TECHNISCHE UNIVERSITÄT MÜNCHEN
Physik Department
Lehrstuhl für Molekulardynamik T38

Impact of Protein conformational Changes on Molecular Docking – Design of a Docking Approach including Receptor Flexibility

Simon Leis
M.Sc.

Vollständiger Abdruck der von der Fakultät für Physik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. Friedrich Simmel
Prüfer der Dissertation:	1. Univ.-Prof. Dr. Martin Zacharias
	2. Univ.-Prof. Dr. Iris Antes

Die Dissertation wurde am 20.09.2012 bei der Technischen Universität München eingereicht und durch die Fakultät für Physik am 28.11.2012 angenommen.

Contents

Abstract

Due to their pivotal role in various biological processes and in many diseases, proteins are frequent candidates for drug design studies investigating the interactions between proteins and their binding partners. Being very flexible molecules, proteins can undergo considerable conformational changes upon forming a complex with a ligand. The magnitude of such flexibility can vary from small reorientations of single side chains at the binding site to global loop or whole domain movement. Hence, the appropriate representation of protein flexibility in molecular docking methods is of ample importance for the success of drug design projects. However, due to its computational complexity, the efficient and accurate inclusion of global protein flexibility still marks a major challenge in this field.

Here, a newly developed approach that efficiently includes receptor flexibility in grid-based protein-ligand docking is presented. The method combines fast grid-based energy computations with an interpolation scheme allowing for continuous shifting between a set of receptor deformations during docking. This ensemble of structures is used to represent the flexibility of the receptor protein and can be of arbitrary size and source. Several alternatives for the choice of structure input are presented, each containing a variable amount of knowledge on already known bound receptor structures. The method is tested on various pharmaceutically relevant receptor-ligand complexes and yields significantly improved docking results in contrast to the conventional rigid receptor docking. In addition, the flexible receptor docking shows the potential for identifying the deformations within a structure ensemble that show the largest similarity towards the actual bound receptor form. The ability to reproduce near-native ligand binding modes in the great majority of tested cases comes at only small additional computational cost and is thus providing a promising approach to be employed in virtual screening applications.

Zusammenfassung

Proteine sind von essentieller Bedeutung für eine Vielzahl biologischer Prozesse und spielen eine wichtige Rolle bei verschiedenen Krankheiten. Dementprechend groß ist das biophysikalische und medizinische Interesse am genauen Verständnis der Funktionsweise von Proteinen und deren Wechselwirkungen mit Bindungspartnern. Da Proteine flexible Moleküle sind, kann es bei der Bindung mit anderen Proteinen oder Ligandenmolekülen zu den unterschiedlichsten Konformationsänderungen kommen. Diese Änderungen reichen von kleinen Verschiebungen einzelner Seitenketten bis hin zur globalen Neuanordnung ganzer Proteindomänen. Das Ziel von molekularen Docking-Methoden ist die realistische Vorhersage der Bindungsenergie und -affinität von Liganden-Rezeptor-Komplexen. Insbesondere die Einbeziehung von Konformationsänderungen des Proteinrezeptors ist dabei ein noch weitgehend ungelöstes Problem.

In dieser Arbeit wird ein neu entwickelter Ansatz vorgestellt, der es erlaubt, die Flexibilität des Proteinrezeptors auf effiziente Art und Weise in Docking Berechnungen zu berücksichtigen. Die Methode kombiniert dabei schnelle gitterbasierte Energieberechnungen mit einem Interpolationsschema, das einen stufenlosen Übergang zwischen verschiedenen Proteinstrukturen eines Ensembles während des Dockings ermöglicht. Umfangreiche Tests der Methode an verschiedenen medizinisch relevanten Protein-Liganden Komplexen zeigen eine deutliche Verbesserung der Dockingergebnisse im Vergleich zum Docking an rigiden Rezeptoren. Ebenso ermöglicht die Methode die Identifikation der Rezeptorstruktur im Ensemble, die der tatsächlichen gebundenen Proteinstruktur am ähnlichsten ist. Da die verbesserte Vorhersagegenauigkeit mit nur geringer zusätzlich benötigter Rechenzeit einher geht, stellt die Methode eine vielversprechene Möglichkeit für Anwendungen im virtuellen Screening von potentiellen Wirkstoffen dar.

Chapter 1

Introduction

1.1 Motivation and Aim of this Thesis

Proteins – often termed *"machines of life"* – are of paramount importance for a vast number of functions inside living organisms: A variety of proteins, for example, give structure and elasticity to our body tissue and skin. Others form an integral part of our immune system, act as transporters through the cell membranes or in the blood, and many of them regulate crucial processes of the cell cycle.

Advanced knowledge of protein structure, dynamics, and their interaction with other proteins or small molecules can be the determining factor for the success of research efforts in pharmaceutical medicine and drug design related areas. Computational approaches to simulate the dynamics of proteins have made considerable headway in recent years, as well as methods with the goal to reliably estimate and predict protein interactions with binding partners (so-called molecular docking). Amongst other insights, it enables the identification of pharmaceutically active ligand molecules and can predict their binding mode and affinity to receptor targets of medical relevance.

However, one characteristic that most proteins have in common, still poses a significant challenge for molecular docking approaches: Upon interacting with other proteins or a ligand molecule, proteins are flexible and can undergo conformational changes of various extent. These changes can range from only minor movement of individual atoms or amino acids to the complete rearrangement of the protein structure, hence, an accurate incorporation of this factor into algorithms is a demanding task. The additional degrees of freedom, introduced by a flexible protein, cause tremendous computational costs and create a major bottleneck of modeling methods that calls for efficient solutions.

Throughout this thesis, several existing models to incorporate protein flexibility into modeling approaches are introduced, along with their drawbacks and limitations. As a contribution to solving this challenging problem, the main goal of this PhD work was the development of a flexible receptor docking algorithm scheme that accurately considers small and global backbone changes in receptor proteins during docking. At the same time, the method should be computationally efficient, thereby allowing for a possible use of the approach in virtual screening efforts.

1.2 Thesis Outline

This introduction chapter briefly summarizes the basics on proteins and the different levels of protein organization. Furthermore, it outlines experimental techniques to resolve protein structures as well as the theory on interactions between proteins and other molecules that lead to the consequential significance for drug design. The following sections of the thesis are set out as follows:

- Chapter 2 addresses the theoretical background of different computational techniques that were used throughout the thesis. In addition, the characteristics of several proteins that were used as test cases for the developed method are introduced.

- In Chapter 3, the importance of receptor flexibility for protein docking is summarized and the efforts that have been made so far for its inclusion into existing methods are discussed. The findings create the motivation for the development of our flexible receptor docking approach as presented in the following chapter.

- Chapter 4 illustrates the theoretical background and development of the newly developed flexible receptor docking approach, named ReFlexIn (Receptor Flexibility by Interpolation). The method is tested in several case studies where normal mode deformations provide the basis for receptor flexibility and the results are presented and discussed.

- In Chapter 5, several other possibilities to employ our new flexible receptor docking approach are shown. Three structure input methods other than normal mode derived deformations are tested, including the use of several bound protein structures, a morphing approach that employs only the unbound protein and one bound structure, and an approach that uses receptor deformations that are derived from NMR experiments.

- Chapter 6 deals with the scenario where the binding site of proteins is not known prior to docking and reviews computational approaches to predict binding sites on proteins. Several methods are compared and their robustness is evaluated on a test set of various proteins that is deliberately challenging in terms of protein flexibility.
- The final chapter summarizes the findings of this thesis and discusses possible future directions in a concluding outlook.

1.3 Protein Structure

The basic structural component that all proteins have in common are amino acids that are covalently bound to form a linear chain polymer. The standard amino acids contain only 5 elements: hydrogen, carbon, oxygen, and nitrogen in all, and sulphur in some amino acids. All amino acids – except proline – share the basic chemical structure of an amino group (NH_2) and a carboxyl group ($COOH$) (as shown in Figure 1.1) and differ only in their side chain rest (R). For example, (R) is a hydrogen atom in glycine, a methyl group (CH_3) in alanine, CH_2OH in serine, and so on. The carbon atom to which the side chain is attached, is also called C_α-atom.

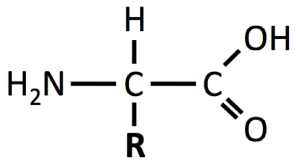


Figure 1.1: *General chemical structure of amino acids containing an amino group (NH_2), a carboxyl group ($COOH$), and a side-chain rest (R).*

Two amino acids are joined by forming the so-called peptide bond where the amino group of one amino acid forms a covalent bond (–CO–NH–) with the carboxyl group of the other amino acid. A condensation reaction leads to the release of a water molecule. Single amino acids are also often referred to as *residues* of a protein.
Corresponding to their properties in water surrounding, the different amino acids can be grouped into polar, non-polar, acidic, and basic amino acids:

- non-polar/hydrophobic: Ala, Val, Met, Leu, Ile, Pro, Trp, Phe
- polar/neutral: Tyr, Thr, Gln, Gly, Ser, Cys, Asn
- basic: Lys, Arg, His
- acidic: Glu, Asp

On the lowest level, one can describe the organization of proteins by their *primary structure*, i.e. the simple sequence of amino acids in a polymer chain (for example "Leu-Ile-Glu-Glu-Pro-Val-His-Ala-..."). The organization of polypeptides in regular structures, such as α-helices or β-sheets (as illustrated in Figure 1.2) is dependent on the specific sequence and is referred to as *secondary structure*. Those patterns are formed and stabilized by hydrogen bonds between amino acids that are only several positions apart from each other within the primary sequence. The two highest protein organization levels are based on interactions between amino acids that are further apart within a single chain (*tertiary structure*, the three-dimensional structure or "fold" of a protein) and the interaction between different protein chains or sub-units that form a multi-subunit complex (*quaternary structure*).
Furthermore, proteins can be divided into three major classes:

1. *Fibrous Proteins*
   Proteins of this class often adopt an elongated rod-like shape and are involved in structural elements outside the cell. Well-known examples for this class of mostly insoluble proteins are e.g. keratins, elastins, and collagens (which also exist in multiple types).

2. *Membrane Proteins*
   The major components of cell membranes are freely movable amphipathic phospholipid molecules and several types of proteins embedded inside the membrane. The latter act as channels for substance flow through the membrane, receptors (recognition and binding of molecules outside the cell), or markers that identify the cell's nature. Due to the hydrophobic nature of the inner membrane region, membrane proteins are only stable within this surrounding and loose their characteristics when being removed for purification or processing. Hence, investigating the structure of this class of proteins is still a major challenge and available structures are rare.

3. *Globular Proteins*
   In contrast to fibrous and membrane proteins, globular proteins are usually soluble in aqueous solution, therefore, more polar residues are found on the protein surface, whereas hydrophobic residues are typically present at buried positions. Even though the term "globular" suggests only compact and spherical shapes, members of this protein class can widely vary with respect to helical/sheet content, elongation, and domain arrangement. Throughout this thesis, only globular proteins are investigated.

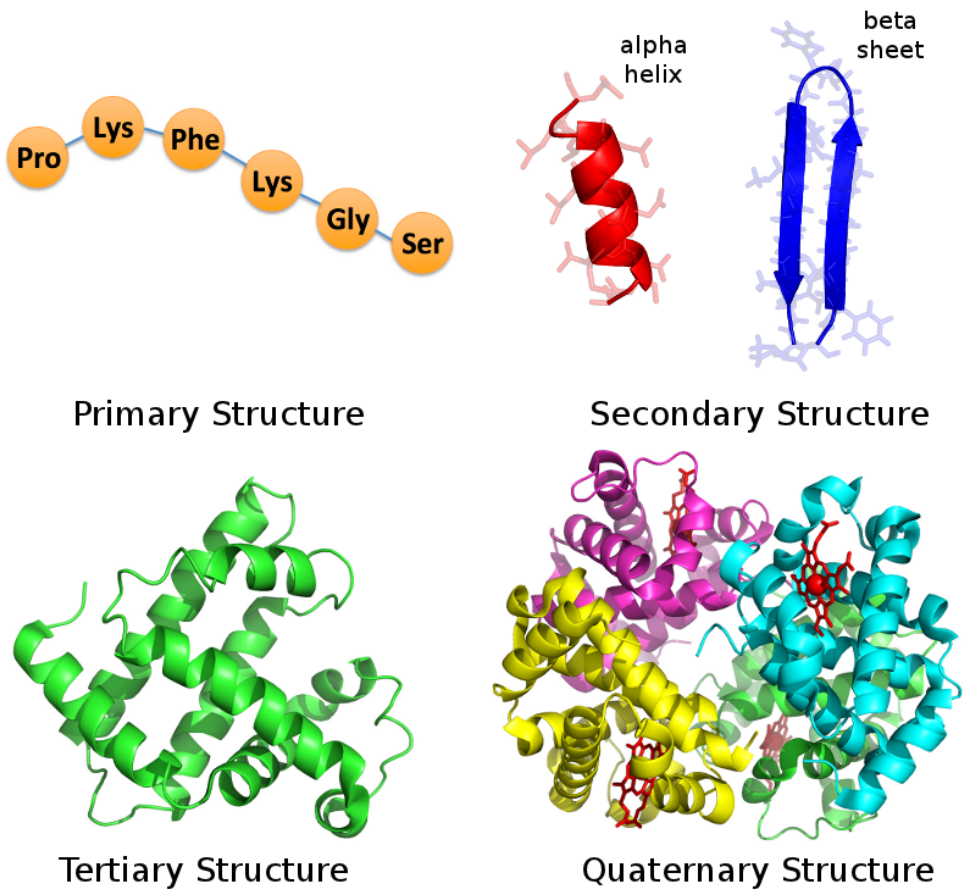


Figure 1.2: *Structural organization of proteins into primary structure (amino acid sequence), secondary structure (motifs like α-helix or β-sheet), tertiary structure (the fold of a chain), and quaternary structure (multi-subunit complex made of several combined chains) - Examples for the tertiary structure is the muscle protein myoglobin (PDB ID: 1MBN) and for the quaternary structure bovine hemoglobin (PDB ID: 1G09) that is present in red blood cells and consists of 4 subunits: 2 α-subunits (green and violet) and 2 β-subunits (yellow and blue), each of them containing a heme group (red sticks with iron atom shown as sphere).*

The number of different proteins within the human proteome (that is the collection of all proteins that can possibly be expressed in the human organism based on the approximately 20,000 genes of the human genome) are estimated to be in the ten-thousands. However, only a marginal fraction of proteins has yet been structurally resolved, meaning their atomistic structure has been determined by experimental research.

Knowledge of the structure of proteins can serve a variety of insight on the function of proteins and their role in biological processes. The availability of these resolved structures is crucial especially for the fields surrounding molecular modeling and computational chemistry. Since the first protein structure was solved by Kendrew in 1958 [1], experiments have been constantly improved in terms of accuracy and efficiency.

1.4 Structure Derivation and Collection

There are two main experimental methods that are commonly used for the structure determination of proteins and other biomolecules at atomic resolution: X-ray crystallography and Nuclear Magnetic Resonance (NMR) spectroscopy. Both methods have different possibilities and limitations.

Instead of using visible light as in typical light microscopy experiments, crystallography employs X-rays with a wavelength of about 1Å. This short wavelength – in contrast to the 5000Å of wavelength for visible light that a standard light microscope would use – allows for the investigation of structures at the atomic level.
The production of the actual protein crystal is the most crucial and demanding task. The crystal is required to be of large purity, order, and size (at least 0.1-0.2 mm in at least 2 dimensions) to produce a significant diffraction at a good resolution. Membrane proteins that are embedded inside the lipid cell membrane in their native state pose an additional challenge [2, 3]. X-ray beams are scattered by the crystal into patterns that can be collected with a CCD (charge-coupled device)-based detector. Several algorithms are applied to the collected data for further integration, merging, and scaling before an electron density map is created. Finally, in a structure refinement step, an atomic model is fitted into the electron density map and the structure is ready for further analysis. Even though the proteins are present in a crystallized (i.e. non-flexible) form and often cooled down to 100-110K to prevent radiation damage of the sample, X-ray crystallography results contain a certain amount of information on protein flexibility. The so-called B-factor is a term that describes the spreading of electron densities for each atom of the derived molecule.

In most models, the B-factor is given by: $B = 8\pi^2 U_i^2$ where U_i^2 is the mean-square displacement of the atom i [4]. The higher the value for the B-factor (given in $Å^2$), the larger the positional deviation of this atom, whereas very low B-factors indicate that the respective atom is found at the same position for each molecule of the crystal.

NMR spectroscopy is used when either the crystal production of a certain protein is infeasible, or the experiment is supposed to gain insight on stability or dynamic processes of a protein [5]. Similar to X-ray crystallography, the purity of the initial protein sample is crucial for a successful NMR experiment. Protein NMR methods take advantage of the fact that the nuclei of atoms align with a strong magnetic field if applied. The alignment of the nuclei magnetic momentum can be changed from their equilibrium orientation by applying radio wave energy of certain frequencies. Nuclear magnetic resonance (NMR) is the term for the back and forth between the unstable and the equilibrium state. This resonance can be detected and atom types and geometries can be deduced by analysing the NMR spectrum (plotting the resonance versus the applied radio wavelengths).
Further details on X-ray crystallography and protein NMR methods can be found in the reviews [6, 7] and the connected literature.

Substantial improvement of experimental methods and the urge for further knowledge on protein structures have fuelled the growth rate of solved structures in the last decades. The growing number of solved structures is very beneficial for biochemistry, drug design, and neighboured disciplines but a central library to hold and provide all existing biomolecular coordinates is crucial for efficient research.
The Protein Data Bank (short PDB, [8]) serves as a central collection point for resolved structures of biomolecules. It is freely accessible under `http://www.pdb.org/`. The files contain information about the included molecules, their experimental origin and technical details, authors of the experiment, protein sequence, and most importantly, the coordinates and atom types of all atoms included in the structure.

Over 8,100 new structures in 2011 only and the actual total of over 80,000 structures demonstrates the spectacular growth in recent years.
Most structures are solved with a resolution between 1.5 and 3Å and the majority are proteins (92.6% of all PDB structures), while the rest consists of nucleic acids and complexes of proteins and nucleic acids. The main experimental methods for structure solving in the PDB are X-ray crystallography(87.6%) and NMR spectroscopy (11,6%).

With the help of structure solving techniques, a thorough understanding of the structure of proteins and other biomolecules has been formed. However, most experimentally solved structures (especially X-ray based), usually fail to thoroughly represent protein flexibility and dynamics and can only depict a single snapshot at a certain moment and state.

1.5 Protein Docking

The increasing amount of solved structures available in recent years has also fuelled the application possibilities and development of docking approaches. Molecular docking is a versatile tool for predicting the formation of stable protein-ligand complexes that are involved in many biological processes. The core question that is solved computationally is: "How does molecule A interact with molecule B?" given the atomic coordinates of molecules A and B. While one of the molecules typically is a protein, several kinds of docking exist, depending on the type of molecule B, for example protein-protein docking, protein-ligand docking, protein-DNA docking, protein-RNA docking, and so fourth. In addition to the basic question of how the complexes A and B arrange geometrically, docking also tries to compute and predict the specific interactions between A and B as well as a quantitative estimate of their strength. Here, several different interactions are taken into consideration: charge-charge interactions, van-der-Waals interactions, desolvation effects, and hydrogen bonds.

The question of whether the binding site of a protein is known before docking is a decisive factor for the complexity of the calculations. If the binding site is known, the search space can be significantly reduced by considering only the area around the known site of interaction. In cases where the binding site of a target protein (or other molecule) is not yet known, the docking can theoretically be applied to the whole system, which, however, massively increases the complexity. Hence, a binding site prediction might be a necessary step, prior to docking. An overview of existing methods for the prediction of binding sites on proteins and their theoretical background as well as an evaluation on different target proteins is given in Chapter 6.

In addition to the six degrees of freedom that have to be sampled when two rigid bodies are docked against each other, the fact that ligands as well as proteins can undergo certain conformational changes, adds a tremendous amount of complexity to the problem and increases the search space for algorithms by several dimensions.

Because ligand molecules are typically much smaller than the receptor proteins and hence, create less complexity when flexible, especially the flexibility of proteins poses a significant problem to docking approaches. The resulting inaccuracies and limitations are further dissected in Chapter 3.

1.6 Protein Flexibility

With the contemporary sophisticated techniques to get insight on the dynamics of proteins, it is well-known that they can undergo a wide range of small to large conformational changes. Achievement of understanding this, however, required several decades of research. More than a century ago, in 1894, Emil Fischer postulated the famous 'lock-and-key principle' [9]. He compared enzymes (which are mostly proteins) to a rigid lock and their substrate to a specific key that fits into this lock. If the shape of the key is not correct, it is not able to open the lock, i.e. the enzyme will not fulfil its function if the substrate does not fit to the geometry of the active site.
Nevertheless, this visionary principle had to be refined soon after science provided evidence that proteins are far from being rigid molecules. Several enzymes have shown to be overly specific but other enzymes turned out to be able to recognize several ligands of diverse structural geometries. Among other things, this led to Koshland's *induced fit* concept where he stated in 1958 that the accommodation of substrates can lead to considerable changes in or around the active site of enzymes [10].

Figure 1.3 gives several examples of such conformational changes and illustrates well the different spatial extent of displacements within the molecules. Chapter 3 of this thesis depicts more examples of protein flexibility and highlights the great importance of including the conformational changes in proteins when considering them in any molecular modelling context.

It is desirable to include possible conformational changes of both the ligand and the receptor structure for a reliable prediction of ligand-receptor binding geometries. However, depending on the number of added degrees of freedom, the computational demand can increase dramatically during ligand-receptor docking. Hence, it is important to analyse in detail which types of conformational changes can occur in proteins upon ligand binding. In many cases, only minor conformational changes have been found comparing crystal structures of unbound and bound protein conformations [11]. However, in general, conformational changes can range from local side-chain adjustments to global motions of entire subunits, as indicated in Figure 1.3.

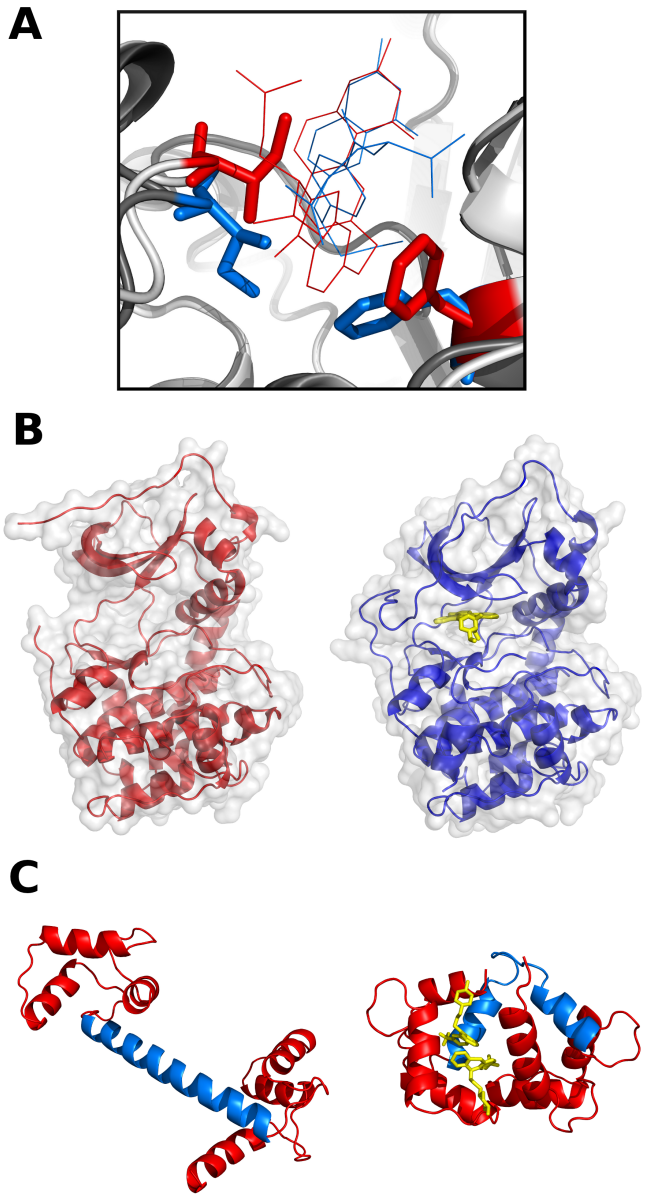


Figure 1.3: *Three examples of protein flexibility with increasing magnitude.*
A: Side-chain readjustments upon binding of different ligands illustrated by aligning two ligand-bound forms of thymidylate synthetase (red: PDB ID 1TSD, blue: 2TSC). Active site side chains Ile79 (left) and Phe176 (right) are shown as stick models. For clarity, the respective ligands are shown as thin lines only.
B: Shown on the left is an unbound state of the Protein Kinase A (PKA, PDB ID 1J3H) in red cartoon and transparent grey surface representation. The widely opened binding site can be clearly seen whereas the right structure illustrates well the closure of the active site upon binding of a small inhibitor molecule (staurosporine, shown as yellow sticks, PDB ID: 1STC).
C: Unbound calmodulin (left, PDB ID: 3CLN) contains a long helix (shown in blue) that extensively rearranges upon ligand binding as shown on the right (PDB ID: 1A29) with two bound trifluoperazine molecules represented by yellow sticks.

Gerstein et al. hierarchically classified motions according to their size, into fragment, domain, and subunit motions [12]. The motion of fragments smaller than domains commonly refers to the motion of surface loops, but also motion of secondary structure elements. The amplitude of the motion is proportional to its characteristic time-scale, and as a consequence, domain motions are in many cases still out of scope for detailed computer simulation methods such as molecular dynamics (MD) simulations [13].

Domain and fragment motions often involve portions of the protein closing around a binding site (also shown in Figure 1.3 B). Hence, they make up for specific mechanisms of induced-fit. Both hinge and shear motions can contribute to up to 70% of the overall motions in both the fragment and the domain class [12, 14]. A more simple way of classification is to discriminate between side-chain and backbone motions. In the light of the aforementioned variety of putative independent small scale and concerted large scale motions, one might expect huge conformational changes on different levels at least for those proteins exhibiting large fluctuations in the absence of a binding partner.

Luque and Freire found that binding sites can feature regions of high flexibility (often most flexible regions of the entire protein) but also regions of high rigidity [15]. Regarding the atomic displacements involved during conformational changes, however, Najmanovich et al. concluded after an extensive database analysis of protein-ligand complexes that in 85% of cases only three or less residues actually change their conformation upon binding [16]. It was also found that certain amino acids exhibit significantly more flexibility than others, e.g. lysine is on average more flexible than phenylalanine. Besides, 94% of $\chi 1$ (first side-chain dihedral angle) and 96% of $\chi 2$ dihedral angles remain in their native conformation upon complex formation. Additionally, the authors observed that only in 12% of the complexes a backbone displacement of more than 2Å takes place upon binding whereas in 75% of the complexes a backbone motion of less than 1Å was found. Consequently, the authors concluded that compared to side-chain flexibility, backbone flexibility is of minor importance. However, it has also been observed that already receptor backbone conformational changes in the range of 1Å can significantly affect receptor-ligand interaction [17]. In this regard, one could as well conclude that in more than 25% of all cases observed, backbone motions might play a decisive role in complex formation. Considering that the set of available experimental structures today is most probably biased towards less flexible proteins (since proteins with highly flexible parts are often harder to crystallize), the number of cases where conformational changes of the receptor is of importance might even be higher.

With different spatial extent of geometrical rearrangement, protein dynamics and motions also happen on very different timescales that can reach from a few picoseconds (1 picosecond = 10^{-12} seconds) to several seconds to minutes [18, 19].
The fastest motions are bond vibrations or stretching and oscillations of atoms, which can occur the picosecond timescale or even below. The rotation of single protein side chains that, for example, allows a ligand to access a binding site on a receptor, typically happens within tens to hundreds of picoseconds. The nanosecond (1 nanosecond = 10^{-9} seconds) regime is covered as soon as protein motions of a larger amplitude happen. The motion of loops or collective motions within a protein (e. g. so-called hinge/shear movements) can take several nanoseconds. Longer timescales include the binding of ligands (several nanoseconds up to micro seconds) and the folding of proteins. Here, also the size of the observed system is important: peptides and small proteins can fold within tens to hundreds of nanoseconds, whereas the folding of larger proteins can take microseconds to seconds.

Due to the restricted information on flexibility that one can get from experiments, research on how to computationally predict and simulate protein flexibility has made tremendous progress in the last decades. Molecular Dynamics (MD) Simulations, for example, give insight into dynamical processes of proteins at the picosecond to multi-nanosecond timescale and are thereby a valuable tool complementary to structural data obtained from X-ray, NMR, or atomic force microscopy experiments [20, 21]. Here, several approximations allow simulating the dynamical trajectory of an arbitrary molecular system, e.g. a protein in solution or with different ligands. However, depending on the size of the investigated system, the calculation of such trajectories is computationally costly which makes simulations of long time-scale events as folding or the binding of drugs infeasible to simulate (if a lab does not own specifically built supercomputers [22]). Also molecular docking approaches struggle with the increased amount of computational complexity that is introduced by protein flexibility.
A thorough discussion of recent approaches to deal with this problem is given in Chapter 3. It is followed by my contribution to break this bottleneck in Chapters 4 and 5, which show the theoretical background and the results of my approach to efficiently model protein flexibility in docking.

Chapter 2

Computational Methods and Materials

2.1 Grid-based Docking with AutoDock

AutoDock is one of the most cited docking programs (together with programs such as FlexX [23], ICM [24], GLIDE [25], or GOLD [26]) and is primarily designed for protein-ligand docking. The AutoDock algorithm was first published in 1990 and initially created by D.S. Goodsell and A.J. Olson at the U.S. Scripps Research Institute [27]. The program has been constantly maintained and updated ever since [28–30]. Currently (2012), AutoDock is available in the Version 4.2 which was used throughout the projects of this thesis. The program suite is written in C and C++, and the full source code is available under the GNU General Public License. The fact that the program can be freely used in academia and industry as well as the availability of the source code is the main motivation for choosing AutoDock for the extensions that were applied as described further below.

Pre-calculation of energy grids in docking has a substantial speed advantage compared to algorithms that sequentially move and evaluate ligand positions. In grid-based approaches, the possible binding energies for each possible ligand atom are pre-computed, hence, when a certain ligand conformation is evaluated, a simple lookup method can be employed. During this lookup phase, a tri-linear interpolation is used to assign the ligand atom interactions to the surrounding receptor atoms for an assumed ligand configuration.

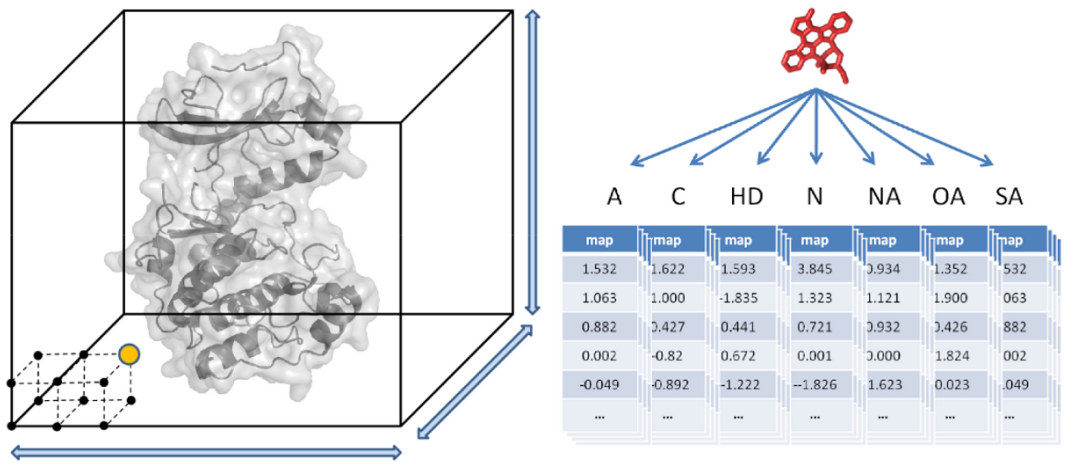


Figure 2.1: *Regular 3D grid as employed by AutoGrid for the calculation of atomic, electrostatic, and desolvation energy grids. The grid has been drawn around the whole protein for clarification, note, that due to computational costs, one usually choses a smaller grid only around the supposed binding site.*

In the AutoDock docking suite, the sub-program *AutoGrid* is responsible for the calculation of the atomic affinity grids as depicted in Figure 2.1. A regular 3D-grid is placed on to either the whole protein receptor or a certain area of the protein where the binding site is known or suspected to be. This grid has a default resolution of 0.375Å, i.e. all grid points have a distance of 0.374Å from their closest neighbour grid points. For each point within the grid, the interaction between all surrounding receptor atoms (within a non-bonded cutoff radius of 8Å) and a probe atom is computed. Therefore, the algorithm returns n separate map files where n is the number of atom types present in the ligand. In addition to the atomic affinity maps, AutoGrid calculates two more map files, namely an electrostatic and a desolvation potential map.

AutoDock Scoring Function

The Autodock 4 force field is parametrized to use Gasteiger partial charges on atoms and distinguishes between polar and non-polar hydrogen atoms in molecules. Hydrogen atoms bound to carbon atoms are non-polar, whereas hydrogen atoms that are bound to electronegative atoms as oxygen or nitrogen are polar. Thus, after adding hydrogens to receptor and ligand molecules and assigning Gasteiger charges, the non-polar hydrogen atoms are removed from the structure and their charges are merged into the charge of the respective bound carbon atom. This united-atom model treats the carbons and the

attached hydrogen atoms as one interaction unit. Hence, in all illustrations of docking results within this thesis, only the polar hydrogens are depicted.
Autodock's force field was parametrized using experimentally derived structures and binding constants from a large amount of protein-ligand complexes [30]. The free energy of binding ΔG is defined as follows:

$$\begin{aligned}\Delta G = &(V_{bound}^{lig-lig} - V_{unbound}^{lig-lig}) \\ &+ (V_{bound}^{prot-prot} - V_{unbound}^{prot-prot}) \\ &+ (V_{bound}^{prot-lig} - V_{unbound}^{prot-lig} + \Delta S_{conf}).\end{aligned} \tag{2.1}$$

ΔS is an estimated contribution due entropy loss upon ligand binding. Each of the six pair-wise energy evaluations consists of the following terms as shown in equation 2.2; a van-der-Waals term, a hydrogen bonding term, an electrostatic term, a desolvation term, and a torsional entropy contribution:

$$\begin{aligned}V = &W_{VDW} \times \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) \\ &+ W_{H-bonds} \times \sum_{i,j} E(t) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) \\ &+ W_{Elec} \times \sum_{i,j} \frac{q_i q_j}{\epsilon(r_{ij}) r_{ij}} \\ &+ W_{Desolv} \times \sum_{i,j} (S_i V_j + S_j V_i) e^{(-r_{ij}^2/2\sigma^2)} \\ &+ W_{Tor} \times N_{tor}.\end{aligned} \tag{2.2}$$

Special weight constants W have been derived by experimental knowledge and are used for an optimal adjustment of the free energy. The van-der-Waals term uses the classic 12-6 Lennard-Jones potential to describe the attraction/repulsion between two non-bonded atoms i (of the receptor) and j (of the ligand) that are at a distance of r_{ij} from each other. The hydrogen bond term includes the function $E(t)$ that accounts for the directionality of a hydrogen bond. The term is favoured if the angle t between the hydrogen acceptor atom, the polar hydrogen, and the acceptor atom is closer to the ideal of 180°. A Coulombic potential gives the electrostatic contribution that depends on the charges q of atom i and j. The calculations for this term are without a distance cutoff and are based on the partial charges assigned as already mentioned above. The desolvation energy term is calculated only for the protein and includes a solvation term S and an atomic fragmental volume V for atoms i in the ligand and j in the receptor.

The gaussian distance constant σ is 3.5Å. Finally, torsional entropy of the ligand due to binding to the receptor is estimated as the product of a weight constant and the number of flexible torsions N_{tor} within the ligand. This rather rudimentary representation of the internal ligand energy adds approximately 0.3 kcal/mol per flexible torsion angle to each binding free energy result which will also be discussed later in this thesis.

Ligand and Receptor Flexibility

Like various other docking programs, AutoDock offers full ligand flexibility. The user can define all, individual, or no torsion angles within the ligand molecule as flexible (freely rotatable) bonds during the docking experiment. However, the number of rotatable bonds within a ligand is not limitless. If larger ligands (with more than 6-8 rotatable bonds) are treated as fully flexible, docking results become significantly harder to cluster and show sampling problems. This will be shown and discussed in Chapter 4. Due to the genetic algorithm's random assignment of various torsion angles, the search space drastically increases for a large amount of flexible torsions.

In contrast to ligand flexibility, the flexibility of the protein receptor is only rudimentary implemented in AutoDock. One or several side chains can be set by the user to be treated flexibly, i.e. the bonds of the selected side chains are allowed to rotate during the docking. However, the large computational effort and resulting sampling problems (equal to a large amount of flexible ligand torsions) allows only for a very limited selection of flexible protein side chains – global flexibility within the receptor is still not satisfactorily modelled. Our approach, as a first step for a better inclusion of receptor flexibility in grid-based protein ligand docking, is presented in Chapter 4.

2.2 Genetic Algorithms

So-called *genetic algorithms* are a combination of mathematical models with well-known processes and semantic terms from genetic biology and its evolution theory. Here, the 'randomness' of nature is used to exhaustively sample a search space within a given problem – i.e. the placement of a ligand during molecular docking.
Most of the termini are borrowed from genetics, such as *generations*, *mutation*, *crossover*, or *selection*. Those concepts and how they are employed in molecular docking will be briefly explained in the following.

As illustrated in Figure 2.2, every docking calculation starts with the so-called initial population. Here, a set of randomly chosen ligand placements is created in which a single member of the set is called *individual*. The size of the population can be set by the user, but the default value is 150. Those initial starting ligand configurations are placed within the 3-dimensional box around the supposed binding site as defined before.

