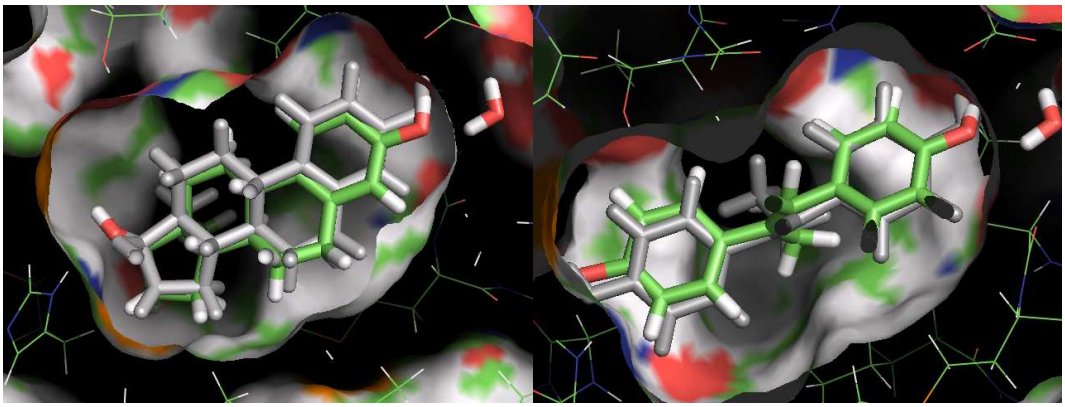

Figure 8.4: Superposition of binding modes from the FMO calculation and and *FlexScreen* for diethylstilbestrol (DIS) and 17α-Estradiol (ESTA), respectively. The grey colored ligand represents the ligand, which position was determined by quantum mechanical calculations and the 'colored' ligand illustrates the binding mode calculated by *FlexScreen*.

successfully find the 'native' binding mode in eight of the ten simulations. The right panel illustrates the calculated binding mode for 17α-Estradiol (ESTA) in complex with ERα, where the *FlexScreen* pose deviates from the crystal structure binding mode (augmented by some hydrogen optimization[51]) by 0.83 Å — the binding mode on the basis of the classical calculation clearly captures the main binding motif of the quantum calculation for the same ligand. Table 8.1 summarizes the deviation between the experimental and the predicted binding mode in the docking simulations for all ligands where crystal structures of the complex are available. We note that there is little difference between the binding modes between the different

	ERα			
	EST 1ERE	DES 3ERD	RAL 1ERR	OHT 3ERT
QSF	0.73	0.63	1.94	0.85
FSF	1.70	0.66	2.23	0.97
	RAR			
	ATRA 2LDB	9CRA 3LDB	CD564 1FCY	bms181156 1FCZ
QSF	0.436	0.918	0.844	1.154
FSF	0.418	0.919	0.845	1.161

Table 8.1: Median RMS deviation (in Å) between the experimental and the calculated binding mode for the ligands where crystal structures are available. The ligand names are followed by the pdb-id of the crystal structure that was used for comparison and the RMSD values for the QSF and FSF models respectively.

scoring functions. Differences in estimated binding energies/affinities thus originate mostly from differences in the representation of the interactions.

These results are a first evidence that the QSF model describes the inter- and intramolecular interactions sufficiently well to reproduce the key features of the binding mode. Using the QSF model we find the correct binding mode in 68% of all simulations: Of 110 independent runs 75 have a RMSD < 2.0 Å to the binding mode determined in the quantum calculation, which demonstrates the reliability of the *FlexScreen* docking protocol. For the FSF model the fraction of reproduced binding modes drops to 55% (of 110 simulations only 61 had a RMSD of less than 2.0 Å), indicating that the two models stabilize slightly different binding conformations. These fluctuations in the binding mode are mirrored in fluctuations of the binding energies obtained in the 10 independent simulations for each ligand. We notice that on average the standard deviation of the energy of the FSF model is almost twice as large as that of the QSF model.

Figure 8.5 (upper panel) illustrates the correlation between the FMO binding energies and the QSF/FSF models. The regression coefficient increases from R=0.71 for the FSF model to R= 0.94 for the QSF model. The binding energies computed in the purely classical model on

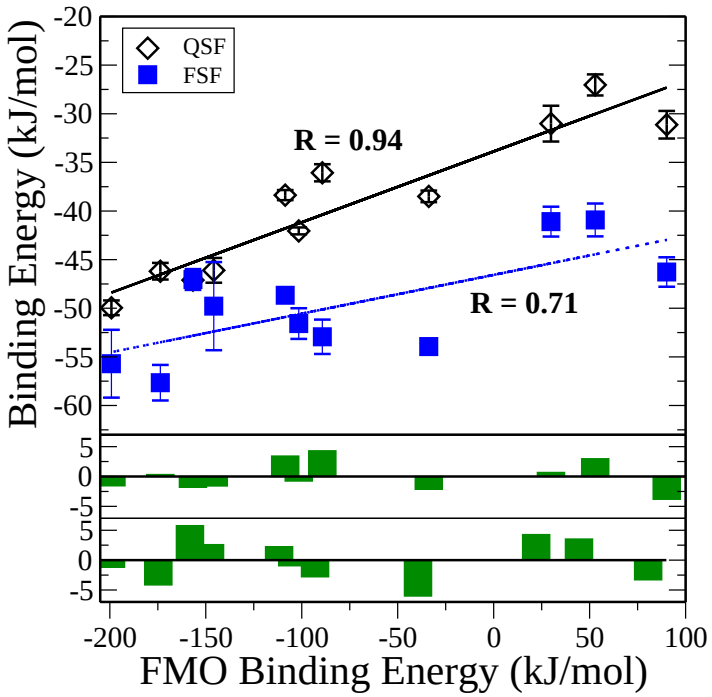

Figure 8.5: Correlation between binding energies in the QSF (diamonds) and FSF (squares) models with the binding energies in the FMO quantum chemical calculation for 11 ligands to the estrogen receptor. The lines indicate least square fits to the data, the correlation coefficients are indicated. The bottom panels give the residual errors of the QSF and FSF model (from top to bottom) to the line of regression.

the basis of the receptor-specific partial charges almost perfectly capture the full quantum-mechanical interactions with the protein environment. It is important to note that no extraneous binding modes arise as a result of the approximate charge model. These data demonstrate that the electrostatic interactions are very well represented in the scoring function derived from the quantum calculation. It should be noted that the ESFF force field, which we used for comparison, gives excellent results for the thymidine kinase [107] (see chapter 7) and dihydrofolate reductase receptors [106]. Of additional interest is also the correlation for a mixed model in which the protein is described with quantum partial charges, while the ligands are treated with a classical force field. Such a mixed model might be used for screening a large ligand database, in which a quantum-chemical calculation for all ligands may be infeasible. A regression coefficient of R = 0.79 indicates that only part of the accuracy gain in the QSF model is retained.

We have also directly compared the computed binding energies in both models with relative binding affinities measured in experiment, which are available relative to EST for 8 of the 11 ligands. Only relative binding affinities are available [92], which were measured by

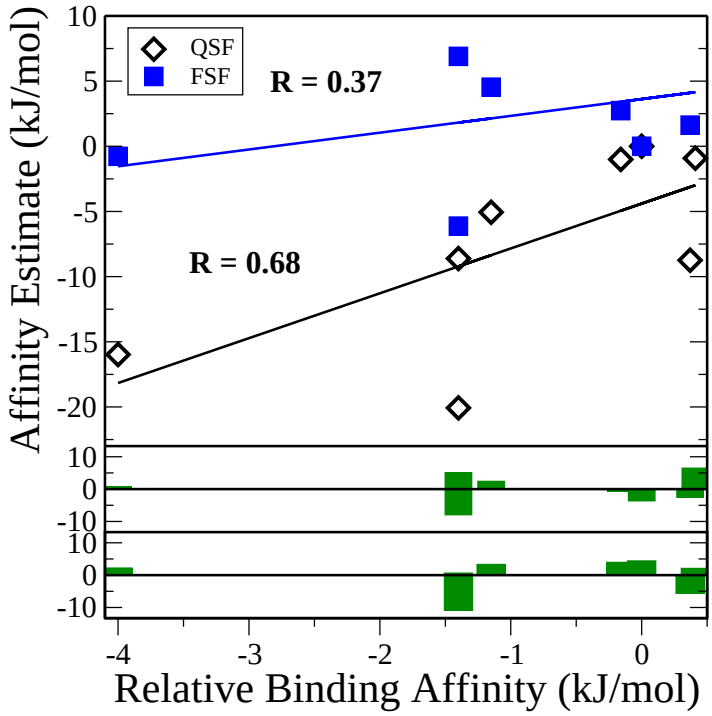

Figure 8.6: Correlation between binding energies in the QSF (diamonds) and FSF (squares) models with experimental relative binding affinities (relative to the ligand EST) for 8 ligands to the estrogen receptor (without solvation energies). The lines indicate least square fits to the data, the correlation coefficients are indicated. The bottom panels give the residual errors of the QSF and FSF model (from top to bottom) to the line of regression.

a solid-phase binding system as a screening assay [60] in which the selected ligand displaces a radio-labeled ligand. The signal detection is based on the fact that ^{3}H emits low energy electrons that have a very short range in solution and therefore only the radioligands bound to receptors triggers a scintillation process. Therefore, the docking results are transformed in relation to the affinity of ligand EST:

$$\Delta\Delta E_{Ligand} = \Delta E_{Ligand} - \Delta E_{EST}. \tag{8.1}$$

Figure 8.6 again illustrates that the quantum-derived scoring function (R=0.68) performs much better than the purely classical scoring function (R=0.37). Including QM parameter into the force field improves also the accuracy of the binding energy in relation to experimental affinities.
Since the models lack de-solvation terms, such comparison cannot be expected to yield quantitative agreement. In order to improve the comparison we have post-scored the QSF/FSF models with a GB/SA type solvation model (see section 5.3) that can approximately account for the differential de-solvation effects of the ligands. Because the receptor pocket of ERα is almost completely filled with the ligands, considered in this study, differential de-solvation

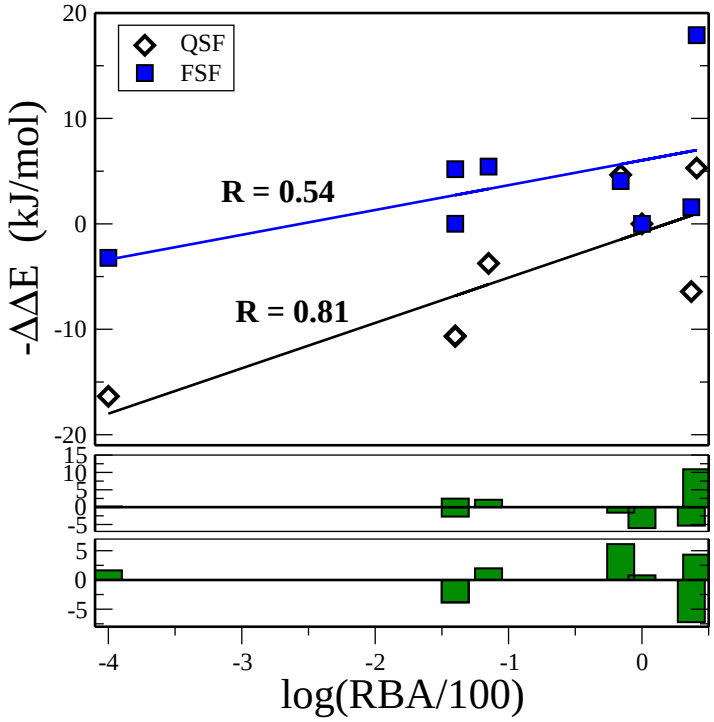

Figure 8.7: Correlation between estimated affinities of the QSF with solvation corrections (diamonds) and the FSF (squares) models with experimental relative binding affinities (relative to the ligand EST) for 8 ligands to the estrogen receptor. The lines indicate least square fits to the data, the correlation coefficients are indicated. The bottom panels give the residual errors of the QSF and FSF model (from top to bottom) to the line of regression.

effects of the receptor are not expected to play an important role. We transform our results for the different ligands correspondingly with the ligand de-solvation energies (listed in table 8.2).

LIG	EST	DES	RAL	OHT	GEN	TAM	ESTA	BISA
QSF	-9.64	-12.03	-14.94	-15.72	-7.44	-19.25	-11.05	-9.32
FSF	-5.27	-5.10	-5.87	-12.51	2.16	-16.91	-6.23	-2.70

Table 8.2: De-solvation energies of the 8 ligands calculated by the GB/SA model for the two different force fields QSF and FSF

Figure (8.7) shows the correlation of the computed results to the experimental affinity values of log(RBA/100). For the QSF model we find a very good correlation to the experimental data ($R = 0.81$), which is by far higher than that of the FSF model ($R = 0.54$). As above we have additionally investigated a mixed model, where partial charges for the receptor were derived from the quantum calculation, while the ligands were parameterized with ESFF. For this model we also find a high correlation ($R = 0.78$) [$R = 0.52$ without the de-solvation energy], data not shown, which is again significantly higher than with ESFF alone.

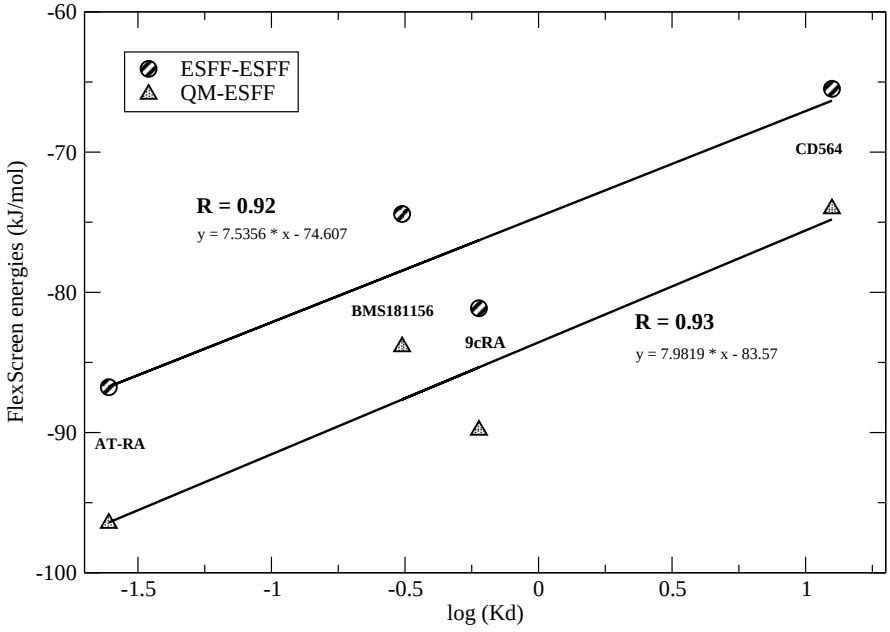

Figure 8.8: Comparison of two docking simulations to the RAR receptor. The receptor is described one time with the FSF and the other time with the QSF force field. The results of the binding energies for four ligands (AT-RA, BMS81156, 9cRA and CD564) are plotted in comparison to experimental results ($\log K_D$).

The results for the RAR receptor are similarly encouraging. As figure 8.8 illustrates we obtain a correlation of R=0.93 and R=0.92 when correlating the experimental affinity with the binding energy obtained with the QSF and FSF, respectively. The correlation of the binding energy with that obtained by the quantum calculation is even higher (R = 0.99 and R = 0.98, respectively). For consistency we have computed the same de-solvation corrections used for the estrogen receptor, which do not affect the the observed correlations. Including these terms we obtain a correlation of R=0.95 for the QSF and R=0.93 for FSF model. The deviation between the computed and the experimental binding modes varies from 0.4-1.6 Å, as detailed in table 8.1.

8.4 Discussion

The development of accurate, yet fast in-silico screening protocols remains a daunting task. While there is a growing consensus that serval docking methods are well suited to predict binding modes of small molecules to well characterized proteins to near-experimental accuracy [175], the development of quantitative methods to estimate the affinity remains a significant challenge. The estrogen receptor, where receptor-ligand binding is mediated by a conserved water molecule for many ligands is a good example for such a system. A comparative study for investigating several docking protocols and scoring functions found that binding modes are well predicted, while affinity estimates remain poor [13]. In the presence of many well characterized ligands, high-dimensional QSAR methods can be used to parameterize the interactions to achieve a high correlation between computed and estimated affinity [153, 164]. Similarly high correlation may be obtained using alternate heuristic schemes [29, 184, 35] or as heuristic post-scoring as the comparative molecular field analysis method [180].
These methods, however, fail to address the fundamental question how to improve scoring functions for in-silico screening tools in the absence of protein-specific experimental data. Many present day docking strategies use knowledge based scoring functions that average over many known protein-ligand complexes in the hope to generate a good ranking also for different ligands. Alternatively one may try to improve force field based scoring functions using quantum chemical models to parameterize a specific protein [133, 129]. Given the complexity of quantum chemical calculations for whole proteins or significant parts of the protein environment and the difficulty to parameterize models that combine classical force field terms with protein specific parameters obtained from other sources, these methods are still in their infancy.
Here, we investigated the binding energies for eleven typical ligands to the binding domain of the human estrogen receptor ER-α. We find that binding energies calculated at the quantum chemical level correlate very well with those obtained with a classical scoring function, provided that the classical scoring function is parameterized with charges obtained from the quantum calculation. This result does not hold, when an all-atom force field is used to obtain charges for the ligands and receptor, indicating that the use of quantum charges may improve

the accuracy of binding energy calculations in high throughput screening applications. This conclusion also carries over, when calculated affinity estimates are correlated with relative binding affinities measured experimentally.
These data demonstrate that electrostatic potentials obtained from quantum calculations beyond the semi-empirical level can be used in conjunction with classical force field based scoring functions to improve the affinity estimates for the particularly complex ERα receptor. We find that the binding modes are equally well predicted with FSF and QSF based simulations, so that the improvement in the affinity estimate stems from the improved treatment of the electrostatic interactions in the QM model.
Even higher correlations for the measured and predicted affinities were also observed for the QM derived scoring function of the RAR receptor, where quantum methods were also advocated in the search for new antagonists [152]. Recent studies of the related retinoic X receptor also required the specific adaptation of the scoring function [156] to obtain adequate results, which an earlier screen of the Available Chemicals Directory (MDL Information Systems, San Leandro, CA), a compound structure database of over 150,000 molecules, using the ICM scoring function resulted in a single novel antagonist [141].

8.5 Conclusion

The development of accurate scoring functions for complex protein-ligand binding interactions remains a difficult task. The increasing availability of quantum calculations for large protein fragments or even entire proteins offers exciting possibilities for the design of protein specific scoring functions to improve force field based scoring functions for in-silico screening [133]. Our data demonstrates that it may be possible to transfer electrostatic potentials from quantum calculations to develop such protein specific interaction based scoring functions. This approach has the advantage that it does not require pre-existing experimental data for the given protein, while still producing adequate models to describe the interactions.
As quantum chemical calculations for large biomolecular systems, such as those of the FMO method exploited here, become more accurate, protein specific interaction based scoring functions may help to capture the specifics of a particular system. Our study demonstrates that such an approach can work for post-semiempirical quantum methods; much work remains to be done before a generic protocol for the derivation of such protein specific functions is established. It complements other work [132, 122], where methods of quantum mechanics are used to improve structure based drug design.
It should be noted that the high cost of quantum chemical calculations at the ab-initio level (HF or higher) is no major deterrent for the derivation of such models. The protein characterization at the quantum level is required only once, while the resulting model can subsequently used for ten-thousands of binding simulations, which dwarf the cost of the quantum calculation to define the scoring function.

Chapter 9

Summary

In the projects reported in this thesis, we have investigated novel simulation protocols for the rational development of drugs. To this end, we have developed a new high-throughput protein-ligand docking method. This method employs the three-dimensional structural information of the protein to identify new compounds that bind with a high affinity to the protein.

In chapter 6, we show that our simulation approach reproducibly determines known binding modes with high accuracy [42]. This evaluation is an important test to prove the validity of the scoring function and the search method. The scoring function approximates the receptor-ligand interactions employing several different energy potential types. A good scoring function should be able to identify a native protein-ligand conformation, i.e. a conformation that is observed experimentally, as the energetic global minimum on the potential energy surface of the protein and the ligand. This can only succeed if the docking tool used is able to locate the global minimum on the potential energy surface with great reliability and speed. For this purpose, we employ the stochastic tunneling search method.
In the evaluation of our approach, we successfully show that both the scoring function and the search method can reproduce known binding modes. These results demonstrate the validity of our docking method and allow us to apply our approach in further studies, in which the ligand affinity and the native protein-ligand conformation are not known.