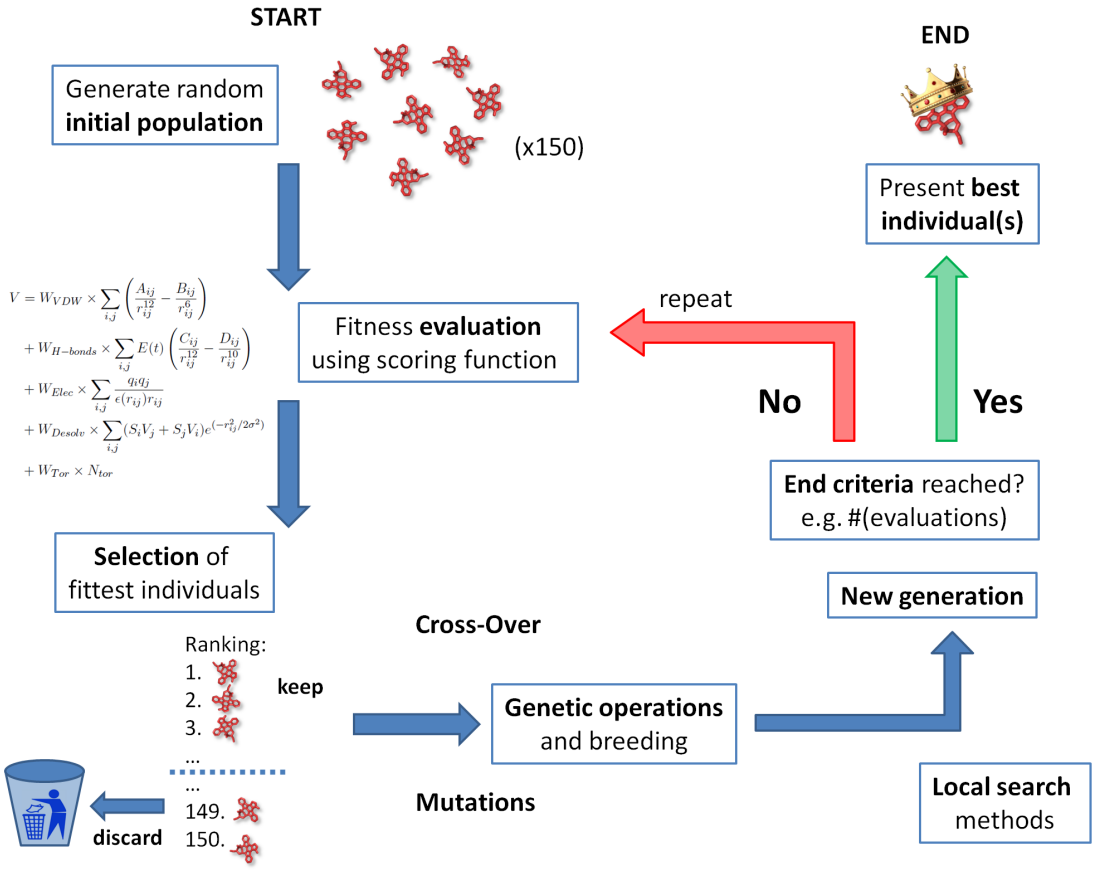


Figure 2.2: *Flow diagram of a genetic algorithm (starting at the upper left) as it is used in the AutoDock protein-ligand docking suite.*

Each ligand placement at a certain position around the receptor molecule can be exactly described by using several variables or *genes*:

- translation (the displacement of the ligand in x/y/z-direction)
  Those are 3 genes: $\mathrm{x}, \mathrm{y}, \mathrm{z}$
- rotation (position of the ligand towards the receptor)
  Those are 4 genes: a unit vector $\vec{Q} = \begin{pmatrix} q_x \\ q_y \\ q_z \end{pmatrix}$ and the rotation angle q_w
- ligand torsions (rotation of ligand bonds that are specified to be flexible)
  Depending on the number of flexible ligand torsions, angle $\theta \in [-180, 180]$ degrees

Altogether, these variables define the *genotype* of the respective ligand conformation. Every genotype (that is basically a collection of vectors or numbers) can be translated into a *phenotype*, which is the actual ligand placement within the search space. These phenotypes are then evaluated by AutoDock's energy score as described above.
According to the calculated energy, a *selection* step is applied to the population of individuals where phenotypes with a better interaction score are kept and phenotypes that yield a score below a certain threshold are discarded. Just as in nature, the fittest individuals are able to pass on their genes to the next generation whereas the weaker individuals successively become extinct.
From the fittest survivors of that generation, successors are created by employing the concepts of gene *cross-over* and *mutation*. In cross-over events, genes between randomly picked individuals are swapped against each other. In addition, Cauchy distribution based mutations are carried out where randomly picked genes are changed within a certain allowed margin. Mutations play a crucial role within the algorithm as they can allow for leaving local energy minima and therefore find better solutions. The mutation rates, step sizes, and probabilities can be adjusted prior to the docking run. In addition, *elitism* can be enabled, by which the specified number of fittest individuals proceed to the next generation unchanged. The steps of evaluation/cross-over/mutation/selection are subsequently repeated until the number of evaluations performed has reached the threshold specified by the user.

With the *Lamarckian Genetic Algorithm (LGA)*, AutoDock combines a genetic algorithm as described above with a local search method. Here, a certain subset of each population undergoes a local search where small translational and (if activated) torsional changes are applied to those individuals, are evaluated, and finally kept if interaction with the receptor has improved. Hence, the LGA perfectly combines effective scanning of the whole conformational space (genetic algorithm) with further optimization of found solutions (local search).

2.3 Structure Deviation Measurement

The RMSD – short for **R**oot **M**ean **S**quare **D**eviation – is frequently used in structural studies to determine how similar two biomolcule structures are (e.g. proteins, small ligand molecules). One has to differentiate between a) using the RMSD to measure the similarity of two structures as, for example, used in this study to illustrate the flexibility induced changes of two protein structures (referred to as $\text{RMSD}_{Protein}$ throughout this thesis) and b) to determine the quality of a ligand placement after docking into a protein binding site (RMSD_{Ligand}). In the latter case, the RMSD is employed as a measurement of how close the docked solution is to a given "real" ligand placement (docking accuracy).

RMSD values have been calculated using Pymol [31] and are given in all-atom RMSD if not indicated otherwise (another common measurement is C-alpha or C_α RMSD where only the C_α atoms of the protein chain are considered for the calculations).
The RMSD between two structures X and Y consisting of N atoms each is calculated as the square root of the average of the squared distances between the corresponding atoms of structure X and structure Y:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \mid x_i - y_i \mid^2}. \tag{2.3}$$

Typical values for RMSD_{Ligand} that are considered as in good agreement (e.g. with experimental ligand configurations) are RMSDs of $\leq$ 2Å. This boundary was also used throughout this thesis to distinguish between "good" and "poor" ligand placements after docking.

2.4 Normal Mode Analysis and Elastic Network Models

One drawback of the calculation of flexible degrees of freedom for a protein receptor molecule is the high computational demand of calculating either harmonic modes at atomic resolution or principal components of MD trajectories. Additionally, the calculation of principal components of motion from an MD simulation can significantly depend on the simulation length and convergence [32, 33].

A different computational approach has emerged in recent years that enables this limitation to be avoided: Elastic Network Models (ENM) of proteins assume that the lowest frequency normal modes (also called soft modes) represent the biologically relevant large-scale movements within a protein [34, 35].
The calculations are based on simplified spring models of proteins and on the hypothesis that the mobility of a protein region is determined by the local density (or the locally available free space) [36, 37].
Harmonic mode analysis of this simple energy function enables identification of possible flexible (soft) collective degrees of freedom of the protein within a few minutes computer time [35]. Tama and Sanejouand found that such approximate mode calculations resulted in predicted soft modes that show significant overlap with observed conformational changes in proteins determined experimentally under different conditions (e.g. different crystal forms or apo vs. bound form of a protein molecule) [38]. In some cases, over 50% of the conformational difference between two structures of a protein, determined for example by X-ray analysis of two crystal forms or as ligand-free and bound forms, could be assigned to a single approximate soft mode of the protein [38].

Figure 2.3 demonstrates the capability of normal mode analysis to account for the conformational changes in the enzyme protein kinase A. In this case, just ten softest modes obtained from a ENM analysis applied to the apo form of the enzyme are sufficient to approximate the backbone conformational changes observed between apo and an inhibitor bound form to >50%.

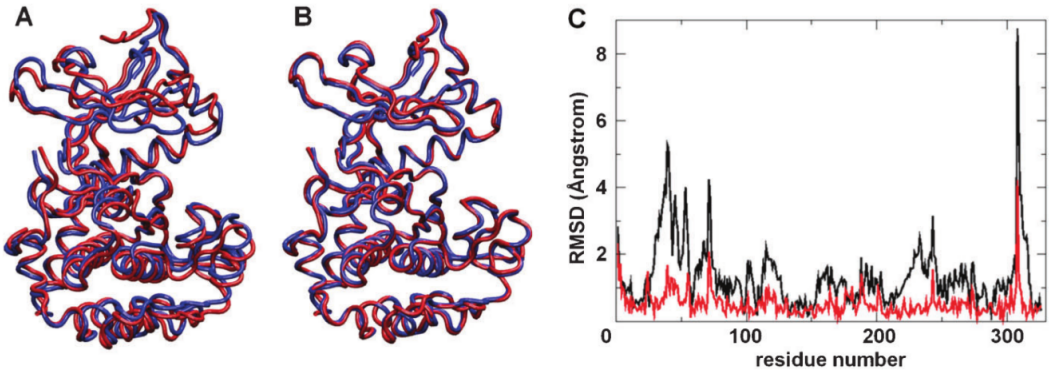


Figure 2.3: *A: Conformational difference between bound (PDB ID 1STC; blue tube representation; complex with staurosporine, ligand not shown) and unbound (PDB ID 1J3H; red tube) structures of the catalytic domain of protein kinase A (PKA). The backbone RMSD between the two structures is 1.7Å .*
B: Superposition of the bound PKA and the apo structure deformed in the 10 softest normal modes obtained from an elastic network calculation on the apo form to give the smallest possible deviation from the bound form.
C: Backbone RMSD of apo and bound PKA (black line) vs. residue number compared to the backbone RMSD for the apo structure optimally deformed in the 10 softest normal modes (red line). The RMSD between the optimally deformed structure and the bound PKA structure is 0.9Å .

Throughout this thesis, normal mode calculation was performed using an elastic network model (ENM) for proteins developed by Hinsen [35]. Here, only the C_α atoms of the proteins are considered for the calculations, hence predicting rather global backbone changes than side chain movements. The normal mode analysis, as employed here, uses a simplified force field which assumes that the protein input structure represents a local energy minimum. This is presumably true for X-ray or NMR derived structures, differently derived input should thus be subjected to an energy minimization beforehand.

The forcefield is based on a pair-wise distance-dependent energy function:

$$V(R_1, ..., R_N) = \sum_{C_{\alpha-pairs}} V_{ij}(R_i - R_j) \tag{2.4}$$

with the pair-wise term

$$V_{ij} = k(R_{ij}^{(0)})(|r| - R_{ij}^{(0)})^2 \tag{2.5}$$

where $R_{ij}^{(0)}$ is the pair's equilibrium (input) distance. The force constant k is distance-dependent:

$$k(r) = \gamma \times \exp(-\frac{|r|^2}{r_0^2}) \tag{2.6}$$

such that short distances are more restrained than long distances. γ is a constant that controls the overall flexibility and an r_0 of 4Å was found to give the best overlap between the softest non-trivial modes and the conformational difference between apo and holo structures in test calculations [39].

The harmonic modes with respect to the above energy function (2.4) can be obtained as follows: The second derivative matrix **H** of the potential energy is calculated. It is of size $3N \times 3N$ where N is the number of (C_α) atoms within the system.
The mass-weighted Hessian Matrix **H*** is given by:

$$\mathbf{H^*} = \mathbf{M}^{-\frac{1}{2}} \times \mathbf{H} \times \mathbf{M}^{-\frac{1}{2}} \tag{2.7}$$

where M is a diagonal $3N \times 3N$ matrix with the atomic masses as the only non-zero elements. The normal modes are the eigenvectors of **H*** and the respective eigenvalues λ_i are the squares of the vibrational frequencies:

$$\nu_i = \frac{\sqrt{\lambda_i}}{2\pi} \tag{2.8}$$

Figure 2.4 illustrates the first softest mode for the unbound form of CDK2 (PDB ID: 1HCL) that was used as one of the test systems for the flexible receptor docking approach.

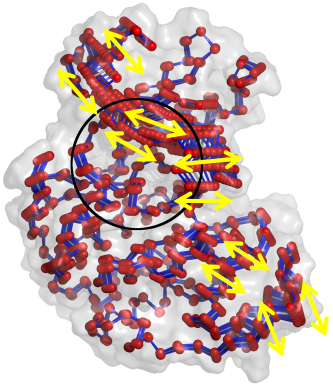


Figure 2.4: *First normal mode deformations of apo cyclin-dependent kinase 2 (CDK2). The C_α atoms are shown as red spheres and are connected by blue bonds. The first mode yields 6 deformations of the apo structure, three into each direction. The ligand binding site is encircled black. Yellow arrows indicate the largest motions within the system.*

In the present thesis, normal modes are employed exclusively to model global low-frequency backbone movements. Further examples for employing of ENM calculations in modeling efforts are discussed in Chapter 3.

2.5 Test Systems

Throughout this thesis, several protein systems have been used to apply and evaluate the introduced approaches. These examples are real-world drug design targets and are of key interest for the research on threatening diseases, as e.g. different protein kinases related to cancer, or proteases that play key roles in the course of infections with the Human Immunodeficiency Virus (HIV). The following paragraphs briefly summarize the basic biological theory and important structural features of the investigated proteins.

Protein Kinases

With over 500 members, the class of protein kinases represents a prominent element of the human proteome [40]. Due to their role for the signal transduction in the cell and the diseases resulting from defective enzymes, protein kinases have become promising drug targets for inhibitor based cancer therapy [41]. The historic timeline of protein kinases as drug targets and the most successful developments have been well reviewed. [42, 43].

The basic function of protein kinases is adding a phosphate group (PO_4^{3-})to their substrate protein, a process called phosphorylation. Here, the phosphate of an ATP molecule is transferred to the free hydroxyl group of an either serine, threonine, or tyrosine residue. Different classes of kinases (serine kinases, tyrosine kinases, and so on) are able to phosphorylate different combinations of those residues. The consequential chemical alteration leads to conformational changes that can ultimately trigger or suppress functions in the substrate protein.

Protein Kinase A (PKA) is one of the most prominent members of the large group of protein kinases. PKA was the first solved kinase structure [44] and is today one of the best studied examples of kinases. PKA is also often referred to as cAMP-dependent kinase because the amount of cyclic adenosine monophosphate, a second messenger molecule, around PKA significantly regulates the activity of PKA.
Like other protein kinases, the main structural features of PKA are two regulatory (R) and two catalytic (C) subunits. The R subunits hold two (or four, depending on the type of eukaryotic cell) binding sites for cAMP, as soon as those binding sites are occupied by cAMP molecules, the two C subunits are released from the complex. On release, the C subunits undergo substantial conformational changes and are now able to phosphorylate available substrate proteins. The structure of the unbound PKA as well as a structure with the inhibitor staurosporine bound at the active site are illustrated in the chapter before (Figure 1.3 B).

Cyclin-Dependent Kinase 2 (CDK2) represents another important member of the kinase family. In general, cyclin-dependent kinases can be involved in the mediation of the cell cycle progression (e.g. CDK1, CDK2, CDK4, CDK6) or in the regulation of the transcription (e.g. CDK7 ord CDK9). The name is derived from its native binding partners, regulatory enzymes called cyclins that play a crucial role in controlling the cell cycle by activating the CDK upon binding and phosphorylation. The complex role of CDKs in the regulation of the cell cycle has been thoroughly reviewed [45, 46] as well as the CDK's role in cancer and the anti-cancer potential of CDK inhibitors [47, 48].

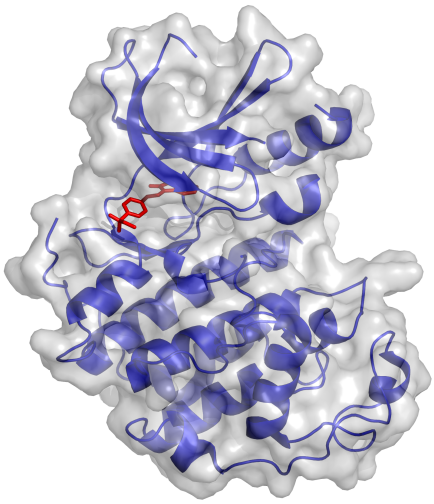


Figure 2.5: *Cartoon representation of the CDK2 tertiary structure (blue) with gray transparent molecular surface (PDB ID 1KE5). The inhibitor molecule (LS1, shown as red sticks) fits inside the cleft between the two lobes of CDK2: one smaller amino-terminal lobe on top of the ligand that contains beta-sheets and the PSTAIRE helix as well as the larger carboxy-terminal lobe below the ligand containing mainly alpha-helices.*

HIV-1 Protease (HIV-1P)

HIV-1 Protease – a member of the aspartyl protease class – plays a crucial role in the HIV replication cycle and is therefore an interesting and well-studied target.
Its specific role is the cleaving of newly synthesized polypeptides into functional proteins that are needed for the HIV virion to become active. If it is possible to find drug molecules that mimic a polypeptide chain and bind tightly to the protease active site (more tightly than its natural polyproteins), the protein's function is blocked, thus preventing the HI-virus from maturing and becoming harmful [49].

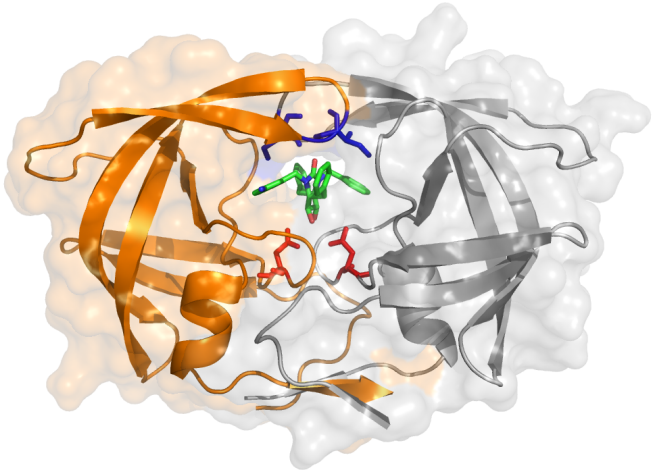


Figure 2.6: *Cartoon representation of the HIV1-Protease with transparent molecular surface (PDB ID 1DMP). The two homodimer chains are shown in orange and gray. The bound ligand (DMQ) is shown in stick representation. Also in stick representation are the catalytic aspartic acids Asp25 in red and the two isoleucine residues Ile50 in blue.*

The first experimental structure of HIV1-Protease was solved in 1989 [50] and since then, several hundred 3D-structures of bound and unbound states have been deposited in the PDB. With the large amount of available structures and the remarkable conformational changes upon ligand binding (see below), the protease is a well-suited test system for our flexible receptor docking approach. The structure of HIV1-Protease in complex with the cyclic urea inhibitor DMQ (PDB ID: 1DMP) is illustrated in Figure 2.6. In total, the homo-dimer is built of 198 residues with each subunit consisting of 99 amino acids. Possible inhibitors bind at the active site that is located in the central part of the protein. The tunnel-like binding site is covered by two β-sheet flaps that have shown to be opened in the unliganded HIV1-Protease and take a closed conformation upon inhibitor binding [51, 52]. In Chapter 5, the large conformational change in the flaps is illustrated further. At the binding site, the two aspartic acid residues Asp25 are highly conserved and mutation studies on those residues have proven their importance for the proteolytic function [53–55]. Under physiological pH conditions, one of the two Asp25 residues exist in the deprotonated form and can therefore act as a hydrogen bond acceptor to an inhibitor.

Chapter 3

Accounting for Target Flexibility during Protein–Ligand Docking

3.1 Introduction

An essential goal of computational drug design is to reduce expenses within the designing process, either by the selection of putative lead-structures from databases of drug-like chemical compounds, or by de-novo design of active substances [56]. Target-based drug design uses available three-dimensional (3D) structural information of the receptor to dock compound libraries (target based virtual screening) or de-novo designed ligands that specifically bind to the target. In recent years, the number of 3D structures of protein molecules has increased significantly. Additionally, for a growing number of protein target sequences with sequence similarity to a known protein structure (template), comparative modelling methods allow for building fairly accurate model structures (depending on the degree of target-template sequence similarity). The rapid growth of structural knowledge on biomolecular drug target molecules forms the basis for the increasing applicability of structure-based ligand-receptor docking and drug design applications. The ultimate goal of structure based docking and drug design is the identification of putative ligands, the prediction of the binding geometry, and the prediction of the binding affinity to a given receptor molecule.

The affinity and specificity of binding reactions is determined by the structural and physicochemical properties of the binding partners and the solution environment. The basis for this specificity was first investigated by Emil Fischer already in 1894 [9]. He addressed the foundations of specific binding by introducing the well-known lock-and-key analogy to describe enzyme-substrate interactions. The basic idea of this concept

is that the substrate has to fit specifically like a key into the binding site (lock) of an enzyme. The lock and key concept was developed further by Koshland, who proposed that a global conformational change of the enzyme hexokinase is necessary to adapt to its substrate [10]. He developed the idea of 'induced fit' recognition, meaning that both partners can structurally differ in their unbound (apo) and bound (holo) conformations. During the association process, the interacting molecules induce conformational changes that are required to achieve high affinity and specificity of binding. It has also been recognized, that in principle all possible molecular binding processes require a certain degree of conformational adaptation.

Binding reactions can be accompanied by a variety of conformational changes. The magnitude of such changes can range from alterations of side-chain conformations [16] at the binding site to global changes of domain arrangements [57–59] and can even involve refolding of protein segments upon association [60]. Based on ideas from statistical physics, the induced-fit concept has been extended suggesting a pre-existing ensemble of several inter-convertible conformational states that are in equilibrium [61]. These states include conformations close to unbound but also near-bound forms. Binding of a partner molecule to the near-bound form stabilizes this structure and shifts the equilibrium towards the bound form [62]. Computational approaches to realistically model and predict ligand-receptor binding geometries should preferably include such conformational changes during receptor-ligand docking simulations. It is the focus of the present chapter to give an overview on available computational strategies to include receptor conformational changes during docking methods and to discuss their strengths and weaknesses.

While considerable progress has been achieved in modelling the conformational flexibility of ligands during docking over the last decade, inclusion of receptor flexibility is still an unsolved problem, especially in case of significant changes in the protein backbone structure upon binding. In the following, recent efforts and progress on including receptor flexibility during docking are presented. Special emphasis is put on global and semi-global receptor conformational changes in protein-ligand docking calculations and on aspects of computational efficiency. Various levels of flexibility have been considered in docking approaches. A schematic illustration of the different levels of protein flexibility that can play a role during association is given in Figure 3.1. Since the great majority of drug targets are proteins, the focus will be on the flexibility of proteins. However, most aspects of conformational flexibility are of general importance also for other types of receptor biomolecules.

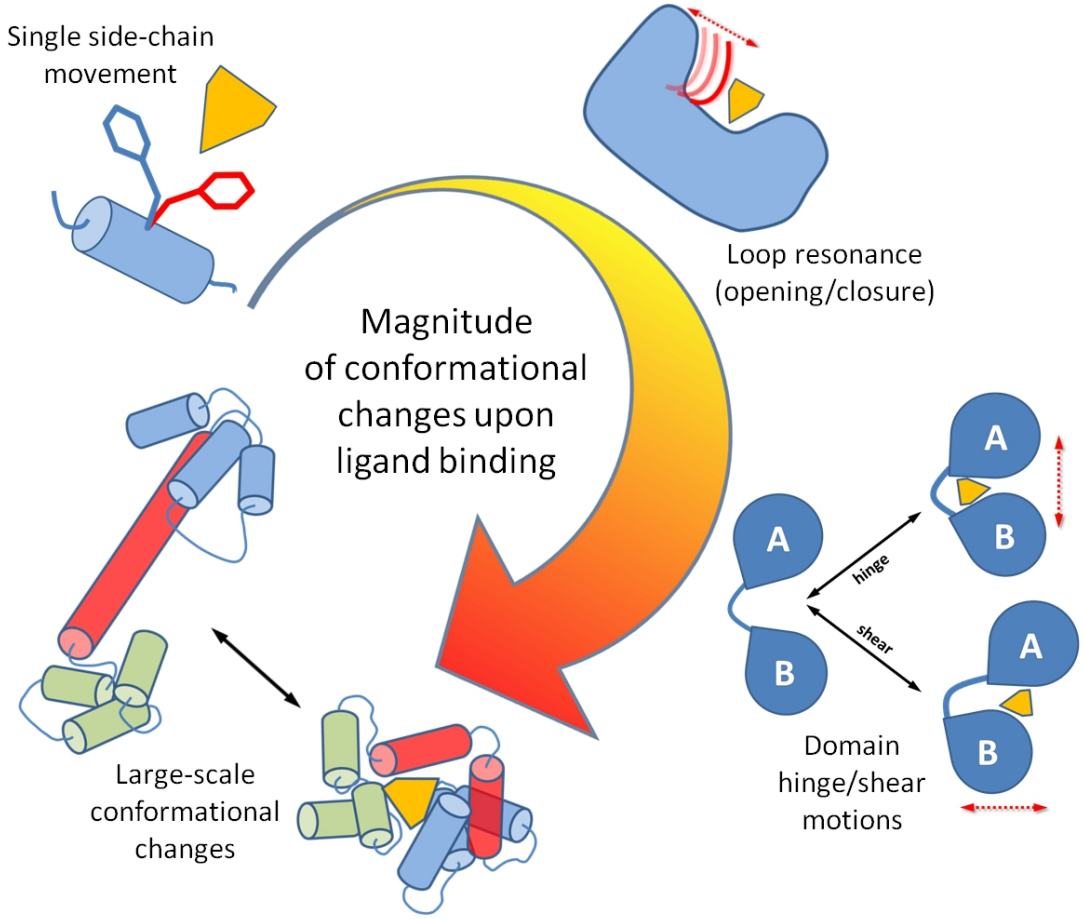


Figure 3.1: *Schematic overview of different conformational changes occurring in a receptor protein upon ligand binding with increasing magnitude following the large arrow from the upper to the lower left. The place and direction of conformational change is illustrated in red, potential ligand molecules are shown as orange triangular structures, protein domains are labelled with letters A and B.*

3.2 Thermodynamic Driving Forces of Binding and the Scoring Problem

Thermodynamics tells us that the driving force for binding of two molecules is given by the free energy change associated with the binding reaction. A free energy change indicates that there are entropic and energetic (enthalpic) components involved in binding reactions [63–66]. On the energetic part, the association process can involve changes in electrostatic interactions, van der Waals interactions and hydrogen bonding. These types of interactions can include contributions within and between the interacting biomolecules but in addition also between the biomolecules and surrounding solvent

molecules (usually water and possibly ions). Besides of internal energy or enthalpy, the binding process can also involve changes in the entropy of the biomolecules and the surrounding solvent molecules [67]. For example, binding of biomolecules can result in a change in flexibility of the binding partners affecting the conformational entropy of both molecules. Association can also lead to a change in the ordering of solvent molecules (resulting in changes of enthalpy and entropy) around the binding partners that influences the association process (hydrophobic effect).

There are two important aspects of computational methods to predict receptor-ligand complexes. Firstly, it is desirable to predict an arrangement as close to reality as possible and, secondly, to score it appropriately with respect to alternative complex geometries. Since several enthalpic and entropic contributions influence the binding affinity, the scoring function needs to realistically account for many contributions to binding affinity. At the same time, docking approaches need to be fast enough to allow for rapid docking and evaluation of many thousand putative complexes. Computational speed is a critical issue for ligand-receptor docking which requires a reasonable compromise between accuracy and speed to calculate a score for a ligand-receptor complex. Current scoring functions to evaluate ligand-receptor complexes range from simple schemes that just account for sterical complementarity to complete force field functions [68–70]. These can include terms that account for sterical and electrostatic interactions but also account for desolvation and hydrophobic contributions to ligand-receptor interaction [67]. The change in conformational entropy of binding partners also influences the binding affinity. However, so far only few approaches have tackled the issue of including conformational entropy effects during docking scoring [71, 72].
Instead of scoring functions based on a molecular mechanics force field, it is also possible to design knowledge-based potentials to evaluate complexes that are extracted from known receptor-ligand complexes. The basic idea of such knowledge-based scoring is to relate the observed frequency of atom-atom (or group-group) contacts to the expected contact frequency at receptor-ligand interfaces to extract favourable and unfavourable atom-atom interactions [73]. Here, expected contact frequency means contacts that are obtained if atoms would be distributed randomly at the interface. The design of an appropriate scoring function for the realistic evaluation of ligand-receptor complexes is still an unresolved issue. The problem of designing a realistic scoring function for computational docking is not at the focus of the current chapter. However, its relation to the problem of how to account efficiently for conformational changes is of interest. Frequently, the difficulties of realistic docking (in terms of minimal deviation from an experimental complex structure) and scoring are considered as separate issues.

However, since scoring functions are often exclusively parameterized using experimental structures of protein-ligand complexes, it is important to generate docking models already in the sampling phase that come as close as possible to the native complex structure. Appropriate treatment of conformational flexibility during docking is, therefore, tightly connected to the improvement of realistic scoring of docked complexes. The design of more rigorous and more specific scoring functions requires at the same time an improvement of the prediction accuracy of binding modes in terms of deviation from the experimental binding geometry. Highly accurate scoring of a ligand placement requires that the complex geometry has also been predicted with high precision. Vice versa, the more errors a scoring function may tolerate, the less specific it becomes. Consequently, there is a direct relation between the robustness or softness of a scoring function and the number of false positives obtained in a virtual screen.

3.3 Accounting for conformational Flexibility in Docking

Inclusion of Ligand Flexibility

Basically all popular docking approaches include ligand flexibility, which is reasonable, because even small changes in dihedral torsion angles around rotatable bonds can significantly change the shape and electrostatic potential around a ligand molecule. Many methods have been developed and applied including soft docking [74, 75], ligand conformational ensembles [76], molecular dynamics [77], or Monte Carlo simulations [27], and various smart advanced sampling strategies [78]. Several current efficient docking programs employ 'build up' or incremental construction approaches [79, 80]. The incremental construction scheme involves splitting the ligand into fragments that are assumed to be conformationally rigid. Subsequently, a starting fragment (anchor) is docked into the target site and the complete ligand is generated by attaching fragments to a growing ligand in the binding pocket. The connection between fragments allows for conformational adjustment of the ligand to optimally fit into the binding site. The efficient inclusion of ligand flexibility during docking has already been extensively reviewed [81].

Accounting implicitly for Receptor Flexibility

As discussed in the previous paragraphs, the analysis of available protein structures indicates in several cases only small differences of receptor structures in complex with a ligand (bound/holo form) compared to the apo (unbound) form. However, even in such cases, docking of ligands to rigid receptor structures may not be successful due

to the sensitivity of sterical interactions with respect to even small conformational adjustments in a receptor binding pocket. The application of a soft interaction scoring function represents one possibility to still keep a rigid receptor structure during docking but allowing for some sterical overlap between receptor and ligand. Various functional forms of soft scoring functions have been suggested that allow for various degrees of sterical overlap between ligand and receptor atoms [74, 75, 82, 83]. Intuitively, such an approach is reasonable since even in a lock-and-key mechanism of ligand-receptor binding, small atom displacements in the receptor pocket might be possible (e.g. due to thermal fluctuations). The thermal fluctuations can effectively reduce sterical overlap between ligand and receptor atoms which can be approximately described by a softening of the sterical interactions. A common method to implicitly account for limited conformational changes in the receptor structure is a broadening or shift of the repulsive part of the van-der-Waals interaction potential (soft-core potential). Thus, slightly increased overlap of atoms upon docking is permitted. One should keep in mind, however, that atomic motions which may lead to a reduction of atomic overlap are usually strongly coupled motions with defined direction meaning that a motion of one atom to reduce overlap also affects the position and interaction of neighboring atoms. In contrast, simple softening of interactions 'reduces' possible overlap independent of neighboring atoms. Nevertheless, soft scoring functions are widely used during docking searches [84]. However, one should keep in mind that uniform softening of sterical interactions can greatly reduce the specificity of interactions resulting in false-positive docking solutions or unrealistic placement of the ligand in the binding pocket [85].

Ligand-Receptor Docking using Molecular Dynamics Simulations

In principle, it is possible to perform ligand-receptor docking searches allowing for conformational changes of both the receptor and ligand structure in all Cartesian or bond rotation degrees of freedom. This can be achieved by using energy minimization (EM), molecular dynamics (MD) or Monte Carlo (MC) simulations (or related simulation methods). Such simulations are typically based on a molecular mechanics force field describing all intra- and inter-molecular interactions of receptor and ligand molecules [69, 86]. In MD simulations, Newton's equations of motion are solved numerically in small time steps (1-2 fs; 1fs = 10^{-15} s) allowing in principle for full ligand and receptor flexibility [87]. In case of docking a ligand into a receptor pocket, one typically starts the simulation from various start placements of the ligand near the expected binding site. For computational efficiency, the flexible parts of the protein are frequently restricted to the vicinity of the binding site and the rest of the protein is kept rigid [88]. However, due to the presence of energy barriers, MD simulations

can be trapped in unrealistic docking sub-states and may require long and computationally demanding simulations to reach a realistic complex structure (conformational sampling problem). Di Nola et al. partly solved this problem by applying a different coupling scheme for the heat bath. In their method, the center of mass of the ligand moves at a considerably higher temperature than the remaining system [77]. Mangoni et al. expanded the approach by allowing for internal motions of the receptor which was coupled to a low-temperature bath [89]. Alternatively, sampling of ligand placement and receptor conformational states can be improved by simulated annealing methods [90] or scaling/rescaling methods of the ligand-receptor interaction potential during MD simulations [91, 92]. Such techniques allow a more rapid convergence to an optimal interface structure with simultaneous adjustment of both side-chain and backbone interface structure. More recently, advanced MD sampling methods have been successfully combined with docking searches that show promising results on test systems [78, 93, 94].

The calculation of the ligand binding affinity to a known receptor binding site and the evaluation of ligand modifications on binding affinity is another application area of molecular dynamics based approaches. A prominent example is the MM/PBSA (Molecular Mechanics/Poisson Boltzmann, Surface Area) method or the MM/GBSA method. The latter method employs the Generalized Born approach to calculate electrostatic interactions instead of the more time consuming PB approach [95]. In both cases, an ensemble of conformations of the receptor-ligand complex generated by MD simulations is evaluated based on a continuum model (Poisson Boltzmann or Generalized Born model) for the surrounding solvent. The application to the complex as well as to the isolated receptor and ligand molecules results in an estimate of the ligand-receptor interaction. Although computationally much more demanding than using standard docking scoring functions, the approach has been employed also in virtual compound screening efforts [96]. Other even more demanding methods are based on thermodynamic perturbation or thermodynamic integration where a ligand or parts of a ligand are created or annihilated during MD simulation (reviewed in [97, 98]). Although mostly applied for the evaluation of a selected number of ligands or modifications of ligands [92] the development of more efficient free energy simulation methods and increasing computer performance may allow for a growing use of such approaches to evaluate ligand-receptor binding affinity [99].

A major drawback of MD simulations applied to ligand-receptor docking is the large computational demand due to the many flexible degrees of freedom of both solute and solvent molecules. The steady increase of computer power and the implementation of smart sampling methods will undoubtedly broaden the applicability of MD simulation methods for docking searches. However, even in case of implicit solvent models and a known receptor binding site, the computational effort can be too large to systematically dock multiple ligands. Therefore, MD methods are currently still more applied to lead structure optimization or refinement of a small set of pre-selected docking poses. Indeed, multistep docking approaches containing an MD-based refinement step are becoming increasingly popular [93, 97].

Treatment of Side-Chain Flexibility and local Protein Backbone Changes

In case of any systematic exploration of ligands and possible binding placements, it is desirable to restrict the flexibility to fewer degrees of freedom that correspond to the most important variables for the binding adjustment. The variation of bond lengths and angle vibrations makes up only for a comparably small contribution to conformational changes. The motions that contribute most are movements of dihedrals (bond rotation). Often, conformational changes upon ligand-protein association are limited to changes in the side-chains that form the binding site. In such cases, the inclusion of side-chain reorientations during docking can result in drastic improvements of the docking performance [100]. It is well known that side-chains possess conformational preferences for a discrete number of dihedral states. Dunbrack and Karplus [101] and others [102–107] compiled the most common values for side-chain dihedral angle states from a large database of protein structures. Initially, these libraries were used for assigning side-chain conformations to a given backbone structure in comparative modelling. The first investigation of the applicability of side-chain rotamer states from a library during docking was undertaken by Leach in 1994 [108]. There exist backbone dependent and independent libraries, small libraries that cover only the most prominent states, and exhaustive libraries providing residence probabilities for each state [109]. Several available docking programs can include side-chain conformational changes at a proposed ligand binding site either at the level of searching for optimal discrete side-chain dihedral angle combinations (rotamers) or upon minimization of side-chain dihedrals during docking. The dead end elimination method and the A* algorithm were tested that allow for the selection of a best possible combination of possible side-chain rotamers in the spatial vicinity of binding cavities [110, 111]. Alternative approaches based on a multi-greedy strategy or a branch and cut algorithm have also been used [112]. Several two-stage docking methods allow for a relaxation

of side-chain conformations allowing for continuous dihedral changes either by energy minimization or Monte Carlo at a final stage of docking for a limited set of selected complexes [113]. Wang et al. proposed a possible refinement of the rotamer method called Rotamer Trials and Minimization (RTMIN) for protein-protein docking [114]. This method allows for sampling of different rotameric states coupled with a subsequent continuous side-chain adjustment and was applied extensively to protein-ligand docking [115]. Side-chain optimization during docking can also be achieved by energy minimization or employing Monte Carlo methods allowing continuous adjustment of side-chain dihedral angles and local adjustment of the protein main chain [24,116–118]. A combination of loop structure prediction and docking was described by Sherman et al. employing several loop copies, which were generated in a pre-sampling run [118]. During docking, a particular copy which exhibits the most favourable interactions with the ligand was selected.

3.4 Conformational Ensemble Methods

To efficiently account for larger conformational changes of the main chain and side-chains near the binding site, it is also possible to represent the binding site by an ensemble of protein conformations [119, 120]. For many proteins of biological or pharmaceutical interest, experimental crystal structures in the apo form and often also in complex with different ligands are available. In such cases, the ensemble of target conformations can be formed by the available experimental structures. Alternatively, computational methods such as MD simulations can be used to generate conformational ensembles [97]. Others generated ensembles using a combination of loop fragments [121]. Distance geometry approaches, as for example implemented in tCONCOORD [122], can also be used to obtain conformational ensembles. The structures are generated by fulfilling a set of upper and lower inter-atomic distances, where the difference between these upper and lower boundaries depends on the estimated interaction strength and sterical hindrance. The resulting structures are usually analyzed by principal component analysis (PCA) to identify a maximally diverse collection of possible conformations. A set of methods, including MD simulations with different solvents and tCONCOORD, has been compared and applied to identify flexible ligand binding pockets on the surface of proteins [123]. The CONCOORD method has also been used to generate conformations in reasonable agreement with bound structures based on data extracted from unbound structure combined with the radius of gyration of the bound ligand-receptor complex structure [124]. In the CONCOORD/PBSA approach, the generated ensembles are successfully used to efficiently sample protein flexibility when

predicting free energy and protein stability changes upon protein-protein binding [125]. Several groups made use of principle components of motion [126] or normal modes to deform a starting receptor structure and thus generate an ensemble. Mustard and Ritchie for example employed distance constraint essential dynamics eigenstructures as input for their protein-protein docking program Hex [127]. Cavasotto et al. used relevant normal modes to generate an ensemble for the cAMP dependent kinase [128]. The authors also introduced a measure of relevance to select a set of mid-frequency modes to be able to cover localized backbone motions. Irrespective of the way an ensemble is obtained, the question arises whether the structural snapshots sufficiently cover the protein's conformational space as a whole. It is not clear how many different structures are necessary. Similar to rotamer library and loop copy approaches, there is one nontrivial problem: If there is no 'correct' or near-correct conformation included in the set of structures, there is no guarantee for improvement over single receptor docking, the prediction accuracy might even drop below the level of docking using a single rigid receptor.

3.5 Structural Ensembles in Docking Calculations

Once an ensemble has been obtained (see Figure 3.2 for different possibilities), ligands can for example be docked successively against every receptor structure. For instance Pang and Kozikowski [129] used 69 snapshots from a short MD-simulation in a docking study; Barril and Morley employed a large set of experimentally derived X-ray structures for successive docking [130]. Moreno and Leon proposed to generate a binding site descriptor from a set of protein conformations rather than from a single structure for input in the program DOCK [131]. Similarly, the relaxed complex scheme by McCammon and coworkers utilizes a set of MD snapshots for docking searches in combination with the AutoDock program as docking engine [132]. This concept was extended by Carlson et al. who proposed the 'dynamic pharmacophore model' [62, 133], where a pharmacophore can be modelled for each structure in the ensemble. In turn, the intersection of all pharmacophore models is called the dynamic pharmacophore, which can then be utilized to identify putative new ligands. The authors tested their approach on HIV-integrase and on HIV-protease. For multiple protein structures from an MD simulation, it was possible to distinguish true inhibitors from drug-like non-inhibitors [36]. However, docking to each individual structure of an ensemble can become highly computationally demanding and requires evaluation of a large set of alternative structures which may increase the chance of obtaining many false positive docking solutions.

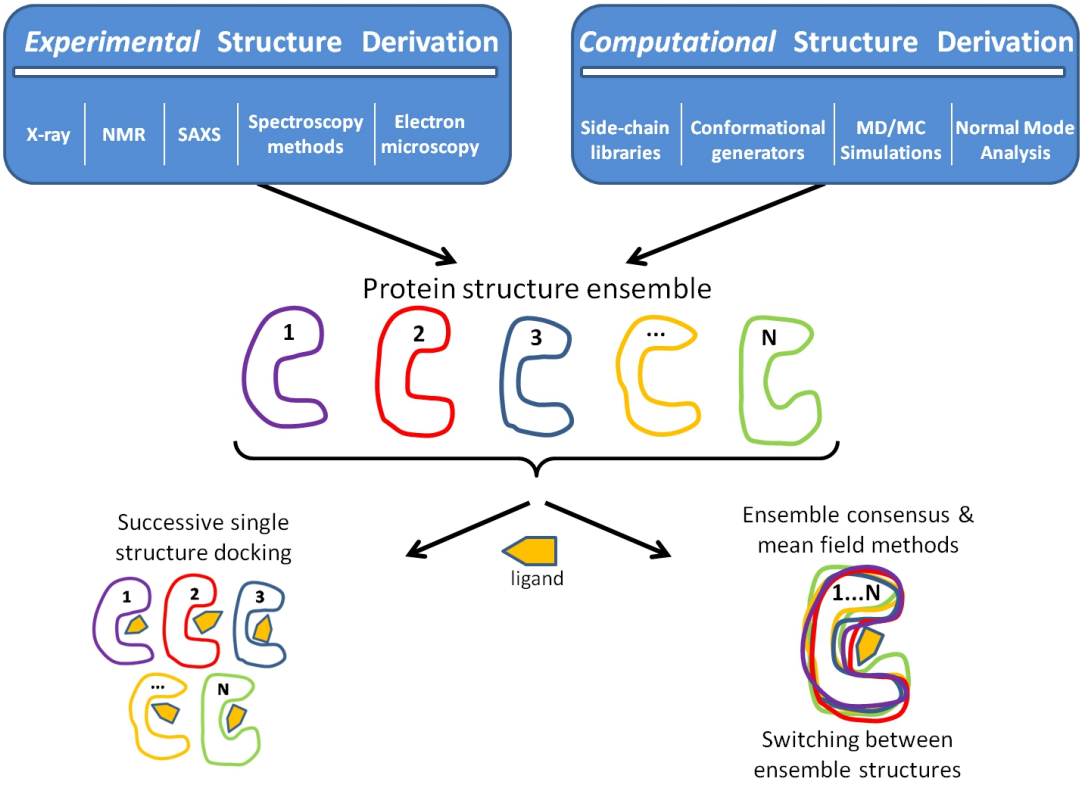


Figure 3.2: *Schematic overview of experimental or computational sources for ensemble protein structures and ensemble docking methods. Experimental or computational methods contribute several (ranging from just 2 to hundreds of) structures in different conformational states. The ensemble is then used for either separate docking to each ensemble structure (lower left) or methods that combine or switch between the conformers within the ensemble (lower right panel).*

One possibility is to calculate a mean field from all the structures in the entire ensemble. For example, discrete sets of side-chain and protein loop conformations can be efficiently combined in a mean field to self-consistently optimize the ligand-receptor interface structure during the docking calculation. In such mean field approach, each alternative side-chain or main chain conformation (conformational copy) is given an initial weight based on the internal energy and the interaction with a ligand (usually a Boltzmann weight). During docking, the weights on each copy can change such that in the finally docked structure the optimal side-chain/main-chain structure can be identified as the copy with highest weight.

Many structure-based drug design and docking programs employ pre-calculated grids for representation of the receptor-ligand interactions. The energy function of the receptor atoms of the binding site are first mapped onto a regular grid and interactions with

ligand atoms are calculated by interpolation from neighbouring grid points. A number of researchers have tried to map not only a single receptor structure onto such grids. Early attempts by Knegtel et al. computed composite energy-weighted, or geometry-weighted grids, from a set of experimental protein structures [37]. For energy-weighting, the authors computed grids for each structure and subsequently averaged over several grids, thereby at each grid point, grids with high negative values get high weights, grids with high positive values get small weights. In case of geometry-weighting, first an average structure is calculated, thereby using the mean structure for regions with low flexibility whereas regions with high flexibility are retained as independent conformational copies. Finally, a grid is calculated for this average structure and used during docking and scoring. Österberg et al. report similar findings, when testing four different ways of combining grids, energy-weighting or Boltzmann-weighting works well, whereas simple averaging can lead to an increase of incorrect docking solutions [134]. To further improve the efficiency, it is possible to represent part of the structural ensemble by a single rigid structure and to treat only the most flexible part of the binding site by explicit alternative structures [135]. Instead of docking ligands to each conformation in an ensemble, it is possible to use mean field grid representations to reduce the computational expense [37]. Several variants have been presented mainly differing in the way of averaging the interaction energy contribution [134, 136].

More sophisticated ensemble docking protocols have recently been suggested to avoid the increased computational effort in sequential docking all putative ligands to all target conformers (reviewed by Totrov and Abagyan [137]). Huang and Zou suggested an ensemble docking method that allows simultaneous optimization of placement and receptor conformation out of a set of protein structures and found significant improvement compared to docking to single structures [138]. Besides of using different experimental structures of a protein it is also possible to employ ensembles of homology models [139]. An interesting new approach for grid-based ensemble docking is the so called 4D-docking method [140]. Here, all structures are sorted by their conformational deviations. During docking, the integer index of the sorted stack is used as a fourth dimension. Hence, the search space can be reduced by consideration of limited conformational deviation, according to the actually considered structure (index neighbours). This allows a faster convergence towards optimal docking placement. Although the use of ensembles of structures can improve docking results it has also been found that the success depends critically on the choice of the ensemble [69, 141].

The underlying assumption of an ensemble docking approach is that conformations sufficiently close to the bound form are included in the ensemble. As indicated in the previous paragraph on generating ensembles, the absence of a receptor structure close to the bound form may deteriorate the docking performance. Recent efforts have therefore also focussed on recipes to obtain optimal conformational ensembles [142, 143].

3.6 Use of collective Modes to describe Global Motions

The application of ensemble approaches, as discussed above, is one promising route to account for the various conformational changes that can occur during ligand binding. However, it has the drawback of a discrete representation of predicted flexibility. The analysis of protein motions by molecular dynamics simulations in recent years has indicated that most of the conformational fluctuations can be described by a few collective degrees of freedom. These degrees of freedom can be obtained by the principle component analysis of the covariance of atomic fluctuations during MD simulations (also termed essential dynamics analysis) [144]. Alternatively, normal mode analysis, that is the analysis of the curvature of the potential energy function around an energy minimum, can also be used to extract collective degrees of large mobility in proteins and other biomolecules. It requires the diagonalization of the second derivative matrix of the energy with respect to the atomic coordinates yielding harmonic or normal (mass-weighted) modes of the biomolecule. To use the softest modes from a normal mode analysis as variables during flexible docking was suggested by Zacharias and Sklenar and first applied to DNA in complex with a minor groove binding ligand [145]. It allowed for fast minimization of the receptor structure during docking and for an estimation of the receptor deformation energy. The calculation of an energy associated with the deformation is based on the eigenvalue of the corresponding mode. The degree of deformation of the receptor in the soft degrees of freedom can be controlled by a penalty potential to limit deformations.
This is only an approximation to the receptor internal energy change, however, sufficient for detecting possible ligands and putative binding sites. It avoids the computationally costly explicit calculation of the internal receptor energy at every docking minimization step. Compared to docking methods that employ ensembles of discrete (rigid) receptor structures, the receptor conformation can change continuously during docking (in the pre-calculated soft normal modes) and has therefore an increased capacity for induced fit adaptation.

Normal mode analysis has also been used to study induced fit binding in case of integrase [146] from human immunodeficiency virus (HIV) and in combination with molecular dynamics to study ligand binding to HIV-protease [147]. However, pre-calculation of harmonic modes of a large receptor molecule requires very extensive energy minimization and is usually performed in the absence of solvent. Energy minimization under these conditions can lead to large deviations from the realistic (experimental) receptor structure and the calculated soft harmonic modes may not correspond to realistic soft degrees of freedom. The calculation of flexible degrees of freedom of a protein molecule can be performed under more realistic conditions using a principal component analysis (PCA) of motions obtained during a molecular dynamics simulation including surrounding waters and ions [144]. Flexible (essential) modes from the MD simulation can then be used as additional variables during docking as described above. This has been explored for the immunosuppressant FK506 to an 'unbound' conformation of a FK506 binding protein FKBP using the program PCrelax [148]. Accounting for relaxation in the pre-calculated soft modes of the receptor significantly improved the docking performance compared to docking to a rigid 'unbound' FKBP receptor structure at a modest increase in computational demand [148]. The approach can in principle lead to a dramatic reduction of the computational complexity to account for receptor flexibility during docking. This may form the basis for systematic docking including approximately for global conformational changes in the receptor. The contribution of global motions obtained from MD simulations has also been investigated to support protein-protein docking efforts [136].

It is also possible to use ENM analysis to identify rigid or flexible units in a protein and to define hinge regions [149]. Sandak et al. proposed the concept of hinge-domain-bending motion to account for global domain motion during docking with promising results [150, 151]. It was shown that with this approach relatively large displacements of the receptor backbone structure can be achieved during docking in directions that overlapped significantly with experimentally observed changes [34, 35, 38, 152]. Inclusion of normal mode minimization during docking was systematically explored by May and Zacharias for protein-protein complexes [39, 153, 154]. Inclusion of up to 5 softest modes for protein partners during docking improved the docking results at a very modest increase of computer time by a factor of 2-3 compared to rigid docking. However, the systematic test on several protein-protein complexes also indicated that in order to achieve realistic docking predictions both side-chain flexibility and global flexibility need to be accounted for simultaneously during docking [39].

One promising route for future developments might be a combined treatment of global flexibility employing soft global degrees of freedom together with several copies of binding site loops for the protein main chain and representation of side-chain flexibility by discrete rotameric states. Such methodology was used to dock several inhibitors to the protein kinase CDK2 (cyclin-dependent kinase 2). Application to different X-ray structures in the apo and various bound forms allowed an evaluation by cross-docking of different inhibitor molecules [153]. Application of the pre-calculated soft modes as flexible variables both with and without inclusion of side-chain flexibility resulted in improved ranking as well as ligand placement during docking. Interestingly, accounting for flexibility in the soft modes alone gave overall similar docking performance as in case of including side-chain flexibility (alone or in combination with normal mode minimization) but at significantly reduced computational costs. Abagyan and co-workers recently applied normal mode deformations to generate conformational variants of receptor structures and used the conformers in a Monte Carlo search during ligand docking [142]. For this application it is necessary to represent each receptor conformation by a potential grid. Kazemi et al. recently presented an interesting variant of the grid representation to account for structural flexibility [155]. The idea of the approach is to deform the geometry of potential grids associated with the receptor structure following possible global conformational changes. The deformed grid can still serve as a basis for calculating interactions with the ligand using interpolation from nearest grid points. In the initial application of the method it was, however, necessary to know the initial and final structure of the receptor in order to determine the necessary deformation.

3.7 Summary and Conclusions

Ultimately, ligand-receptor docking approaches should be able to reliably predict placement, conformation and affinity of ligands bound to target receptor molecules. For rigid receptor structures, at least the prediction of a near native binding geometry is often possible. However, many proteins including some of the most prominent drug targets, like protein kinases and HIV protease undergo significant conformational changes upon complex formation with substrates or inhibitors. Significant progress has been achieved within recent years in particular to better account for changes in side-chain conformation during structure based drug design and docking. However, efficient and sufficiently accurate inclusion of local but especially more global backbone conformational changes remains a challenge [69, 156].

Since one is typically interested to screen large numbers of putative ligands, it is important to find an optimal compromise between required accuracy and feasibility in terms of currently available computational resources. Current methods to tackle this problem range from brute force and time consuming but very detailed molecular dynamics simulation approaches to conformational ensemble methods and methods that try to restrict the flexible degrees of freedom to a subset of most relevant degrees of freedom. It should be emphasized that even if computational resources are available, the realistic modelling of conformational changes and adaptation is also limited by the accuracy of the underlying force field. Hence, inclusion of too many degrees of freedom may also degrade the performance of docking and scoring approaches. Methods that are based on identifying relevant soft degrees of freedom of a given protein receptor structure are computationally rapid and may allow for approximate inclusion of global flexibility during screening of large databases of putative ligands. Combinations of such approaches with modelling of side-chain flexibility on top of continuous soft mode backbone motion could be promising routes for future developments. Another promising effort is the design of carefully prepared conformational ensembles either based on experimental or modelled structures. Such methods could be valuable for rapid screening of large numbers of putative ligands followed by subsequent refinement steps.

As discussed above, there is also a close relation between the accuracy of scoring a ligand receptor complex and the inclusion of conformational changes during docking. A soft scoring function that tolerates inaccurate placement of a ligand in a binding pocket or allows a large degree of overlap between ligand and receptor atoms can only be of limited specificity. Future efforts to improve scoring of ligand receptor complexes may also include entropic contributions due to changes of receptor and ligand flexibility during the binding process [70, 71].

Chapter 4

Efficient Inclusion of Receptor Flexibility in grid-based Docking using ENM-derived Deformations

4.1 Introduction

Proteins and other bio-molecules are flexible and can undergo significant conformational changes upon complex formation with binding partners. As previously shown, such structural alterations can range from very small motions like single side chain fluctuations [16] to large global movement of loops or whole protein domains [57–59]. Pharmacologically relevant examples of such flexible targets include dihydrofolate reductase [157, 158], thymidylate synthase [159], HIV-1 protease [160, 161], reverse transcriptase [162], and the large group of protein kinases [163]. The importance of receptor flexibility for drug design has been extensively reviewed [69,86,164–167] and it has been shown that already small conformational changes of the receptor backbone in the range of 1Å can significantly affect receptor-ligand interactions [17]. The flexible modelling of ligands has made considerable progress during the last decade and has been included by now in the majority of structure-based drug design and docking programs [81].

Efficient and appropriate inclusion of global receptor flexibility, however, is still an unsolved problem, especially in the case of significant changes of the protein backbone structure upon ligand binding. During virtual screening of large drug-like compound libraries, the target protein structure is typically kept rigid or flexibility is allowed only for a few selected amino acid side chains. Several approaches that consider side chain flexibility have been proposed [108, 112, 168] and many available docking

programs include options to vary side chain dihedral angles during conformational search. Ultimately, it is desirable to account not only for local side chain changes but also for more global backbone conformational changes during the docking simulations. Several methods have been developed to tackle this problem as introduced before in Chapter 6.

In principle, it is also possible to perform ligand-receptor docking searches allowing for conformational changes in both the receptor and ligand structure in all Cartesian or bond rotation degrees of freedom. To this end, one can employ energy minimization (EM), Monte Carlo (MC), or Molecular Dynamics (MD) techniques. Different methods that combine such simulation techniques with docking have already been proposed (see reviews [93, 97]). The major drawback of these methods is the immense computational demand, induced by the many degrees of freedom of both solute and solvent during the simulations. To account for larger conformational changes near the binding site, it is also possible to represent the receptor by an ensemble of rigid structures and to use each ensemble structure separately for sequential docking [129, 130, 141]. The structures enclosed in such ensembles can be derived either computationally (e.g. by MC/MD simulations or using appropriate structural modelling methods) or from different experimental structures of a given target. In addition to the structure acquisition, separate docking of many structures increases the computational effort with the number of input structures, making high-throughput studies hardly feasible. In order to avoid this, more sophisticated methods have been suggested recently [137]. These include, for example, approaches that apply an ensemble average, select a consensus receptor out of the ensemble, or employ normal mode calculations for ensemble generation [37, 83, 134, 135, 138, 140, 142]. Others employ homology model ensembles instead of experimental structures [139].

In the case of global conformational changes, it is often possible to describe the conformational change by one or a few collective degrees of freedom of the whole receptor structure. Such collective degrees can be obtained from normal mode analysis (NMA) or from principal component analysis (PCA) of MD trajectories and have shown to produce a significant overlap with observed conformational changes in proteins [38]. Also, by the use of Elastic Network Model (ENM) calculations, it is possible to identify rigid or flexible units and to define hinge regions in proteins [149]. The inclusion of deformations along collective modes during docking has been used in docking studies and also for generating ensembles of structures [86, 128, 142, 145].

To accelerate docking searches, many of the popular docking methods map the energy function on to a regular 3D-grid around the ligand binding site. Interaction with ligand atoms can then be calculated very rapidly by interpolation from the nearest grid points. A drawback of this method is that a potential grid usually represents only one rigid receptor structure. Approaches to combine several receptor structures and represent it by one average or consensus grid have been developed [134]. However, in most cases, ensembles of structures are represented by one grid per conformer. It is then possible to either switch between grids or to perform individual docking simulations for each representative receptor structure. In addition, methods that deform the potential grid towards a given bound protein structure have been proposed [155], but these methods implicitly require the availability of a bound structure as direction for the grid deformation.
To allow for continuous deformation along selected degrees of freedom and still profit from the computational benefits of representing the receptor by a potential grid, I have developed a method that interpolates between different receptor grid representations during the docking search. The method, named ReFlexIn, was implemented in the popular protein-ligand docking software AutoDock. An application to the unbound (apo) structure of Protein Kinase A and the Cyclin-Dependent Kinase 2 demonstrates that for several test cases, significant improvement of docking performance compared to rigid receptor docking was achieved at very moderate additional computational cost. Table 4.1 lists the PDB IDs of the used receptor-ligand complexes and the deviation of the bound kinases structures versus the respective unbound form.

4.2 Modified LGA Implementation for flexible Receptor Docking

The AutoDock docking program employs a grid representation for all interaction potentials with ligand atoms. Prior to the actual docking procedure, the auxiliary program AutoGrid pre-computes regular 3D potential grids with a user-defined spacing and position at the ligand binding site of the receptor. It assigns the calculated interaction energies for each ligand atom type to each grid point as well as the desolvation and electrostatic potential energy. Interactions between a receptor and a ligand atom can then be calculated from the eight nearest grid points by using a tri-linear interpolation and a lookup function scheme. With no need for many pairwise atomic evaluations, this lookup table based method is highly efficient. At the same time, however, the usage of regular 3D-grids drastically impairs any use of receptor flexibility.

PKA binders

PDB ID	ligand name	$RMSD_{Protein}$ vs. apo
1BX6	BA1	1.90 Å
1FMO	ADN	1.61 Å
1JBP	ADP	1.55 Å
1STC	STU	1.51 Å
1YDT	IQB	1.92 Å
2ERZ	HFS	1.10 Å
3DND	LL2	1.49 Å
2DNE	LL1	1.52 Å
3MVJ	XFE	1.80 Å

CDK2 binders

PDB ID	ligand name	$RMSD_{Protein}$ vs. apo
1AQ1	STU	0.87 Å
1E1V	CMG	0.66 Å
1E9H	INR	0.84 Å
1FVV	107	0.72 Å
1G5S	I17	0.66 Å
1GZ8	MBP	0.59 Å
1JSV	U55	0.57 Å
1KE5	LS1	0.54 Å

Table 4.1: *Ligand test sets of PKA and CDK2 binders used in this chapter including the binding site C-alpha RMSD apo versus bound structure.*

All docking runs were performed using AutoDock's Lamarckian genetic algorithm (LGA) which employs variables (so-called genes) for translational, rotational, and torsion angle variation of the ligand [28]. In order to be able to consider several receptor structures and thereby introduce receptor flexibility to the docking, an additional conformational variable/gene – termed lambda – is integrated to the genetic algorithm of AutoDock.

Lambda can (in the present case employing seven structures in the ensemble) take values between 1.0 and 7.0. For the example of PKA, a lambda value of 1.0 represents the grid for the receptor deformation with the most opened binding site, 4.0 the undeformed apoPKA, and 7.0 being the last (most closed) of the seven PKA deformations (see also Figure 4.1). Maximal deformations were chosen such that complete closing of the ligand binding site (and a similar opening magnitude) was possible.

During the conformational search of the genetic algorithm, the lambda variable can be freely mutated (between 1.0 and 7.0, in this case), thus allowing for a variation of the receptor potential. To set this approach apart from already existing sequential docking methods (see Chapter 3) that dock ligands into a discrete set of receptor structures one after the other, or switch between discrete structures, an interpolation step is included. This allows the variation of the receptor potential to be not only discrete but continuous: non-integer values for lambda trigger a linear interpolation between the two nearest potential grids (see also Figure 4.2).

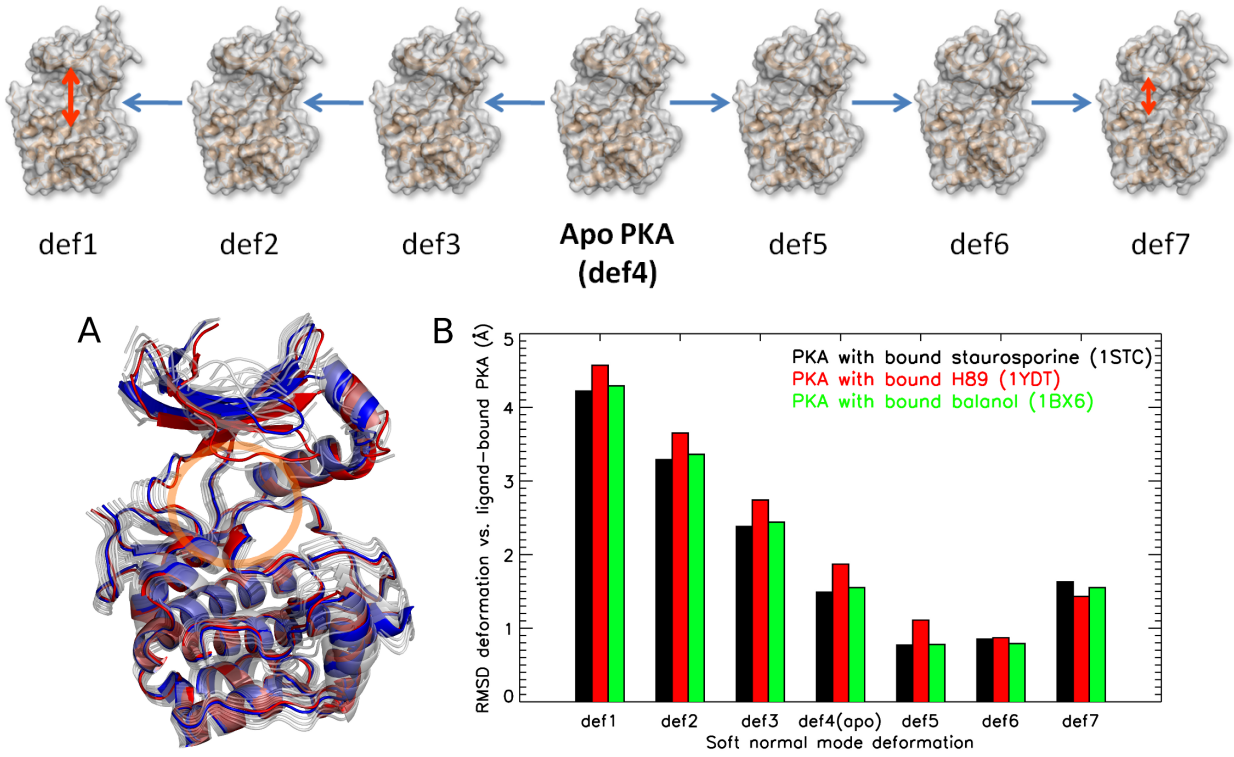


Figure 4.1: *Top: Seven deformations of the apo PKA structure (shown in the middle, def4) illustrate the closing and opening of the binding site. Here, deformation 1 represents the most open, and deformation 7, a completely closed binding site.*
A: Cartoon representation of the unbound PKA (blue, PDB ID: 1J3H), aligned with a ligand-bound PKA structure (red, PDB ID: 1STC, ligand: staurosporine, not shown). Deformations in the first soft normal mode ($RMSD_{Protein}$ between each deformation $\sim$ 1Å of the unbound PKA are shown in transparent gray cartoon. The PKA binding site is highlighted by an orange circle.
B: Backbone RMSD ($RMSD_{Protein}$) between the normal mode deformations and bound PKA receptor structures. The exact $RMSD_{Protein}$ values are also given in the legend of Figure 4.6.

It should be emphasized that it is also possible to combine variations in the lambda variable with Monte Carlo based sampling, instead of using the LGA method.
Benefiting from the usage of energy grid maps for interaction calculations, one is now also able to consider intermediate structures that might include only marginal changes but can possibly make the difference.

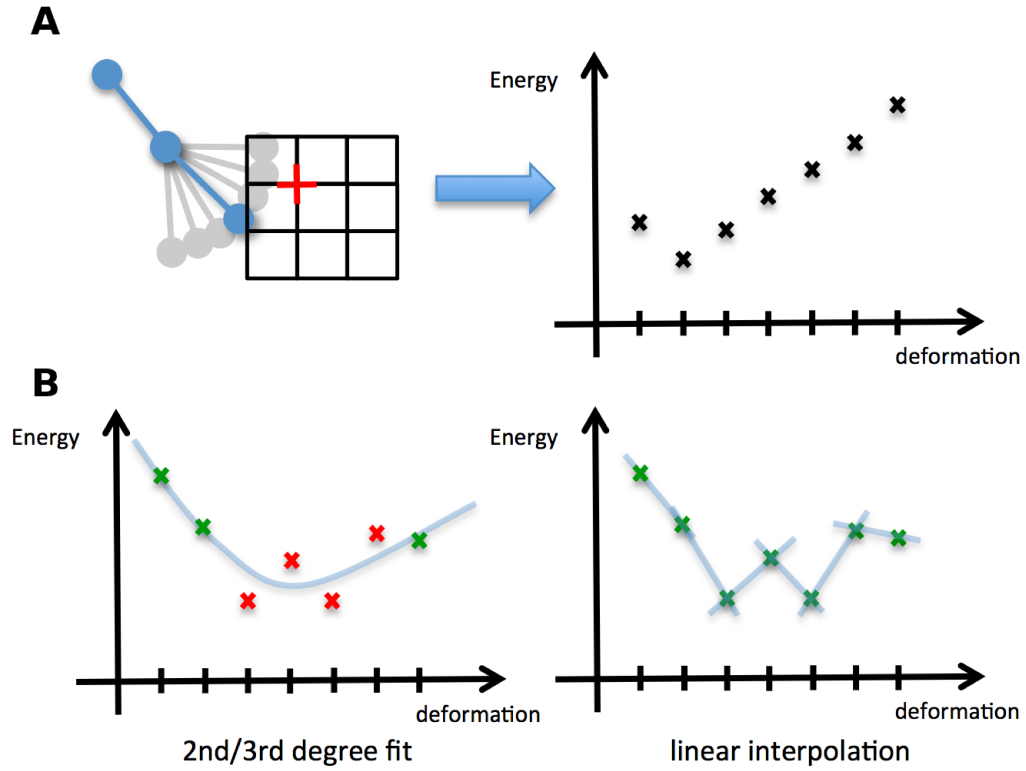


Figure 4.2: *A: Schematic translation of geometrical changes of atoms (blue) into interaction energy differences (plot on the right). Shown is a potential interaction energy change that occurs when the energies are calculated at one certain grid point (red) for different deformations (grey).*
B: Two possible ways of fitting energy curves to cover intermediate energies. When a second or third degree fit is applied (left), the intermediate energies between the red coloured values are poorly represented. A single linear interpolation between every deformation as shown on the right improves the coverage of intermediate state energies.

Normal mode calculation were performed as described previously in Chapter 2. Protein side chains were treated as rigid bodies that are allowed to move following their respective backbone pseudo atoms. This separation of backbone and side chain motion has the advantage that the AutoDock option of sampling rotameric states of selected side chains is still possible and can be combined with global backbone changes. In AutoDock, mobile receptor side chains are treated explicitly (similar to ligands) separate from the grid representation of the rest of the receptor (interactions between ligand and the flexible side chain are calculated explicitly from pairwise atom-atom distances) [30]. To combine this option with global motions of the rest of the receptor, one would move the side chain as one unit that performs the same motion as the backbone atom it is connected to. In such a way the side chain can undergo rotameric changes but still stays connected to the protein backbone (preserving the bond length and bond angles to the backbone).

In the present study, 7 receptor structures, deformed to various degrees along a soft normal mode direction, were used (see Figure 4.1). The amount of 7 deformations was found to be sufficient for a reasonable interpolation between intermediate structures in terms of conformational deviation between single neighboured structures. Tests with a smaller amount of e.g. 5 or 3 deformations resulted in too large gridpoint energy differences between deformations that made the interpolation fail. However, using more than 7 deformations did not significantly improve the results.

To evaluate the performance of our flexible receptor method, docking was performed using the X-ray structure of cAMP-dependent protein kinase catalytic subunit (PKA) and the cyclin-dependent kinase 2 (CDK2), both in the un-liganded apo form. For PKA, two side chain residues that are missing in the crystal structure near the ligand binding site (Lys72 and Glu127) were added using Swiss-PDBViewer version 4.01 [169] yielding low energy conformations reasonably close to the conformation in the bound form. The flexible receptor method was initially tested on the docking of three well-known PKA inhibitors of different size and flexibility: staurosporine, H89, and balanol. The ligand structures were extracted from their respective PKA-bound crystal structure (staurosporine: PDB ID 1STC [170], H89: PDB ID 1YDT [171], and balanol: PDB ID 1BX6 [172]). Docking, including ligand flexibility, was performed with variable numbers of fixed or variable dihedral torsion angles. For further testing, the size of the PKA ligand test set was increased and the docking on CDK2 was tested with a ligand test set of comparable size.

Docking Parameter Details

Receptor and ligand structures were prepared using AutoDock Tools version 1.5.4 and for both rigid and flexible receptor docking, the AutoDock version 4.2.3 was used with the same settings for all docking runs. Energy grids were produced using 50x50x50 point grid of 0.375Å spacing centred at the kinase active site.
In docking runs treating the ligand rigidly, the conformation from the known native complex structure was used. For the genetic algorithm, the following settings were used (explanations as comments):

```
ga_pop_size=150 \\number of individuals in population
ga_num_evals=2,500,000 \\number of evaluations until terminating GA
ga_num_generations=27,000 \\maximum number of generations
ga_elitism=1 \\number of top individuals to survive to next generation
ga_mutation_rate=0.02 \\rate of gene mutation
ga_crosover_rate=0.8 \\rate of crossover
ga_run=100 \\number of separate dockings
```

The variables for the genetic algorithm were kept at standard values, except ga_num_evals and ga_run: 2.5 million evaluations was set to be the exit criteria of the genetic algorithm. Testing significantly lower and higher numbers for this variable showed that this is the best agreement between runtime and reproducibility of results.
For dockings of a rigid ligand to a rigid receptor, 10 separate runs were sufficient. However, as soon as the ligand contains flexible bonds, or the receptor is treated flexible using ReFlexIn, the increased search space reduces the clustering accuracy of the results. This typical behaviour is shown and discussed below but for the sake of consistency, 100 separate docking runs were carried out for each test.

Sources for Deformation Structures

The ReFlexIn approach is not restricted to a single source of deformations that are used for the interpolation structure ensemble. While the results shown in this chapter are retrieved by only using receptor deformations generated by an elastic network model, a number of different deformation sources that differ in complexity and the amount of knowledge on bound protein structures have been tested. These results are presented in the following chapter.

The only criterion that has to be fulfilled for the retrieval of deformation structures in the context of this thesis is "efficiency". Hence, methods as e.g. Molecular Dynamics Simulations for structure generation are disregarded due to the high computational effort that is necessary to capture relevant conformational changes by this method.

The different sources used within this thesis are listed below (more details can be found in the respective chapters).

- Normal Mode Analysis (this chapter)
  The normal mode approach uses the least amount of information (namely none) on bound structures of the target protein. Only the unbound receptor structure is employed for the NMA calculations and the created deformations of the first mode are taken as input for the flexible receptor docking.

- Different bound protein structures (Chapter 5)
  Here, a set of bound receptor structures is employed and used as deformations to interpolate between during docking. This approach uses the largest amount of information on bound protein structures.

- NMR Structures (Chapter 5)
  An NMR structure that is available for HIV-1 Protease was taken as basis for the flexible receptor input. The experimental structure (PDB ID 1BVE) contains 23 models of the HIV-1 Protease in complex with a ligand. The models with the least conformational differences were removed, leaving 15 deformations for the structure ensemble of the flexible receptor docking.

- Structure Morphing (Chapter 5)
  Here, the deformations for the flexible docking ensemble are received by simple structure morphing. Therefore, only one bound receptor structure is necessary and a linear morphing is generated between this bound and the respective unbound (apo) structure of the protein. Five intermediate structures are generated using the USCF Chimera program [173] employing the corkscrew method with linear interpolation rate. To remove possible atomic clashes, the intermediate structures have been energy minimized for 500 minimization steps in the same program.

4.3 Results for Protein Kinase A (PKA)

The enzyme PKA undergoes significant backbone conformational changes upon complex formation with substrates and inhibitors (see Figure 4.1). The softest calculated normal mode based on an ENM overlaps well with the conformational difference between apo (ligand-free) and holo (ligand-bound) forms. For example, the overall backbone $\text{RMSD}_{Protein}$ of the PKA structure in complex with H89 (PDB ID 1YDT) and the apo form is ∼1.9Å. A best possible deformation of the apoPKA in the softest mode yields a structure with an $\text{RMSD}_{Protein}$ of ∼0.8Å. A similar overlap between the softest normal mode and the observed conformational difference between apo and holo receptor form was found for staurosporine and the balanol bound to PKA (albeit not at the same receptor deformation, Figure 4.1 B). To include deformability in the softest normal mode during docking, potential grids not only for the apo structure but also for the normal mode deformations have been calculated. The deformation in the mode was performed in three steps in the two opposite directions (step length ∼1.0Å $\text{RMSD}_{Protein}$ between neighboring deformed structures) along the soft collective mode resulting in seven structures (including the undeformed apoPKA structure). During docking, an additional variable lambda was used to control quasicontinuous deformations along the deformation direction. This was achieved by linear interpolation between potential grids representing each deformed receptor structure. No knowledge of the ligand bound form of the receptor was included – the bound PKA structures were only used to evaluate the ligand placement predictions. In a first set of docking simulations, three different inhibitors (staurosporine, H89, and balanol, see Figure 4.3) were docked to apoPKA in the bound ligand conformation either using rigid apoPKA or including receptor deformability (variable lambda). The results of 100 independent docking runs were compared (each with 2.5×10^6 GA evaluation steps).

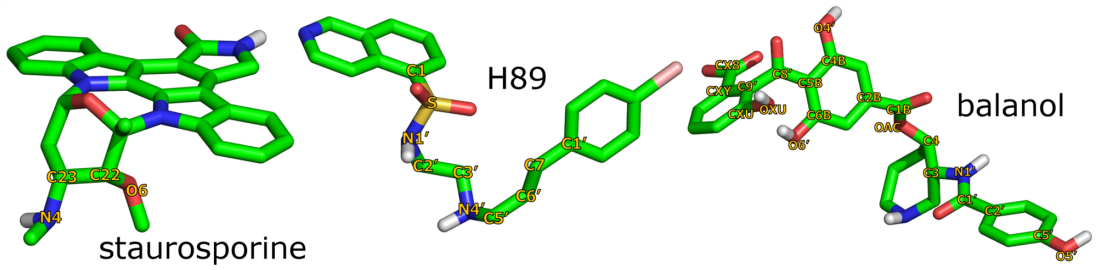


Figure 4.3: *Stick models of the ligands used throughout this study. Carbon atoms are colored green, oxygen red, nitrogen blue, hydrogen white, sulfur yellow, and bromine light pink.*

Rigid Receptor Apo Docking vs. Flexible Receptor Docking

For all 100 runs and for all three ligands, docking to rigid apoPKA failed to identify ligand placements closer than 2.5 Å relative to the known binding placement and orientation (Table 4.2). In contrast, with the flexible receptor option enabled, significant improvement in ligand placement, relative to the known bound form was observed (Table 4.2, Figure 4.4 and 4.5). The deviation from the ligand placement ($RMSD_{Ligand}$) in the known complex (after superposition of the receptor structures) dropped from 2.8 to 1.3 Å (for staurosporine), from 2.6 to 1.3 Å for H89, and from 6.6 to 1.7 Å for balanol comparing docking to rigid versus flexible receptor, respectively.

	rigid apo docking					flexible receptor docking				
staurosporine	0flex	2flex	—	—	—	0flex	2flex	—	—	—
	2.83	2.82	—	—	—	1.30	1.32	—	—	—
H89	0flex	2flex	4flex	5flex	8flex	0flex	2flex	4flex	5flex	8flex
	2.56	2.19	2.23	2.77	2.85	1.33	1.33	1.48	2.22	1.60
balanol	0flex	5flex	6flex	12flex	13flex	0flex	5flex	6flex	12flex	13flex
	6.64	4.85	4.06	4.35	3.82	1.72	2.43	2.58	5.50	3.90

Table 4.2: *Best ligand RMSD (given in Å) results out of 100 separate dockings are compared for the ligand docking into only the rigid PKA (left) versus the docking using the ReFlexIn approach (right). The allowed ligand flexibility is given by the number of freely rotatable torsions during docking (0–13 flex.).*

To test the influence of including ligand flexibility, increasing numbers of dihedral torsion angles of the ligands were allowed to vary during docking. For the sets with few flexible dihedral angles only bonds near the termini of the ligands were chosen (resulting in modest possible Cartesian coordinate changes upon bond rotation). For increasing ligand flexibility, dihedral rotations around bonds within the core part of the ligands were also allowed to rotate. Due to the planar ring geometry of staurosporine, the maximum number of rotatable bonds is only 2, and in this case, docking with a flexible ligand gave very similar results to docking with a rigid staurosporine structure (Figure 4.5 A). Also in the case of the H89 and balanol, ligand docking results with 0–8 rotatable ligand bonds showed significant improvement when docking to a deformable receptor instead of only using the rigid apoPKA (Figures 4.5 B and 4.5 C). However, the distribution of docking predictions for the deformable apoPKA became less distinguishable from rigid docking with increasing number of mobile dihedral angles.

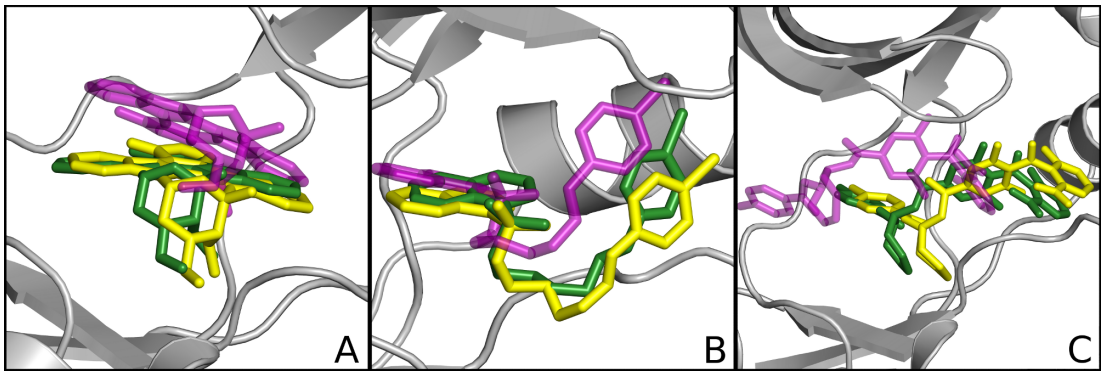


Figure 4.4: *Comparison of rigid apoPKA docking versus flexible receptor docking results for the ligands staurosporine (A), H89 (B), and balanol (C). The apoPKA (PDB ID 1J3H) is shown in grey cartoon. Yellow sticks represent the native ligand placements based on the respective crystal structure of the complexes. Transparent purple sticks are the best ligand RMSD docking results obtained for rigid receptor docking to the apoPKA, the best docking results including receptor flexibility are shown in green. In the case of staurosporine, two torsion dihedrals were mobile, eight for H89, and balanol was kept rigid.*

For low and intermediate ligand flexibility of balanol (0–6 flexible torsions), the approach yields still better results in terms of best RMSD_{Ligand} with respect to the native ligand placement than docking to rigid apoPKA (Table 4.2). However, when mobility of all dihedral torsion angles of balanol is allowed, neither docking to a rigid nor to a deformable PKA receptor was successful. It is of interest to note that even in the case of using rigid apoPKA, most of the 100 balanol docking trials for each case gave a different ligand placement and conformation (after quite extensive sampling of 2.5×10^6 genetic algorithm evaluations).

Holo Docking

To check if this result is due to a sampling problem of relevant ligand placements, the same set of docking searches was used for ligand docking to the holo (bound form) of the PKA receptor (extracted from the corresponding co-crystal structure). Similar to rigid docking to the apo structure, for low and intermediate ligand flexibility (0–6 flexible torsions), single (or few) clusters of ligand placements close to the native geometry were found as best scoring results (see Fig A.1 in the Appendix). The energy scoring of these solutions (for flexible ligands) was $\sim$ 2–3 kcal/mol lower than docking to the rigid apo form. However, for large numbers of flexible dihedral angles, there are no well-defined clusters of solutions similar to docking to the apo structure (Figure A.1 in

the Appendix and Figure 4.5). Apparently, allowing high degrees of ligand flexibility results in a rapidly increasing variety of putative binding placements, and the scoring function is unable to distinguish between realistic and incorrect placements (for docking to both the apo form and the holo form of the receptor).