It is therefore important to develop a methodology that further improves the accuracy of the calculated binding energies and the binding modes for ligand database screens. As one way out of this dilemma, we attempted to improve our force field based scoring functions by augmenting it with receptor-specific data obtained from quantum-chemical calculations. With our study of the estrogen receptor and of the retinoic acid receptor [43], presented in chapter 8, we compare scoring functions derived from purely classical models with those derived from quantum-chemical calculations with respect to experimental affinities. Because the electrostatic interactions in proteins and protein-ligand complexes are very complicated, the assignment of protonation states and partial charges is one major determinant of the stability

of particular complexes. The choice of these parameters is one of the most important ingredients of a molecular-mechanics type scoring function. We have therefore attempted to augment scoring functions derived from standard force fields, which describe many properties of proteins and ligands sufficiently well, by receptor-specific partial charges and protonation states from quantum mechanical calculations (QM). Employing results of the fragment molecular orbital (FMO) method, receptor specific scoring functions (here based on quantum derived partial charges) were integrated into our protein model. We then compared the docking results using scoring functions derived from purely classical models with their hybrid cousins: We find that the highest accuracy can be obtained by incorporating the QM derived partial charges into the ligand and also into the receptor model. By integrating QM derived partial charges into the receptor model only, we still see a significant increase of accuracy compared to our standard receptor model. These results are very important for high-throughput screening docking tools: Even though QM calculations for the receptor structure are very time consuming, this effort has to be invested only once, before the actual screening of the ligand begins. Due to the separability of scoring function parameterization and the actual docking calculation, the accuracy of the screen improves with little increase in the overall computational cost.

In chapter 7, we show the advantage of our flexible protein model in the docking study of the thymidine kinase enzyme (TK). Using only one rigid protein structure for docking simulations generates an artificial constraint on particular ligands. Ligands which have a binding mode similar to the native ligand of the used protein structure benefit from this constraint and appear to dock better in the simulation than dissimilar ligands with equally good experimental affinities.
However, a protein is a dynamic structure: Due to the thermal activation energy, a protein is constantly subject to slight conformational changes in the backbone or, even more important, in the protein side chains. If a ligand and a protein form a complex, the ligand is able to induce small conformational changes to the protein structure without high energetic cost (usually in the order of kT). As a consequence, compared to the ligand-free structure, we observe a slightly different three-dimensional structure for the protein when bound to the ligand. Often, these conformational changes can be traced back to conformational changes of the side chains only.
In the TK study (see chapter 7), we investigate if advantages can be observed using a protein model with flexible side chains in large-scale docking simulations. We analyze the conformational changes of a flexible side chain upon docking with different ligands and compare these ligand and side chain orientations with the known binding conformations obtained from crystal structures. We observe very similar structural conformations in our model compared to the crystal structures. These results prove that in the case of the TK enzyme our protein model with several flexible side chains can well approximate the dynamic protein that accommodates to different ligands by different conformations.
In large database screen, we compare the docking results of 10000 ligands with 10 ligands of known high affinity to TK. The database screens are performed using a rigid enzyme struc-

ture and flexible protein models. Here, we demonstrate the superiority of our flexible protein models also quantitatively. Due to being less biased to only a few ligands, our flexible protein models identify the active compounds with higher accuracy by comparing the energetic rank of the active compounds with the ligands of the database.
Additionally, in this study, we observe the importance of crystal water molecules for protein-ligand docking. Crystal water molecules are water molecules which are strongly stabilized in the protein. These molecules often can not be neglected in docking studies. Their presence or absence may determine the final protein-ligand conformation. Observations like these also guide our present research endeavors. Recently, we have developed an extension to our approach that explicitly incorporates crystal water molecules as movable and removable parts of the system (not included in the thesis).

Through our studies, we have successfully developed a high-throughput docking approach which allows us to identify well binding ligands in large databases. The inclusion of protein flexibility by allowing side chain flexibility proved to be an advantage over using only one rigid protein structure for our docking simulations.

Outlook

Our approach has proved to be useful, but we also observe difficulties which have not yet been addressed. Sometimes, the ligand-induced conformational changes of the protein can not be approximated solely by side chain flexibility. For example, docking Staurusporine into the ligand-free protein kinase structure does not work with our present methodology because backbone movements must also be taken into account. We are presently starting to work on challenging cases like these and integrate local backbone movements into our approach.
Another field for improvements is the scoring function. For the studies presented in this thesis, the solvent contribution was only, if at all, taken into account for the ligands. For more opened cavities this approach does not prove to be successful.
In the current version, the solvation energies of the ligand-protein complex are calculated during the docking simulation. For each atom, the solvent accessible surface area (SASA) is determined and the solvation energy of the complex is calculated as the sum over all atoms of the SASA per atom multiplied by an atom type specific factor. These SASA atom-type specific parameters can not be derived analytically. At the moment, we are searching for the optimal parameters in reference to our current force field.

Other parts of the scoring function also need to be improved making further research necessary:

1. Pi-stacking: Aromatic rings are stacked together by an overlap of p-orbitals. This interaction type can not be calculated by usual attractions of a van der Waals interaction. The interaction differs in strength and geometry. Therefore, a new potential type has to be implemented that takes the pi-stacking interaction into account.

2. Transition metal interaction: It is often the case that the ligand coordinates with transition metals. These interactions depend on the geometric coordination group of the transition metal. Such effects are not considered in our scoring function. Integrating and parameterizing these interactions into a scoring function is very challenging and will be a study in its own.

Our approach is well suited for protein-ligand systems for which solely the structural information of the protein is available. If no native inhibitor is known, it is not possible to search for well-binding compounds by similarity. Very recently, for example, a new strategy for fighting the HIV-virus was discovered. Zhou et al. [185] structurally resolved one of the viruses' flexible proteins that does not alter during the mutation. This protein is important for the virus to attach to cells in the organism. With discovering the protein's structure the first step has been done and the search for inhibitors can start. These inhibitors can hopefully be used to prevent the virus from multiplying. With our studies we aim to contribute to research efforts like these.

Appendix A

Analytical Calculation of Electrostatic Energies

For cases in which ligands are buried in the protein, the whole protein-ligand complex can be approximated as a sphere, surrounded by water. Under this assumption eq. 4.16 can be calculated analytically [82, 178, 59] and expressed in Legendre polynomials. With a minor approximation eq. 4.16 can be written in the following form [178]

$$\begin{aligned} E_{elec} &= \sum_i \left\{ -\frac{q_i^2}{2R} \frac{\frac{1}{\epsilon_u} - \frac{1}{\epsilon_v}}{1 + \frac{\epsilon_u}{\epsilon_v}} \left(\frac{\epsilon_u}{\epsilon_v} + \frac{1}{1 - \left(\frac{r_i^2}{R}\right)} \right) \right\} \\ &\quad + \sum_{i>j} \left\{ \frac{q_i q_j}{\epsilon_u |\mathbf{r_i} - \mathbf{r_j}|} - \frac{q_i q_j}{R} \frac{\frac{1}{\epsilon_u} - \frac{1}{\epsilon_v}}{1 + \frac{\epsilon_u}{\epsilon_v}} \left(\frac{\epsilon_u}{\epsilon_v} + \frac{1}{\sqrt{1 - 2\frac{r_i r_j}{R^2}\cos\theta + \left(\frac{r_i r_j}{R^2}\right)^2}} \right) \right\}, \end{aligned} \quad \text{(A.1)}$$

r_i, r_j being the distance from the origin (center of the sphere), R the radius of the assumed sphere with $r_i, r_j < R$ and enclosed angle θ between $\mathbf{r_i}$ and $\mathbf{r_j}$.

In order to apply this approximation, we first try to find the best suitable sphere for each specific receptor-ligand complex. We start with an independent calculation of the solvation energies for a charge of one electron at different positions in the protein cavity, using the program APBS [8]. The different positions are usually the atom positions of a ligand and of surrounding side chains. Then, by changing the origin and the radius of the virtual sphere, we determine the sphere that best approximates the calculated solvation energies (least square error).

It takes several hours to numerically calculate the solvation energies, but it has to be done only once for the specific protein. After a suitable sphere is identified by an optimization procedure, the electrostatic interaction and self-energies can be calculated very quickly using the analytical eq. A.1.

The method works well for closed pockets, but it is difficult to generalize for open pockets and cases in which the ligand docks far outside the cavity.

In figure A.1, we compare the solvation energies at the cavity of the receptor-ligand complex (pdb: 1KI2) with the analytical solvation energies of the sphere model. Compared are not directly the solvation energies, but the solvation energies over distance. The black dots represent the ideal analytical sphere solution. As one can see, the main behavior, predicted by the analytical calculations, is well represented. On the other hand, outliers can also be observed.

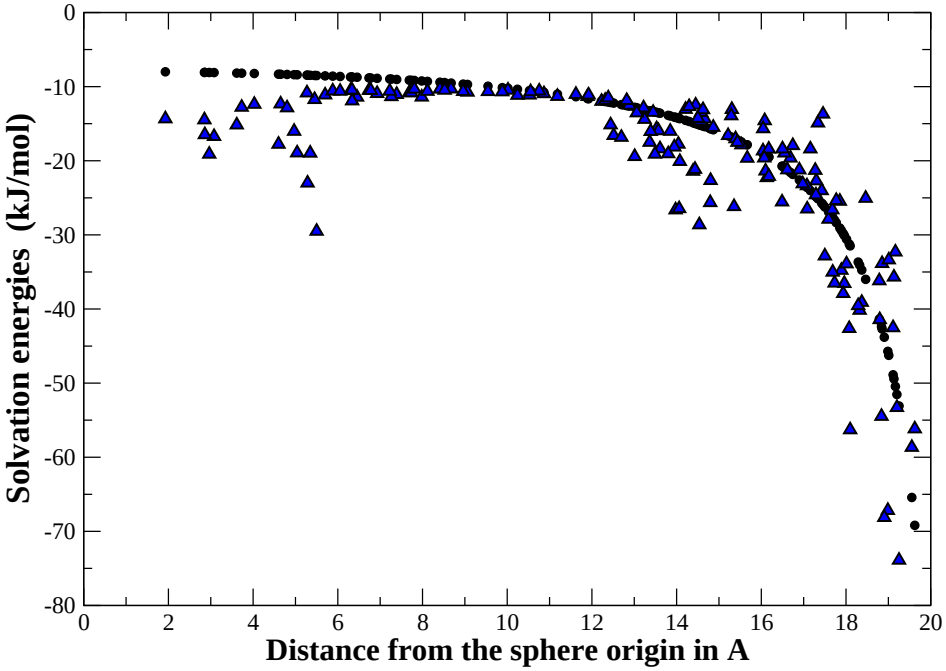

Figure A.1: Comparison of the analytical solvation energies (as in eq. A.1) of 157 point charges (black dots) in the cavity with the solvation energies calculated numerically with the program APBS [8] (blue triangles). The 157 point charges always have the charge 1e and are positioned at the atoms of a docked ligand and of representative surrounding side chains. As one can see, the main behavior, predicted by the analytical calculations, is well represented. On the other hand, outliers also can be observed.