AutoDock Binding Energies and Receptor Structure Deformation

The results also indicate that for fully flexible ligands, the docking search is still undersampled and may also point toward deficiencies in realistic scoring of the generated conformers and placements. Figure 4.5 dissects the 100 docking results for their respective AutoDock binding energy, after which the program scores the ligand placements. In the staurosporine case (Figure 4.5A), the calculated binding score for the lowest RMSD_{Ligand} placement was also significantly more favorable compared with docking to a rigid receptor.

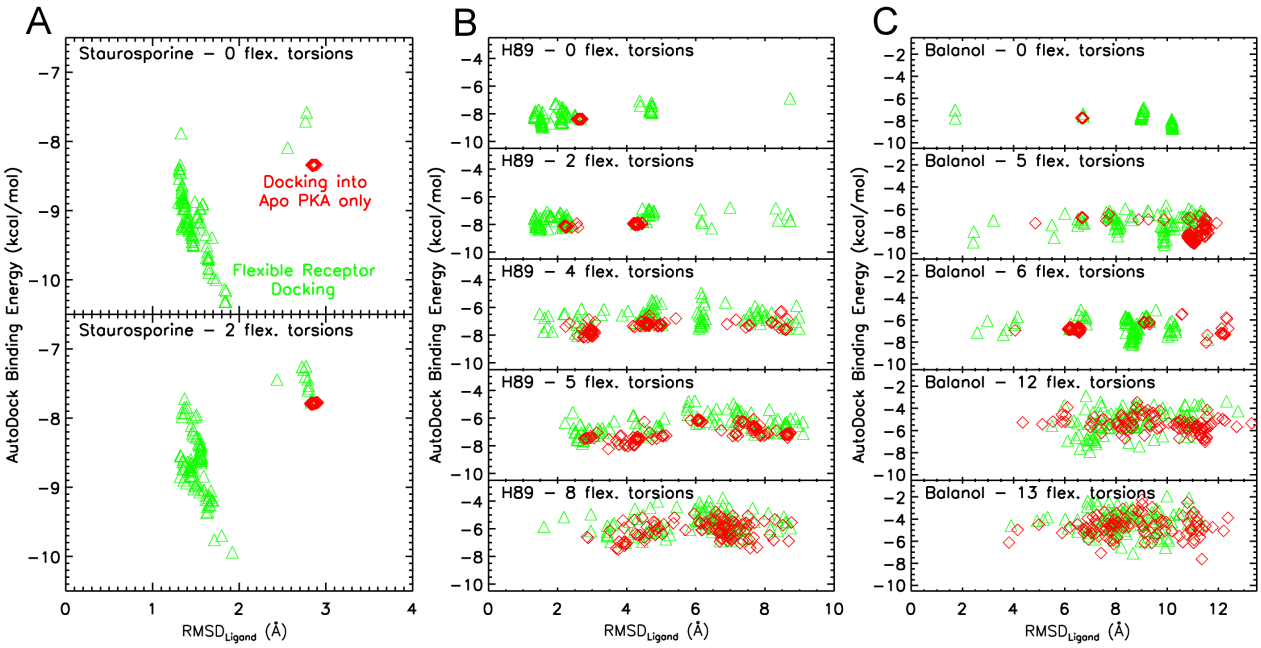


Figure 4.5: *Calculated AutoDock scoring energy for 100 separate docking results versus $RMSD_{ligand}$. Results are shown for the ligands staurosporine (A), H89 (B), and balanol (C), each with different numbers of mobile dihedral torsion angles. Docking results obtained in case of a rigid apoPKA are indicated by red diamonds; green triangles indicate results obtained with a deformable PKA receptor docking.*

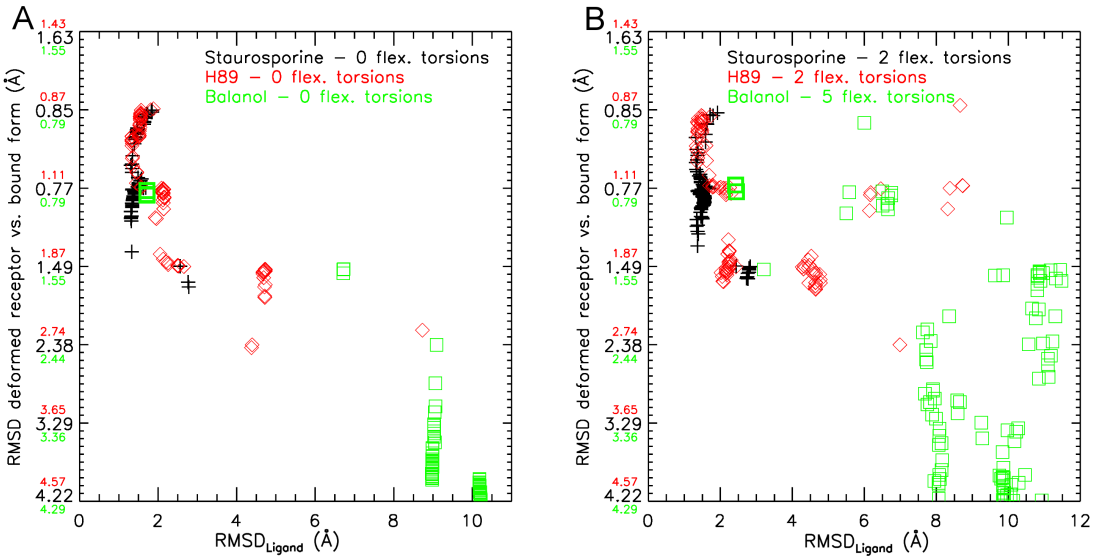


Figure 4.6: *Deviation of finally obtained PKA structures normal mode deformations from the native receptor conformation versus obtained $RMSD_{Ligand}$ for 100 separate docking runs. The $RMSD_{Protein}$ value on the y-axis indicates the most favorable final deformation selected by the flexible docking algorithm (color codes for the $RMSD_{Protein}$ versus the different ligand-bound PKA structures, black (crosses): staurosporine bound (PDB ID 1STC), red (diamonds): H89 bound (1YDT), and green (boxes): balanol bound (1BX6)). Comparing with Figure 4.1, the most opened deformation (def7) is on the bottom, the most closed deformation on the top (def1). Results are shown for no (A) and moderate (B) ligand flexibility.*

Especially low RMSD$_{Ligand}$ solutions (<2Å) yield much lower energies than the rigid apo docking (approaching more closely the score in the case of docking to the holo receptor form, Figure A.1 in the Appendix). Qualitatively similar observations were made but to a lesser extent for H89 and balanol docking (Figure 4.5). When comparing the docking scoring results, one needs to keep in mind that in case of a flexible ligand, a torsional free energy of the ligand of ∼0.3 kcal/mol per flexible dihedral torsion is added to the AutoDock score (subtraction of this contribution would lower the score substantially in docking cases of ligands with flexible dihedral torsions, e.g. 3.9 kcal/mol for the fully flexible balanol).

Inspection of the obtained final lambda values for each docking run (flexible receptor) indicates a strong correlation of the near-native ligand placement and predicted receptor deformation (Figure 4.6). All docking runs with a near-native placement of the ligands in the receptor binding pocket also resulted in a receptor deformation that resembled the bound structure of PKA more closely than the apoPKA structure.

The majority of staurosporine and H89 docking results yielding a low RMSD_{Ligand} placement were also found in a receptor deformation that was closest to the crystal bound receptor structure. Balanol docking frequently resulted in unreasonable ligand RMSDs, due to the fact that the larger ligand found interactions in the wider opened deformations that were scored favorably by the AutoDock scoring function (see also Figure 4.5C). However, the low RMSD_{Ligand} results for balanol were also found in a receptor deformation very similar to the balanol-bound PKA.

Hence, not only the ligand placement was significantly improved but also the PKA receptor structure moved significantly closer toward the bound form without including any information on the bound structure. In none of the docking runs, could any ligand placement close to experiment be observed, that coupled with a receptor deformation in the incorrect direction (e.g. more open PKA). Figure 4.6A shows examples where H89 or balanol yielded unreasonable binding modes with RMSD values larger than 4 or even 8Å . This emphasizes that chances to identify a near-native ligand placement are indeed small without accounting for backbone conformational changes toward the bound case during docking to apoPKA. Although docking including receptor flexibility resulted in solutions in better agreement with the native complex geometry, these solutions did not always score better than solutions with larger deviations from the native structure (Figure 4.5).
However, in a realistic docking screening one typically retains not only the best scoring solutions but also clusters of solutions within a threshold of the score. For the docking searches including limited ligand flexibility, it is possible to identify few clusters of solutions including those in close agreement with the native structure (both in terms of ligand and receptor deviation). These solutions can then be rescored using alternative and more sophisticated scoring functions than the score used in the initial screen and the inclusion of solutions that come close to the native structure increases the chance to identify such solutions as the most realistic placement in a rescoring step.

Extended Ligand Test Set

The initial test set of 3 ligands was extended by 6 additional PKA ligands for which a receptor-bound crystal structure exists (PDB IDs of the receptor-ligand crystal structures from which the ligands have been extracted are given in Table 4.1 with the respective ligand names). This test set contains a variety of ligand sizes and geometrical shapes, for example, the planar ligand staurosporine (STU) or large ligands (IQB, BA1, ADP) as well as smaller ligands (XFE, HFS, LL1). Their chemical structure is shown below in Figure 4.7 (all chemical structures in this thesis have been drawn using the program 'Marvin Sketch' by ChemAxon [174]).

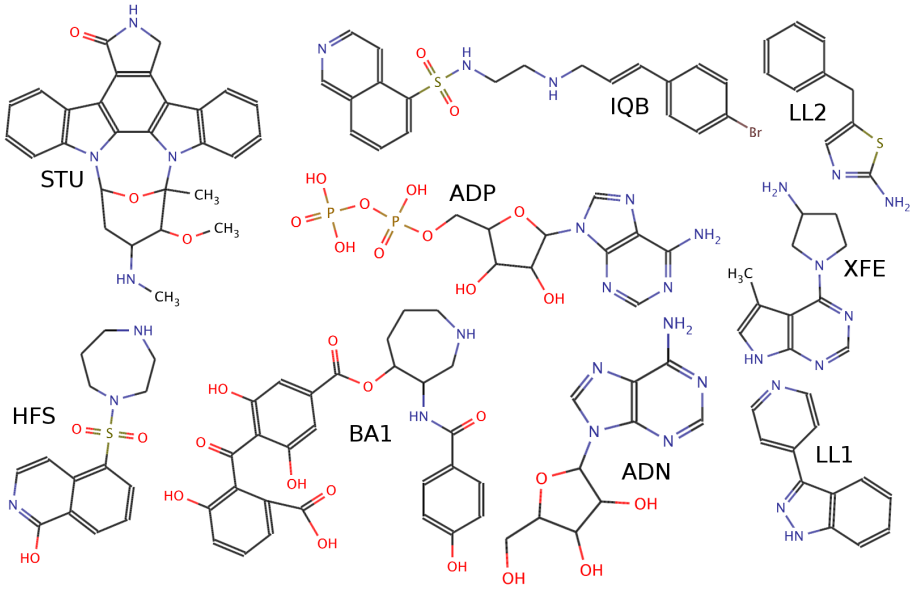


Figure 4.7: *Chemical structures of the PKA ligand test set.*

With the knowledge on the effect of result spreading of highly flexible ligands as discussed above, in these further tests, only moderate ligand flexibility was used. Depending on the size and total number of possible flexible torsions per ligand, only one or two torsions have been allowed flexible in each ligand. The results of 100 separate docking for each ligand are shown in Table 4.3 and distinguished by ligand flexibility for the rigid receptor docking into the unbound PKA form (apo docking) and the flexible receptor docking employing the set of normal mode deformations.

When only the apo structure of PKA is taken into consideration for a rigid receptor docking of the ligands, the overall docking performance is poor and yields unreasonable ligand placements with large RMSD_{Ligand} values. Only the ligands HFS (when treated

	ligand rigid				ligand flexible			
	flex.docking		apo docking		flex.docking		apo docking	
	best RMSD	low.en. RMSD	best RMSD	low.en RMSD	best RMSD	low.en. RMSD	best RMSD	low.en RMSD
ADN	4.03	5.38	7.22	7.23	3.67	4.56	7.02	7.03
ADP	6.12	8.97	3.67	3.74	3.35	7.80	7.26	5.91
STU	1.30	1.83	2.83	2.86	1.32	1.92	2.82	2.83
IQB	1.33	1.57	2.56	2.64	1.33	1.47	2.19	2.41
HFS	1.49	2.41	1.50	5.21	2.01	5.14	4.63	5.11
LL2	3.76	5.26	5.24	5.25	3.84	5.79	5.70	6.08
LL1	1.48	6.59	5.96	5.97	1.75	6.10	5.99	6.10
XFE	2.21	3.57	2.23	6.45	1.93	2.83	1.91	3.08

Table 4.3: *Best RMSD, i.e. the lowest yielded $RMSD_{Ligand}$ out of 100 separate docking runs (given in Å), and the RMSD values of the docking solution with the lowest AutoDock binding energy for different ligand and receptor flexibilities.*

rigid) and XFE (both rigid and flexible) yield already low RMSD solutions that cannot be further improved by the flexible receptor docking.

By using the normal mode based ReFlexIn approach, the ligand placements are significantly improved towards the experimental placement in most cases: for both rigid and flexible docked ligands, the best $RMSD_{Ligand}$ values that can be obtained are significantly lower for seven out of eight ligand cases when the flexible receptor approach is used. Four out of eight ligands even yield a ligand RMSD that is lower than 2Å. The biggest improvement was found for the ligand LL1 where the apo docking RMSD of around 6Å could be reduced to 1.48 and 1.75Å (for rigid and flexible ligand). The ligands STU, IQB, and HFS (flexible ligand) already showed reasonable $RMSD_{Ligand}$ but could be improved even more by the flexible receptor docking. Table 4.3 also lists the ligand RMSD values of the best scored docking solutions (lowest energy RMSD). Here, the AutoDock scoring energy itself is insufficient to reliably distinguish between good and bad ligand placement. Only the ligands STU and IQB yield ligands best energy RMSDs that are close to the best ligand RMSD.

This issue also becomes visible in Figure 4.8, where, for example ligand LL1 yields significantly improved RMSD values below 2Å that are, however, scored at the same level as poor ligand placements with an $RMSD_{Ligand}$ of ≈ 6Å.

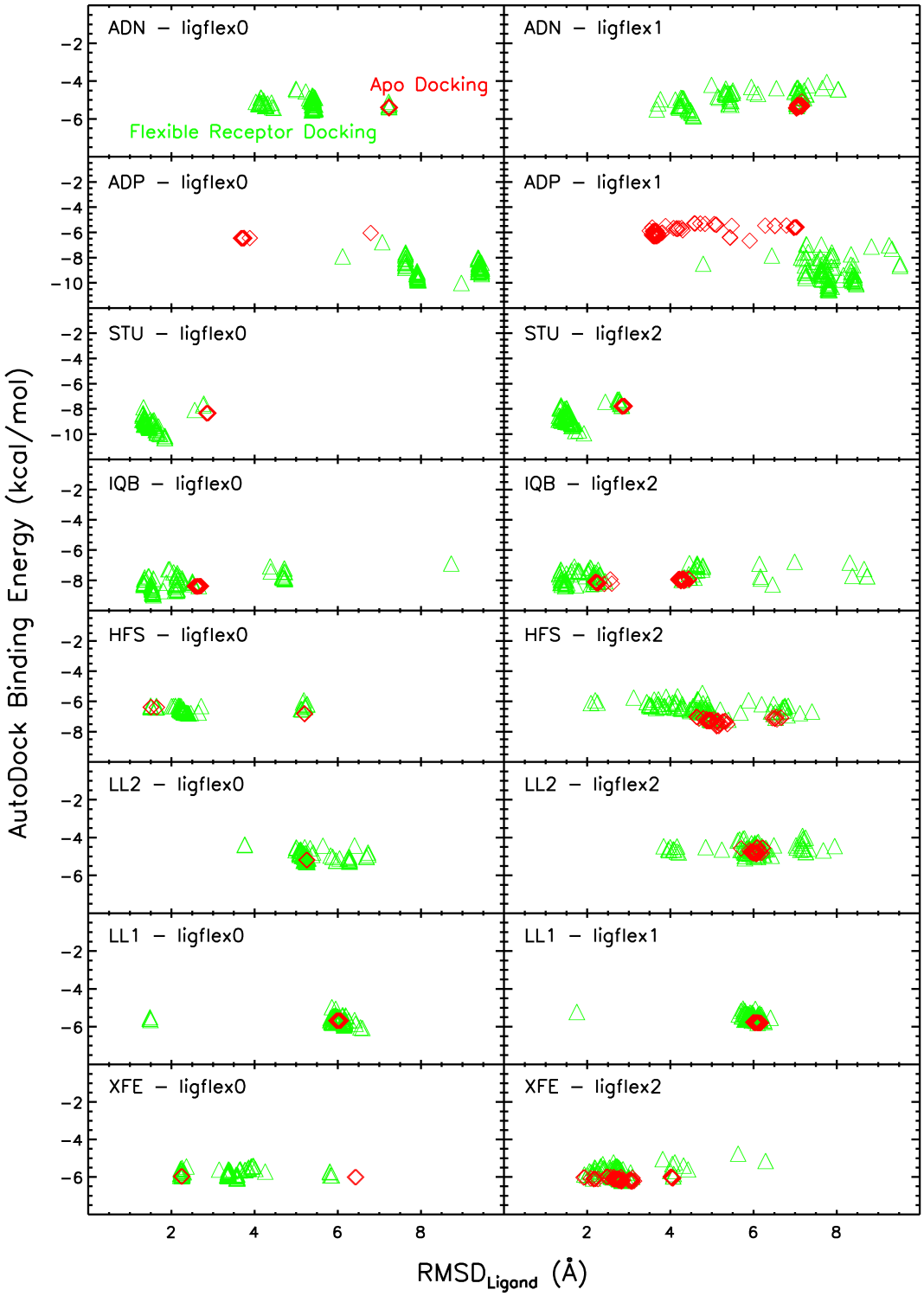


Figure 4.8: *Calculated Autodock Binding energies plotted against the yielded $RMSD_{Ligand}$ for 100 separate docking runs of rigid receptor apo docking (red diamonds) versus flexible receptor docking (green triangles). Plots on the left show the results for the ligand being rigid, plots on the right with the ligand having moderate flexibility of 4 rotatable torsions.*

4.4 Results for Cyclin-Dependent Kinase 2 (CDK2)

Another protein of the kinase class, the cyclin-dependent kinase 2 – or short CDK2 – was tested in the normal mode deformation based flexible receptor docking approach. This protein has already shown promising results for the representation of the conformational changes upon ligand binding by normal mode deformations [86, 175].

In addition to the apo CDK2 structure (PDB ID 1HCL), eight known structures of different ligands bound to CDK2 were extracted from the PDB as listed in Table 4.1. One of the ligands inside this test set (STU) was also present in the test set for PKA as described above. The set contains ligands of different sizes and their chemical structures are shown in Figure 4.9. As in the previous PKA section, the maximal ligand flexibility was reduced to 1 or 2 flexible torsions per ligand to ensure for a reasonable clustering of results.

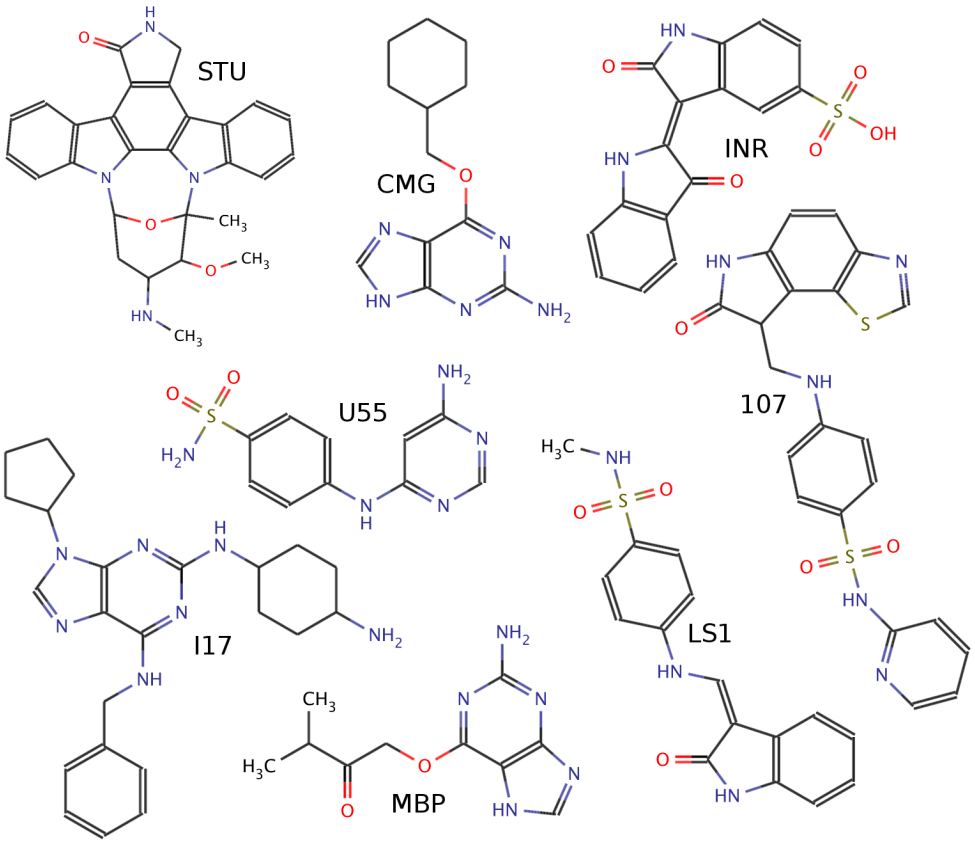


Figure 4.9: *Chemical structures of the CDK2 ligand test set.*

	ligand rigid				ligand flexible			
	flex.docking		apo docking		flex.docking		apo docking	
	best RMSD	low.en. RMSD	best RMSD	low.en RMSD	best RMSD	low.en. RMSD	best RMSD	low.en RMSD
STU	4.55	6.31	9.38	9.39	6.06	7.57	6.41	7.55
CMG	1.56	10.14	7.96	10.17	1.73	9.85	5.95	9.90
INR	6.65	7.07	6.59	6.73	2.14	6.78	4.88	6.72
107	2.16	5.76	2.53	2.84	2.46	2.46	8.97	9.00
I17	0.93	1.00	8.66	8.67	1.12	9.38	4.40	4.61
MBP	0.77	7.96	0.63	7.95	0.77	7.96	0.91	7.96
U55	6.77	7.63	7.50	7.61	2.07	7.65	6.62	7.53
LS1	7.53	7.72	7.99	8.00	1.98	7.18	7.64	8.30

Table 4.4: *Best RMSD, i.e. the lowest yielded $RMSD_{Ligand}$ out of 100 separate docking runs (given in Å), and the RMSD values of the docking solution with the lowest AutoDock binding energy for different ligand and receptor flexibilities.*

Table 4.4 shows the results of both the rigid receptor apo docking as well as the results for the flexible receptor docking approach.

Apo Docking vs. Flexible Receptor Docking

As in the PKA docking, the resulting best-RMSD ligand placements yielded by the rigid receptor apo docking are, in most cases, far from the native experimental structure. Only the docking of the ligand MBP yields very low $RMSD_{Ligand}$ values below 1Å and the ligand 107 just below 3Å whereas the other rigid receptor dockings result in incorrect ligand placements. Those results show a considerable deviation from the experimental placement by more than 6Å.

With the flexible receptor docking enabled, the results show a substantial improvement. When the ligand is treated rigid, the outperformace of our docking approach stays moderate. Only the ligands CMG and I17 yield a significant improvement of docking results where the best apo docking results of around 8Å ligand RMSD are reduced to native-like placements with an RMSD of 1.56 and 0.77Å, respectively. In the case of I17, these best-RMSD values even yield the best energy scoring and are energetically favoured over the apo docking results by approximately 1.5 kcal/mol.

In the dockings where the CDK2 ligands are allowed to have moderate flexibility of 1-2 rotatable bonds, the best-RMSD values again improve. In six of the eight test cases, the flexible receptor approach is able to capture ligand placements that are significantly closer to the native ligand conformation, compared to the rigid apo docking. The improvement is in the range of 2-6Å (e.g. ligand 107 that yields a best RMSD_{Ligand} of 8.97Å in the apo docking versus 2.46Å using the flexible docking approach).

However, as shown also in the previous results for PKA, the correct docking solutions cannot be reliably distinguished from poor results by just looking at the AutoDock binding energy/score (see also Figure 4.10). The flexible ligand INR, for example, yields a low RMSD solution of approximately 2Å but this solution is in the same scoring regime (± 0.5 kcal/mol) as solutions with an RMSD of around 7Å.
Cases like the rigid ligand U55, with a whole cluster of solutions that yield the same RMSD_{Ligand} values, but result in a different scoring energy at the same time, is an indication for inconsistencies within AutoDock's scoring function.

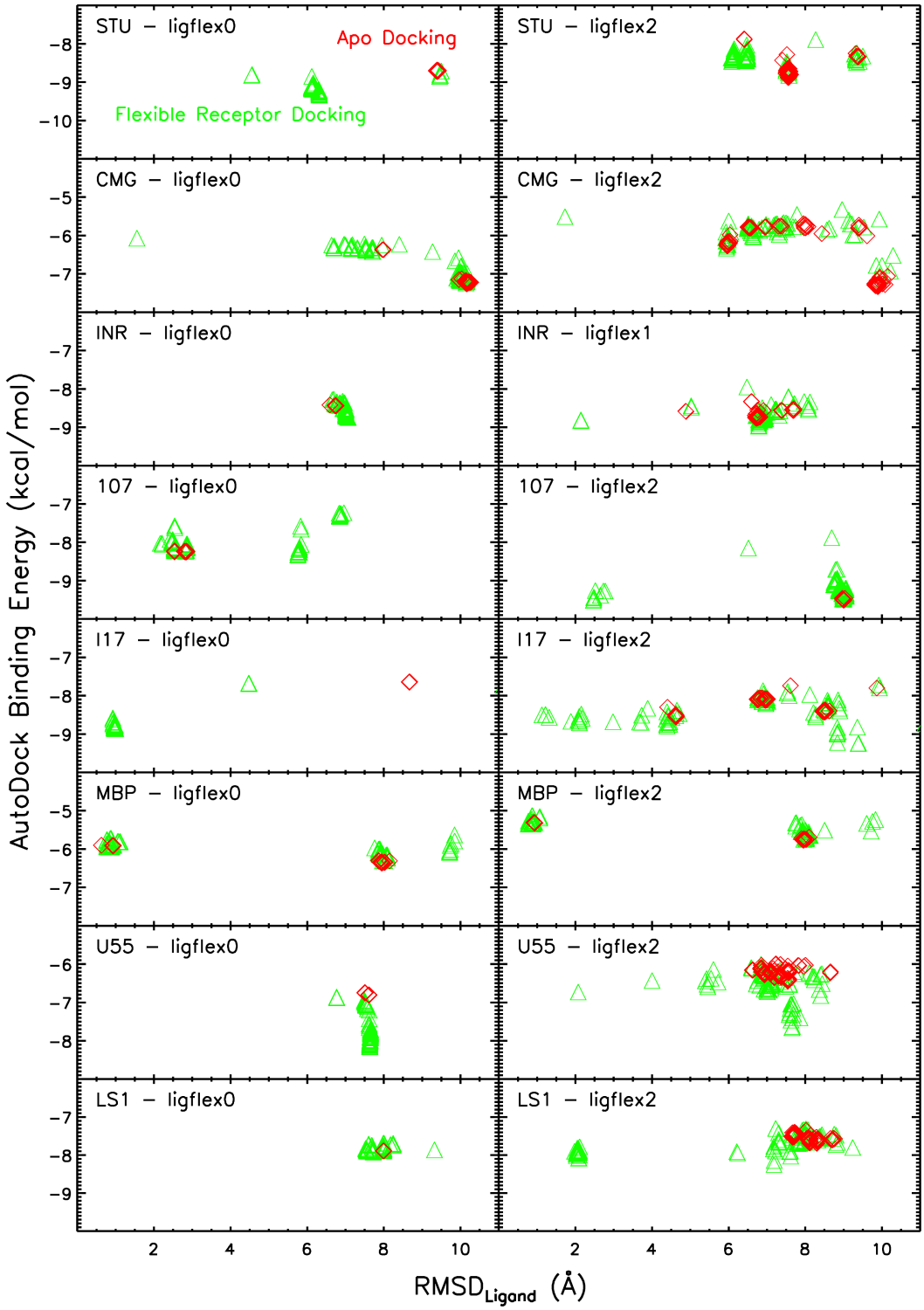


Figure 4.10: *Calculated Autodock Binding energies plotted against the yielded $RMSD_{Ligand}$ for 100 separate docking runs of rigid receptor apo docking (red diamonds) versus flexible receptor docking (green triangles). Plots on the left show the results for the ligand being rigid, plots on the right with the ligand having moderate flexibility of 4 rotatable torsions.*

Receptor Structure Deformation (Lambda)

Figure 4.11 shows two examples of ligands where the solutions with a low ligand RMSD favourably bound to the normal mode deformations with the lowest deviation towards the actual bound structure.

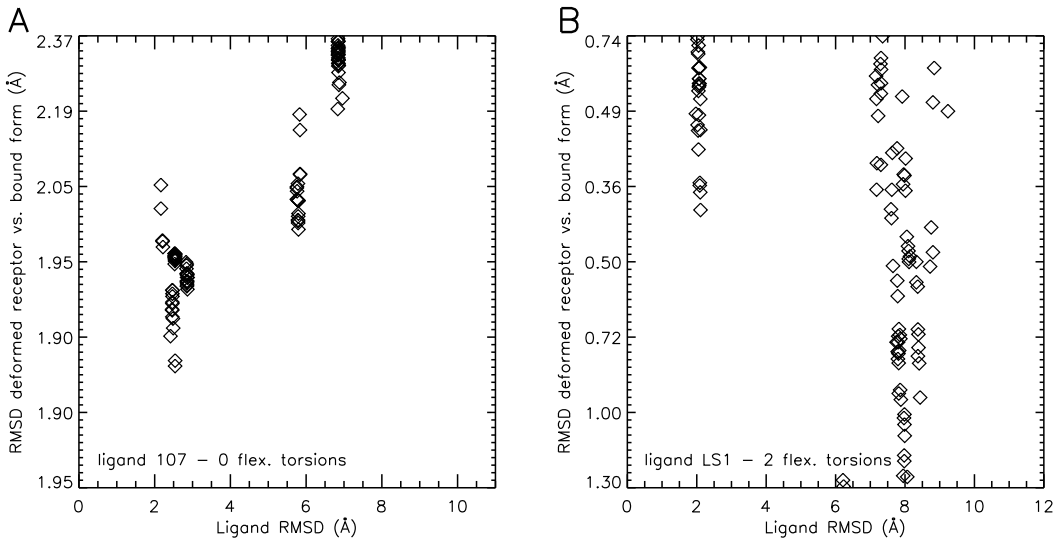


Figure 4.11: *Lambda values for the CDK2 docking of the ligands 107 (A) and LS1 (B). Given on the y-axis is the protein C-alpha RMSD between the deformation and the actual bound receptor.*

For the ligand 107, the second and third normal mode deformation of the unbound apo structure yield the protein structures that are closest to the real receptor form with an $\text{RMSD}_{Protein}$ of both 1.9Å (see Figure 4.11 A). This is only a very minor deviation as the protein RMSD between the bound structure and the apo form (deformation 4) is 1.95Å . Many results with a ligand RMSD around 2Å are found in the intermediate region between deformation 3 and 4, whereas the high ligand RMSD results have a larger lambda value, thus, being docked to receptor deformations with a larger deviation towards the true bound form. No poor RMSD solutions are found for small lambdas (better agreement of the deformation with the true bound form) and vice versa.
Similar observations can be made for the flexible LS1 ligand as shown in Figure 4.11 B. Here, deformation 5 and 6 are the closest towards the actual bound structure and several good ligand placements are found that docked into those structures or adjacent intermediate states. However, there are also docking results yielding large ligand RMSD of around 8Å that are docked to the 'correct' deformation. Here, there is not such a good differentiation as for the 107 ligand.

Minimization of Structure Deformations derived by Normal Mode Analysis

Typical for ENM-derived deformations are possible bond distortions that can occur on regions of high flexibility (e.g. exposed loop regions) resulting in slightly expanded atom-atom distances. These distortions are most frequently observed in the outer normal mode deformations, as in the present case for PKA and CDK2, yielded for the deformations with the most opened structures (deformations 6 and 7).
To rule out possible worsening of docking results triggered by these distortions, additional docking runs have been performed, where the normal mode derived deformations have been subjected to structure energy minimization, prior to energy grid calculations docking. To this end, the six deformations surrounding the apo structure have been minimized for 500 steps using the minimization routine of Chimera with standard options. Here, 500 steps of steepest descent minimization is followed by another 500 steps of the conjugate gradient minimization (step size 0.2Å).

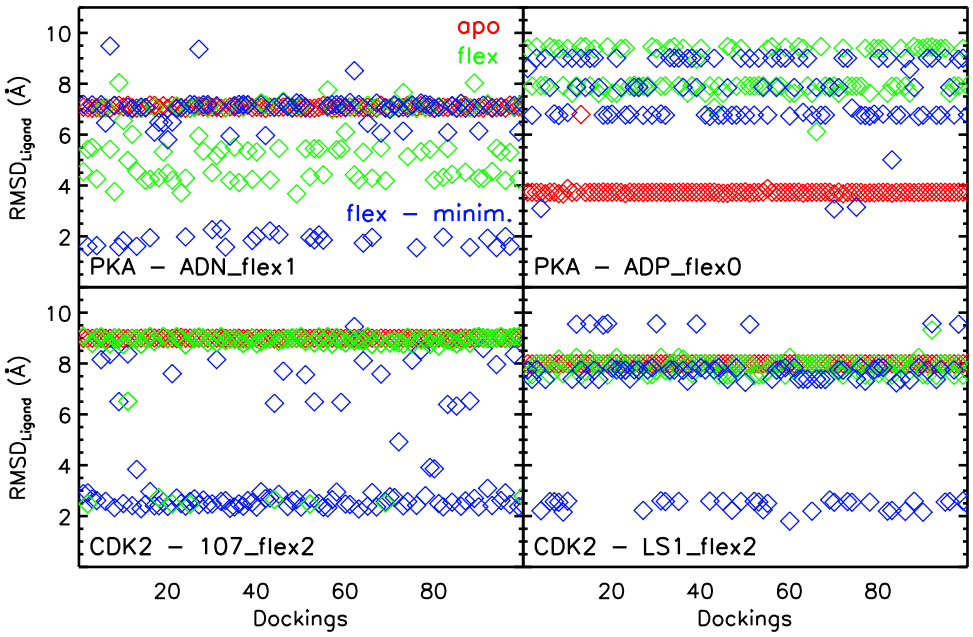


Figure 4.12: *Effect of minimizing the normal mode derived deformations prior to docking shown for two PKA ligands (top) and two CDK2 ligands (bottom). The ligand RMSDs of 100 docking results are shown for the rigid receptor apo docking (red), the flexible receptor docking of the raw normal mode deformations (green) and the flexible receptor docking with minimized deformation structures.*

Tests with a flexible docking using the minimized deformation structures for PKA and CDK2 show that the minimization does not worsen the docking results but only few results can be significantly improved. Those cases are shown in Figure 4.12.

For the PKA ligand ADN (upper left plot in Figure 4.12), the poor ligand placement was already improved by approximately 3.5Å using the flexible receptor approach with the untouched normal mode deformations. When the deformations are minimized prior to docking, the docking results can be improved even further (below 2Å RMSD_{Ligand}). The docking of the PKA ligand ADP showed apo docking results that could not be improved by flexible receptor docking. All docking results yielded considerably worse values than the apo docking. However, when the normal mode deformations are minimized, three of the 100 dockings are able to yield ligand conformations that are slightly better than the apo docking results.
The flexible receptor docking of the untouched normal mode deformations for the CDK2 ligand 107 yielded several docking results with a favourable ligand RMSD. The great majority of values, however, can be found at the same level as the apo docking. Using the minimized structure ensemble, the opposite is the case: only several high RMSD solutions are found, whereas the majority of the 100 dockings yields ligand placements that are close to the native ligand conformation.
The last example shows CDK2 ligand LS1. Here, no reasonable ligand placement was found for either the rigid receptor docking, or for the flexible receptor docking. Only when the normal mode deformations are minimized, approximately one third of the results are significantly improved.

4.5 Summary and Conclusions

ReFlexIn, a new method to combine backbone receptor flexibility in terms of a global collective degree of freedom with the computationally efficient grid-based representation of the receptor potential has been designed and implemented in AutoDock.
Results from several rigid receptor docking tests indicate that a simplified rigid consideration of the receptor protein in docking fails in most cases to lead to reasonable docking results. Applying the presented flexible receptor docking approach, however, can lead to a significant improvement of the ligand placement quality.

Figure 4.13 summarizes the results of all kinase dockings and illustrates this improvement well. In this Figure, the best achieved RMSD_{Ligand} values for all eight ligands of each test set are shown and ordered by descending RMSD values for the flexible receptor docking.

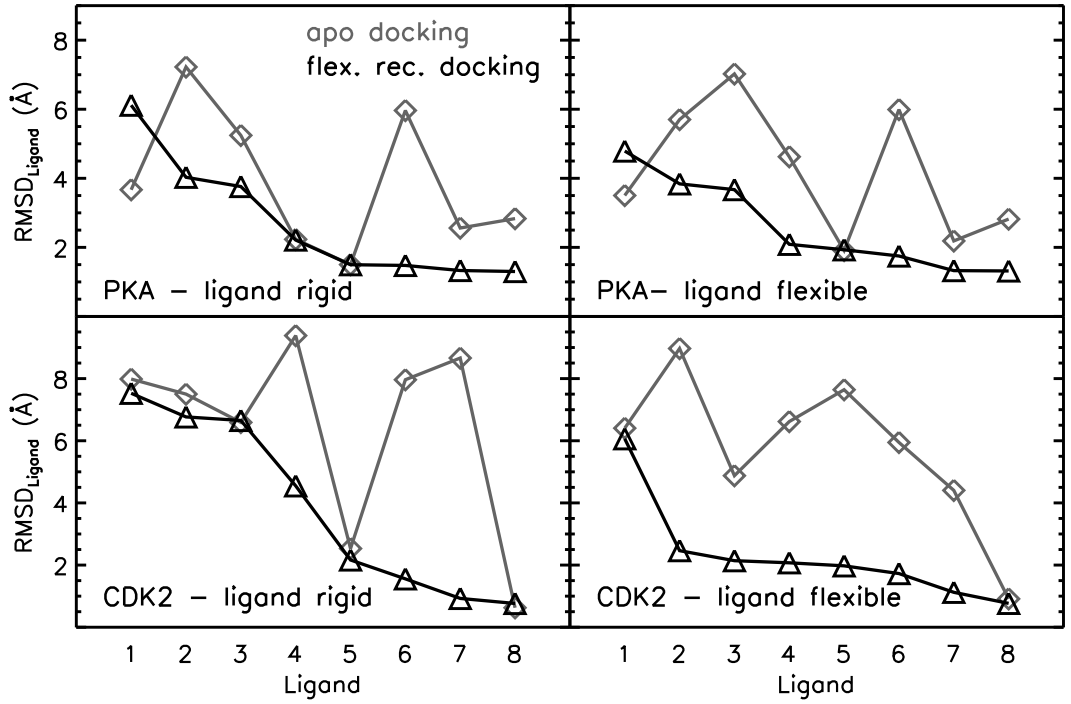


Figure 4.13: *Best achieved $RMSD_{Ligand}$ values of flexible receptor docking (black triangles) versus rigid receptor apo docking (grey diamonds). Shown are the results for docking of rigid and flexible ligand molecules (left and right column, respectively) into the kinases PKA (upper plots) and CDK2 (lower plots).*

The improvement of docking results when using the flexible receptor approach is especially remarkable when considering the fact that no knowledge of the bound form of the receptor was used (the deformations of the receptor ensemble are solely created by ENM calculations of the unbound protein structure).

Observed improvements come at small additional computational costs (approximately 50% increase in CPU time compared with rigid docking). No improvement compared with docking to a rigid apoPKA structure was observed, if more than 4–6 ligand dihedral angles were allowed to vary during docking. Interestingly, a very similar result was found when docking to the holo form of the receptor. Only in the case of limited ligand flexibility (up to 4–6 flexible dihedral torsion angles), clusters of solutions with small deviation from the native complex structure could be identified.
For larger numbers of flexible dihedrals, a range of solutions with different deviations from the native placement and very similar scores were obtained (see Figure Figure 4.5). It has also been pointed out to limit the ligand flexibility for successful ligand–receptor docking using AutoDock [30]. As discussed above, this might be in part due to the scoring function approximations. A more realistic evaluation of the ligand internal energy might be a route to further improve the docking method. Additionally, a pregeneration of appropriate ligand conformations (e.g. using conformational generators as e.g. FROG [176]) with reasonable internal energy and employing those few conformations at low ligand flexibility could be used which may reduce the amount of different (possibly unrealistic) ligand conformations generated during the docking search.

In the present work, the possible energetic changes for deforming the receptor structure along a soft normal mode were neglected. This is reasonable, because such global changes (opening/closing motions) of enzymes can often be observed in molecular dynamics simulations indicating that the associated energy changes are within the thermally accessible regime per degree of freedom of $1RT$(0.59 kcal/mol) which is in small relation to the binding energy of a ligand. It should be emphasized that if an estimate of the receptor energy change upon deformation is available, it can be included at basically no cost during the docking search. The approximate inclusion of receptor flexibility through interpolation between grids is therefore well suited for systematic virtual screening efforts at basically the same computational cost as using a rigid receptor structure.

Compared with approaches that perform Monte-Carlo switches between discrete receptor structures [142], the continuous deformation of the receptor along preselected directions in this method allows for a smooth receptor potential variation which may in turn require fewer representative receptor structures. The very modest increase in computational demand compared with docking to a rigid receptor makes the approaches well suited for systematic virtual screening applications.

One should keep in mind that this approach was kept simple and used only the first mode that is derived by ENM calculations. Results could be further improved by taking into account additional modes that might also capture the conformational changes of a receptor better than only the first mode. Future work could include more modes (adding another additional gene to the algorithm) even though this would again increase the complexity by another level and hence, increase the runtime.

It should also be emphasized that the approach is not restricted to normal mode derived deformations. It is, for example, possible to represent the structures used for the interpolation between grids by different bound forms of a receptor, different model structures, or different experimental structures of a target receptor structure. This possibility was further considered and is presented and discussed in the following chapter.

Chapter 5

Different Deformation Sources for Flexible Receptor Docking

5.1 Introduction

The previous chapter demonstrated that our flexible receptor docking approach in combination with elastic network model derived deformations has the potential to account for the conformational changes in receptor proteins and to improve the docking of various ligands. The methodology is not restricted to deformations along a normal mode direction of a protein. Some conformational changes in a protein upon ligand binding cannot be properly sampled by means of normal mode deformations. For these cases, The ReFlexIn method relies on different sorts of structure input.

In the present chapter the approach extended to an, in principle, arbitrary set of receptor conformations that can be ordered according to some measure of structural similarity between neighboring conformations. The "reaction coordinate" for deformations is then represented by a hypothetical path connecting all structures in the ensemble. The interpolation scheme allows for a smooth interpolation between the "end-points" of the ensemble along the pathway.

The method is applied to HIV-1 (Human Immunodeficiency Virus 1) protease, a member of the aspartyl protease class. Inhibiting drug molecules, that mimic a polypeptide chain, bind tightly to the protease active site and can block the protein's function, thus preventing the HI-virus from maturation. Since the HIV-1 protease structure undergoes significant conformational changes upon ligand binding, it represents a challenging target for docking approaches. While the flexibility of protease inhibitors has

been included in different computational docking studies, the efficient and accurate consideration of the protease flexibility is still a challenging task.
It is also possible to combine molecular dynamics (MD) simulations with molecular docking to predict ligand-receptor interactions treating both the ligand and the receptor conformation flexible. Such approach has been applied to the HIV-1 protease system [177]. However, since it is necessary to run one or several computationally demanding MD simulations for each putative ligand candidate, the approach is prohibitively expensive if one needs to screen hundreds or even thousands of putative drug candidates. Huang et al. introduced an ensemble docking method that allows simultaneous optimization of placement and receptor conformation out of a set of protein structures and found significant improvement compared to docking to single structures [138]. This method was successfully applied by the same authors to an ensemble of NMR (nuclear magnetic resonance)-derived structures of HIV-1 protease [178]. The application of ensembles of HIV-1 protease crystal structures and from MD simulations has also been shown to improve docking results [179].

In the present chapter, the ReFlexIn receptor-ligand docking approach was used for docking a series of putative ligands to HIV-1 protease into an ensemble of different bound conformations, as well as to a set of structural models obtained by morphing the protease apo structure towards one bound structure.

5.2 Bound Receptor Structure Ensemble

The conformational changes involving the flap regions within the binding site of HIV-1 Protease upon binding cannot be properly sampled by means of normal mode analysis. Hence, for this target protein, the same deformation interpolation scheme as in the previous examples (PKA, CDK2) was used, but instead of employing deformations derived by ENM, different bound forms of the HIV-1 Protease were used as intermediate steps for the flexible docking. The seven bound forms are derived from the crystal structures with the PDB codes 1AJV, 1DMP, 1G2K, 1HVH, 1HWR, 2UPJ, and 7UPJ, all comprising HIV-1 Protease with different bound ligand molecules. Figure 5.1 demonstrates the large conformational changes of the closing flap regions as well as the minor side chain differences between different bound forms.

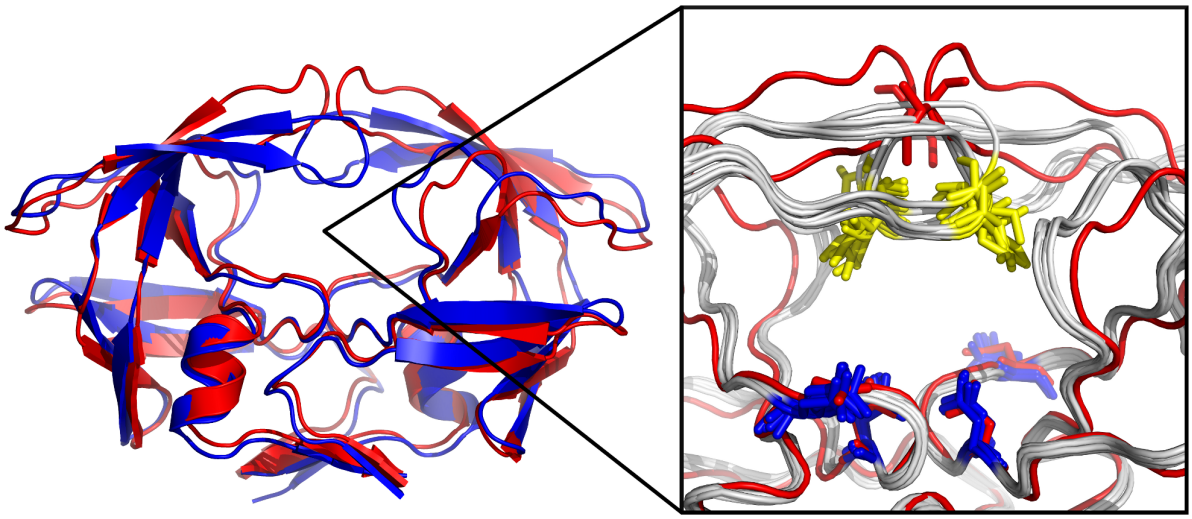


Figure 5.1: *Cartoon representation of unbound (red, PDB ID 3HVP) aligned to a bound receptor structure of HIV-1 Protease (blue, PDB ID 7UPJ). Upon ligand binding, the closure of the two flaps, narrowing the binding site is clearly visible.*
Picture detail: close-up view of the binding site. Apo HIV-1 Protease (red) aligned to 7 different bound forms of HIV-1 Protease shown in grey (ligands not shown). Four flexible side chains (yellow: Ile'50 residues and blue: Asp'25 of both monomers of the homo-dimeric protease structure) at the substrate and inhibitor binding site are represented as stick structures (coloured red in case of the apo structure).

true binders			foreign binders			non-binders
PDB ID	ligand name	$RMSD_{Protein}$ vs. apo	PDB ID	ligand name	$RMSD_{Protein}$ vs. apo	
1AJV	NMB	1.68 Å	1AJX	AH1	1.79 Å	ARA, CEL,
1DMP	DMQ	1.67 Å	1BV9	XV6	1.77 Å	GAL, LSN,
1G2K	NM1	1.66 Å	1G35	AHF	1.78 Å	OLM, SAM,
1HVH	Q82	1.53 Å	1OHR	1UN	1.16 Å	ZAF, ZED,
1HWR	216	1.62 Å	1PRO	A88	1.67 Å	ADN, 107,
2UPJ	U02	1.80 Å	1QBU	846	1.68 Å	U55
7UPJ	INU	1.75 Å	1T7K	BH0	1.95 Å	

Table 5.1: *The three different ligand test sets employed to test the flexible receptor docking approach. True and foreign binder ligand test sets with the PDB ID of the receptor ligand complex, the corresponding ligand identifier, and the C_α-binding site RMSD of the respective bound protein structure vs. the apo HIV-1 Protease (PDB ID 3HVP)*

In Table 5.1, the test sets of different ligands and their respective bound protein structure are listed. Throughout this chapter, ligand molecules will be differentiated between groups of true, foreign, and non-binders.

True binders are ligands that are taken from the crystal receptor structures that are actually within the ensemble that is used for the flexible receptor docking.

Foreign binders are HIV-1 Protease binding ligands as well, but are taken from different HIV-1 Protease crystal structures whose bound protein form shows small conformational differences towards the receptors of the true binders.

Non-binders are molecules for which no affinity to bind to HIV-1 Protease is known. However, the non-binders have been chosen such that they contain the same atom types and are of approximately the same size as the true and foreign binders. The non-binder test set is explained in more detail in the respective section below.

Firstly, only the true binders will be considered, foreign and non-binders at the respective later sections. Figure 5.2 shows the chemical structures of the seven true binder ligands.

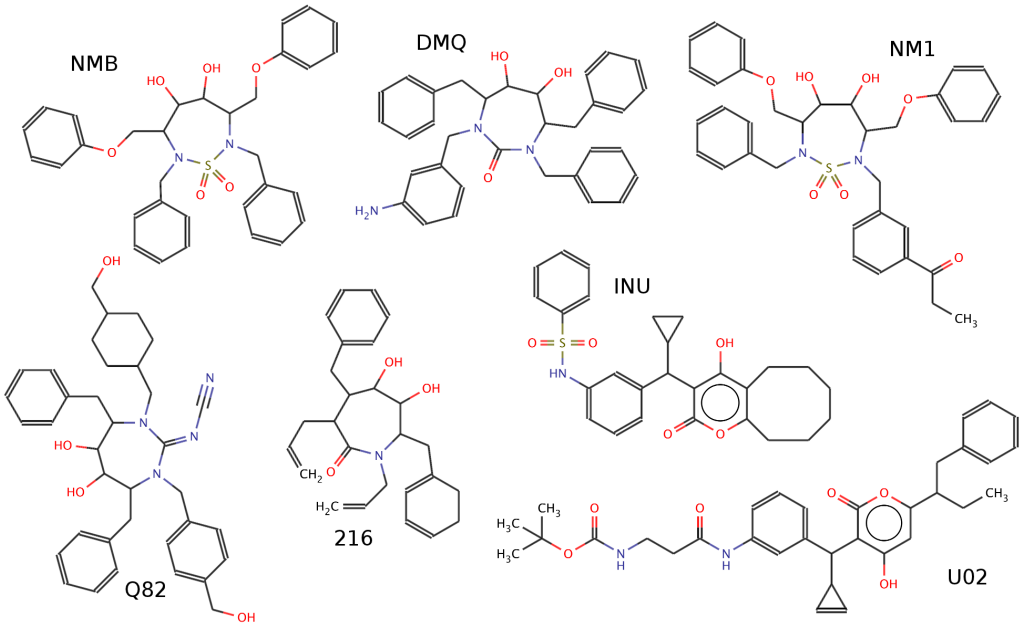


Figure 5.2: *Chemical structures of the HIV-1 Protease true binder test set.*

Ordering of the Bound Structures

The interpolation routine that enables us to access also intermediate states between two structures has the requirement that the 'deformations' of the ensemble are ordered in a way such that similar structures are neighbours.

While elastic network models return deformations that are regularly arranged around the input structure, the seven picked bound forms of HIV-1 Proteases must be manually ordered.
Here, the ordering is solved as follows: all-atom $\mathrm{RMSD}_{Protein}$ values have been calculated between each possible set of two bound structures inside the ensemble, where only the atoms in a radius of 13Å around the binding site have been accounted for.
All possible permutations to form a pathway going through the seven structures (this equals 7!=5040) are being examined for their total sum of RMSD differences and finally, the shortest path that includes the sequence with the smallest possible sum of RMSD has been determined. This order – called deformation 1 to deformation 7 – with given PDB codes of the receptor-ligand complex is given on the right.

def 1 (1HVH)
↕ 0.365Å
def 2 (1DMP)
↕ 0.328Å
def 3 (1HWR)
↕ 0.468Å
def 4 (2UPJ)
↕ 0.402Å
def 5 (1AJV)
↕ 0.17Å
def 6 (1G2K)
↕ 0.455Å
def 7 (7UPJ)

Rigid Receptor Holo Docking

Serving as a first test to validate the results and the scoring function of the used docking program, the ligands of each protein-ligand complex have been removed from the receptor and separately re-docked to their respective native receptor crystal structure (holo docking).
All docking runs in this chapter have been executed using the same options as for the dockings presented in the previous chapter (see section 4.2) The underlined values in Table A.1 of the Appendix as well as Figure 5.3 show how many of 100 separate holo docking runs yield correct docking results. Here, the threshold value for a correct ligand position is an RMSD_{Ligand} below 2Å towards the native ligand position in the crystal structure.
For six out of the seven considered receptor-ligand pairs, the holo docking is able to successfully reproduce the correct placement of the corresponding ligand in all of the 100 separate docking runs. Only the ligand of 1HVH, Q82 fails in holo docking. Inspecting the docking results further shows that this ligand is favourably placed in a position that is the exact upside-down projection of the native placement.
When the ligand is treated as flexible, the holo docking performance does not drop significantly. In 88-100% of the docking runs, the correct ligand position is found, even Q82 can be successfully re-docked to its native receptor structure in one case.

Rigid Receptor Cross Docking

In addition to docking the ligands to their original crystal protein receptor, a cross-docking has been performed where all ligands have been docked to all HIV-1 Protease structures within the receptor ensemble.
Due to the lack of a reference structure of those ligands in a 'foreign' bound form of the protease, the calculation of the ligand RMSD is done as follows: all receptor structures including their ligands have been aligned to one common protease form (apo HIV1-Protease, PDB ID 3HVP), taking into consideration only the 13Å area around the binding site for alignment. For calculation of the RMSD_{Ligand} values, the ligand positions from the aligned structures have been taken as reference for the correct solution.
The results from Table A.1 and Figure 5.3 indicate that there are several ligands as for example NMB, DMQ, or NM1, that yield correct binding conformations in almost all foreign receptor structures. The Q82 ligands is docked in the same upside-down placement as mentioned before, however, it is placed correctly in several different receptor conformations (1AJV, 1G2K).

Rigid Receptor Apo Docking

Here, the 7 ligands of the true binder test set were docked to the apo (unbound) form of HIV-1 Protease (PDB ID 3HVP). The rows denoted 'apo' in Table A.1 list the number of docking runs that found a good ligand placement ($\text{RMSD}_{Ligand} < 2$ Å) and demonstrate that the apo docking returns very poor results. Only 16 of 100 docking runs for the NM1 inhibitor taken from the 1G2K HIV-1 Protease structure yielded an RMSD below 2Å. The rest of the docked ligands cannot be found at the correct binding site, and preferentially bind in other cavities of the apo form that is – compared to the bound HIV1-Protease structure – in a more opened conformation.

Flexible Receptor Docking

In the following, flexible receptor docking means that during the docking, the algorithm is able to switch and interpolate between the seven bound receptor structures of HIV-1 Protease that are ordered as described above. Table 5.2 shows the results of the flexible receptor docking and the rigid receptor apo docking. Listed are the best RMSD_{Ligand} values and the RMSD_{Ligand} at the lowest predicted AutoDock Binding energy (the best scored solution) of 100 separate docking runs for each ligand.

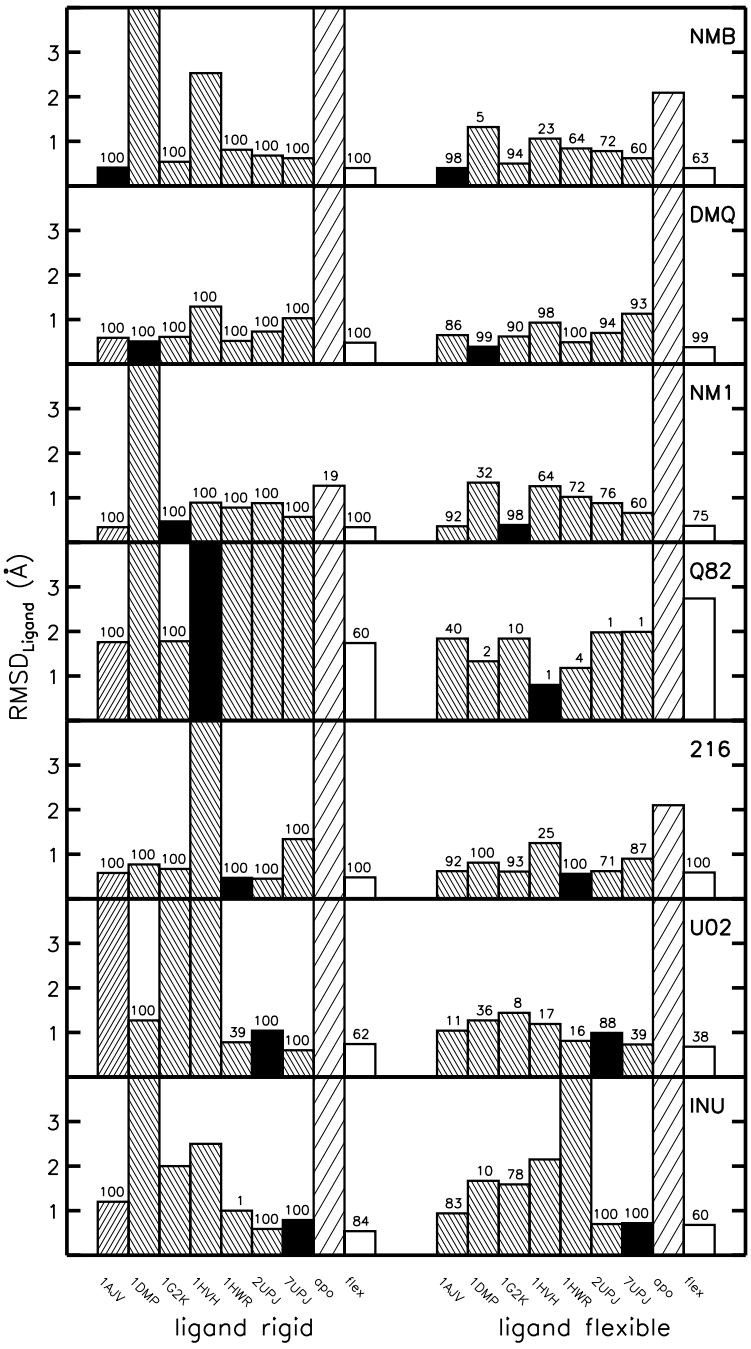


Figure 5.3: *Docking results in terms of deviation of the docked ligand from the native placement for the rigid receptor cross docking, apo docking, and flexible receptor docking of the HIV-1 protease true binder test set. Bar height indicates the best RMSD solution found in 100 separate docking runs (values are cut at 4Å). Narrow shaded bars are the dockings of the ligand denoted in the upper right corner of each plot into the rigid receptor structure as indicated in the bottom. Black filled bars stand for the holo docking, i.e. re-docking of ligands into their original bound protein structure. Wide shaded bars show the results of the rigid receptor apo docking, non-shaded bars for the flexible receptor docking. Numbers on top of bars indicate the number of dockings per 100 separate docking runs that yield ligand RMSD values below 2Å .*

	ligand rigid				ligand flexible			
	flex.docking		apo docking		flex.docking		apo docking	
	best RMSD	low.en. RMSD	best RMSD	low.en RMSD	best RMSD	low.en. RMSD	best RMSD	low.en RMSD
NMB	0.40	0.42	7.54	7.74	0.40	0.40	2.09	7.58
DMQ	0.48	0.51	7.73	7.74	0.38	0.42	6.14	7.87
NM1	0.34	0.35	1.27	10.34	0.37	0.42	6.45	7.38
Q82	1.74	5.50	5.40	5.59	2.74	5.60	5.42	5.97
216	0.48	0.48	6.30	6.45	0.59	0.71	2.10	6.73
U02	0.74	1.08	9.82	9.86	0.68	0.99	5.57	9.23
INU	0.54	0.60	9.06	9.07	0.68	0.83	6.02	6.50

Table 5.2: *Best RMSD, i.e. the lowest yielded $RMSD_{Ligand}$ out of 100 separate docking runs (given in Å), and the RMSD values of the docking solution with the lowest AutoDock binding energy for different ligand and receptor flexibilities.*

While only one result of the rigid apo docking shows an $RMSD_{Ligand}$ smaller than 2 Å , the ReFlexIn based docking performs significantly better with very low best-RMSD values, irrespective of the ligand flexibility.
In addition, the results with the lowest binding energies are equal or very close to the best-RMSD solutions. This becomes also visible in Figure 5.4, where the $RMSD_{Ligand}$ is plotted against the AutoDock binding energy for each of the 100 separate docking runs. For example, the flexible receptor docking of the ligand NMB with four rotatable bonds yields three large RMSD clusters around 1, 6, and 9Å . However, the low RMSD solutions exhibit considerably better binding energies, and can be clearly distinguished from the high-RMSD solutions by a 3-4 kcal/mol difference.

In all cases, the flexible receptor docking yields results with significantly better $RMSD_{Ligand}$ values compared to the rigid receptor docking to the apo form. Additionally, the binding energy of the flexible receptor docking is considerably lower (approx. 5-8 kcal/mol) which makes the low-RMSD results clearly distinguishable. Allowing the ligand to be moderately flexible results in a larger number of different RMSD clusters, as already observed before. Comparing the results of the rigid ligands to the ligands having 4 flexible torsions, the flexible ligands results tend to adopt slightly higher binding energies. This is due to AutoDocks treatment of internal ligand energy where to each result the value of approx. 0.3 kcal/mol per flexible torsion is added to the binding free energy estimate.

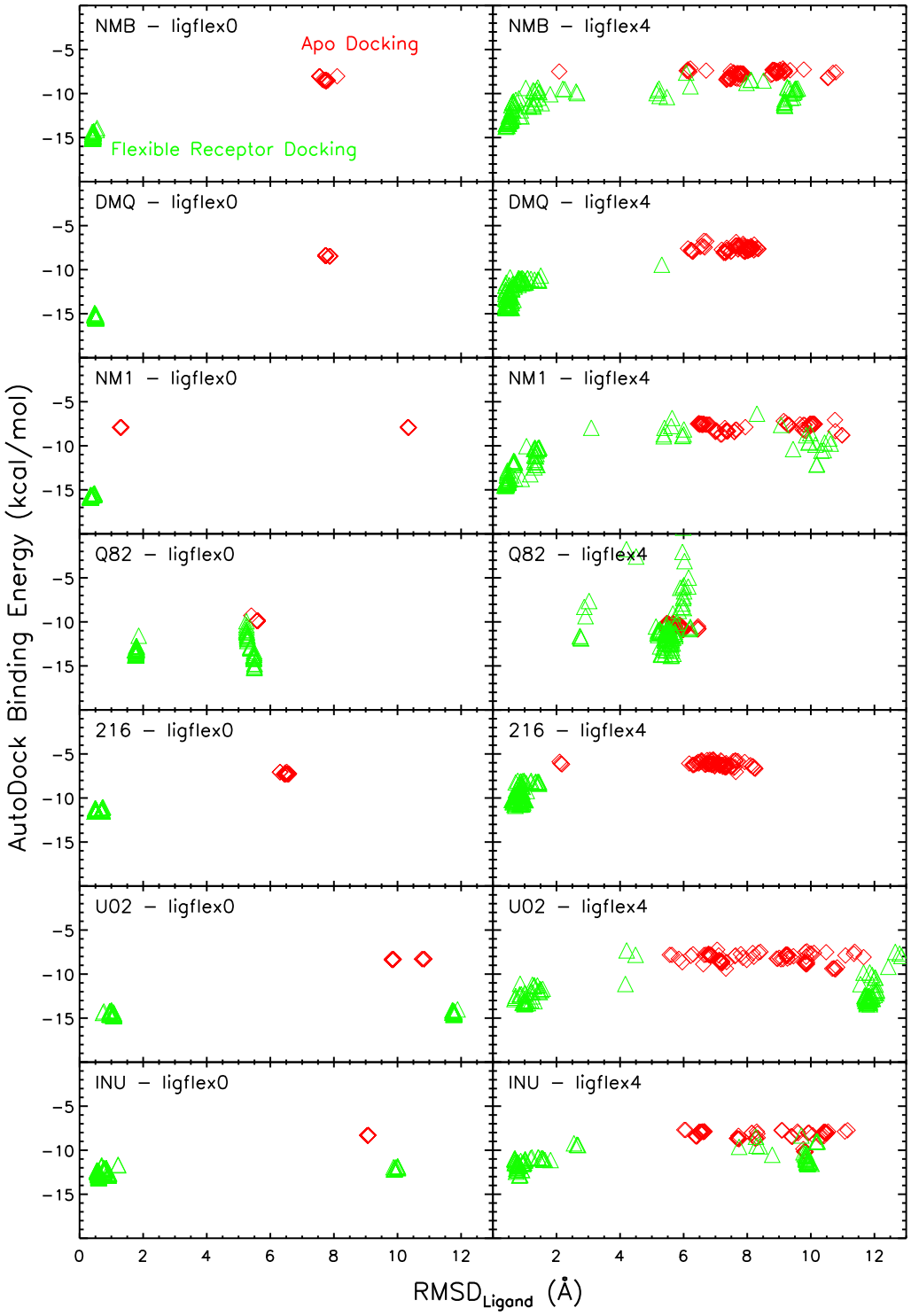


Figure 5.4: *Calculated Autodock Binding energies plotted against the yielded $RMSD_{Ligand}$ for 100 separate docking runs of rigid receptor apo docking (red diamonds) versus flexible receptor docking (green triangles) using seven bound HIV-1 Protease structures for interpolation. Plots on the left show the results for the ligand being rigid, plots on the right with the ligand having moderate flexibility of 4 rotatable torsions.*

Receptor Structure Deformation (Lambda)

For each separate docking, the flexible receptor docking algorithm additionally returns the lambda value for the respective docking result. This value indicates into which bound structure – or intermediate between bound structures for non-integer values – the ligand was docked.

When using the bound structures as the input for the flexible docking, one interesting test is to check whether the ligands have been docked to the correct (the original bound structure from which they were extracted) or a near-correct receptor conformation.

Figure 5.5 shows the distribution of lambda values for the flexible receptor docking runs and specifies exactly how many of the 100 separate solutions have been assigned to a correct receptor structure. Considering only the shaded areas (i.e. the correct lambda values), in the great majority of cases one can can find very low RMSD solutions with a correct or near-correct ligand placement. In these cases the ligand was docked into a receptor structure that is geometrically very similar to the original receptor structure, thus the algorithm is correctly selecting the structure out of the ensemble. The ligand from 1DMP achieves the best results for rigid and flexible ligand with 100 and 96% of correct lambda values, respectively. Again, ligand Q82 fails with low RMSD solutions in the correct lambda area and solutions with an RMSD_{Ligand} or around 2Å adopting lambda values between 5 and 6. This is in agreement with the results of the cross docking above (see Table A.1) where the docking of ligand Q82 into deformation 5 and 6 (1AJV and 1G2K) yielded the best results.

For the ligands NMB, DMQ, 216, and INU, the shaded areas are mostly occupied by solutions that exhibit very low RMSD_{Ligand} values. Only very few high-RMSD solutions have been assigned to a correct lambda value (one for DMQ flexible, four for NM1 flexible, two for INU flexible).

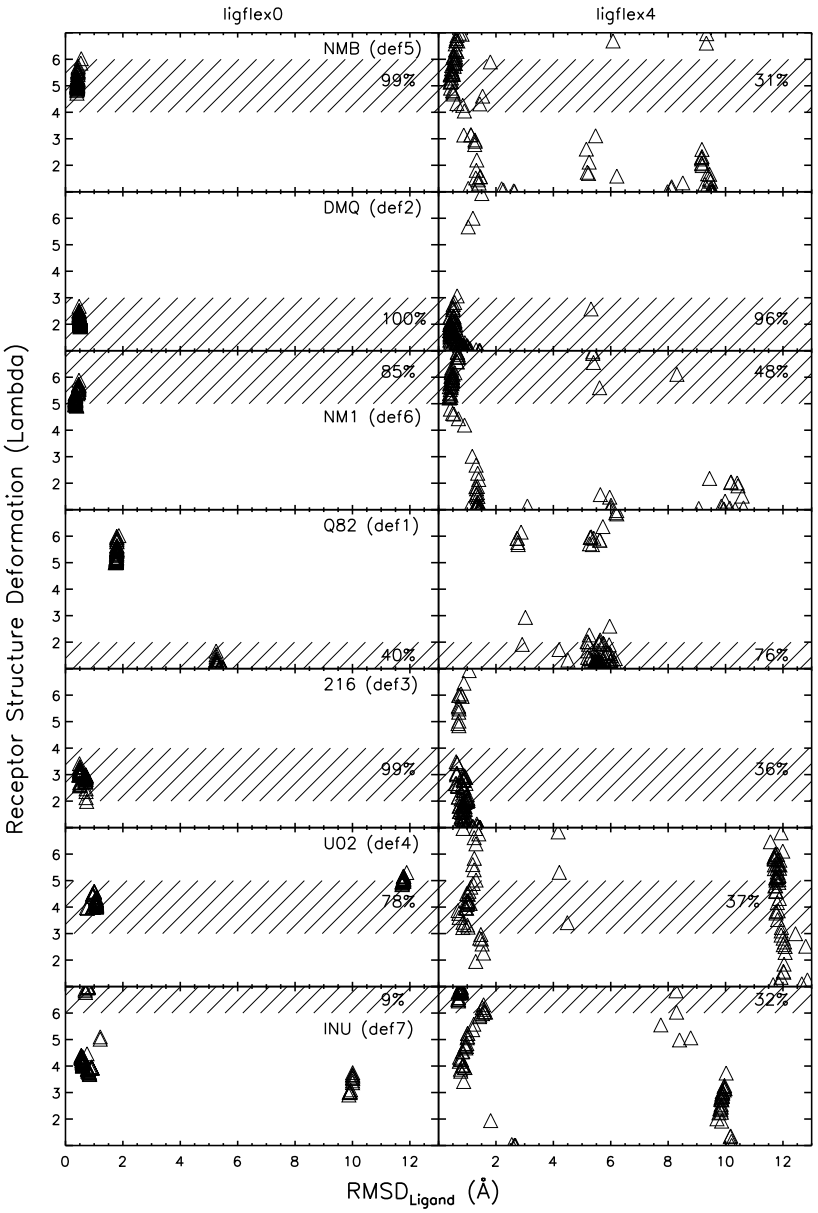


Figure 5.5: *HIV-protease structure vs. $RMSD_{Ligand}$ for flexible receptor docking applied to the true binder test set. Variation of the parameter lambda corresponds to a continuous deformation of the HIV protease receptor structure beginning at the bound form PDB ID 1AJV (lambda=1) and following the order 1DMP, 1G2K, 1HVH, 1HWR, 2UPJ and ending with 7UPJ (lambda=7) which represent a minimum RMSD pathway. $RMSD_{Ligand}$ corresponds to the deviation of the 100 docked ligand coordinates from the native structure after best superposition of the receptor structure. Correlation between ligand placement and receptor structure are shown for both treating the ligand as rigid bound conformer of flexible (allowing bond rotation). The shaded areas are drawn to indicate the expected receptor deformation close to the conformation found in the crystal structure of the ligand in complex with HIV-1 protease. Numbers inside the shaded area give the percentage of correctly assigned lambda value for all 100 docking runs.*

Flexible Receptor Docking with the correct Receptor not included

To address the question of whether an approach that includes the actual correct bound receptor structure of a ligand in the set of deformations contains too much information on the bound form, a flexible receptor docking is tested, where the respective correct bound form of the protein is not included in the ensemble of bound structures. In this case, the set of deformations (bound structures) contains only six protein structures per flexible docking run.

The results are shown in Table 5.3 and indicate that leaving out the correct receptor structure has only very moderate effect on the RMSD_{Ligand} values. Only the docking of the rigid ligand U02 fails to find any solutions with an $\text{RMSD}_{Ligand} < 2.0\text{Å}$. Still, for the rest of the dockings with the correct bound structure not in the ensemble, moderate to large amounts of low-RMSD solutions are found.

	ligand rigid			ligand flexible		
	bound-all	bound-csm	apo	bound-all	bound-csm	apo
NMB	100	100	0	63	44	0
DMQ	100	100	0	99	99	0
NM1	100	99	19	75	83	0
Q82	60	96	0	0	1	0
216	100	100	0	100	98	0
U02	62	0	0	38	22	0
INU	84	84	0	60	40	0

Table 5.3: *Percentages of docking results with an $RMSD_{Ligand} < 2.0Å$ when comparing the flexible receptor dockings with all bound HIV1-Protases structures used (bound-all) and using all bound but the correct structures (bound-csm, csm stands for 'correct structure missing')*

Flexible Receptor Docking of foreign HIV1-Protease Ligands

Seven known HIV1-Protease binders different from the ligands taken from the first set of bound structures have been chosen (see also Table 5.1). In the following, those ligands will be referred to as 'foreign ligands' or 'foreign binders' as they are known to be actual HIV1-Protease binders, but are taken from bound structures that exhibit small conformational differences compared the 'true binder' receptor structures.

The foreign binder test set are: ligand AH1 from PDB ID:1AJX, ligand XV6 from PDB ID:1BV9, ligand AHF from PDB ID:1G35, ligand 1UN from PDB ID:1OHR, ligand A88 from PDB ID:1PRO, ligand 846 from PDB ID:1QBU, and ligand BH0 from PDB ID:1T7K. The chemical structures of the ligands are shown in Figure 5.6.

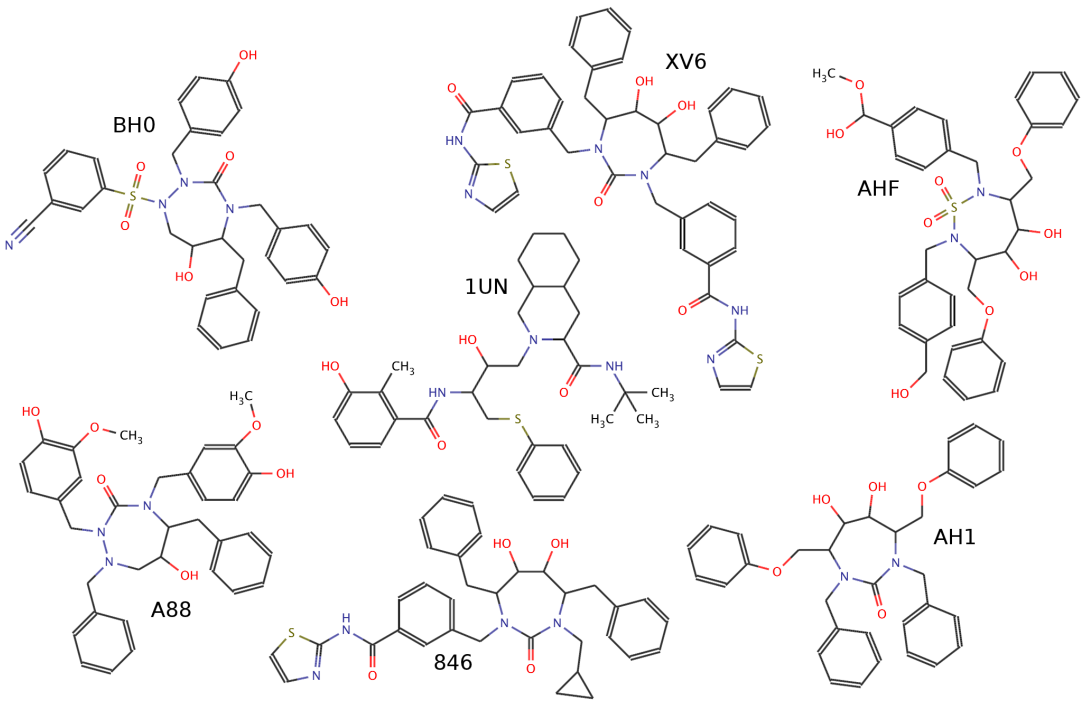


Figure 5.6: *Chemical structures of the HIV-1 Protease foreign binder test set.*

Figure 5.7 shows the comparison of docking results for rigid receptor cross docking versus the flexible receptor method.

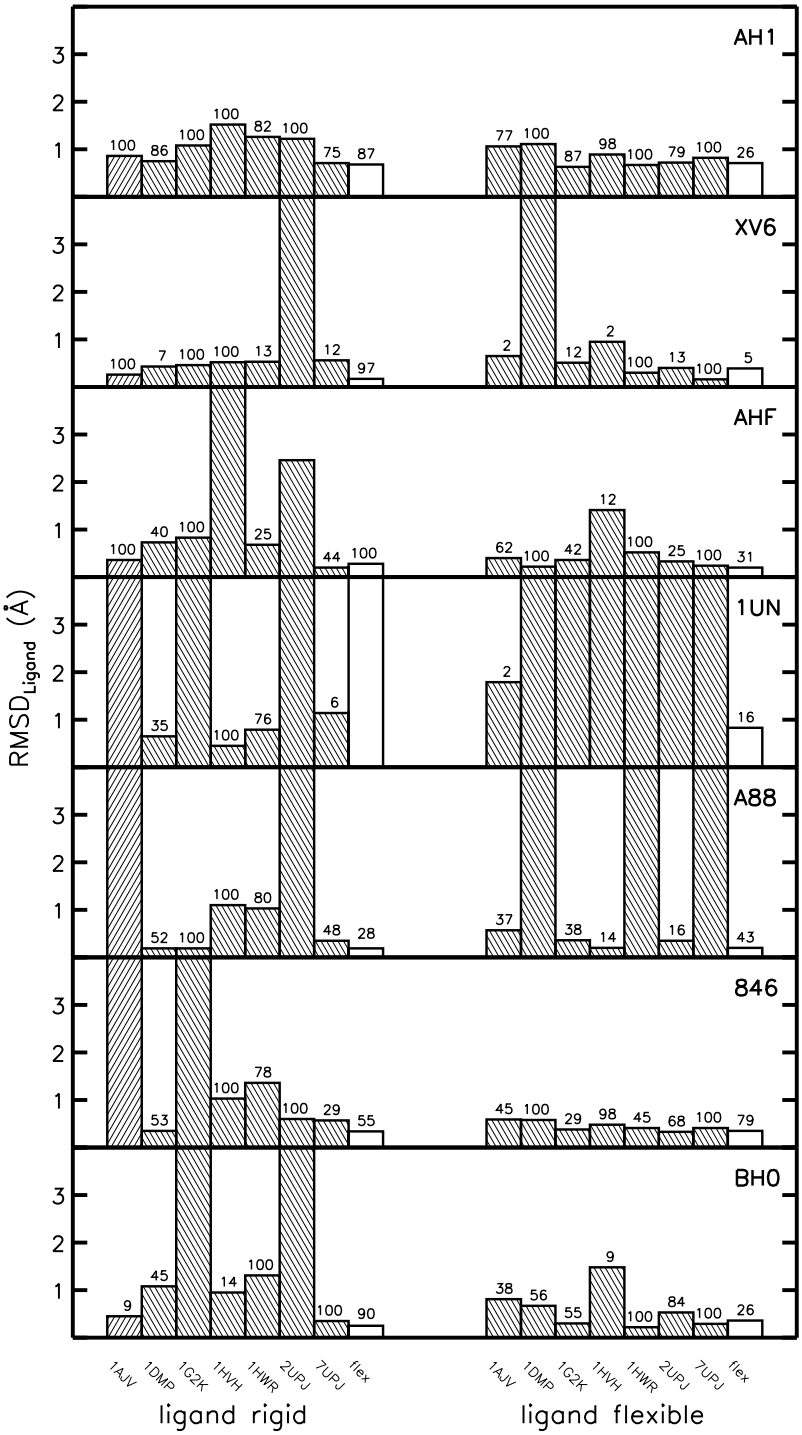


Figure 5.7: *Docking results in terms of deviation of the docked ligand from the native placement for the rigid receptor cross docking, apo docking, and flexible receptor docking of the HIV-1 protease foreign binder test set. Bar height indicates the best RMSD solution found in 100 separate docking runs (values are cut at 4Å). Narrow shaded bars are the dockings of the ligand denoted in the upper right corner of each plot into the rigid receptor structure as indicated in the bottom. Wide shaded bars show the results of the rigid receptor apo docking, non-shaded bars for the flexible receptor docking. Numbers on top of bars indicate the number of dockings per 100 separate docking runs that yield ligand RMSD values below 2Å .*

With this cross docking being conducted on a rigid receptor, one does not have to distinguish between best RMSD and best energy RMSD because the energy differences of the 100 separate docking runs are very low, in the range of $\sim$ 0.1 kcal/mol. (see also the constant energy levels for the rigid receptor apo docking results in Figure 5.4).
The differences between best and worst energy for the flexible receptor docking are between 0.19 and 1.48 kcal/mol. Therefore the best energy RMSDs are also given in brackets after the best RMSD_{Ligand} in the appendix Table A.2.
For rigidly treated ligands 846 and BH0 and flexibly treated ligands AH1 and BH0, the best energy RMSDs are considerably larger. However, for those cases, better RMSD solutions are found that are also scored very high. For example, at only 0.15 to 0.3 kcal/mol higher score, solutions are in the range of the best RMSD values.