Appendix B

Tables

Table B.1: Results of the docking test set of chapter 6: median RMSD (root mean squared deviation of non hydrogen atoms), the number of runs yielding a RMSD < 2.0 Å (out of 10), root mean energy fluctuation, RMSD results of Glide, Gold and FlexX (in Å).

Name	RMSD(Å)	# < 2.0Å	E. fluc	Glide	Gold	FlexX
1a28	0.33	10	0.1	-	-	-
1a4q	0.74	10	0.1	-	-	-
1a6w	0.88	10	0.5	-	-	-
1abe	0.38	10	0.2	0.17	0.86	1.16
1abf	0.58	10	1.0	0.20	-	1.27
1aoe	0.51	10	0.1	-	-	-
1apt	1.04	9	7.2	0.58	1.62	1.89
1apu	0.73	10	3.1	-	-	-
1aqw	7.76	0	4.3	-	-	-
1atl	3.93	2	7.3	0.94	-	2.06
1b58	0.96	10	0.7	-	-	-
1bma	1.03	9	1.28	9.31	-	13.41
1byb	0.79	10	12.2	10.49	-	1.62
1c1e	5.14	0	1.0	-	-	-
1c5c	0.43	10	1.5	-	-	-
1c5x	1.88	10	0.1	-	-	-
1c83	0.48	10	0.1	-	-	-
1cbs	0.38	10	0.0	1.96	-	1.68
1cil	1.63	9	2.2	3.82	-	3.85
1coy	0.49	10	0.5	0.28	0.86	1.06
1d0l	1.11	10	9.2	-	-	-
1d3h	0.47	10	0.2	-	-	-
1ejn	0.40	10	0.3	-	-	-

Name	RMSD(Å)	# < 2.0Å	E. fluc	Glide	Gold	FlexX
1eta	8.43	0	2.6	2.92	11.21	8.46
1f3d	0.68	10	0.3	-	-	-
1fen	0.56	10	0.1	0.66	-	1.39
1flr	0.58	10	0.6	-	-	-
1glp	0.33	10	2.3	0.34	-	0.47
1glq	1.28	10	4.1	0.29	1.35	6.43
1hfc	2.49	0	2.9	2.24	-	2.51
1hpv	1.04	10	1.7	-	-	-
1hsb	0.41	8	4.6	-	-	-
1hsl	0.79	10	1.1	1.31	0.97	0.59
1hvr	0.64	10	0.3	1.50	-	3.35
1hyt	1.08	9	1.3	0.28	1.10	1.62
1ida	1.14	10	1.4	11.88	12.12	11.95
1jap	1.40	10	1.3	-	-	-
1kel	6.40	0	2.2	-	-	-
1lcp	0.56	10	0.2	1.98	-	1.65
1lic	1.08	10	2.3	4.87	10.78	5.07
1lna	2.49	0	9.6	0.95	-	5.40
1lst	0.60	10	3.6	0.14	0.87	0.71
1mld	0.63	10	0.2	0.32	-	1.45
1mmq	0.66	10	0.4	0.92	-	0.52
1mrg	0.50	10	0.0	0.30	-	0.81
1mrk	0.86	10	0.4	1.20	1.01	3.55
1mts	0.57	10	0.2	-	-	-
1nco	0.34	10	0.9	6.99	-	5.85
1phd	0.95	10	0.0	1.22	0.85	0.65
1phg	0.28	10	0.0	4.32	1.35	4.74
1ppc	1.38	9	2.1	7.92	-	3.05
1pph	2.00	5	1.9	4.31	-	4.91
1qbr	0.46	10	0.6	-	-	-
1qbu	0.50	10	0.8	-	-	-
1rds	0.59	10	1.2	3.75	4.78	4.89
1rnt	1.02	10	4.7	0.72	-	1.90
1rob	1.16	10	7.3	1.85	3.75	7.70
1slt	0.87	9	2.1	0.51	0.78	1.63
1snc	6.19	0	5.4	1.91	-	7.48
1srj	7.23	0	0.9	0.58	0.42	2.36
1tmn	0.83	10	4.4	2.80	1.68	0.86
1tng	0.36	10	0.1	0.19	-	1.93

Name	RMSD(Å)	# < 2.0Å	E. fluc	Glide	Gold	FlexX
1tnh	0.82	10	0.5	0.33	-	0.56
1tni	2.80	0	0.7	2.18	-	2.71
1tnl	2.73	0	0.1	0.23	-	0.71
1tyl	7.37	0	0.9	1.06	-	2.34
1ukz	6.21	0	0.7	0.37	-	0.94
1wap	0.26	10	0.3	0.12	-	0.57
1xid	3.22	1	5.0	4.30	0.92	2.01
2ak3	0.45	10	0.3	0.71	5.08	0.91
2cmd	0.55	10	0.4	0.65	-	3.75
2cpp	0.37	10	0.7	0.17	-	2.94
2ctc	1.64	10	0.8	1.61	0.32	1.97
2fox	0.84	8	10.4	-	-	-
2gbp	0.66	10	1.6	0.15	-	0.92
2qwk	0.97	10	2.2	-	-	-
2tmn	1.22	10	1.1	0.58	-	5.16
2tsc	1.52	10	0.6	-	-	-
3ert	0.58	10	5.4	-	-	-
3tpi	0.38	10	0.8	0.49	0.80	1.07
4dfr	0.94	10	1.7	1.12	1.44	1.40
5abp	0.43	10	1.2	0.21	-	1.17
6rnt	5.83	0	4.3	2.22	1.20	4.79
7tim	1.25	10	8.7	0.14	0.78	1.49

Table B.2: Comparison of the docking results (RMSD in Å) of *FlexScreen* and AutoDock for a subset of the 83 complexes. For 14 complexes, we did the calculation with AutoDock by ourselves, and 11 complexes are from a previous study [23]. These are labeled by '*' (2 complexes are the same)

Name	*FlexScreen*	AutoDock
1a4q	0.74	1.40
1aqw	0.88	3.84
1abe	0.38	0.16*
1abf	0.58	0.48*
1apt	1.04	1.89*
1apu	0.73	9.10*
1c1e	5.14	9.63
1cil	1.63	5.81*
1mrk	0.86	1.35
1phg	0.28	3.52*
1rds	0.59	4.71
1rnt	1.02	2.04
1snc	6.19	1.97
1srj	7.23	1.85
1tng	0.36	0.62*
1tni	2.80	1.97/2.61*
1tnl	2.73	0.44/0.41*
1ukz	6.21	2.73
2ak3	0.45	5.73
2cpp	0.37	3.40*
2ctc	1.64	1.09
2fox	0.84	1.83
2tsc	1.52	4.46
5abp	0.43	0.48*
6rnt	5.83	1.57

Table B.3: Median binding energies and affinities (in kJ/mol) for the ligands investigated in the study of chapter 8. QM designates the quantum mechanical calculated binding energy, QSF and FSF the median binding energies of the 10 independent docking simulations without and with ($\triangle G$) the solvation correction described in the chapter 8. RBA is the measured relative binding affinity as in [51].

	QM	QSF	FSF	MIXED	QSF+$\triangle G$	FSF+$\triangle G$	RBA
EST	-158	-47	-47	-39	-57	-52	100
DES	-112	-38	-49	-36	-50	-54	236
RAL	-148	-46	-50	-42	-61	-56	69
OHT	-175	-46	-58	-35	-62	-71	257
GEN	-39	-38	-54	-38	-45	-52	4
TAM	44	-27	-41	-26	-46	-58	4
OHC	-199	-50	-56	-38			
CLO	22	-31	-41	-28			
ESTA	-105	-42	-52	-39	-53	-58	7
BISA	80	-31	-46	-30	-40	-49	0.01
BISF	-93	-36	-53	-39			

Appendix C

Parameter for the Scoring Function

C.1 Lennard-Jones Parameter

In this section, the atom type parameters for the Lennard-Jones interaction energies are annotated. As described in section 5.3, the Lennard-Jones interaction energies are expressed as

$$E_{ij}^{LJ} = 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]. \tag{C.1}$$

	OPLSAA parameter		**AutoDock parameter**	
Atom type	$\sigma_{\mathbf{ii}}$ (/**nm**)	$\epsilon_{\mathbf{ii}}$ (/$\frac{\mathbf{kJ}}{\mathbf{Mol}}$)	$\sigma_{\mathbf{ii}}$ (/**nm**)	$\epsilon_{\mathbf{ii}}$ (/$\frac{\mathbf{kJ}}{\mathbf{Mol}}$)
C	0.35	0.276	0.356	0.63
H	0.225	0.126	0.178	0.08
O	0.312	0.711	0.285	0.84
S	0.355	1.05	0.356	0.84
N	0.325	0.711	0.29	0.67
I	0.35	1.81	0.35	1.81
Cl	0.40	0.494	0.40	0.494
F	0.25	3.01	0.25	3.01
Br	0.41	0.377	0.41	0.377
Li	0.21	0.076	0.21	0.076
Na	0.33	0.012	0.33	0.012
K	0.44	0.002	0.44	0.002
Mg	0.17	3.66	0.17	3.66
Ca	0.24	1.88	0.24	1.88
Zn	0.18	3.66	0.18	3.66
else	0.35	0.276	0.35	0.276

Table C.1: Two sets of Lennard-Jones parameters per atom type: OPLSAA and AutoDock

The table above lists all the Lennard-Jones parameters per atom type which are used for the presented studies. We use two parameter sets: either from the OPLSAA force-field [75] or from the program AutoDock [114].
If interactions of different atom types are considered, the following equations are used to calculate the Lennard-Jones parameter for mixed atom types:

$$\varepsilon^{ab} = \sqrt{\varepsilon^{aa}\varepsilon^{aa}} \tag{C.2}$$

$$\sigma^{ab} = \sqrt{\sigma^{aa}\sigma^{bb}}. \tag{C.3}$$

C.2 Hydrogen bond parameters

The hydrogen bond energies are calculated, as explained in section 5.3, with

$$E_{ij}^{HB} = \cos\Theta_{ij}\left(\frac{\tilde{R}_{ij}}{r_{ij}^{12}} - \frac{\tilde{A}_{ij}}{r_{ij}^{10}}\right) + \sin\Theta_{ij}\ \ 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]. \tag{C.4}$$

Depending on the angle Θ_{ij}, the van der Waals potential crosses over to a hydrogen bond specific term. Therefore, we distinguish two different hydrogen bond types and classify them according to the hydrogen bond donor atom: Either oxygen, nitrogen or sulphur.
We employ the same parameters the program AutoDock [114] also uses. These are listed in the following table:

Atom type	$\mathbf{\tilde{R}_{ij}}$ $(/\frac{\mathbf{kJ\,nm^{12}}}{\mathbf{Mol}})$	$\mathbf{\tilde{A}_{ij}}(/\frac{\mathbf{kJ\,nm^{10}}}{\mathbf{Mol}})$
N	2.315e-7	7.696e-6
O	2.315e-7	7.696e-6
S	1.247e-6	2.394e-5

Table C.2: Parameters for the hydrogen bond potential

The angle Θ_{ij} is calculated as described in section 4.1.3.