Binders versus Non-Binders

To test the potential of the ReFlexIn method to distinguish between binders and non-binders, a test set of eleven ligands that have no known binding affinity to HIV-1 Protease was established. Eight non-binder molecules are taken from a ligand set that Fanfrlík and colleagues employed for the evaluation of a rescoring scheme for the docking of several ligands (binders and nonbinders) to HIV-1 Protease [180]. Again, non-binder ligands were chosen such that included atom types are comparable to the true and foreign binders. The selected ligands are ARA, CEL, GAL, LSN, OLM, SAM, ZAF, and ZED as shown in Figure 5.8. Ligand molecules were built from scratch using the PRODRG server [181]. 500 energy minimization steps in Chimera were applied to each ligand for adequate relaxation and removal of unfavourable energies evoked from molecule building. Three additional non-binder molecules were chosen from the known PKA and CDK2 ligands as used in chapter 4 (ligands from PDB IDs 1FMO, 1FVV, and 1JSV).

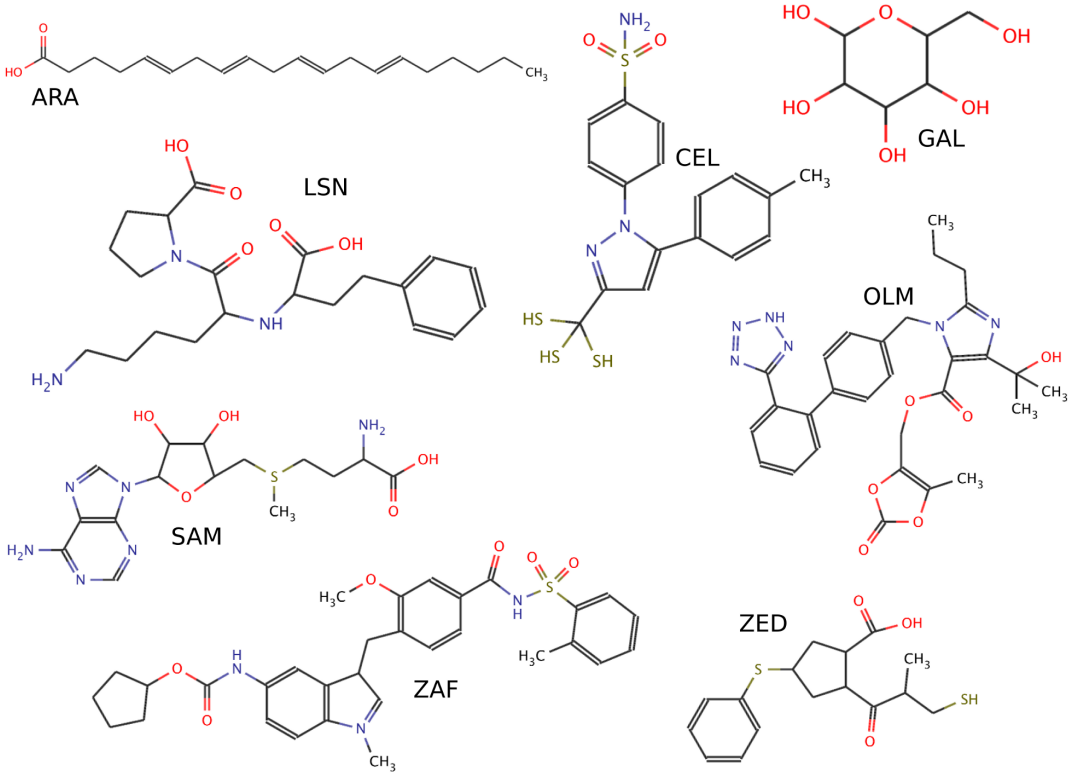


Figure 5.8: *Chemical structures of the HIV-1 Protease non-binder test set.*

In this testing scenario, all ligands were docked flexible with 4 rotatable bonds per ligand. AutoDock accounts for internal ligand energy by adding approximately 0.3 kcal/mol for each rotatable bond, as discussed before. Hence, for a consistent binding energy comparison of different results, it is important to consider examples of identical ligand flexibility .

The best AutoDock binding energies from 100 separate flexible receptor docking runs of several flexible ligands (binders and non-binders) are shown in Figure 5.9. This figure should only show the trend, hence, the energies are not assigned to the respective ligand but are sorted from favourable to unfavourable energy values.

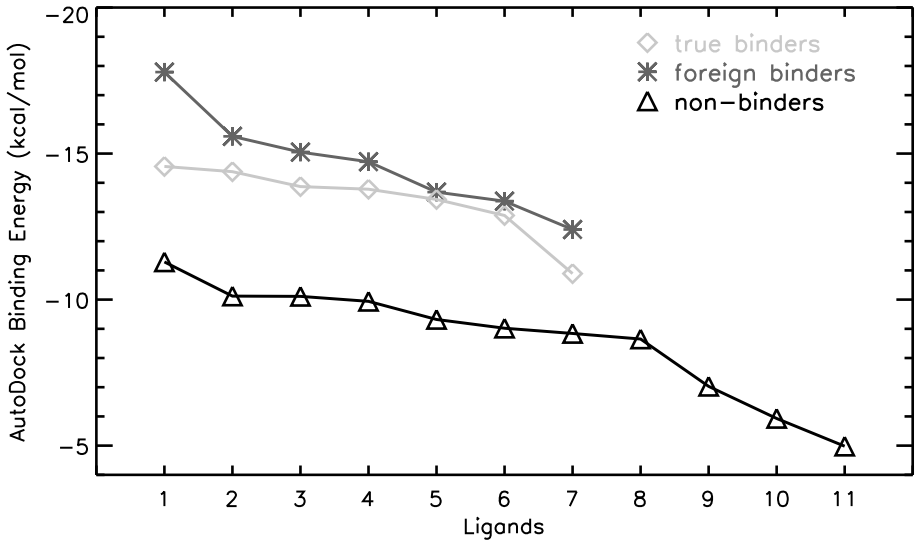


Figure 5.9: *AutoDock Binding energies for the flexible receptor docking approach when docking true binders (diamonds), foreign HIV1-Protease binders (stars), and non-binders (triangles).*

Figure 5.9 demonstrates that the flexible docking approach is able to distinguish between ligands that have a high potential to bind to the active site of HIV-1 Protease and non-binding molecules. The binding energies for the true and foreign binders that are effective HIV1-Protease binding ligands range between -17.8 and -10.9 kcal/mol. Non-binder molecules yield significantly worse energies up to -5 kcal/mol. Only one non-binder (ZAF) yields an energy value that is better scoring than the worst scoring binder molecule (-11.3 kcal/mol of ZAF vs. -10.9 kcal/mol of DMQ). These findings imply a significantly decreased predicted binding affinity for the non-binders.

The bottom plot of Figure 5.10 gives the best AutoDock binding energies in better detail for each of the flexible receptor docking of each ligand of the true, foreign, and non-binder test set. In addition to the flexible receptor docking results, the best binding energies of the ligand docking into the separate rigid deformations (bound structures) are given. The dotted line in each subplot of Figure 5.10 represents the lowest binding energy yielded by the non-binders and thus serves as a reference point to check for binding energy differentiation between protease binders and non-binders. Already with this strict criterion (the other non-binders show significantly higher binding energies), most true and foreign binders can be distinguished from the decoys, both in the rigid receptor dockings into deformation 1 to 7 as well as in the flexible receptor docking. Here, 13 out of 14 binders are clearly discriminated from the non-binders by at least 1.5 kcal/mol. Comparison of the performance of the flexible receptor docking over the rigid receptor docking is aided by the cross symbols in Figure 5.10. As they represent the best binding energies for the true and foreign binders for flexible receptor docking to the single deformation docking plots, it is observed that the flexible docking is generally able to pick the lowest (most favourable) binding energies.

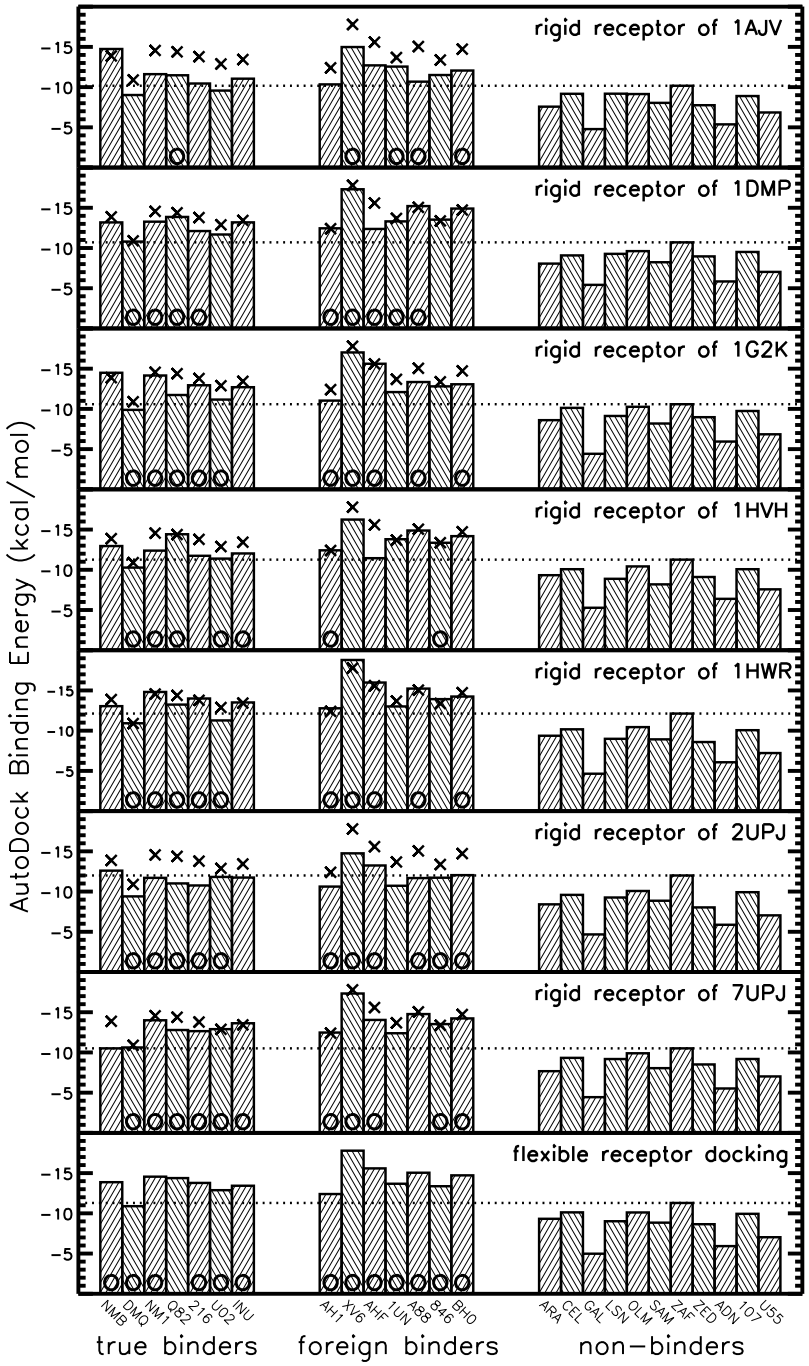


Figure 5.10: *Best scoring results for docking of true, foreign and non-binders upon docking to 7 different rigid receptor structures and using flexible receptor docking (using interpolation between the 7 receptor structures). The dotted line marks the best binding energy obtained for the non-binder molecules (corresponds to the limit for distinguishing binders from non-binders). The cross symbols indicate the results of flexible receptor docking and are included for comparison for each docking to a rigid protease receptor (upper seven rows). The circle symbols are present at the bottom of each bar, if the respective best energy docking solution also yields a* $RMSD_{Ligand}$ *below* 2Å *.*

5.3 NMR-derived Structures

For HIV1-Protease in complex with the DMP323 inhibitor, a set of 28 NMR structures is available under the PDB accession code 1BVE [182] as shown in Figure 5.11. This solution structure might give an idea about apparent structural changes in the ligand-bound receptor structure and will be used as another source of deformations to be employed in our interpolation routine.

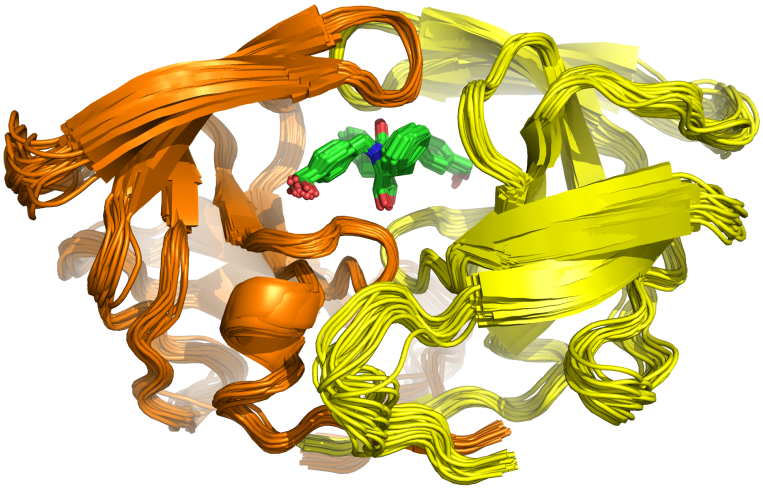


Figure 5.11: *NMR structures of HIV1-Protease, PDB ID 1BVE. Subunits shown in orange and yellow cartoon, ligand in green stick model.*

To this end, the 28 structures have been separated and stripped from their ligands. In contrast to the 7 bound HIV1-Protease structures as considered in the previous sections, the ordering of 28 structures is too complex to be solved by a permutation-based distance matrix (the factorial of 28 yields approx. $3{\times}10^{29}$ permutations).
Instead, the mean structure of all 28 receptor conformations has been calculated using the NMR-structure superimposition program *suppose* [183]. Subsequently, all-atom RMSDs have been calculated between this mean structure and all 28 NMR configurations. Here, the same criterion has been used as previously, i.e. considering only atoms that are in 13Å range of the inhibitor binding site.

	ligand rigid				ligand flexible			
	flex.docking		apo docking		flex.docking		apo docking	
	best RMSD	low.en. RMSD	best RMSD	low.en RMSD	best RMSD	low.en. RMSD	best RMSD	low.en RMSD
NMB	1.40	9.56	7.54	7.74	1.60	8.57	2.09	7.58
DMQ	1.24	1.63	7.73	7.74	1.81	2.12	6.14	7.87
NM1	1.75	10.70	1.27	10.34	1.43	9.49	6.45	7.38
Q82	0.95	1.00	5.40	5.59	1.66	5.81	5.42	5.97
216	1.23	1.82	6.30	6.45	1.56	1.69	2.10	6.73
U02	1.44	12.00	9.82	9.86	1.36	11.31	5.57	9.23
INU	1.81	10.17	9.06	9.07	1.34	10.01	6.02	6.50

Table 5.4: *Best RMSD, i.e. the lowest yielded $RMSD_{Ligand}$ out of 100 separate docking runs (given in Å), and the RMSD values of the docking solution with the lowest AutoDock binding energy for different ligand and receptor flexibilities.*

Several structures that showed only very small binding site RMSD differences towards the mean structure in the range of 0.01 – 0.05Å have been removed from the set, leaving a final number of 15 "deformations" that are ordered according to the distance to the mean structure and are employed as deformations for the ensemble of the following flexible receptor docking.

The results comparing the flexible receptor to the rigid receptor docking performance are shown in Table 5.4 and Figure 5.12. Except in the case of rigid ligand NM1, the flexible receptor docking based on NMR structures is able to outperform the rigid docking using only the apo form protease. The majority of best RMSD solutions for the flexible receptor docking are well below the best RMSD values yielded for the rigid receptor docking, comparable to the previously presented appraches. Good results are found for the ligand DMQ which bears resemblance to the original ligand DMP323 of the experimental NMR structure. Surprisingly, rigidly treated ligand Q82, which showed good results in the previous approaches only very rarely, yields near-native placements for all 100 separate docking runs. Those conformations, however, yield a score that is worse than the score of solutions with a larger RMSD (as e.g. seen for the cluster of Q82 with 4 flexible torsions). Figure 5.12 supports that, in contrast to the bound structure and morphing flexible docking approach, the scoring of the present results is not able to safely distinguish good $RMSD_{Ligand}$ results.

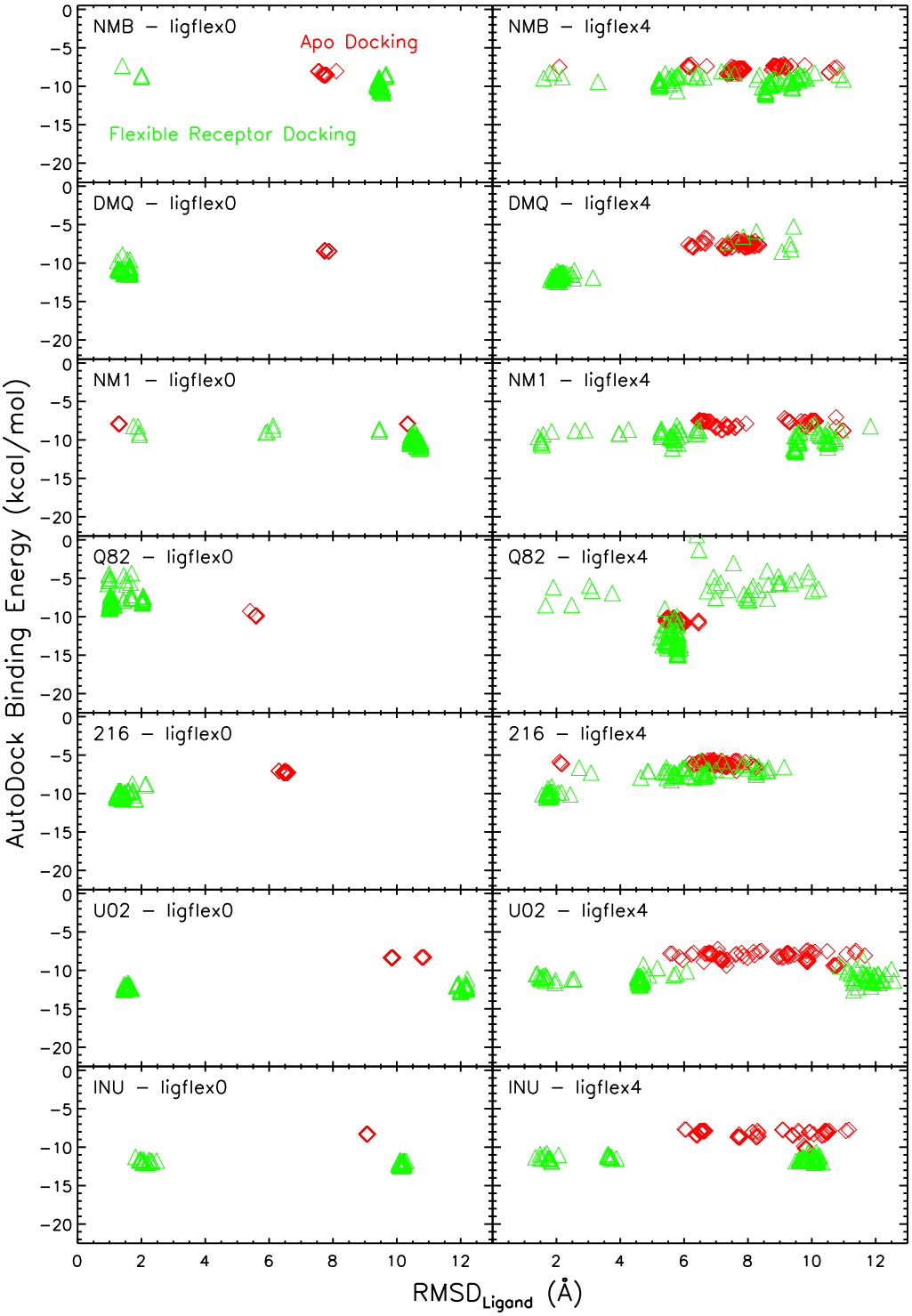


Figure 5.12: *Calculated Autodock Binding energies plotted against the yielded $RMSD_{Ligand}$ for 100 separate docking runs of rigid receptor apo docking (red diamonds) versus flexible receptor docking (green triangles) using the NMR structures from PDB ID 1BVE. Plots on the left show the results for the ligand being rigid, plots on the right with the ligand having moderate flexibility of 4 rotatable torsions.*

5.4 Morphing between bound and unbound Structures

Another possible source of deformation structures can be the morphing between an unbound(apo) and a bound protein structure. Because only those two structures are necessary, such an approach might be especially fruitful if there is a lack of several resolved bound structures of a protein.
A morphing approach has been tested for the HIV1-Protease receptor. The input for the morphing are the PDB ID 3HJV (apo form) and a bound protease structure with the PDB ID 2UPJ. Five intermediate structures have been created using the linear corkscrew morphing approach of the UCSF Chimera program [173]. To remove possible atomic overlaps and resulting unphysical potential energies, a short minimization step for each of the generated intermediate steps was applied. The resulting structures are depicted in Figure 5.13. The computational effort to generate these intermediate

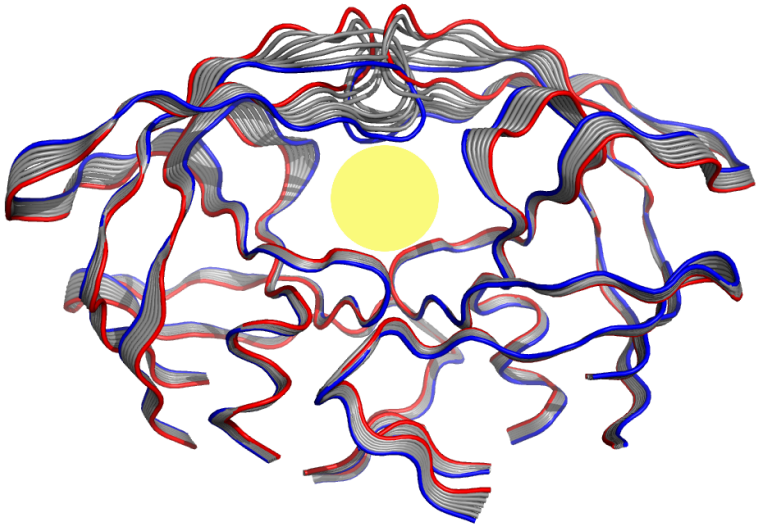


Figure 5.13: *Tube representation of the morphing intermediate structures (grey) between the apo HIV-1 protease (red, PDB ID 3HJV) and one bound structure (blue, PDB ID 2UPJ). The location of the inhibitor binding site is marked with a yellow circle.*

structures is approximately two seconds on a single CPU and can thus be neglected. Generated intermediate steps are now being employed as the deformations in the flexible receptor docking. Here, the naming convention is deformation 1 for the most opened (the apo) state and deformation 7 for the bound protein structure.

	ligand rigid				ligand flexible			
	flex.docking		apo docking		flex.docking		apo docking	
ligand	best RMSD	low.en. RMSD	best RMSD	low.en RMSD	best RMSD	low.en. RMSD	best RMSD	low.en RMSD
NMB	0.50	0.73	7.54	7.74	0.76	1.50	2.09	7.58
DMQ	0.76	0.86	7.73	7.74	0.82	1.23	6.14	7.87
NM1	0.44	0.69	1.27	10.34	0.73	0.73	6.45	7.38
Q82	5.25	5.26	5.40	5.59	5.20	7.03	5.42	5.97
216	0.58	0.62	6.30	6.45	0.92	1.03	2.10	6.73
U02	1.05	1.09	9.82	9.86	0.99	1.10	5.57	9.23
INU	1.06	1.27	9.06	9.07	0.73	0.99	6.02	6.50

Table 5.5: *Best RMSD, i.e. the lowest yielded $RMSD_{Ligand}$ out of 100 separate docking runs (given in Å), and the RMSD values of the docking solution with the lowest AutoDock binding energy for the morphing approach.*

The $RMSD_{Ligand}$ results for the flexible receptor docking are given in Table 5.5. This table gives the best $RMSD_{Ligand}$ and the $RMSD_{Ligand}$ of the solution with the lowest AutoDock binding energy.

With both no or moderate ligand flexibility, the flexible receptor docking proves to find RMSD values well below 2Å for all ligands but Q82. When the ligand is treated rigid, for five of the seven ligands, the best binding energy RMSDs closely agree with the very low $RMSD_{Ligand}$. When the ligand is allowed to be moderately flexible (4 rotatable bonds), the flexible receptor docking results are comparably good, still staying well below 2Å. The best energy RMSDs are slightly larger compared to the rigid ligand docking but still both are better than the apo docking results and smaller than 2Å. Even though this morphing approach uses the information of only one bound structure of HIV-1 Protease unlike seven bound structures in the approach presented before, the docking results are comparably good. The flexible receptor docking clearly outperforms the rigid receptor apo docking and yields large clusters of low-RMSD solutions as shown in Figure 5.14. This Figure also depicts well that the lowest energy solutions from 100 separate docking runs of each ligand are almost exclusively (ligand Q82 fails) found in the low-RMSD range ($RMSD_{Ligand} < 2$Å). This approach could become helpful in cases where only one bound structure of the receptor is yet known; the missing intermediate structures can be very quickly computed.

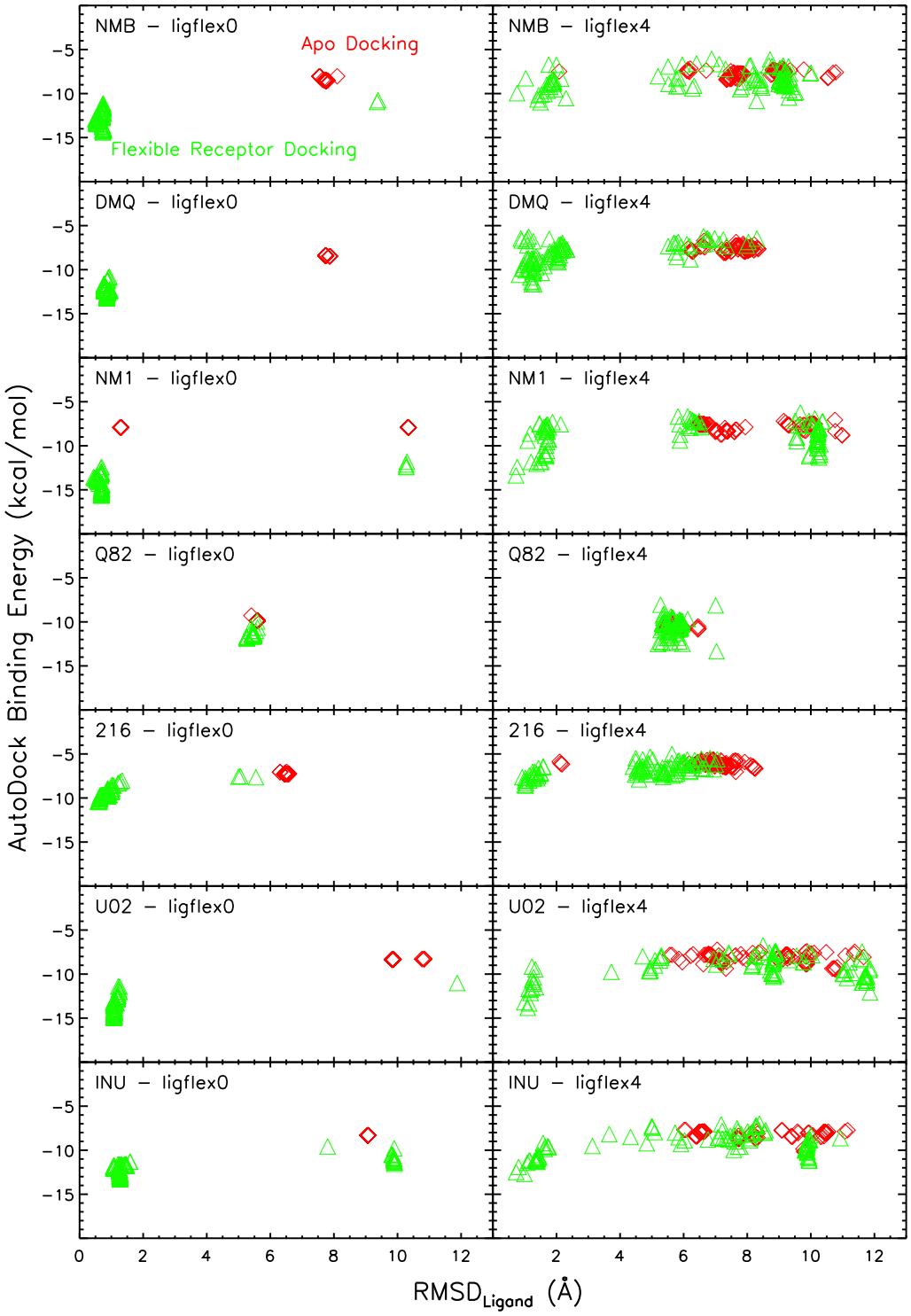


Figure 5.14: *Calculated Autodock Binding energies plotted against the yielded $RMSD_{Ligand}$ for 100 separate docking runs of rigid receptor apo docking (red diamonds) versus flexible receptor docking (green triangles) using the morphing structures between the apo HIV-1 Protease and bound form PDB ID 2UPJ. Plots on the left show the results for the ligand being rigid, plots on the right with the ligand having moderate flexibility of 4 rotatable torsions.*

Receptor Structure Deformation (Lambda)

Also for this approach, the final lambda values of each docking result have been collected and are shown in Figure 5.15. In contrast to the previous seven bound structures approach, where the lambda value changed towards the number of the actual bound form, here, the lambda values are noticeably accumulated around values of 5 to 7. With deformation number 7 being an actual protease bound form (PDB ID 2UPJ), this implies that the flexible receptor docking tends to choose the deformations that are in close conformational resemblance to a real bound structure. As observed before, the results are better clustered when the ligand is docked rigidly. Here, the best RMSD results are found for high lambda values and larger RMSD results are nicely separated at slightly lower lambda values. This can be observed well e.g. for the ligands NM1, 216, and INU. A larger sampling space for ligands with 4 rotatable bonds leads to a larger spreading of results, however, results with a low RMSD_{Ligand} are still found at high lambda values (see ligands INU, U02, and 216).

Effect of bound Structure Choice for Morphing

The above results are based on the morphing from the apo structure towards the bound receptor structure of 2UPJ. With different bound structures available, the question rises as to what extent the choice of that target structure has an impact on the results. To this end, the flexible receptor docking is tested with input structures that are derived from the morphing towards four different bound structures. As shown in Table 5.1, the two bound forms with the largest deviation towards the apo structure are 7UPJ and 2UPJ. 1HWR and 1HVH are the structures with the least deviation from the apo structure. Testing of the respective two outer extremes seemed sufficient, as the intermediate morphing steps as well as the interpolation between those deformations adequately samples enough. The results are shown in Table A.3 in the Appendix.

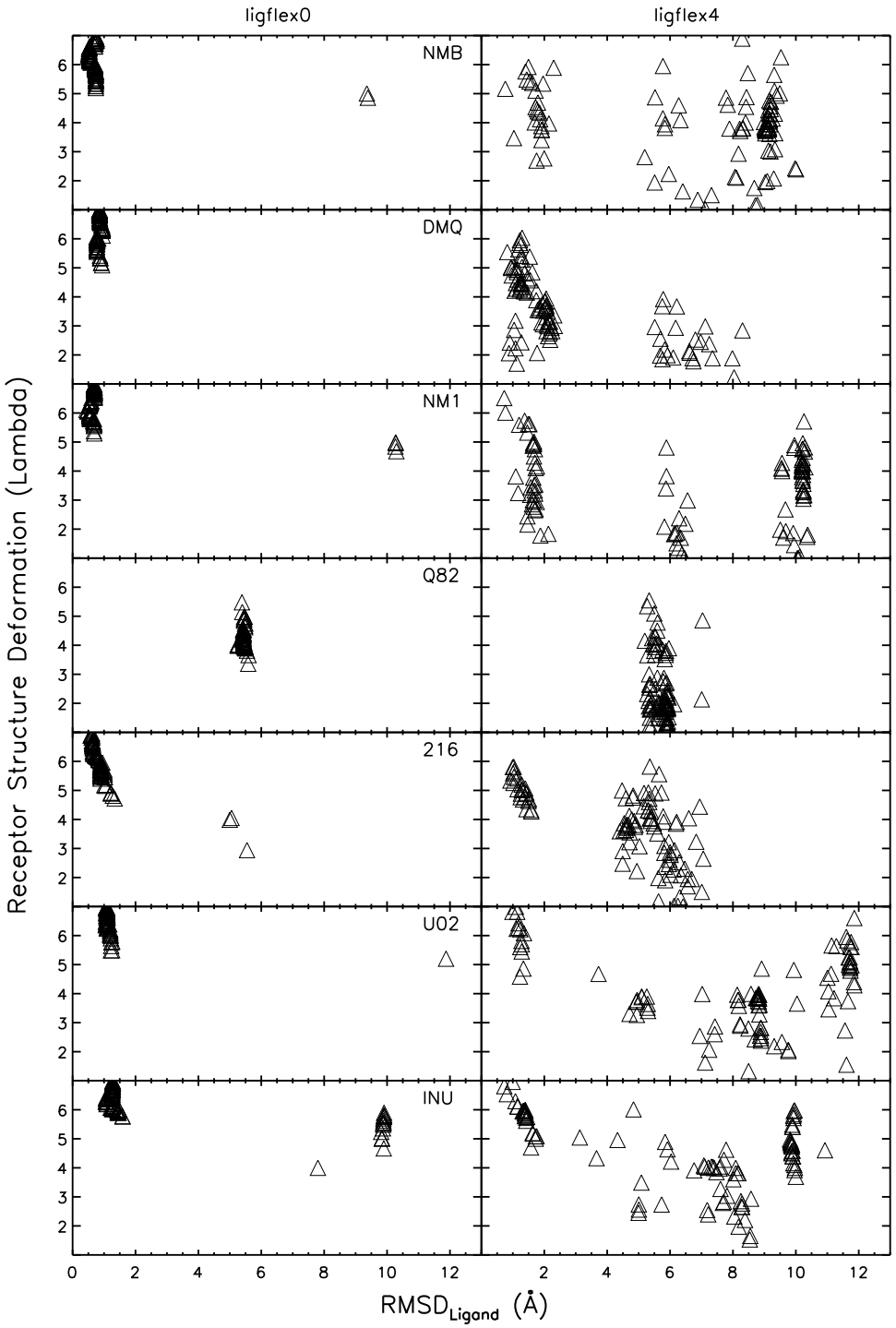


Figure 5.15: *HIV-protease structure vs. $RMSD_{Ligand}$ for flexible receptor docking applied to the true binder test set. Variation of the parameter lambda corresponded to a continuous morphing of the HIV-1 protease receptor structure beginning at the apo form (PDB ID 3HJV, lambda=1) and ending with one bound form (PDB ID 2UPJ, lambda=7). $RMSD_{Ligand}$ corresponds to the deviation of ligand coordinates from the native structure after best superposition of the receptor structure. Correlation between ligand placement and receptor structure are shown for both treating the ligand as rigid bound conformer or flexible (allowing bond rotation).*

Binders versus Non-Binders

The set of non-binder molecules as described in the sections above has also been tested against flexible receptor docking employing morphing structures. In principle, only the ligand from 2UPJ (the target structure for the morphing from the apo protein), U02, should be denoted as a true binder and all other binders as foreign binders. However, the naming of the ligand test sets was kept to stay consistent with the naming in the section above.

Only two results for the non-binders are below -10 kcal/mol: The rigid receptor docking of the ligand ZAF into deformation 6 and 7 (-10.02 and -10.43 kcal/mol respectively). For the deformations that are closest to the real bound structures (deformations 6 and 7), the binding energies clearly distinguish the majority of true and foreign binders from the non-binder molecules.

Comparing the flexible docking of the true binders to the docking into the rigid deformation structures shows that the flexible receptor docking yields significantly better binding energies for the ligands DMQ, NM1, Q82, U02, and INU. This out-performance is also visible for the docking into deformation 6 and 7 which are the closest to actual bound structures. 6 of the 7 true binders yield better binding energies than the non-binder molecules and can thus be clearly separated from them.

The flexible receptor docking of the foreign binder molecules yields slightly worse binding energies when compared to the rigid receptor docking into deformation 6 and 7. Only two of the seven foreign binders are clearly distinguishable from the non-binders with an energy difference to the best non-binder energy of > 1 kcal/mol.

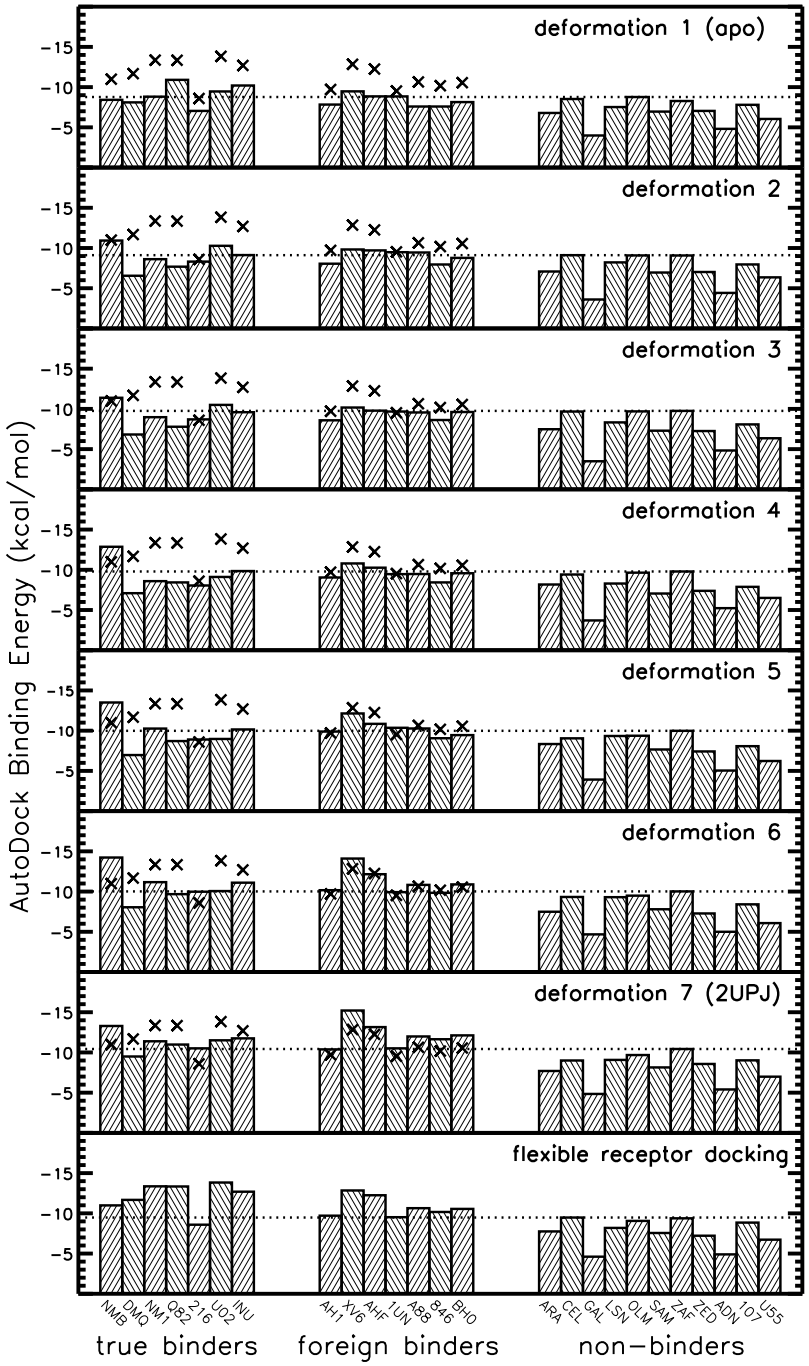


Figure 5.16: *Best scoring results for docking of true, foreign and non-binders to apo (first row), 5 morphing intermediate receptor structures (row 2 to 6) and the bound receptor structure PDB ID 2UPJ and using flexible receptor docking (using interpolation between the 7 different receptor structures). The dotted line marks the best binding energy obtained for the non-binder molecules. The cross symbols indicate the results of flexible receptor docking and are included for comparison for each docking to a rigid protease receptor.*

5.5 Summary and Conclusions

Several possibilities of structure input for the presented new flexible docking approach have been tested on the pharmaceutically relevant target HIV-1 Protease. For the great majority of docking various ligands into the unbound structure of HIV-1 Protease, incorrect ligand placements far from any native ligand binding modes are observed. In the most simple approach, a set of seven bound protease structures that exhibit small conformational differences at the active site is employed.
With the flexible receptor docking being able to continuously switch between bound protease structures, ligand placements can be significantly improved. Best-RMSD solutions as well as the solutions with the best scoring are found in a near-native position. Those docking solutions, which resulted in a good placement of the ligand, also selected a receptor "deformation" close (but often not exactly) to the native bound structure corresponding to the particular ligand. In such cases, the ligand was docked into a receptor structure that is geometrically very similar to the true receptor structure, thus the algorithm is correctly selecting the structure out of the ensemble.
Tests, where the original bound structure that belongs to a ligand is left out of the flexible receptor ensemble, suggest that the presence of the original bound structure in the ensemble has only a minor effect on the quality of the docking results. It should be noted here, that the flexible docking with a set of bound structures was also tested on Protein Kinase A that was originally used to test the normal mode deformations based approach. The results of these tests (not shown here) are comparably good.

Another set of HIV-1 Protease ligands that were crystallized in bound structures different from the flexible docking ensemble structures (termed 'foreign binders') is used both for rigid receptor cross docking and flexible receptor docking. Even though the flexible receptor docking relies on different bound receptor structures, it is able to reach very good ligand placements and outperforms most of the rigid receptor cross dockings. Ligand molecules that have no demonstrated binding affinity to HIV-1 Protease (non-binders) are used to check for ReFlexIn's capability to distinguish between true and non binding ligands. Our findings show that the flexible receptor docking results for non-binders are scored significantly lower compared to true binder energies. This allows in principle for the simultaneous identification of ligands that may bind to putative intermediate structures in between the ensemble structures.

In an NMR-structure based approach, the ensemble used for flexible docking consists only of receptor structures that were solved by a single ligand NMR experiment. This approach is able to yield ligand placements with an RMSD ≤ 2Å, however, the low-RMSD results cannot be clearly distinguished from the poor-RMSD results by means of AutoDock binding energy.

With only using the knowledge of a single bound receptor structure and the unbound structure of the protein, the morphing approach as presented above shows comparably good results. Best-RMSD as well as best energy solutions of the flexible receptor docking (using the intermediate structures from the morphing between the unbound protease structure and a bound protease form) are in very good agreement with the native ligand placements from experiments. The algorithm favours docking the ligands into intermediate structures that are closer to the actual bound form, hence giving an orientation towards a correct bound structure. It is shown that the choice of the bound form for the morphing has only a minor effect and still results in improved ligand dockings with enabled receptor flexibility. The morphing approach shows the capability of binder/non-binder differentiation for approximately half of the tested binder molecules.

Chapter 6

In silico Prediction of Binding Sites on Proteins

The previous chapters dealt with the docking problem where the actual binding site of a ligand is already known. Hence, the actual docking calculations can be restricted to the approximate area around this known active site. The situation is different when – earlier in the course of a drug design project – one deals with a protein receptor for which the active site is not identified yet. This might be the case if, for example, the protein-ligand complex cannot be solved due to technical obstacles and only the unbound form of a protein is available.
To fill this gap, a wide range of computational tools has been developed that try to predict active sites (or "hot spots") on proteins while tackling the problem of accuracy versus efficiency. Several most commonly used approaches are introduced here and are evaluated for the binding site prediction capacity on a test set of proteins with various extent of conformational changes upon complex formation.

6.1 Introduction

Biomolecules and many other organic ligands can bind to proteins with high affinity at specific sites on the protein surface. The question of what distinguishes such recognition sites from other surface regions of proteins has been the subject of intense experimental and theoretical research [184, 185]. In recent years, the possibility to predict putative binding regions on the surface of protein molecules has become increasingly important. Together with the rapidly growing structural knowledge of proteins of biological and medical importance, such prediction methods become more applicable and can be helpful for rational drug design and to elucidate the function of a protein molecule.

Both these applications, function prediction as well as rational drug design, require a reliable method for identifying and characterizing the ligand-binding sites of a protein.

Knowing the location of the functional sites (e.g. substrate or ligand-binding sites of enzymes or receptor proteins) on the protein surfaces prior to experiment, makes it possible to design inhibitors or antagonists and to introduce targeted mutations aimed at improving the protein function. It is also possible to apply these methods to assist in modeling the three-dimensional (3D)-structure of protein-ligand complexes [86, 186, 187].

The availability of 3D structures of many proteins in complex with proteins or other types of ligands (lipids, nucleic acids or drug-like molecules) allows the systematic comparison of protein surfaces involved in interactions [185, 188–201]. Comparative studies of the amino acid distribution and physicochemical features of protein-protein interfaces [188–195] and proteins in complex with small organic drug-like ligands [196–201] made it possible to characterize recognition sites. Furthermore, often interface residues around binding sites are evolutionary more conserved than other surface regions. A variety of computational methods have been developed that try to integrate this information for predicting putative binding sites in proteins.

The realistic prediction of putative ligand or protein binding sites has not only important implications for rational drug design but could also have an impact on a better understanding of protein-protein interaction networks. The possibility to identify and to characterize putative protein binding sites on proteins can help to elucidate the number and kind of protein interaction partners. *In silico* methods to predict protein-protein interaction sites can also be used to predict the propensity of proteins to aggregate [202, 203] or to bind non-specifically to many different partners [204]. Recent approaches to predict not only binding sites on proteins but also which partner protein may bind could potentially be useful to predict protein interaction networks [205].

First, this chapter gives an overview of the geometric and physicochemical properties of protein binding sites for small drug-like ligands (in the following: protein-ligand complexes) based on the analysis of known 3D structures. Recent approaches for predicting putative protein-protein interfaces and binding sites for drug-like ligands will be discussed in the second part, followed by an analysis of the robustness of ligand binding site prediction, with respect to conformational changes or inaccuracies in the protein structure. Finally, challenging future issues will be discussed.

6.2 Comparison of Protein-Protein and Protein-Ligand Interaction Regions

Complexes of proteins are non-covalent protein assemblies that fold separately and associate under certain physiological conditions. Examples of protein-protein complexes are antigen-antibody, enzyme-inhibitor, and many signal transduction and cell cycle protein complexes [206]. The majority of known protein-protein complex structures have been determined by X-ray crystallography which requires stable complex structures that can form well ordered crystals. Based on known complex structures, the geometric and physicochemical properties of protein-protein interfaces have been characterized in detail [185, 188–194, 206]. In the following, an overview of the main results will be given. It is important to indicate that the analysis of interface properties of protein-protein binding sites is restricted to sufficiently stable protein-protein complexes (that can form well ordered crystals). Rules derived for these complexes may differ from interfaces formed during transient interactions with a short lifetime. Most protein-protein complexes bury a surface area in the range of 1200-2000Å^2 which is much larger than the buried surface area upon binding small drug-like molecules of a few hundred Å^2 depending on the size of the ligand [196]. The comparison of known protein-protein complex structures indicates that protein-protein interfaces are in many cases overall flat in shape with the exception of several enzyme-inhibitor complexes [192–195] where the inhibitor site often forms a convex surface fitting to the concave shape of the enzyme active site. This contrasts to binding sites for enzyme substrates or other small organic ligands that are usually very non-planar allowing contacts to the ligand from many different sides of the binding pocket [196–202].

Protein interface regions clearly differ on average from the rest of the protein surface in terms of physicochemical properties and geometric characteristics [191–195]. However, the interactions between proteins are very diverse. It is therefore not possible to distinguish a binding site from the rest of the protein surface based on a single surface attribute [195]. Interface residues in protein-protein complexes can be divided into two distinct regions, the 'core' and the 'rim' region, based on the solvent accessibility in the complex [193, 194]. The 'core' region contains residues that have at least one fully buried interface atom (i.e. zero accessibility after complex formation) and usually contain mostly non-polar residues surrounded by the more polar 'rim' region, which contains residues that are at least partially solvent exposed even in the complex. The composition of amino acid residues at specific protein-protein interfaces differs from the rest of the protein surface. Interface regions are enriched in aliphatic (Leu, Val, Ile,

Met) and aromatic (His, Phe, Tyr, Trp) residues, and depleted in charged residues (Asp, Glu, Lys) with the exception of arginine [189–195]. The higher abundance of Arg at interfaces compared to Lys has been attributed to formation of cation-interactions [192] and the greater capacity of the guanidinium group in Arg to form hydrogen bonds (compared to Lys) [194, 206]. The role of arginine-arginine pairing and its contribution to protein-protein interactions was recently investigated by Vondrácek and coworkers employing computational approaches [207].

One way to characterize the relative contributions of interface residues to the binding free energy, is to determine the change in affinity upon mutation of interface residues to alanine. Substitution of residues by alanine (alanine-scanning mutagenesis) corresponds (except for glycine) to the removal of side chain atoms from the interface and its effect on binding strength [208–212]. Interestingly, for most protein-protein complexes analysed by alanine scanning mutagenesis only a fraction of substitutions showed a substantial effect on binding affinity [208, 209]. This finding has led to the concept of 'hot spots' on protein surfaces that are responsible for most of the interaction between proteins [209, 212] and methods for *in silico* alanine-scanning have been developed [125, 210, 213, 214]. Several methods to predict protein-protein interaction sites aim at identifying such 'hot spots' on protein surfaces (reviewed in [212]).

It is important however to keep in mind that the binding affinity between two proteins is determined by interacting pairs of residues or even higher order motifs and not only by individual amino acids on just one partner. Hence, a given contacting pair (e.g. of two polar or charged residues) at an interface may overall contribute little to binding, for example, because the desolvation of the two polar residues upon binding offsets the interaction energy between the residues. Nevertheless, substitution of one of the polar or charged residues by alanine may result in a significant drop of binding affinity (because alanine cannot form polar contacts), which may lead to the erroneous conclusion that the region is a hot spot. The substitution of a residue with zero contribution to binding energy can still result in a large drop of binding affinity if it creates an unfavorable contact with another residue.

Similar to protein-protein interaction sites, high affinity binding cavities for small drug-like ligands are often less polar (low desolvation penalty) or more hydrophobic compared to the rest of the protein surface [196–202]. However, due to the smaller size of organic drug-like molecules compared to proteins, the buried surface-area upon small molecule protein ligand interaction is generally smaller than in the case of protein-protein inter-

actions. In order to achieve strong interactions through a sufficiently large number of favorable protein-ligand contacts, high-affinity binding sites are usually strongly concave pockets or cavities on the surface of proteins or sometimes partially buried [196]. Algorithms for predicting protein-protein interfaces are, in many aspects, similar to methods for predicting binding regions for small drug-like molecules. However, there are also some important differences due to the distinct general architecture of these types of binding sites [200, 215].

6.3 Approaches to predict small Molecule Binding Sites

Similar surface properties and sequence conservation as used to predict protein-protein interfaces can also be used to identify putative interaction sites for small drug-like ligands. However, the predictive power of each property may differ from predictions of protein-protein interfaces due to the difference in architecture of high affinity binding sites for organic ligands compared to protein-protein interfaces (see previous paragraphs). Burgoyne and Jackson [215] compared the predictive power of different surface properties and found that it is in general easier to identify putative protein-ligand interfaces compared to protein-protein interfaces. Binding cleft detection and desolvation properties, as well as sequence conservation and to some degree electrostatic potential, have been identified as the strongest signals for predicting protein-ligand interfaces [215].

Since ligand binding sites involve in most cases the presence of a concave binding cleft on the protein surface (in contrast to the more flat protein-protein interfaces) the detection of binding pockets or protein cavities deserves special attention. Presumably, the better performance of binding site prediction for small drug-like ligands compared to protein binding site prediction is due to the importance of a concave binding site in the latter case. Apparently, such concave regions are less frequently found on protein surfaces than flat or slightly curved surfaces typical for protein-protein interfaces. Several algorithms based on different detection principles have been designed in recent years. Only the principles of the most common methods will be explained here, since pocket detection methods and explanations of algorithmic design have been reviewed in detail in [216]. In the following, the basic algorithmic ideas of pocket detection of the most common available methods are summarized. An full-size overview on available (mostly web-accessible) ligand binding site prediction methods including the respective web-links is given in [217], the methods that are employed in the algorithm evaluation section of this chapter are briefly summarized in Table 6.1.

Principles of Pocket Detection

Various algorithms to identify surface clefts in proteins have been reviewed and explained in detail by Laurie & Jackson [216]. One can distinguish between geometry-based and energy-based detection methods. The latter methods define favourable cleft regions based on energetic evaluations, the former based on sterical considerations. Many methods employ a regular 3D-grid and move probes along grid lines to define accessible and inaccessible or energetically favourable and unfavourable positions. Alternatively, probes placed on the solvent accessible surface of the protein can be used in combination with a variety of algorithms to define pocket regions. For example, the PASS program [218] filters out highly accessible surface probes and creates additional layers of surface probes on top of surface probes located in clefts. The procedure is repeated until all clefts are filled with probes. In addition, structural motifs typical for binding pockets have also been used to define binding sites [219]. The principles of the most common ligand binding site prediction methods are briefly explained in Table 6.1. For a more detailed explanation the reader is referred to the original literature and for a comparison of the pocket detection methods to the review by Laurie & Jackson [216].

In the following, six of the most common web-accessible programs are considered and subsequently used to compare the performance on several proteins in bound, unbound and in the form of modelled structures. The LigSite method is based on a regular 3D-grid placed around the protein [220]. A probe is moved along the x, y, and z directions and the cube diagonals of the grid. A grid point that is counted as part of a pocket is assigned if the grid line contains points before and after the point that overlapped with the protein. A web-accessible extension of the method (LIGSITEcsc) includes the degree of surface conservation and has been shown to improve the performance [224]. The Q-SiteFinder [223] is an energy-based binding site predictor that clusters grid points of favourable (van der Waals) interactions with the protein to define a putative binding site. The CASTp algorithm uses an entirely different principle to detect binding pockets [222]. In the CASTp method, a Delaunay triangulation of the protein is performed (meaning the entire protein shape is approximated by triangles). A pocket can be detected, based on the direction of norm vectors associated with triangles for a set of neighboring triangles. A web-accessible server for this method is available (see Table 6.1). The Mark-Us method is another binding site prediction tool based on the SCREEN algorithm [225], which is based on a large set of physicochemical, structural, and geometric descriptors extracted from known complexes. Finally, the web-accessible Fuzzy-Oil-Drop (FOD) method [226] identifies primarily hydrophobic patches on protein surfaces to assign putative binding regions.

Method	Description
LIGSITE [220]	On a regular grid around the protein, lines are drawn from each grid point along the x/y/z-axis as well as the cubic diagonals. Segments of lines that are enclosed by protein from both sides are considered as cavities.
ConSurf [221]	Identifying functional sites on proteins by determining the conservation of sequence homologues.
CASTp [222]	Uses alpha shape theory and triangulation methods to predict pockets.
Q-SiteFinder [223]	Energetically based method: clusters of protein surface regions that show favorable Van-der-Waals interactions with a methyl-group are collected and ranked.
LIGSITEcsc [224]	In extension to the traditional LIGSITE method, the Connolly surface area is calculated and grid points are scanned for surface-solvent-surface events. Additionally, the top three predicted pockets are re-ranked according to sequence conservation.
Screen/Mark-Us [225]	Cavities are geometrically determined via the difference between the molecular surface and the probe-specified molecular envelope and statistically analysis.
Fuzzy-Oil-Drop-Model [226]	Analyzes the protein for regions with high hydrophobic deficiency, i.e. the difference between observed and idealized hydrophobicity distribution declared by the 'Fuzzy Oil Drop Model'.

Table 6.1: *Short descriptions of the working mode for the protein-ligand binding site prediction methods tested here. For a full overview of available methods and further references, see [217].*

6.4 Robustness of Ligand Binding Site Prediction with Respect to Protein conformational Changes

Proteins can undergo conformational changes upon ligand binding that may influence the steric accessibility of a binding cleft and can interfere with the ability of an algorithm to identify a potential binding site. For many proteins of biological and pharmaceutical importance, no 3D-structure is available but very frequently a structure of a protein with similar sequence can be used to generate a homology model of the target protein. Depending on the degree of target-template similarity such homology modeled structures frequently include structural inaccuracies that may interfere with the prediction of putative ligand binding sites. It is of importance to check the performance of prediction methods under realistic conditions where only the unbound structure or a model structure of the target protein is available. This corresponds to an often realistic scenario of a rational drug design project where for a given protein target of interest only a sequence but no 3D-structure is available.

In order to obtain an impression on the performance of several of the most recent web-accessible binding site prediction methods, the application to several protein structures in bound and unbound conformation or even generated homology modeled variants for some of the proteins are compared. The pairs of bound and unbound structures show varying degrees of structural similarity and are listed in Table 6.2.
The proteins correspond to typical targets for rational drug design. Most ligand binding site prediction methods have been tested on bound and unbound protein structures described in the original publications but the test set and conditions may vary for each case. A direct comparison of available methods applied to the same targets can be useful to obtain a hint of the performance of each method and may indicate overall trends. It should be emphasized that it is not the purpose of the chapter to provide a comprehensive benchmark test or to provide a quantitative evaluation of the prediction results. The selected target structures do not represent unsolved problems of drug-design but well-known examples to give interested researchers an overview of the available methods and their performances by comparing the predictions with the known binding sites.

PDB ID	Molecule	State	Ligand	RMSD(Å)
2ANO	E.coli dihydrofolate reductase	bound	Inh. MS-SH08-17	∅
5DFR	E.coli dihydrofolate reductase	apo	—	0.7
5DFR_2KGK	DHFR based on 2KGK structure	homology	—	1.6
5DFR_3IA5	DHFR based on 3IA5 structure	homology	—	1.3
3BNZ	Thymidylate synthase	bound	8A inhibitor	∅
1NJB	Thymidylate synthase	apo	—	0.8
1EB2	Trypsin inhibitor complex	bound	BPO	∅
1BTY	Trypsin inhibitor complex	apo	—	0.3
1BTY_1GVL	Trypsin based on 1GVL structure	homology	—	0.8
1BTY_1L2E	Trypsin based on 1L2E structure	homology	—	0.9
1FKS	FK506 binding protein	apo	—	1.3
1FKS_2VCD	1FKS based on 2VCD structure	homology	—	2.3
1FKS_2KE0	1FKS based on 2KE0 structure	homology	—	2.6
1D6O	FK506 binding protein	apo	—	0.3
1D7H	FK506 binding protein	bound	DMSO	∅
1APB	Arabinose binding protein	bound	arabinose	∅
1ANF	maltose-binding protein (MPB)	bound	maltose	∅
1OMP	maltose-binding protein (MPB)	apo	—	3.8
1OMP_2FNC	MPB based on 2FNC structure	homology	—	2.4
1OMP_2GHA	MPB based on 2GHA structure	homology	—	2.7
1R2D/1Y2D	BCL-X L bound vs. unbound	—	—	3.7
1M47/1PY2	IL-2 bound vs. unbound	—	—	2.9
1T4E/1Z1M	MDM2 bound vs. unbound	—	—	2.2

Table 6.2: *Protein structure test set. Homology models were generated using the Modeller program with default settings [227] based on a template structure (indicated as second PDB ID in the name given in the first column) with a sequence identity of 30% to 50%. The main chain RMSD is given relative to the respective bound structures.*

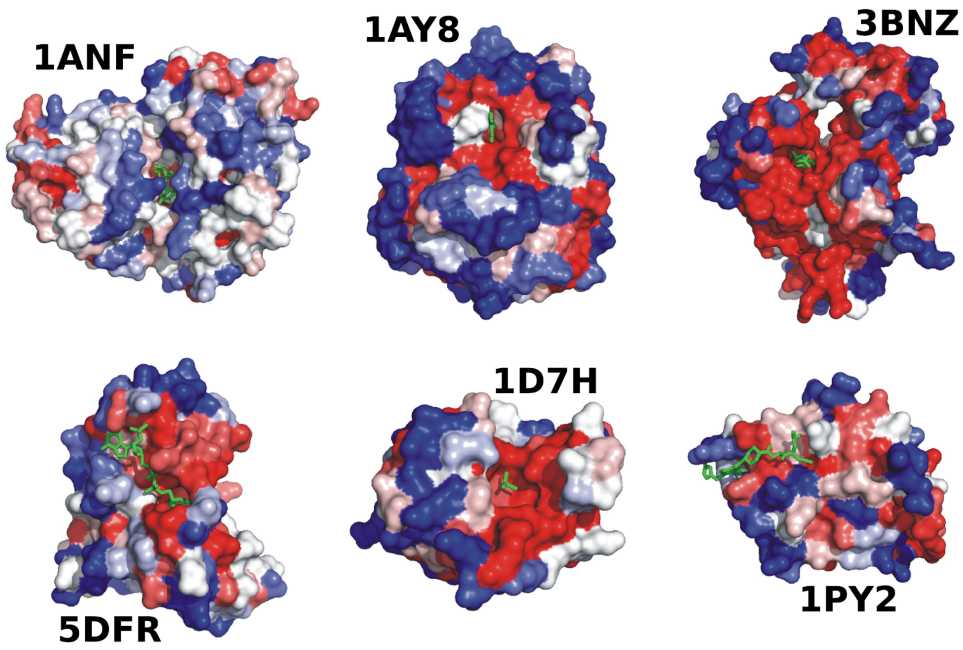


Figure 6.1: *Result of ConSurf-Server application to six ligand binding protein structures (labeled with PDB ID). The sequence conservation of protein surface residues is color-coded with increasing conservation from blue to red. Bound ligands are indicated as stick models (green).*

The ConSurf-method provides a map of the sequence conservation of residues extracted from a multiple alignment of proteins homologous to the target protein [221]. Although the predicted regions of high sequence conservation frequently overlapped with the binding sites for ligands on the target proteins (Figure 6.1), the conserved surface regions in the example cases extend often beyond the ligand binding site or include parts of the protein surface that are far off the binding site. Hence, the specificity of sequence conservation alone may in general not be sufficient to exactly locate the putative ligand binding site (compare ligand position and red colored protein surfaces in Figure 6.1). It should be emphasized that conserved regions not overlapping with the known ligand binding site can be of other functional importance (for example a binding site for another protein). In addition to ConSurf, the programs CASTp [222], Q-SiteFinder [223], LIGSITEcsc [224], Mark-Us [225] and Fuzzy-Oil-Drop FOD, [226] were applied on the test proteins (see above and Table 6.1). The web-accessible methods were employed using default parameters.

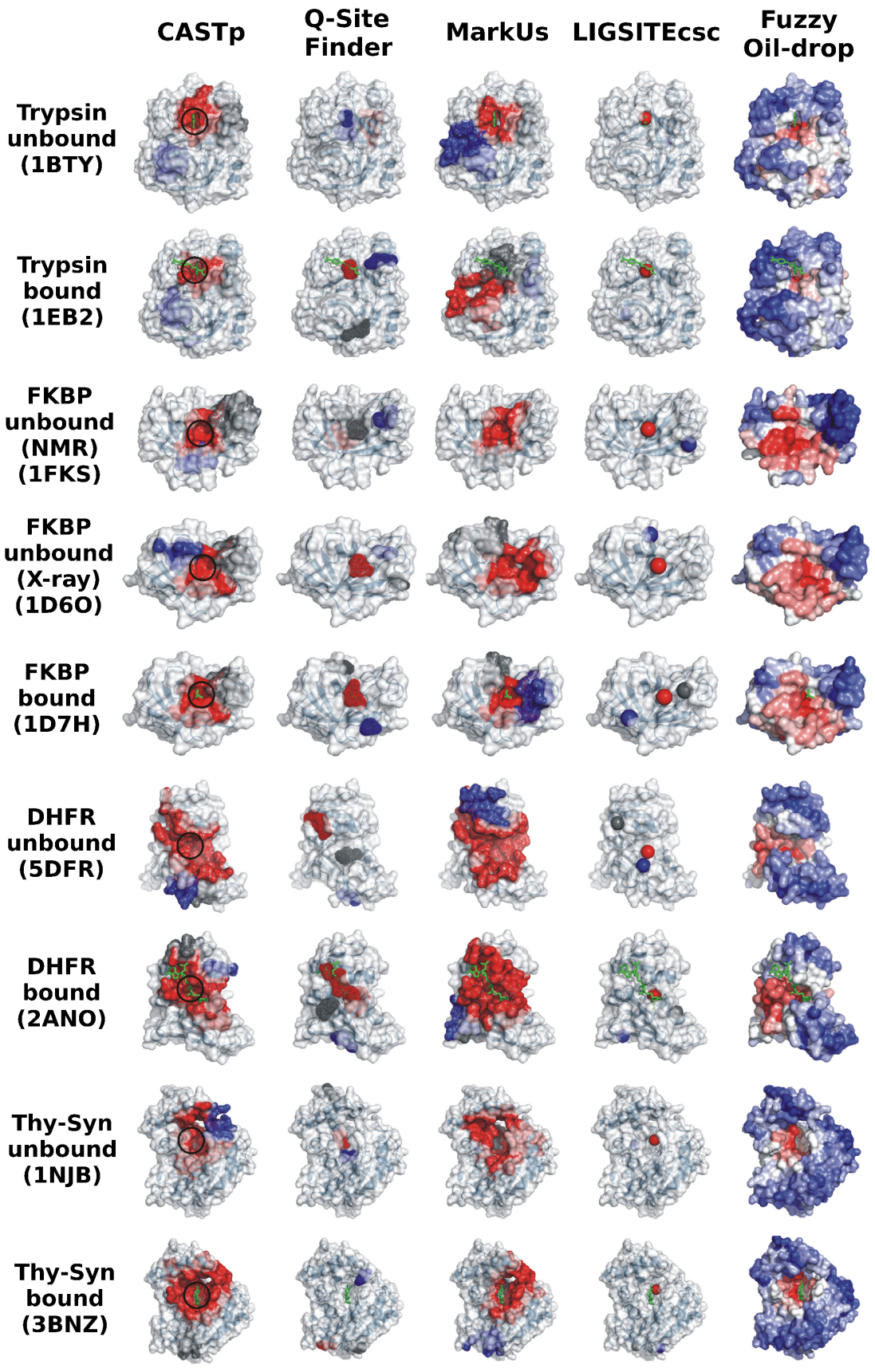
In the case of using bound structures and for the four protein cases illustrated in Figure 6.2, all tested methods performed very well in identifying the native binding pocket as the top ranking or one of the top ranking solutions. The most likely site predicted by LIGSITEcsc or Q-SiteFinder was close or overlapped with atoms of the ligand in all cases. Only in the case of Thymidylate synthase (Thy_Syn), Q-SiteFinder scored a position close to the binding site at rank 3. For the CASTp, Mark-Us, and FOD methods that encode likely binding sites as B-factors in the PDB-file, the predicted binding regions overlapped or completely included the known binding region in all cases. However, sometimes the predicted top ranked regions significantly extended beyond the known binding region lowering the specificity.

Despite the success of the tested approaches to predict binding regions that overlap with the known binding site, it is important to consider differences in the prediction results that depend on the topology of the binding region. The LIG-SITEcsc server located the binding region well in all cases. However, it uses a one-sphere prediction which in case of elongated ligands does not give information about a possible orientation of a bound ligand. The Q-SiteFinder returns a large set of spheres which gives a better representation of the complete binding site. In contrast to the sphere-based LIGSITEcsc prediction, the Fuzzy-Oil-Drop-Model assigns a hydrophobicity distribution to the protein surface. It achieves a high sensitivity (overlap with binding site) but the large size of the predicted surface patch may lower the specificity of the prediction and makes it more difficult to define the exact binding site. Examples 1FKS, 1D6O, and 2ANO in Figure 6.2 illustrate this problem where protein regions quite distant from the known ligand binding site are included in the prediction.

Interestingly, the overall performance was only slightly worse in the case of employing unbound structures as target proteins. This was even the case for DHFR and Thy_Syn for which the conformational difference between bound and unbound structures near the binding site is more significant compared to Trypsin or FKBP. For example, LIGSITEcsc predicted in every unbound structure a pocket that formed at least part of the binding site for the full ligand (Figure 6.2).

For the homology modeling, templates with a sequence identity between 30-50% with respect to the target protein (Table 6.2) are selected. This degree of sequence similarity is typically considered as yielding reliable models with an overall realistic structure [227]. Homology modeling was performed with the Modeller program ([227,228]). Two models were generated for four proteins based in each case on two different template proteins. No information on the known structure of the target proteins was included during the comparative modeling step. The RMSD (main chain) between modeled structures and the corresponding native structures was case depended (1-3Å, Table 6.2). Remarkably, for several of the homology modeled protein structures (e.g. one Trypsin model, one DHFR model, both FKBP models and both MBP (maltose-binding protein) models, the prediction methods performed qualitatively almost as well as for the native protein cases (shown for Q-Site-Finder and LIG-SITEcsc in Figure 6.3).
In case of the FKBP protein models, the RMSD with respect to the bound native conformation was > 2Å (Table 6.2) but still preserved a detectable pocket. Similar for MBP, the RMSD of the models exceeded 2Å but this concerned mostly the global arrangement of the two domains that encompass the binding site. The modeled MBP structures contained an even more open binding cleft (similar to the unbound conformation) that is detected by the prediction programs.

Figure 6.2 *(facing page)*: *Results of five ligand binding site prediction servers on four target proteins in bound and unbound conformations. The predicted binding regions are either shown as colored molecular surfaces (CASTp, SCREEN and FOD) with increasing probability from blue to red for a ligand binding site or as colored probes (Q-SiteFinder, LIGSITEcsc). Up to three predicted binding sites are shown (red highest score followed by grey and blue). The location of the binding site is encircled black at the most left column. Ligand molecules in the bound structures are shown as stick models (green).*


CASTp
Q-Site Finder
MarkUs
LIGSITEcsc
Fuzzy Oil-drop
Trypsin unbound (1BTY)
Trypsin bound (1EB2)
FKBP unbound (NMR) (1FKS)
FKBP unbound (X-ray) (1D6O)
FKBP bound (1D7H)
DHFR unbound (5DFR)
DHFR bound (2ANO)
Thy-Syn unbound (1NJB)
Thy-Syn bound (3BNZ)

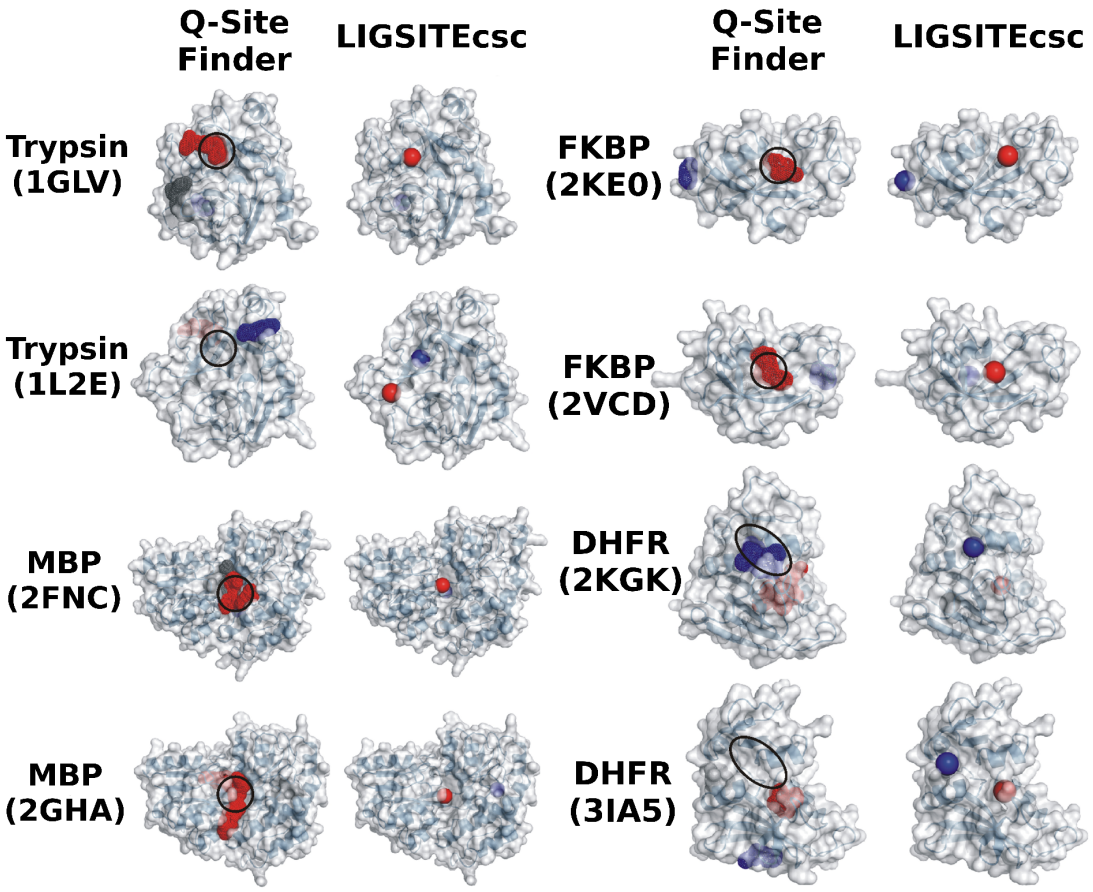


Figure 6.3: *Performance of Q-SiteFinder and LIGSITEcsc on homology modeled protein structures. The name of target protein and the PDB-entry of the template used for homology modeling are indicated. Coloring and labeling scheme is the same as in Figure 6.2.*

Interestingly, for one trypsin model and one DHFR model the prediction of a binding cleft was less precise (Figure 6.3) although the conformational difference of the models from the corresponding bound structure was < 2Å . This indicates that overall deviation of a model from the native structure is not necessarily a good measure for the usefulness of a model to identify putative binding sites. The degree of change and the type of conformational change near the binding cleft is decisive. Even large changes near the binding cleft (in case of MBP) may result in a detectable pocket (for example a more open pocket) but even small changes that result in closure of the pocket can interfere with the ability to detect the pocket.

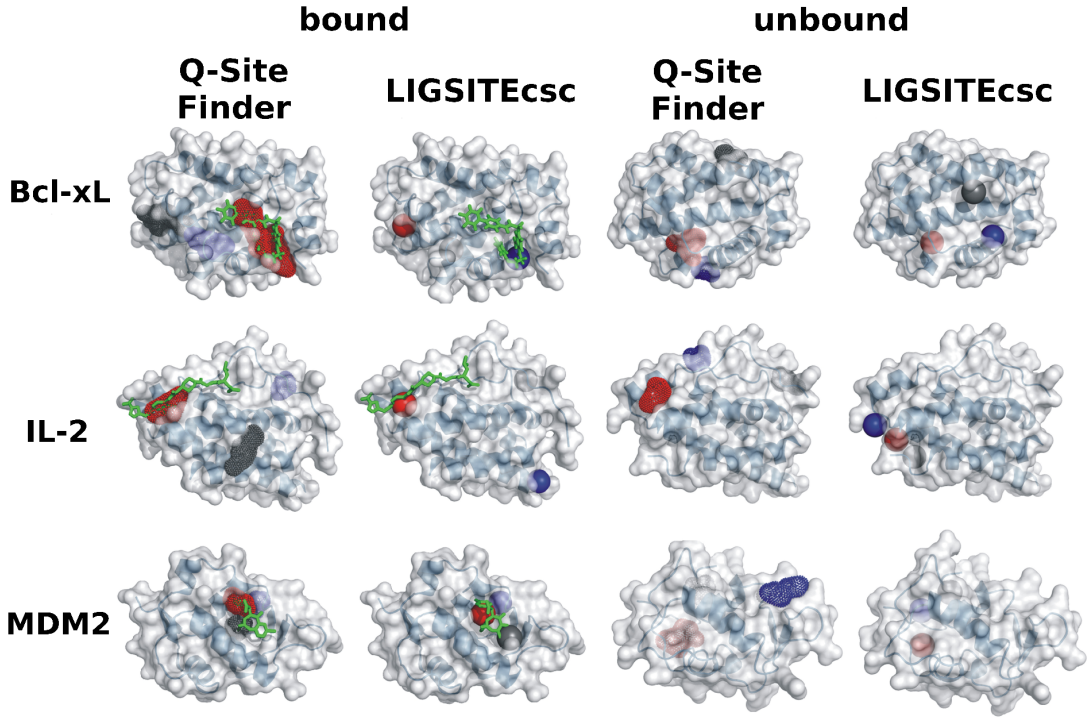


Figure 6.4: *Application of Q-SiteFinder and LIGSITEcsc to detect inhibitor binding sites near protein-protein interaction sites for three proteins in bound and unbound conformations. Coloring and labeling scheme is the same as in Figure 6.2 and 6.3.*

Finally, also the prediction methods on the three protein-protein inhibitor cases mentioned above that have been studied previously [229] are tested. In these cases, the deviation between proteins in unbound and inhibitor bound conformation exceeded the RMSD for the above discussed cases (Table 6.2). For the bound forms, Q-SiteFinder and LIGSITEcsc detected pockets that are at least part of the native pockets in all cases (Figure 6.4).

For the unbound BCL-X L, LIGSITEcsc was able to predict two pockets very close to the native binding pocket (rank 2 and 3) and one (rank 3) pocket close to the native inhibitor in the case of IL-2. Q-SiteFinder was successful in the case of IL-2 with a detected pocket close to the prediction in the bound form as top ranking prediction and one other predicted pocket close to a second binding regime of the inhibitor as also described by Fuller et al. [229]. None of the three top ranked predicted pockets overlapped with the native binding sites in the other two cases (using standard parameters). It was, however, possible to hit regions close to the native binding site if predicted binding sites of lower rank were included (not shown).

Fuller et al. [229] also indicated recognisable ligand binding pockets using Q-SiteFinder applied to the unbound state of BCL-XL considering, however, a larger number of predicted putative pocket sites.
The three examples indicate the limits of current pocket detection if the RMSD (main chain) between bound and un-bound structures reaches 2.5Å or may even exceed 3Å (BCL-XL case, Table 6.2) and if binding pockets are largely closed in the unbound state. Here, methods that allow for conformational changes, like molecular dynamics simulations, could become useful to identify transient pockets [230] albeit at much larger computational costs compared to current prediction methods that require seconds or minutes to perform a prediction.