Appendix D

Used programs and definitions

LaTeX

For the thesis the LATEX typeset and the Kile-1.8 integrated environment for LATEX was used. Accessible via http://www.latex-project.org/ and http://kile.sourceforge.net/

PDB-database

The Protein-Data-Bank (PDB) database [12] is an information portal for macromolecular structures. It can be accessed via http://www.pdb.org

RMSD

The Root Mean Square Deviation (RMSD) is used to compare two data sets with each other. In our examples, one data set represents experimental results and the other results of docking simulations. The RMSD is defined as

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \tilde{x}_i)^2}. \tag{D.1}$$

PyMOL

For all three-dimensional representations of ligands and proteins the program PyMOL [32] was used.

APBS

The adaptive Poisson-Boltzmann solver (APBS) [8] was used to calculate numerically electrostatic solvation energies under the continuum solvation assumption.

MOE

The Molecular Operating Environment (MOE) was used for the two-dimensional representation of protein-ligand complexes.

InsightII

The InsightII molecular modeling environment [70] was used for most of the proteins and ligands preparations. These preparations consisted of constructing small molecules, of the protonation of molecules, of assigning partial charges and also of optimizing the geometries of chemical compounds.

XmGrace

Graphs were plotted using the program XmGrace. XmGrace is developed at the Weizmann Institute in Israel and runs under GNU General Public License.
http://plasma-gate.weizmann.ac.il/Grace

MSMS

MSMS is a program to calculate molecular surfaces. It is developed by M. Sanner and is freely available.
In previous *FlexScreen* versions, the molecular surface calculation was not included and MSMS was used for that purpose. MSMS was also used for comparing the accuracy of molecular surface calculations.

Abbreviations

Amino acids

ALA	Alanine
ARG	Arginine
ASN	Asparagine
ASP	Aspartic acid
CYS	Cysteine
GLN	Glutamine
GLU	Glutamic acid
GLY	Glycine
HIS	Histidine
ILE	Isoleucine
LEU	Leucine
LYS	Lysine
MET	Methionine
PHE	Phenylanaline
PRO	Proline
SER	Serine
THR	Threonine
TRP	Tryptophan
TYR	Tyrosine
VAL	Valine

TK	Thymidine Kinase

Thymidine Kinase ligands

dT	deoxythymidine
idu	5-iododeoxyuridine
hpt	6-(3-hydrody-propyl-thymine)
ahiu	5-iodouracil anhydrohexitol nucleoside
mct	(North)-methanocarba-thymidine
hmtt	6-[6-hydroxymethy-5-methyl-2,4-dioxo-hexahydro-pyrimidin-5-yl-methyl]-5-methyl-1H-pyrimidin-2,4-dione
acv	aciclovir
gcv	ganciclovir
pcv	penciclovir
dhbt	6-[3-hydroxy-2-(hydroxymethyl)propyl]-5-methylpyrimidine-2,4(1H,3H)-dione

ERα	Estrogen receptor of subtype α

Estrogen ligands

EST	17β-Estradiol
DES	Diethylstilbestrol
RAL	Raloxifene
OHT	4-hydroxytamoxifen
GEN	Genistein
TAM	Tamoxifen
OHC	4-hydroxyclomifene
CLO	Clomifene
ESTA	17-Estradiol
BISA	Bisphenol A
BISF	Bisphenol F

RARγ	Retinoic acid receptor of subtype γ

Retinoic acid receptor ligands

AT-RA	all-trans retinoic acid
9C-RA	9-cis retinoic acid
CD564	-
BMS181156	-

MC	Monte Carlo
STUN	Stochastic tunneling method
SA	Simulated Annealing
RMS(D)	Root means square (deviation)
MM	Molecular mechanics
MD	Molecular dynamics
QM	Quantum mechanics
FMO	Fragmental molecular orbital method
PDB	Protein Data Bank
FSF	Force field based scoring function
QSF	Quantum based scoring function
PMF	Potential of mean force

List of Figures

1.1 Illustration of the activation energy with and without the catalytic activity of an enzyme. . . . 6
1.2 Chemical reaction involving an enzyme. . . . 6
1.3 Illustration of drug activity upon an enzyme. . . . 7
1.4 Signal pathway of hormones 8
1.5 Conformational change of a receptor induces an opening of an ion-channel . . 9
1.6 Conformational change of a receptor structure enables enzymatic reactions . 10
1.7 Illustration of side chain flexibility 15

2.1 Illustration of the thermodynamic cycle approach 23

3.1 Effective potential energy surface transformed with the stochastic tunneling method (STUN) 29

4.1 Illustration of different types of bonded interactions 32
4.2 Lennard-Jones potential 34
4.3 Description of the hydrogen bond potential 37
4.4 The solvent-accessible-surface area (SASA) 39
4.5 Comparison of numerical calculated solvation energies with highly accurate reference values calculated by finite-difference Poisson-Boltzmann solvers . . 43

5.1 Process flow of a complete docking simulation 45
5.2 Illustration of our ligand representation 47
5.3 A Ligand and a narrow cavity representation in 2 dimensions 51
5.4 Two different representations of the protein-ligand complex 1STC 52

6.1 Influence of a not-implemented hydrogen bond counting routine on the resulting protein-ligand conformation 57
6.2 Histogram of RMS deviations for all investigated protein ligand complexes . . 59
6.3 Illustration of the computed and experimental binding mode of 1hls and 1hsi. 60
6.4 Docking failures with open cavities: Predicted and experimental binding modes of 1bma 60
6.5 Ligand-protein complexes that are stabilized by crystal water 62

7.1 Histogram of the resulting ligand binding energies after a screen to a rigid and a flexible model of the TK enzyme 67
7.2 Illustration of the importance of a flexible protein model: New important binding motif due to the conformational change of the side chain GLN125 68
7.3 Histogram of the three changeable dihedral angles of GLN125. 69
7.4 Crystal water molecules in an overlay of several different aligned TK crystal structures 70
7.5 Histogram of the resulting ligand binding energies after a screen to a flexible model of the TK enzyme with the additional co-factor SO_4 72

8.1 Illustration of the ligand binding site of ERα 77
8.2 Illustration of the binding pocket of RAR in complex with AT-RA 78
8.3 2 dimensional representation of the ligand-receptor interaction binding site of ERα 78
8.4 Superposition of binding modes from the FMO calculation and *FlexScreen* for diethylstilbestrol and 17α-Estradiol 81
8.5 Correlation between binding energies of the QSF and FSF models with the binding energies of the FMO quantum chemical calculation for 11 ligands to the estrogen receptor. 82
8.6 Correlation between binding energies of the QSF and FSF models with experimental relative binding affinities (relative to the ligand EST) for 8 ligands to the estrogen receptor (without solvation energies). 83
8.7 Correlation between estimated affinities of the QSF with solvation corrections and the FSF models with experimental relative binding affinities (relative to the ligand EST) for 8 ligands to the estrogen receptor. 84
8.8 Comparison of the calculated and the experimental binding energies for four ligands. The receptor is described one time with the FSF and the other time with the QSF force field. 85

A.1 Comparison of analytical solvation energies with the solvation energies calculated numerically with the program APBS 94

List of Tables

5.1 Comparison of docking results with and without the pre-optimization method 54

7.1 Comparison of different database screens: The success to identify known high affinity ligands is quantified 66
7.2 Averaged cross docking results for the ten active compounds: Root mean square deviations to the native structure 70

8.1 Median RMS deviation (in Å) between the experimental and the calculated binding mode 81
8.2 De-solvation energies of the 8 ligands calculated by the GB/SA model for the two different force fields QSF and FSF 85

B.1 Docking results of the test set of chapter 6. 95
B.2 Comparison of the docking results (root mean square deviations to the native binding mode) of *FlexScreen* and AutoDock for a subset of the 83 complexes of chapter 6 98
B.3 Median binding energies and affinities (in kJ/mol) for the ligands investigated in the study of chapter 8. Energies and affinities are listed according to the used force field: QSF, FSF, mixed 99

C.1 Two sets of atom type specific parameters for the Lennard-Jones potential: OPLSAA and AutoDock 101
C.2 Parameters for the hydrogen bond potential 102

Bibliography

[1] R. Abagyan and M. Totrov. Biased probability monte carlo conformation searches and electrostatic calculations for peptides and proteins. *J. Molec. Biol.*, 235:983–1002, 1994.

[2] R. Abagyan and M. Totrov. High-throughput docking for lead generation. *Curr. Opin. Chem. Biol.*, 5:375–382, 2001.

[3] A. Anderson, R. O'Neil, T. Surti, and R. Stroud. Approaches to solving the rigid receptor problem by identifying a minimal set of flexible residues during ligand docking. *Chemistry & Biology*, 8:445–57, 2001.

[4] I. Andricioaei and M. Karplus. On the calculation of entropy from covariance matrices of the atomic fluctuations. *J. Chem. Phys.*, 115:6289–92, 2001.

[5] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.

[6] J. Aqvist, C. Medina, and J.-E. Samuelsson. A new method for predicting binding affinity in computer aided drug design. *Protein Eng.*, 7:385–391, 1994.

[7] H. Ashbaugh and M. Paulaitis. Effect of solute size and solute-water attractive interactions on hydration water structure around hydrophobic solutes. *J. Am. Chem. Soc.*, 123:10721–10728, 2001.

[8] N. Baker, M. Holstand, and F. Wang. Adaptive multilevel finite element solution of the Poisson-Boltzmann equation; ii: Refinement at solvent accessible surfaces in biomolecular systems. *J. Comput. Chem.*, 21:1343–52, 2000.

[9] P. Bash, U. Singh, R. Langridge, and P. Kollman. Free energy calculations by computer simulation. *Science*, 236:564–568, 1987.

[10] C. Baxter, C. Murray, D. Clark, D. Westhead, and M. Eldridge. Flexible docking using tabu search and an empirical estimate of binding affinity. *Proteins*, 33:367–382, 1998.

[11] M. Beato, P. Herrlich, and G. Schutz. Steroid hormone receptors: many actors in search of a plot. *Cell*, 83:851–857, 1995.

[12] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–42, 2000.

[13] C. Bissantz, G. Folkers, and D. Rognan. Protein-based virtual screening of chemical databases. 1. evaluation of different docking/scoring combinations. *J. Med. Chem*, 43:4759–4767, 2000.

[14] H. Böhm. Prediction of binding constants of protein ligands: A fast method for the prioritization of hits obtained from de novo design or 3D database search programs. *J. Comput. Aided Mol. Des.*, 12:309–323, 1998.

[15] H.-J. Böhm, G. Klebe, and H. Kubinyi. *Wirkstoffdesign.* Spektrum Akademischer Verlag GmbH, 2002.

[16] D. Boobbyer, P. Goodford, P. McWhinnie, and R. Wade. New hydrogen-bond potentials for use in determining energetically favourable binding sites on molecules of known structure. *J. Med. Chem.*, 32:1083, 1989.

[17] M. Born. Volumen und Hydratationswärme der Ionen. *Z. Phys.*, 1:45–48, 1920.

[18] G. Brady and K. Sharp. Decomposition of interaction free energies in proteins and other complex systems. *J. Mol. Biol.*, 254:77–85, 1995.

[19] G. Brady, A. Szabo, and K. Sharp. On the decomposition of free energies. *J. Mol. Biol.*, 263:123–5, 1996.

[20] B. Brandsdal, J. Aqvist, and A. Smalås. Computational analysis of binding of p1 variants to trypsin. *Protein Sci.*, 10:1584–1595, 2001.

[21] N. Brooijmans and D. Kuntz. Molecular recognition and docking algorithms. *Rev. Biophys. Biomol. Struct.*, 32:335–373, 2003.

[22] 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. *J. Comp. Chem.*, 4(2):187–217, 1983.

[23] B. D. Bursulaya, M. Totrov, R. Abagyan, and C. L. Brooks. Comparative study of several algorithms for flexible ligand docking. *J. Comput.-Aided Mol. Des.*, 17(11):755–763, 2003.

[24] H. Carlson. Protein flexibility and drug design: how to hit a moving target. *Curr. Opin. Chem. Biol.*, 6:447–52, 2002.

[25] H. Carlson and J. McCammon. Accommodating protein flexibility in computational drug design. *Molecular Pharmacology*, 57:213–18, 2000.

[26] P. Chambon. A decade of molecular biology of retinoic acid receptors. *FASEB J.*, 10:940–954, 1996.

[27] D. E. Clark1, C. E. Poteet-Smith, J. A. Smith, and D. A. Lannigan. Rsk2 allosterically activates estrogen receptor alpha by docking to the hormone-binding domain. *EMBO Journal*, 20:3484–3494, 2001.

[28] H. Claußen, C. Buning, M. Rarey, and T. Lengbauer. FlexE: Efficient molecular docking considering protein structure variations. *J. Mol. Biol.*, 308:377–395, 2001.

[29] P. Cozzini and T. Dottorini. Is it possible docking and scoring new ligands with few experimental data? Preliminary results on estrogen receptor as a case study. *Eur. J. Med. Chem.*, 39:601–609, 2004.

[30] C. Cramer. *Computational Chemistry*. John Wiley & Sons Ltd, 2002.

[31] R. Cramer III, D. Patterson, and J. Bunce. Comparative molecular field analysis (CoMFA). 1. effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.*, 110:5959, 1988.

[32] W. DeLano. The pymol molecular graphics system. http://www.pymol.org, 2002.

[33] R. DeWitte and E. Shakhnovich. Smog: De novo design method based on simple, fast, accurate free energy estimates. 1. methodology supporting evidence. *J. Am. Chem. Soc.*, 118:11733–11744, 1996.

[34] N. Dharmin, V. Ronald, H. Menzo, A. Bout, A. Smitt, and P. Sillevis. Treatment of malignant gliomas with a replicating adenoviral vector expressing herpes simplex virus-thymidine kinase. *Cancer Res*, 61:8743–8750, 2001.

[35] T. Dottorini and P. Cozzini. Probing the binding of ligands to estrogen receptor using an empirical system. *Intl. J. Quant. Chem.*, 106:641–646, 2006.

[36] J. Drews. Drug discovery: a historical perspective. *Science*, 287:1960–1964, 2000.

[37] D. Eisenberg and A. D. McLachlan. Solvation energy in protein folding and binding. *Nature*, 319:199–203, 1986.

[38] R. Engh and R. Huber. Accurate bond and angle parameters for X-ray protein structure refinement. *Acta. Cryst.*, 47:392–400, 1991.

[39] M. Farhat, M. Lavigne, and P. Ramwell. The vascular protective effects of estrogen. *FASEB. J.*, 10:615–624, 1996.

[40] E. Fermi. *Thermodynamics*. Dover Publications, 1937.

[41] A. Finkelstein and O. Ptitsyn. *Protein Physics. A Course of Lectures (Soft Condensed Matter, Complex Fluids and Biomaterials Serie)*. Elsevier Books, 2002.

[42] B. Fischer, S. Basili, H. Merlitz, and W. Wenzel. Accuracy of binding mode prediction with a cascadic stochastic tunneling method. *Proteins*, 68:195–204, 2007.

[43] B. Fischer, K. Fukuzawa, H. Merlitz, and W. Wenzel. Receptor specific scoring functions derived from quantum chemical models improve affinity estimates for in-silico drug discovery. *Proteins*, (in press), 2007.

[44] B. Fischer, H. Merlitz, and W. Wenzel. Increasing diversity in in-silico screening with target flexibility. *In Proceedings, Lect Notes Comput Sc*, 3695:186–197, 2005.

[45] E. Fischer. Einfluss der Konfiguration auf die Wirkung der Enzyme. *Ber. Dtsch. Chem. Ges.*, 27:2985, 1894.

[46] P. Flory. Thermodynamics of high polymer solutions. *J. Chem. Phys.*, 10:51–61, 1942.

[47] V. Fock. Näherungsmethoden zur Lösung des Quantenmechanischen Mehrkörperproblems. *Z. Physik*, 61:126, 1930.

[48] X. Fradera and J. Mestres. Guided docking approaches to structure-based design and screening. *Curr. Topics Med. Chem.*, 4:687–700, 2004.

[49] R. Friesner and V. Gullar. Ab initio quantum chemical and mixed quantum mechanics/molecular mechanics (qm/mm) methods for studying enzymatic catalysis. *Ann. Rev. Phys. Chem.*, 56:389–427, 2004.

[50] R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Reparsky, E. H. Knoll, M. Shelley, J. K. Perry, D. E. Shaw, P. Francis, and P. S. Shenkin. Glide: A new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *J. Med. Chem*, 47:1739–1749, 2004.

[51] K. Fukuzawa, K. Kitaura, M. Uebayasi, K. Nakata, , T. Kaminuma, and T. Nakano. ab-initio quantum mechanical study of the binding energies of human estrogen receptor with its ligands: an application of the fragment molecular orbital method. *J. Comp. Chem.*, 26:1–10, 2005.

[52] A. M. Gao, D. W. Zhang, J. Z. Zhang, and Y. Zhang. An efficient linear scaling method for ab initio calculation of electron density of proteins. *Chem. Phys. Lett.*, 304:293–297, 2004.

[53] J. Gasteiger and M. Marsili. Iterative partial equalization of orbital electronegativity - rapid access to atomic charges. *Tetrahedron*, 36:3219–88, 1980.

[54] B. Gidas. Nonstationary markov chains and the convergence of the annealing algorithm. *J. Stat. Phys.*, 39:73–131, 1985.

[55] H. Gohlke, M. Hendlich, and G. Klebe. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.*, 295:337–356, 2000.

[56] P. Goodford. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.*, 28:849–857, 1985.

[57] A. Gringautz. *Introduction to Medicinal Chemistry. Drugs: How They Act and Why.* VCH, 1996.

[58] H. Gronemeyer and V. Laudet. Transcription factors 3: nuclear receptors. *Protein Profile*, 2:1173–1308, 1995.

[59] T. Grycuk. Deficiency of the coulomb-field approximation in the generalized born model: An improved formula for born radii evaluation. *J. Chem. Phys.*, 119:4817–26, 2003.

[60] J. Häggblad, B. Carlsson, P. Kivelä, and H. Siittari. Scintillating micro-titration plates as platform for determination of estradiol binding constants for hER-HBD. *BioTechniques*, 18:146–151, 1995.

[61] T. Halgren. Merck molecular force field: I. basics, form, scope, parameterization and performance of mmff94. *J.Comput.Chem.*, 17, 1996.

[62] J. Hammersley and D. Handscomb. *Monte Carlo Methods.* Fletcher & Son Ltd., 1964.

[63] M. Hasenbusch. Monte Carlo Simulationen in der statistischen Physik. lecture notes.

[64] R. Head, M. Smythe, T. Oprea, C. Waller, S. Green, and G. Marshall. Validate: A new method for the receptor-based prediction of binding affinities of novel ligands. *J. Am. Chem. Soc.*, 118:3959–3969, 1996.

[65] T. Herges and W. Wenzel. In silico folding of a three helix protein and characterization of its free-energy landscape in an all-atom force field. *PRL*, 94:0181011–4, 2005.

[66] T.-A. Herges. *Entwicklung eines Kraftfeldes zur Strukturvorhersage von Helixproteinen.* PhD thesis, Universität Dortmund, 2003.

[67] D. Hoffmann, B. Kramer, T. Washio, T. Steinmetzer, M. Rarey, and T. Lengauer. Two-stage method for protein-ligand docking. *J. Med. Chem.*, 42:4422–4433, 1999.

[68] W. Hong and M. Sporn. Recent advances in chemoprevention of cancer. *Science*, 278:1073–1077, 1997.

[69] M. Iafrati, R. Karas, M. Aronovitz, S. Kim, T. Sullivan, D. Lubahn, T. O'Donnell, K. Korach, and M. Mendelsohn. Estrogen inhibits the vascular injury response in estrogen receptor α deficient mice. *Nature Med.*, 3:545–548, 1997.

[70] InsightII. Software package, Accelrys Cambridge, UK, 2000.

[71] J. Jackson. *Classical Electrodynamics.* John Wiley & Sons, 1998.

[72] E. Jensen. Steroid hormones, receptors and antagonists. *Ann. NY. Acad. Sci.*, 761:1–17, 1995.

[73] G. Jones, P. Wilett, R. Glen, A. Leach, and R. Taylor. Development and validation of a genetic algorithmfor flexible docking. *J. Mol. Biol.*, 267:727–748, 1997.

[74] G. Jones, P. Willett, and R. C. Glen. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.*, 245:43–53, 1995.

[75] W. Jorgensen and N. McDonald. Development of an all-atom force field for heterocycles. Properties of liquid pyridine and diazenes. *J. Mol. Struct.*, 424:145–155, 1997.

[76] W. L. Jorgensen. The many roles of computation in drug discovery. *Science*, 303:1813–1818, 2004.

[77] W. L. Jorgensen and C. Ravimohan. Monte carlo simulation of differences in free energies of hydration. *J. Chem. Phys.*, 83:3050, 1985.

[78] S. Kearsley, D. Underwood, R. Sheridan, and M. Miller. Flexibases: a way to enhance the use of molecular docking methods. *J. Computer-Aided Mol. Design*, 14:251–63, 1994.

[79] E. Kellenberger, J. Rodrigo, P. Muller, and D. Rognan. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins*, 57:225–242, 2004.

[80] J. Kendrew, G. Bodo, H. Dintzis, R. Parrish, H. Wyckoff, and D. Phillips. Three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature*, 181:662–6, 1958.

[81] S. Kirkpatrick, C. Gelatt, and M. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[82] J. Kirkwood. Theory of solutions of molecules containing widely separated charges with special application to zwitterions. *J. Chem. Phys.*, 2:351–61, 1934.

[83] K. Kitaura, E. Ikeo, T. Asada, T. Nakano, and M. Uebayasi. Fragment molecular orbital method: An approximate computational method. *Chem.Phys.Letters*, 313:701, 1999.

[84] K. Kitaura, T. Sawai, T. Asada, T. Nakano, and M. Uebayasi. Pair interaction molecular orbital method: An approximate computational method for molecular interactions. *Chem.Phys.Letters*, 312:319, 1999.

[85] D. B. Kitchen, H. Decornez, J. R. Furr, and J. Bajorath. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nature Reviews Drug Discovery*, 3:935–949, 2004.

[86] G. Klebe. The use of composite crystal-field environments in molecular recognition and the de novo design of protein ligands. *J. Mol. Biol.*, 237:212–35, 1994.

[87] G. Klebe and H. Gohlke. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew. Chemie (Intl. Ed.)*, 41:2644, 2002.

[88] M. J. Klein. Principle of detailed balance. *Phys. Rev.*, 97:1446–1447, 1954.

[89] R. Knegtel and M. Wagener. Efficacy and selectivity in flexible database docking. *Proteins*, 37:334–345, 1999.

[90] P. Kollman and L. Allen. The theory of the hydrogen bond. *Chem. Rev.*, 72:283–303, 1972.

[91] K. Korach, S. Migliaccio, and V. Davis. *Principles of Pharmacology-Basic Concepts and Clinical Applications.*, chapter Estrogens. Chapman and Hall, 1994.

[92] G. J. M. Kuiper, J. Lemmen, B. C. J. Corton, S. Safe, P. van der Saag, B. van der Burg, and J.-A. Gustafsson. Interaction of Estrogenic Chemicals and Phytoestrogens with Estrogen Receptor beta. *Endocrinology*, 139(10):4252–4263, 1998.

[93] I. Kuntz, J. Blaney, S. Oatley, R. Langridge, and T. Ferrin. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.*, 161:269–288, 1982.

[94] J. Kurie. The biological basis for the use of retinoids in cancer prevention and treatment. *Curr. Opin. Oncol.*, 11:497–502, 1999.

[95] J. E. Ladbury. Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. *Chem. Biol.*, 3:973–980, 1996.

[96] A. Leach. *Molecular Modelling: Principles and Applications*. Pearson Educated Limited, 2001.

[97] B. Lee and F. Richards. The interpretation of protein structures: estimation of static accessibility. *J.Mol.Biol.*, 55:379–400, 1971.

[98] C. Lipinski. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods.*, 44:235–49, 2000.

[99] F. London. Zur Theorie und Systematik der Molekularkräfte. *Zeitschrift für Physik*, 63:245–279, 1930.

[100] R. Lotan. Retinoids in cancer chemoprevention. *FASEB J.*, 10:1031–1039, 1996.

[101] H. Luo and K. Sharp. On the calculation of absolute macromolecular binding free energies. *PNAS*, 99:10399–404, 2002.

[102] F. Lyko and R. Brown. DNA methyltransferase inhibitors and the development of epigenetic cancer therapies. *J. Natl. Cancer Inst.*, 19:1498–506, 2005.

[103] P. D. Lyne. Structure-based virtual screening: an overview. *Drug Discovery Today*, 7:1047–55, 2002.

[104] J. McCammon. Computer-aided molecular design. *Science*, 238:486–491, 1987.

[105] D. McQuarrie. *Statistical mechanics*. University Science Books, U.S., 2000.

[106] H. Merlitz, B. Burghardt, and W. Wenzel. Application of the stochastic tunneling method to high throughput database screening. *Chem. Phys. Lett.*, 370:68–73, 2003.

[107] H. Merlitz, T. Herges, and W. Wenzel. Fluctuation analysis and accuracy of a large-scale in-silico screen. *J. Comput. Chem*, 25:1568–1575, 2004.

[108] H. Merlitz and W. Wenzel. High throughput in-silico screening against flexible protein receptors. *Lecture Notes in Computer Science*, 3045:465–472, 2004.

[109] H. Merlitz and W. Wenzel. Impact of receptor flexibility on in-silico screening performance. *Chem. Phys. Lett*, 390:500, 2004.

[110] N. Metropolis and U. Stanislaw. The monte carlo method. *JASA*, 44:335, 1949.

[111] G. Milne, M. Nicklaus, J. Driscoll, S. Wang, and D. Zaharevitz. National cancer institute drug information system 3d database. *J. Chem. Inf. Comput. Sci.*, 34:1219, 1994.

[112] J. Mitchell, R. Laskowski, A. Alex, and J. Thornton. Bleep - potential of mean force describing protein-ligand interactions: I. generating potential. *J. Comput. Chem.*, 20:1165–1176, 1999.

[113] D. Moras and H. Gronemeyer. The nuclear receptor ligand-binding domain: structure and function. *Curr. Opin. Cell. Biol.*, 10:384–391, 1998.

[114] G. Morris, D. Goodsell, R. Halliday, R. Huey, W. Hart, R. Belew, and A. Olson. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.*, 19:1639–1662, 1998.

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

[116] H. Muirhead and M. Perutz. Structure of haemoglobin. A three-dimensional Fourier synthesis of reduced human haemoglobin at 5.5 å resolution. *Nature*, 199:633–8, 1963.

[117] C. W. Murray, C. A. Baxter, and A. D. Frenkel. The sensitivity of the results of molecular docking to induced fit effects: Application to thrombin, thermolysin and neuraminidase. *J. Comput.-Aided Mol. Design*, 13:547–562, 1999.

[118] J. N. Murrell and A. D. Jenkins. *Properties of Liquids and solutions*. John Wiley & Sons, 1994.

[119] T. Nakano, T. Kaminuma, T. Sato, Y. Akiyama, M. Uebayasi, and K. Kitaura. Fragment molecular orbital method: Application to polypeptides. *Chem.Phys.Letters*, 318:614, 2000.

[120] T. Nakano, T. Kaminuma, T. Sato, K. Fukuzawa, Y. Aikyama, M. Uebayasi, and K. Kitaura. Fragment molecular orbital method: Analytical energy gradients. *Chem.Phys.Letters*, 351:475, 2002.

[121] A. Nicholls and B. Honig. A rapid finite difference algorithm, utilizing successive over-relaxation to solve the Poisson-Boltzmann equation. *J. Comp. Chem.*, 12:435–45, 1991.

[122] E. Nikitina, D. Sulimov, V. Zayets, and N. Zaitseva. Semiempirical calculations of binding enthalpy for protein-ligand complexes. *Int.J.Quantum Chem.*, 97:747–763, 2004.

[123] J. W. M. Nissink, C. Murray, M. Hartshorn, M. L. Verdonk, J. C. Cole, and R. Taylor. A new test set for validating predictions of protein-ligand interaction. *Proteins*, 49:457–471, 2002. the results and the data set is downloadable at www.ccdc.cam.ac.uk.

[124] L. Onsager. Electric moments of molecules in liquids. *J. Am. Chem. Soc.*, 58:1486–93, 1936.

[125] B. Oostenbrink, J. Pitera, M. van Lipzig, J. Meerman, and W. van Gunsteren. Simulations of the estrogen receptor ligand-binding domain: affinity of natural ligands and xenoestrogens. *J. Med. Chem.*, 43:4594–4605, 2000.

[126] F. Osterberg. Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock. *Proteins*, 46:34–40, 2002.

[127] N. Ota, C. Stroupe, J. F. da Silva, S. Shah, M. Mares-Guia, and A. Brunger. Non-Boltzmann thermodynamic integration (nbti) for macromolecular systems: relative free energy of binding of trypsin to benzamidine and benzylamine. *Proteins*, 37:641–653, 1999.

[128] G. Patrick. *An introduction to Medicinal Chemistry*, volume 3. Oxford University Press, 1995.

[129] M. B. Peters, K. Raha, and K. M. Merz Jr. Quantum mechanics structure based drug design. *Curr. Op. Drug. Disc.*, 9:370–379, 2006.

[130] J. Ponder and D. Case. Force fields for protein simulations. *Adv. Prot. Chem.*, 66:27–85, 2003.

[131] V. Pophristic and L. Goodman. Hyperconjugation not steric repulsion leads to the staggered structure of ethane. *Nature*, 411:565–568, 2001.

[132] K. Raha and K. Merz. A quantum mechanics scoring function: Study of zinc-ion mediated ligand binding. *J. Am.Chem.Soc.*, 126:1020–1021, 2004.

[133] K. Raha and K. M. Merz. Large-scale validation of a quantum mechanics based scoring function: Predicting the binding affinity and the binding mode of a diverse set of protein-ligand complexes. *J. Med. Chem.*, 48:4558–4575, 2005.

[134] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe. A fast flexible docking method using an incremental construction algorithm. *J. Mol Biol.*, 261:470–89, 1996.

[135] H. Reiss, H. L. Frisch, and J. L. Lebowitz. Statistical mechanics of rigid spheres. *J. Chem. Phys.*, 31:369–380, 1959.

[136] J.-P. Renaud, N. Rochel, M.Ruff, V. Vivat, P. Chambon, H. Gronemeyer, and D.Moras. Christal structure of the rar-gamma ligand binding domain bound to all-trans retinoic acid. *Nature*, 378:681, 1995.

[137] H. Risken and H.-D. Vollmer. Thermodynamik und Statistik. Vorlesungsmanuskript, 2002.

[138] B. Roux and T. Simonson. Implicit solvent models. *Biophys. Chem.*, 78:1–20, 1999.

[139] M. Scarsi, J. Apostolakis, and A. Caflisch. Continuum electrostatic energies of macromolecules in aqueous solutions. *J. Phys. Chem. A*, 101:8098–8106, 1997.

[140] M. Schaefer and C. Froemmel. A precise analytical method for calculating the electrostatic energy of macromolecules in aqueous solution. *J. Mol. Biol.*, 216:1045–66, 1990.

[141] M. Schapira, B. Raaka, H. Samuels, and R. Abagyan. In silico discovery of novel retinoic acid receptor agonist structures. *BMC Struct. Biol.*, 1:1, 2001.

[142] T. Schlick. *Molecular Modeling and Simulation: An Interdisciplinary Guide.* Springer, 2000.

[143] G. Schneider and H. Boehm. Virtual screening and fast automated docking methods. *Drug Discovery Today*, 7:64–70, 2003.

[144] A. Schug, B. Fischer, A. Verma, H. Merlitz, W. Wenzel, and G. Schön. Biomolecular structure prediction with stochastic optimization methods. *Adv Eng Mater*, 7:1005–1009, 2005.

[145] A. Schug, T. Herges, and W. Wenzel. Reproducible protein folding with the stochastic tunneling method. *Phys. Rev. Letters*, 91:158102–05, 2003.

[146] A. Schug, W. Wenzel, and U. Hansmann. Energy landscape paving simulations of the trp-cage protein. *J. Chem. Phys.*, 122:194711, 2005.

[147] T. Schulz-Gasch and M. Stahl. Scoring functions for protein-ligand interactions: a critical perspective. *Drug Discovery Today: Technologies*, 1:103–110, 2004.

[148] K. Sharp. *Virtual Screening in Drug Discovery.* Taylor & Francis, 2005.

[149] S. Shi, L. Yan, Y. Yang, J. Fisher-Shaulsky, and T. Thacher. An extensible and systematic force field, ESFF, for molecular modeling of organic, inorganic, and organometallic systems. *J. Comput. Chem.*, 24:1059–1076, 2003.

[150] B. Shoichet. Virtual screening of chemical libraries. *Nature*, 432:862, 2004.

[151] B. Shoichet and I. Kuntz. Matching chemistry and shape in molecular docking. *Protein Eng.*, 6(7):723–732, 1993.

[152] C. Silva, P. Almeida, and C. Taft. Density functional and docking studies of retinoids for cancer treatment. *J. Mol. Model.*, 10:38–43, 2004.

[153] W. Sippl. Receptor-based 3D QSAR analysis of estrogen receptor ligands - merging the accuracy of receptor-based alignments with the computational efficiency of ligand-based methods. *J. Comp.-Aided Mol. Design*, 14:559–572(14), 2000.

[154] M. Stahl and M. Rarey. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.*, 44:1035–1042, 2001.

[155] T. Stahl. 75 Jahre Penicillin - ein Grund zum Feiern! *NOVO*, 67, 2003.

[156] J. L. Stebbins, D. Jung, M. Leone, X.-K. Zhang, and M. Pellecchia. A structure-based approach to retinoid x receptor-alpha inhibition. *J. Biol. Chem.*, 281:16643–16648, 2006.

[157] W. Still, A. Tempczyk, R. Hawley, and T. Hendrickson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.*, 112:6127–6129, 1990.

[158] Y. Su and E. Gallicchio. The non-polar solvent potential of mean force for the dimerization of alanine dipeptide: the role of solute-solvent van der waals interactions. *Biophys. Chem.*, 109:251–260, 2004.

[159] C. Tanford and J. Kirkwood. Theory of protein titration curves. I. general equations for impenetrable spheres. *J. Am. Chem. Soc.*, 79:5333–39, 1957.

[160] R. Taylor, P. Jewsbury, and J. Essex. A review of protein-small molecule docking methods. *J. Computer-Aided Mol. Design*, 16:151–166, 2002.

[161] M.-J. Tsai and B. O'Malley. Molecular mechanisms of action of steroid/thyroid receptor superfamily members. *Annu. Rev. Biochem.*, 63:451–486, 1995.

[162] R. Turner, B. Riggs, and T. Spelsberg. Skeletal effects of estrogens. *Endocr. Rev.*, 15:275–300, 1994.

[163] V. Vasilyev and A. Bliznyuk. Application of semiempirical quantum chemical methods as a scoring function in docking. *Theor. Chem. Acc.*, 112:313–317, 2004.

[164] A. Vedani, M. Dobler, and M. A. Lill. Combining protein modeling and 6D-QSAR: Simulating the binding of structurally diverse ligands to the estrogen receptor. *J. Med. Chem.*, 48:3700–3703, 2005.

[165] C. M. Venkatachalam, X. Jiang, T. Oldfield, and M. Waldman. Ligandfit: a novel method for the shape directed rapid docking of ligands to protein active sites. *J. Mol. Graphics Modell.*, 21:289–307, 2003.

[166] G. Verkhivker, K. Appelt, S. Freer, and J. Villafranca. Empirical free energy calculations of ligand-protein crystallographic complexes. i. knowledge-based ligand-protein interaction potentials applied to the prediction of human immunodeficiency virus 1 protease binding affinity. *Protein Eng.*, 8:677–691, 1995.

[167] M. Vieth, J. D. Hirst, A. Kolinski, and C. L. Brooks III. Assessing energy functions for flexible docking. *J. Comp. Chem.*, 19:1612–1622, 1998.

[168] M. Vieth, J. D. Hirst, A. Kolinski, and C. L. Brooks III. Assessing search strategies for flexible docking. *J. Comp. Chem.*, 19:1623–1631, 1998.

[169] H. Vogel. *Gerthsen Physik*. Springer, 1999.

[170] J. Vogt, R. Perozzo, A. Pautsch, A. Prota, P. Schelling, B. Pilger, G. Folkerts, L. Scapozza, and G. Schulz. Nucleoside binding site of herpes simplex type 1 thymidine kinase analyzed by X-ray crystallography. *Proteins*, 42:545–553, 2000.

[171] R. Wade, K. Clark, and P. Goodford. Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 1. Ligand probe groups with the ability to form two hydrogen bonds. *J. Med. Chem.*, 36:140–147, 1993.

[172] R. Wade, K. Clark, and P. Goodford. Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 2. Ligand probe groups with the ability to form more than two hydrogen bonds. *J. Med. Chem.*, 36:148–156, 1993.

[173] W. Walters, M. Stahl, and M. Murcko. Virtual screening - an overview. *Drug Discovery Today*, 3:160–178, 1998.

[174] R. Wang, Y. Lu, and S. Wang. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.*, 46:2287–2303, 2003.

[175] G. L. Warren, C. W. Andrews, A.-M. Capelli, B. Clarke, J. LaLonde, M. H. Lambert, M. Lindvall, N. Nevins, S. F. Semus, S. Senger, G. Tedesco, I. D. Wall, J. M. Woolven, C. E. Peishoff, and M. S. Head. A critical assessment of docking programs and scoring functions. *J. Med. Chem.*, 49:5912–5931, 2006.

[176] F. Weinhold. A new twist on molecular shape. *Nature*, 411:539–541, 2001.

[177] W. Wenzel and K. Hamacher. Stochastic tunneling approach for global minimization of complex potential energy landscapes. *Physical Review Letters*, 82:3003–07, 1999.

[178] P. Werner and A. Caflisch. A sphere-based model for the electrostatics of globular proteins. *J. Am. Chem. Soc.*, 125:4600–08, 2003.

[179] M. Williams and J. Ladbury. *Protein-Ligand Interactions*, chapter Hydrogen Bonds in Protein-Ligand Complexes. WILEY-VCH Verlag GmbH & Co. KGaA, 2003.

[180] P. Wolohan and P. Reichert. CoMFA and docking study of novel estrogen receptor subtype selective ligands. *J. Comp. Aided Mol. Design*, 17:313–328, 2003.

[181] G. Wu, D. H. Robertson, C. Brooks III, and M. Vieth. Detailed analysis of grid-based molecular docking: A case study of cdocker. *J. Comp. Chem.*, 24:1549–1562, 2003.

[182] C. Wurth, U. Kessler, J. Vogt, G. Schulz, G. Folkerts, and L. Scapozza. The effect of substrate binding on the conformation and structural stability of herpes simplex virus type 1 thymidine kinase. *Prot. Sci.*, 10:63–73, 2001.

[183] H. Xu and D. Agrafiotis. Retrospect and prospect of virtual screening in drug discovery. *Curr. Topics Med. Chem.*, 2:1305–20, 2002.

[184] J.-M. Yang and T.-W. Shen. A pharmacophore-based evolutionary approach for screening selective estrogen receptor modulators. *Proteins*, 59:205–220, 2005.

[185] T. Zhou, L. Xu, B. Dey, A. Hessell, D. V. Ryk, S.-H. Xiang, X. Yang, M.-Y. Zhang, M.B., Zwick, J. Arthos, D. Burton, D. Dimitrov, J. Sodroski, R. Wyatt, G. Nabel, and P. D. Kwong. Structural definition of a conserved neutralization epitope on HIV-1 gp120. *Nature*, 445:732–37, 2007.

Acknowledgments

A Ph.D. thesis is the culmination of years of study. Throughout these years, I have received much support from family, teachers, friends. But, I would like to express my sincere appreciation to those people who made this thesis possible and enriched my life in the past few years.

First and foremost I would like to thank my thesis-advisor Priv.-Doz. Dr. Wolfgang Wenzel for the opportunity to conduct my doctoral studies under his supervision. Thank you for your help and support and the guidance into the highly interesting interdisciplinary topic.

I would also like to gratefully acknowledge Ass. Prof. Dr. Holger Merlitz for his kind help and support and his excellent prior work in this project.

Thanks also to Dr. Kaori Fukuzawa and Prof. Dr. Shigeno Tanaka for a successful collaboration and all their help.

In my daily work I have been blessed with congenial and cheerful colleagues who created a friendly and stimulating atmosphere which allowed me to work and to make progress even in difficult times. I am therefore very grateful to Aina Quintilla, Murthy Shrinivasa Gopal, Abhinav Verma, Dr. Konstantin Klenin, Dr. Mathias Hettler, Dr. Alexander Schug, Dr. Daria Kokh, Dr. Horacio Sanchez, Dr. Ali Reza Samanpour and Dr. Isma Tejero Villagrasa. Particular thanks to Birgit Riedel and Markus Armbruster for their encouragement and grim sense of humor in difficult times.

In addition, I would like to thank all my colleagues at the Institute für Nanotechnologie in the Forschungszentrum Karlsruhe for their support and assistance. The Forschungszentrum Karlsruhe provided excellent working conditions and possibilities to participate in interesting conferences Europe wide. Thanks to the DFG and the Forschungszentrum Karlsruhe for funding my doctorate.

I wish to thank my family for their support and encouragement during the time of my thesis. I am very grateful to Adelheid Röben for all her support and the daunting task of proofreading my thesis.