6.5 Summary and Conclusions

The availability of an increasing number of protein-protein and protein-ligand complexes has resulted in an improved understanding of the properties of binding sites. In recent years, this knowledge has been used to design many computational prediction tools to identify putative ligand and protein binding sites on proteins. There are significant differences in the architecture of protein-protein interfaces and high affinity sites for binding small drug-like ligands. The latter require, in the majority of cases, a strongly concave binding pocket to maximize the number of close contacts in order to achieve strong interaction. In the case of proteins, the much larger buried interface area allows a wider distribution of interactions and the exclusion of water from a larger interface area may also contribute to the enhanced interaction at hot spot residues located at the interface. It might be especially useful to focus the design of drugs to interfere with protein-protein interactions to those proteins with defined clefts at the protein-protein interface. Further developments in the area of binding site prediction could also aim at predicting not only where a ligand could potentially bind but also which type of ligand might be suitable for a given binding pocket.

It is expected that conformational differences between bound and unbound proteins affect the ability of prediction methods to locate potential protein or ligand binding sites. Our test on a limited set of proteins in unbound and bound conformations indicates that several of the available web-accessible methods tolerate a certain degree of conformational difference. Encouragingly, for deviations of proteins in bound versus unbound structure of up to 1.3Å of the backbone (∼2Å for heavy atoms), most tested programs identified the native ligand binding site as top ranking or among the top ranking predicted pockets. Even for larger deviations or for some of the homology modeled structures with main chain RMSD up to ∼2.5Å pocket close to the known site could be identified as a potential ligand binding position. It needs to be emphasized that the number of evaluated cases in the present chapter does not represent a comprehensive test set. However, it may form a starting point for more systematic and exhaustive studies including more methods and employing larger sets of proteins including homology modeled structures with varying degrees of structural accuracy. The results also indicate that if ligand binding involves structural changes for pocket opening beyond an RMSD of ∼2Å new methods may be required that allow for conformational adjustment during the pocket detection phase.

Chapter 7

Summary and Outlook

In this thesis, I presented a novel approach, named ReFlexIn, that allows to efficiently include receptor flexibility in grid-based protein-ligand docking. The popular docking program AutoDock is expanded with a function that allows for a better representation of protein flexibility compared to the classic docking where only one single rigid receptor structure is considered. Instead, a set of conformationally different receptor structures can be employed in a flexible receptor docking that is able to continuously switch between the different deformations. In contrast to most existing ensemble docking methods, the single structures are not used for sequential docking but during the same docking run. An interpolation scheme allows for a smooth, continuous deformation of the receptor structure along a series of snapshots, thus also rendering the approach capable of accounting for intermediates between neighboured structures.

The structural input for the presented method can be derived from different sources. For pharmaceutically relevant targets of the protein kinase family, an efficient generation of normal mode deformations proves to be able to reproduce target structures with better resemblance to the native bound protein structures, only employing existing unbound structures of the targets as input. Using the generated deformations as the structure ensemble with ReFlexIn, docking results can be significantly improved for the great majority of the tested ligands compared to the rigid receptor docking. For this approach, no knowledge of bound structures is necessary, as only an apo form of the protein target is needed. In addition, the computational demand of the ENM calculations can be neglected and the flexible receptor docking comes at a very moderate increase of runtime compared to a single rigid receptor docking run.

The ReFlexIn docking approach was also evaluated in various tests on HIV1-Protease which is a well-studied inhibitor candidate for HIV-related research. For this target, several other possibilities of representing the flexible receptor are presented and tested. In one approach, receptor flexibility is represented by employing several bound protease structures. This flexible receptor docking yields significantly better results compared to docking into the rigid apo form alone and, in addition, performs at least equally well and in some cases slightly better than the best result of a rigid receptor cross docking into each of the bound structures separately. Since this best performing bound structure is not known in advance, it is beneficial for the result quality as well as the efficiency to use our approach instead of running separate dockings to each individual receptor structure.

Another approach models receptor flexibility by employing putative intermediate structures generated by morphing between an unbound and bound receptor structure. This approach is able to achieve a considerable improvement of docking results for various ligand molecules, with only the necessity for one bound and unbound receptor structure and minimal computational effort for intermediate structure generation.
In addition to the test set of real HIV-1 Protease binders, a collection of molecules with no protease affinity was tested with both ReFlexIn variants (bound structures only and morphing) to investigate for the method's capability to distinguish between binder and non-binder molecules. With flexible docking enabled, most of the true binder ligands yield better binding energies than the non-binders, however, this effect is slightly weaker for the morphing approach. The results indicate that the usage of our flexible receptor docking approach appears to maintain the ability to identify correct binding modes (for ligand placement and receptor structure) as well as the possibility to discriminate between binders and non-binders.

Additionally, one could also further explore the use of structure sources for the flexible receptor ensemble other than the ones presented in this thesis (ENM-derived deformations, several bound protein structures, a morphing approach, and the use of experimental NMR structures). Several other different sources for deformations are possible: for example the use of homology modelled proteins or snapshots that are extracted by MD or Replica Exchange MD simulation runs. In the latter case, putative important states or conformations of proteins could be sampled (given a sufficient runtime of the simulations) that cannot be created using more efficient methods like ENM or morphing between two structures. Even though the production of such snapshots means larger computational effort, they could be a promising input for various cases.

One limitation of the presented method is based on the scoring problem that many approaches have in common: due to the lack of an universal scoring function that allows scoring conformations with very good accuracy and speed at the same time, the docking on different protein-ligand systems using different docking software can give results of varying quality [231, 232].
Some of the results presented have shown that the scoring function as implemented in AutoDock also does not always work flawlessly. The binding energy that is taken into account for scoring ligand placements is in some cases not able to distinguish between docking results that have a good and a poor ligand RMSD. One example is the HIV-1 Protease ligand U02 in Figure 5.4 of chapter 5 where two large clusters of solutions are found that have very different ligand RMSD (approximately 1Å vs. 12Å) but yield a binding energy of nearly the same level.
Two pathways for future improvement are possible: one could add a re-scoring step after a complete docking run where each of the separate protein-ligand conformations are re-evaluated using a scoring function different from the AutoDock scoring function (for example the DSX protein-ligand scoring function [233]). However, this approach would only increase the scoring quality after a docking is already finished. The other option would be to replace AutoDock's fit evaluation with the other scoring function. With this, the scoring of each ligand placement would be considered using another scoring function already within the runtime of the genetic algorithm for the decision about the fitness of a ligand placement and its subsequent acceptance or rejection.

When using AutoDock, one has to face another bottleneck, namely the fact that from all common docking programs, AutoDock is one of the slowest. In a study on testing six different docking programs on protein kinases, Tuccinardi et al. showed that AutoDock calculations can consume up to 10-fold the CPU-time compared with other major docking routines [234]). Nevertheless, AutoDock docking results yielded in the present work as well as the fact that it is freely available for academic use and modification substantiate its application.
During the course of this study, the source code for AutoDock Vina was made public [235]. This docking suite was created at the same institute as AutoDock (The Scripps Research Institute), yet it still has several differences compared to the original AutoDock, for example, a different scoring function and new optimization algorithms. It uses, however, the same structure formats as AutoDock4 and also employs pre-computation of interaction grids, hence, a transfer of the flexible receptor docking approach into AutoDock Vina might be a promising attempt for future studies.

Appendix

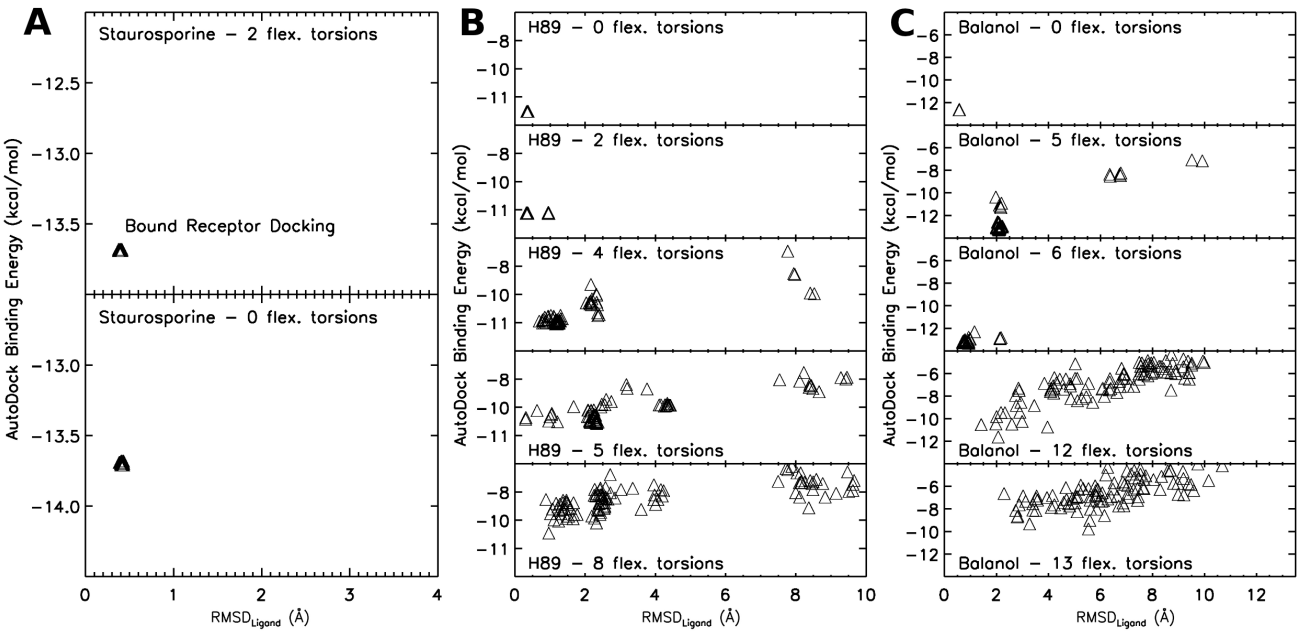


Figure A.1: *Calculated AutoDock scoring energy for 100 separate holo docking (re-docking of ligands into the bound receptor structure) results versus $RMSD_{ligand}$. Results are shown for the ligands staurosporine (A), H89 (B), and balanol (C), each with different numbers of mobile dihedral torsion angles.*

ligand rigid

receptor→ ligand↓	1AJV	1DMP	1G2K	1HVH	1HWR	2UPJ	7UPJ	apo
NMB	0.41	9.36	0.54	2.53	0.81	0.68	0.62	7.54
	(100)	(0)	(100)	(0)	(100)	(100)	100	0
DMQ	0.59	0.51	0.61	1.29	0.52	0.73	1.03	7.73
	(100)	(100)	(100)	(100)	(100)	(100)	(100)	(0)
NM1	0.34	10.56	0.47	0.89	0.78	0.88	0.57	1.27
	(100)	(0)	(100)	(100)	(100)	(100)	(100)	(19)
Q82	1.76	5.24	1.78	5.49	5.65	5.28	5.32	5.40
	(100)	(0)	(100)	(0)	(0)	(0)	(0)	(0)
216	0.58	0.77	0.67	5.38	0.47	0.45	1.34	6.30
	(100)	(100)	(100)	(0)	(100)	(100)	(100)	(0)
U02	11.77	1.27	11.94	11.71	0.78	1.04	0.60	9.82
	(0)	(100)	(0)	(0)	(39)	(100)	(100)	(0)
INU	1.20	9.83	2.00	2.50	1.00	0.59	0.79	9.06
	(100)	(0)	(0)	(0)	(1)	(100)	(100)	(0)

ligand flexible

receptor→ ligand↓	1AJV	1DMP	1G2K	1HVH	1HWR	2UPJ	7UPJ	apo
NMB	0.40	1.32	0.50	1.06	0.84	0.78	0.62	2.09
	(98)	(5)	(94)	(23)	(64)	(72)	(60)	(0)
DMQ	0.65	0.39	0.62	0.93	0.49	0.70	1.13	6.14
	(86)	(99)	(90)	(98)	(100)	(94)	(93)	(0)
NM1	0.36	1.34	0.39	1.26	1.02	0.88	0.66	6.45
	(92)	(32)	(98)	(64)	(72)	(76)	(60)	(0)
Q82	1.84	1.33	1.84	0.80	1.18	1.98	1.99	5.42
	(40)	(2)	(10)	(1)	(4)	(1)	(1)	(0)
216	0.62	0.81	0.61	1.25	0.56	0.62	0.90	2.1
	(92)	(100)	(93)	(25)	(100)	(71)	(87)	(0)
U02	1.04	1.27	1.44	1.19	0.81	0.99	0.73	5.57
	(11)	(36)	(8)	(17)	(16)	(88)	(39)	(0)
INU	0.94	1.67	1.59	2.15	4.56	0.70	0.72	6.02
	(83)	(10)	(78)	(0)	(0)	(100)	(100)	(0)

Table A.1: *Cross docking RMSD results (percentage of docking results with an $RMSD_{Ligand}$ < 2.0Å). Underlined RMSD values represent the holo docking cases where the ligands were redocked to their original bound structure of HIV1-Protease.*

ligand rigid

receptor→ ligand↓	1AJV	1DMP	1G2K	1HVH	1HWR	2UPJ	7UPJ	flex-rec.
AH1	0.86 (100)	0.75 (86)	1.08 (100)	1.52 (100)	1.26 (82)	1.22 (100)	0.71 (75)	0.68 (0.68) (87)
XV6	0.26 (100)	0.43 (7)	0.46 (100)	0.52 (100)	0.53 (13)	11.97 (0)	0.56 (12)	0.17 (0.46) (97)
AHF	0.36 (100)	0.73 (40)	0.83 (100)	11.08 (0)	0.68 (25)	2.46 (0)	0.20 (44)	0.28 (0.37) (100)
1UN	9.42 (0)	0.65 (35)	9.39 (0)	0.45 (100)	0.79 (76)	9.41 (0)	1.14 (6)	9.40 (9.40) (0)
A88	9.55 (0)	0.19 (52)	0.19 (100)	1.10 (100)	1.03 (80)	9.54 (0)	0.35 (48)	0.19 (0.20) (28)
846	10.59 (0)	0.35 (53)	10.63 (0)	1.03 (100)	1.36 (78)	0.60 (100)	0.57 (29)	0.34 (10.64) (55)
BH0	0.45 (9)	1.08 (45)	9.46 (0)	0.95 (14)	1.31 (100)	9.37 (0)	0.35 (100)	0.25(9.46) (90)

ligand flexible

receptor→ ligand↓	1AJV	1DMP	1G2K	1HVH	1HWR	2UPJ	7UPJ	flex-rec.
AH1	1.06 (77)	1.11 (100)	0.63 (87)	0.89 (98)	0.67 (100)	0.72 (79)	0.82 (100)	0.71 (8.56) (26)
XV6	0.65 (2)	11.98 (0)	0.51 (12)	0.95 (2)	0.30 (100)	0.40 (13)	0.16 (100)	0.39 (0.39) (5)
AHF	0.40 (62)	0.22 (100)	0.36 (42)	1.41 (12)	0.52 (100)	0.33 (25)	0.24 (100)	0.20 (0.41) (31)
1UN	1.79 (2)	9.43 (0)	4.06 (0)	4.04 (0)	9.35 (0)	4.00 (0)	9.44 (0)	0.83 (9.40) (16)
A88	0.57 (37)	9.56 (0)	0.36 (38)	0.20 (14)	9.54 (0)	0.35 (16)	9.55 (0)	0.20 (0.21) (43)
846	0.59 (45)	0.58 (100)	0.38 (29)	0.48 (98)	0.41 (45)	0.33 (68)	0.41 (100)	0.35 (0.41) (79)
BH0	0.81 (38)	0.67 (56)	0.30 (55)	1.48 (9)	0.22 (100)	0.53 (84)	0.29 (100)	0.36 (9.52) (26)

Table A.2: *RMSD results (best RMSD in Å, bottom line in brackets: percentage of docking results with $RMSD_{Ligand} < 2.0$Å) for cross docking of foreign HIV1-Protease ligands. For flexible receptor docking, the best $RMSD_{Ligand}$ is followed by the best energy RMSD in brackets.*

ligand rigid

	Flexible receptor docking		Apo docking	
	best RMSD	low.en. RMSD	best RMSD	low.en. RMSD
NMB	0.50(98)|0.62(99)|0.81(56)|0.80(51)	0.54(0.66|0.86|9.4)	7.54(1)	7.74
DMQ	0.76(100)|0.86(100)|0.52(100)|1.14(100)	0.77(1.31|0.53|1.15)	7.73(0)	7.74
NM1	0.44(96)|0.55(94)|0.58(85)|0.87(36)	0.54(0.56|0.65|10.39)	1.27(19)	10.34
Q82	5.25(0)|5.31(0)|0.90(2)|5.42(0))	5.26(5.36|5.48|5.43)	5.40(0)	5.59
U02	1.05(99)|0.87(80)|0.72(23)|0.98(61)	1.08(0.88|11.97|1.00)	9.82(0)	9.86
INU	1.05(82)|0.73(97)|0.88(12)|2.62(0)	9.90(0.76|10.04|8.52)	9.06(0)	9.07

ligand flexible

	Flexible receptor docking		Apo docking	
	best RMSD	low.en. RMSD	best RMSD	low.en. RMSD
NMB	0.76(22)|0.76(21)|0.82(12)|0.80(19)	9.14(1.09|6.20|9.20)	2.09(1)	7.58
DMQ	0.82(53)|0.88(58)|0.58(47)|0.91(65)	1.09(2.03|3.20|1.08)	6.14(0)	7.87
NM1	0.73(36)|0.72(37)|0.80(21)|0.80(30)	6.56(10.36|10.36|10.24)	6.45(0)	7.38
Q82	5.20(0)|5.36(0)|5.29(0)|2.83(0)	7.01(5.67|5.63|5.50)	5.42(0)	5.97
216	0.92(24)|1.04(33)|0.59(27)|1.09(10)	1.43(7.22|6.25|4.57)	2.10(4)	6.73
U02	0.99(14)|1.36(2)|1.04(12)|1.12(4)	11.69(7.07|7.56|6.68)	5.57(0)	9.23
INU	0.73(26)|0.72(45)|0.84(11)|2.70(0)	8.09(1.64|9.86|5.00)	6.02(0)	6.50

Table A.3: *Best RMSD in Å (i.e. the lowest yielded $RMSD_{Ligand}$ out of 100 separate docking runs) and the RMSD values of the docking solution with the lowest AutoDock binding energy for different ligand and receptor flexibilities. Values are for morhing apo→2UPJ, values in brackets are for morphs from apo to 7UPJ, 1HWR, and 1HVH, respectively.*

List of Publications

Parts of this thesis have been published in peer-reviewed journals and a book chapter.

- Simon Leis and Martin Zacharias,
  "ReFlexIn: A flexible Receptor Protein-Ligand Docking Scheme evaluated on HIV-1 Protease",
  PLOS ONE, 7(10): e48008, **2012**

- Simon Leis and Martin Zacharias,
  "Accounting for target flexibility during ligand-receptor docking",
  in the book "Physico-Chemical and Computational Approaches to Drug Discovery",
  RSC Drug Discovery Series, pp. 223-243, ISBN:978-1849733533, **2012**

- Simon Leis and Martin Zacharias,
  "Efficient inclusion of Receptor Flexibility in grid-based Protein-Ligand Docking.",
  Journal of Computational Chemistry, 32(16) pp. 3433-3439, **2011**

- Simon Leis, Sebastian Schneider, and Martin Zacharias,
  "In silico prediction of binding sites on proteins.",
  Current Medicinal Chemistry, 17(15), pp. 1550-62., **2010**

Danksagung

Bei Herrn Prof. Zacharias bedanke ich mich herzlich für die Möglichkeit, meine Doktorarbeit an seinem Lehrstuhl in einer solch guten Arbeitsatmosphäre anfertigen zu können. Seine Tür steht für alle Mitarbeiter stets offen und seine immer gute Laune und Bereitschaft sich Zeit zu nehmen sind bemerkenswert. Für alles Gelernte: Vielen Dank.

Der Prüfungskommission, Frau Prof. Antes und Herrn Prof. Simmel, danke ich für die Zeit, die sie in die Begutachtung meiner Arbeit und den Vorsitz meiner Prüfung investiert haben.

Den beiden Sekretärinnen und guten Seelen des Lehrstuhls, Jill Seidlitz und Sonja Ortner, danke ich herzlich für Korrekturlesen, Unterstützung jeglicher Art und das Fachsimpeln über die Uni, den Sport und das Leben. Allen ehemaligen und aktuellen Mitgliedern der Arbeitsgruppe danke ich für eine gute Zusammenarbeit, eine tolle Zeit auf Konferenzen und bei Unternehmungen, fachliche und nichtfachliche Gespräche, sowie hilfreiches Feedback bei Seminaren. Besonderer Dank geht an Piotr Setny für seine Unterstützung im Büro und die vielen Runden im Echinger See.

Diese Arbeit wäre nie entstanden ohne den Rückhalt, den ich durch meine Familie und meine Freunde erfahren habe. Für Eure Unterstützung, Geduld und Liebe danke ich Euch herzlich.

List of Figures

Bibliography

[1] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and G. N. D. C. Phillips. A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature*, 181(4610):662–666, 1958.

[2] M. Caffrey. Membrane protein crystallization. *J Struct Biol*, 142(1):108–132, 2003.

[3] M. Caffrey. Crystallizing membrane proteins for structure determination: use of lipidic mesophases. *Annu Rev Biophys*, 38:29–51, 2009.

[4] E. E. Lattman and P. J. Loll. *Protein Crystallography – A Concise Guide*. The Johns Hopkins University Press, 2008.

[5] R. Ishima and D. A. Torchia. Protein dynamics from NMR. *Nat Struct Biol*, 7(9):740–743, 2000.

[6] A. Wlodawer, W. Minor, Z. Dauter, and M. Jaskolski. Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *FEBS J*, 275(1):1–21, 2008.

[7] D. Wishart. NMR spectroscopy and protein structure determination: applications to drug discovery and development. *Curr Pharm Biotechnol*, 6(2):105–120, 2005.

[8] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res*, 28(1):235–242, 2000.

[9] E. Fischer. Einfluss der Configuration auf die Wirkung der Enzyme. *Ber. Dt. Chem. Ges.*, 27:2985–2993, 1894.

[10] D. Koshland, W. Ray, and M. Erwin. Protein structure and enzyme action. *Federation Proceedings*, 17(4):1145–1150, 1958.

[11] K. Gunasekaran and R. Nussinov. How different are structurally flexible and rigid binding sites? Sequence and structural features discriminating proteins that do and do not undergo conformational change upon ligand binding. *Journal of Molecular Biology*, 365(1):257 – 273, 2007.

[12] M. Gerstein and W. Krebs. A database of macromolecular motions. *Nucleic Acids Res*, 26(18):4280–4290, 1998.

[13] T. Hansson, C. Oostenbrink, and W. van Gunsteren. Molecular dynamics simulations. *Current Opinion in Structural Biology*, 12(2):190 – 196, 2002.

[14] W. G. Krebs, V. Alexandrov, C. A. Wilson, N. Echols, H. Yu, and M. Gerstein. Normal mode analysis of macromolecular motions in a database framework: Developing mode concentration as a useful classifying statistic. *Proteins: Structure, Function, and Bioinformatics*, 48(4):682–695, 2002.

[15] I. Luque and E. Freire. Structural stability of binding sites: consequences for binding affinity and allosteric effects. *Proteins*, Suppl 4:63–71, 2000.

[16] R. Najmanovich, J. Kuttner, V. Sobolev, and M. Edelman. Side-chain flexibility in proteins upon ligand binding. *Proteins: Structure, Function, and Bioinformatics*, 39(3):261–268, 2000.

[17] I. Halperin, B. Ma, H. Wolfson, and R. Nussinov. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins*, 47(4):409–443, 2002.

[18] J. A. McCammon. Protein dynamics. *Reports on Progress in Physics*, 47(1):1, 1984.

[19] K. Teilum, J. G. Olsen, and B. B. Kragelund. Functional aspects of protein flexibility. *Cell Mol Life Sci*, 66(14):2231–2247, 2009.

[20] M. Karplus. Molecular dynamics of biological macromolecules: a brief history and perspective. *Biopolymers*, 68(3):350–358, 2003.

[21] J. D. Durrant and J. A. McCammon. Molecular dynamics simulations and drug discovery. *BMC Biol*, 9:71, 2011.

[22] R. O. Dror, R. M. Dirks, J. P. Grossman, H. Xu, and D. E. Shaw. Biomolecular simulation: a computational microscope for molecular biology. *Annu Rev Biophys*, 41:429–452, 2012.

[23] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe. A fast flexible docking method using an incremental construction algorithm. *J Mol Biol*, 261(3):470–489, 1996.

[24] R. Abagyan, M. Totrov, and D. Kuznetsov. ICM – a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.*, 15:488–506, 1994.

[25] R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, D. E. Shaw, P. Francis, and P. S. Shenkin. GLIDE: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem*, 47(7):1739–1749, 2004.

[26] G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol*, 267(3):727–748, 1997.

[27] D. S. Goodsell and A. J. Olson. Automated docking of substrates to proteins by simulated annealing. *Proteins*, 8(3):195–202, 1990.

[28] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson. Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry*, 19(14):1639–1662, 1998.

[29] R. Huey, G. M. Morris, A. J. Olson, and D. S. Goodsell. A semiempirical free energy force field with charge-based desolvation. *J Comput Chem*, 28(6):1145–1152, 2007.

[30] G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, and A. J. Olson. Autodock4 and autodocktools4: Automated docking with selective receptor flexibility. *J Comput Chem*, 30(16):2785–2791, 2009.

[31] *The PyMOL Molecular Graphics System, Version 1.2r2 Schrödinger, LLC.* http://pymol.org.

[32] M. A. Balsera, W. Wriggers, Y. Oono, and K. Schulten. Principal component analysis and long time protein dynamics. *The Journal of Physical Chemistry*, 100(7):2567–2572, 1996.

[33] O. F. Lange and H. Grubmüller. Can principal components yield a dimension reduced description of protein dynamics on long time scales? *J Phys Chem B*, 110(45):22842–22852, 2006.

[34] I. Bahar, A. R. Atilgan, and B. Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold Des*, 2(3):173–181, 1997.

[35] K. Hinsen. Analysis of domain motions by approximate normal mode calculations. *Proteins*, 33(3):417–429, 1998.

[36] K. L. Meagher and H. A. Carlson. Incorporating protein flexibility in structure-based drug discovery: using HIV-1 protease as a test case. *J Am Chem Soc*, 126(41):13276–13281, 2004.

[37] R. M. Knegtel, I. D. Kuntz, and C. M. Oshiro. Molecular docking to ensembles of protein structures. *J Mol Biol*, 266(2):424–440, 1997.

[38] F. Tama and Y. H. Sanejouand. Conformational change of proteins arising from normal mode calculations. *Protein Eng*, 14(1):1–6, 2001.

[39] A. May and M. Zacharias. Energy minimization in low-frequency normal modes to efficiently allow for global flexibility during systematic protein-protein docking. *Proteins*, 70(3):794–809, 2008.

[40] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam. The protein kinase complement of the human genome. *Science*, 298(5600):1912–1934, 2002.

[41] D. J. Matthews and M. E. Gerritsen. *Targeting Protein Kinases for Cancer Therapy*. John Wiley & Sons, 2010.

[42] P. Cohen. Protein kinases–the major drug targets of the twenty-first century? *Nat Rev Drug Discov*, 1(4):309–315, 2002.

[43] M. Vieth, J. J. Sutherland, D. H. Robertson, and R. M. Campbell. Kinomics: characterizing the therapeutically validated kinase space. *Drug Discov Today*, 10(12):839–846, 2005.

[44] D. R. Knighton, J. H. Zheng, L. F. T. Eyck, V. A. Ashford, N. H. Xuong, S. S. Taylor, and J. M. Sowadski. Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science*, 253(5018):407–414, 1991.

[45] A. Kamb. Cell-cycle regulators and cancer. *Trends Genet*, 11(4):136–140, 1995.

[46] D. O. Morgan. Cyclin-dependent kinases: Engines, clocks, and microprocessors. *Annual Review of Cell and Developmental Biology*, 13(1):261–291, 1997.

[47] G. I. Shapiro and J. W. Harper. Anticancer drug targets: cell cycle and checkpoint control. *J Clin Invest*, 104(12):1645–1653, 1999.

[48] E. A. Sausville. Complexities in the development of cyclin-dependent kinase inhibitor drugs. *Trends Mol Med*, 8(4 Suppl):S32–S37, 2002.

[49] N. E. Kohl, E. A. Emini, W. A. Schleif, L. J. Davis, J. C. Heimbach, R. A. Dixon, E. M. Scolnick, and I. S. Sigal. Active human immunodeficiency virus protease is required for viral infectivity. *Proc Natl Acad Sci U S A*, 85(13):4686–4690, 1988.

[50] M. Miller, M. Jaskólski, J. K. Rao, J. Leis, and A. Wlodawer. Crystal structure of a retroviral protease proves relationship to aspartic protease family. *Nature*, 337(6207):576–579, 1989.

[51] D. I. Freedberg, R. Ishima, J. Jacob, Y.-X. Wang, I. Kustanovich, J. M. Louis, and D. A. Torchia. Rapid structural fluctuations of the free HIV protease flaps in solution: relationship to crystal structures and comparison with predictions of dynamics calculations. *Protein Sci*, 11(2):221–232, 2002.

[52] R. Ishima, D. I. Freedberg, Y. X. Wang, J. M. Louis, and D. A. Torchia. Flap opening and dimer-interface flexibility in the free and inhibitor-bound HIV protease, and their implications for function. *Structure*, 7(9):1047–1055, 1999.

[53] P. L. Darke, C. T. Leu, L. J. Davis, J. C. Heimbach, R. E. Diehl, W. S. Hill, R. A. Dixon, and I. S. Sigal. Human immunodeficiency virus protease. bacterial expression and characterization of the purified aspartic protease. *J Biol Chem*, 264(4):2307–2312, 1989.

[54] S. Seelmeier, H. Schmidt, V. Turk, and K. von der Helm. Human immunodeficiency virus has an aspartic-type protease that can be inhibited by pepstatin A. *Proc Natl Acad Sci U S A*, 85(18):6612–6616, 1988.

[55] J. Mous, E. P. Heimer, and S. F. L. Grice. Processing protease and reverse transcriptase from human immunodeficiency virus type I polyprotein in Escherichia coli. *J Virol*, 62(4):1433–1436, 1988.

[56] G. Schneider and U. Fechner. Computer-based de novo design of drug-like molecules. *Nat Rev Drug Discov*, 4(8):649–663, 2005.

[57] A. M. Lesk and C. Chothia. Mechanisms of domain closure in proteins. *J Mol Biol*, 174(1):175–191, 1984.

[58] M. Gerstein and C. Chothia. Analysis of protein loop closure. two types of hinges produce one motion in lactate dehydrogenase. *J Mol Biol*, 220(1):133–149, 1991.

[59] M. Gerstein, A. M. Lesk, and C. Chothia. Structural mechanisms for domain movements in proteins. *Biochemistry*, 33(22):6739–6749, 1994.

[60] R. B. Best and G. Hummer. Unfolding the secrets of calmodulin. *Science*, 323(5914):593–594, 2009.

[61] C. J. Tsai, S. Kumar, B. Ma, and R. Nussinov. Folding funnels, binding funnels, and protein function. *Protein Sci*, 8(6):1181–1190, 1999.

[62] H. A. Carlson and J. A. McCammon. Accommodating protein flexibility in computational drug design. *Mol Pharmacol*, 57(2):213–218, 2000.

[63] A. V. Finkelstein and J. Janin. The price of lost freedom: entropy of bimolecular complex formation. *Protein Engineering*, 3(1):1–3, 1989.

[64] G. Klebe and H. J. Böhm. Energetic and entropic factors determining binding affinity in protein-ligand complexes. *J Recept Signal Transduct Res*, 17(1-3):459–473, 1997.

[65] W. P. Jencks. On the attribution and additivity of binding energies. *Proc Natl Acad Sci U S A*, 78(7):4046–4050, 1981.

[66] P. R. Andrews, D. J. Craik, and J. L. Martin. Functional group contributions to drug-receptor interactions. *J Med Chem*, 27(12):1648–1657, 1984.

[67] A. R. Leach. *Molecular Modelling, Principles and Applications*. Pearson Education Limited, 2001.

[68] S. A. Adcock and J. A. McCammon. Molecular dynamics: survey of methods for simulating the activity of proteins. *Chem Rev*, 106(5):1589–1615, 2006.

[69] C. B-Rao, J. Subramanian, and S. D. Sharma. Managing protein flexibility in docking and its applications. *Drug Discov Today*, 14(7-8):394–400, 2009.

[70] S.-Y. Huang and X. Zou. Advances and challenges in protein-ligand docking. *Int J Mol Sci*, 11(8):3016–3034, 2010.

[71] N. Trbovic, J.-H. Cho, R. Abel, R. A. Friesner, M. Rance, and A. G. Palmer. Protein side-chain dynamics and residual conformational entropy. *J Am Chem Soc*, 131(2):615–622, 2009.

[72] S.-Y. Huang and X. Zou. Inclusion of solvation and entropy in the knowledge-based scoring function for protein-ligand interactions. *J Chem Inf Model*, 50(2):262–273, 2010.

[73] I. Muegge. PMF scoring revisited. *J Med Chem*, 49(20):5895–5902, 2006.

[74] F. Jiang and S. H. Kim. "Soft docking": matching of molecular surface cubes. *J Mol Biol*, 219(1):79–102, 1991.

[75] J. Apostolakis and A. Caflisch. Computational ligand design. *Comb Chem High Throughput Screen*, 2(2):91–104, 1999.

[76] D. M. Lorber and B. K. Shoichet. Flexible ligand docking using conformational ensembles. *Protein Sci*, 7(4):938–950, 1998.

[77] A. Di Nola, D. Roccatano, and H. J. C. Berendsen. Molecular dynamics simulation of the docking of substrates to proteins. *Proteins: Structure, Function, and Bioinformatics*, 19(3):174–182, 1994.

[78] Z. Huang, C. F. Wong, and R. A. Wheeler. Flexible protein-flexible ligand docking with disrupted velocity simulated annealing. *Proteins*, 71(1):440–454, 2008.

[79] A. Steffen, A. Kämper, and T. Lengauer. Flexible docking of ligands into synthetic receptors using a two-sided incremental construction algorithm. *J Chem Inf Model*, 46(4):1695–1703, 2006.

[80] G. Schneider and H.-J. Böhm. Virtual screening and fast automated docking methods. *Drug Discov Today*, 7(1):64–70, 2002.

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

[82] R. Abagyan and M. Totrov. High-throughput docking for lead generation. *Curr Opin Chem Biol*, 5(4):375–382, 2001.

[83] C. N. Cavasotto and R. A. Abagyan. Protein flexibility in ligand docking and virtual screening to protein kinases. *J Mol Biol*, 337(1):209–225, 2004.

[84] I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin. A geometric approach to macromolecule-ligand interactions. *J Mol Biol*, 161(2):269–288, 1982.

[85] A. M. Ferrari, B. Q. Wei, L. Costantino, and B. K. Shoichet. Soft docking and multiple receptor conformations in virtual screening. *J Med Chem*, 47(21):5076–5084, 2004.

[86] A. May, F. Sieker, and M. Zacharias. How to efficiently include receptor flexibility during computational docking. *Current Computer - Aided Drug Design*, 4(2):143–153, 2008.

[87] M. Karplus. Molecular dynamics simulations of biomolecules. *Acc Chem Res*, 35(6):321–323, 2002.

[88] Z. R. Wasserman and C. N. Hodge. Fitting an inhibitor into the active site of thermolysin: a molecular dynamics case study. *Proteins*, 24(2):227–237, 1996.

[89] M. Mangoni, D. Roccatano, and A. D. Nola. Docking of flexible ligands to flexible receptors in solution by molecular dynamics simulation. *Proteins*, 35(2):153–162, 1999.

[90] J. A. Given and M. K. Gilson. A hierarchical method for generating low-energy conformers of a protein-ligand complex. *Proteins*, 33(4):475–495, 1998.

[91] R. N. Riemann and M. Zacharias. Refinement of protein cores and protein-peptide interfaces using a potential scaling approach. *Protein Eng Des Sel*, 18(10):465–476, 2005.

[92] Y. Pak and S. Wang. Application of a molecular dynamics simulation method with a generalized effective potential to the flexible molecular docking problems. *The Journal of Physical Chemistry B*, 104(2):354–359, 2000.

[93] H. Alonso, A. A. Bliznyuk, and J. E. Gready. Combining docking and molecular dynamic simulations in drug design. *Med Res Rev*, 26(5):531–568, 2006.

[94] I. Antes. Dynadock: A new molecular dynamics-based algorithm for protein-peptide docking including receptor flexibility. *Proteins*, 78(5):1084–1104, 2010.

[95] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *Journal of the American Chemical Society*, 112(16):6127–6129, 1990.

[96] N. Okimoto, N. Futatsugi, H. Fuji, A. Suenaga, G. Morimoto, R. Yanai, Y. Ohno, T. Narumi, and M. Taiji. High-performance drug discovery: computational screening by combining docking and molecular dynamics simulations. *PLoS Comput Biol*, 5(10):e1000528, 2009.

[97] C. F. Wong and J. A. McCammon. Protein flexibility and computer-aided drug design. *Annu Rev Pharmacol Toxicol*, 43:31–45, 2003.

[98] Y. Deng and B. Roux. Computations of standard binding free energies with molecular dynamics simulations. *J Phys Chem B*, 113(8):2234–2246, 2009.

[99] J. Michel and J. W. Essex. Prediction of protein-ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations. *J Comput Aided Mol Des*, 24(8):639–658, 2010.

[100] J. A. Erickson, M. Jalaie, D. H. Robertson, R. A. Lewis, and M. Vieth. Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. *J Med Chem*, 47(1):45–55, 2004.

[101] R. L. Dunbrack and M. Karplus. Backbone-dependent rotamer library for proteins. application to side-chain prediction. *J Mol Biol*, 230(2):543–574, 1993.

[102] J. W. Ponder and F. M. Richards. Tertiary templates for proteins. use of packing criteria in the enumeration of allowed sequences for different structural classes. *J Mol Biol*, 193(4):775–791, 1987.

[103] M. D. Maeyer, J. Desmet, and I. Lasters. All in one: a highly detailed rotamer library improves both accuracy and speed in the modelling of sidechains by dead-end elimination. *Fold Des*, 2(1):53–66, 1997.

[104] S. C. Lovell, J. M. Word, J. S. Richardson, and D. C. Richardson. The penultimate rotamer library. *Proteins*, 40(3):389–408, 2000.

[105] P. Tuffery, C. Etchebest, S. Hazout, and R. Lavery. A new approach to the rapid determination of protein side chain conformations. *J Biomol Struct Dyn*, 8(6):1267–1289, 1991.

[106] Z. Xiang and B. Honig. Extending the accuracy limits of prediction for side-chain conformations. *J Mol Biol*, 311(2):421–430, 2001.

[107] Z. Xiang, P. J. Steinbach, M. P. Jacobson, R. A. Friesner, and B. Honig. Prediction of side-chain conformations on protein surfaces. *Proteins*, 66(4):814–823, 2007.

[108] A. R. Leach. Ligand docking to proteins with discrete side-chain flexibility. *J Mol Biol*, 235(1):345–356, 1994.

[109] R. L. Dunbrack and F. E. Cohen. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci*, 6(8):1661–1681, 1997.

[110] J. Desmet, M. D. Maeyer, B. Hazes, and I. Lasters. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, 356(6369):539–542, 1992.

[111] A. R. Leach and A. P. Lemon. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. *Proteins: Structure, Function, and Bioinformatics*, 33(2):227–239, 1998.

[112] E. Althaus, O. Kohlbacher, H.-P. Lenhof, and P. Müller. A combinatorial approach to protein docking with flexible side chains. *J Comput Biol*, 9(4):597–612, 2002.

[113] L. Schaffer and G. M. Verkhivker. Predicting structural effects in HIV-1 protease mutant complexes with flexible ligand docking and protein side-chain optimization. *Proteins*, 33(2):295–310, 1998.

[114] C. Wang, O. Schueler-Furman, and D. Baker. Improved side-chain modeling for protein-protein docking. *Protein Sci*, 14(5):1328–1339, 2005.

[115] J. Meiler and D. Baker. ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. *Proteins*, 65(3):538–548, 2006.

[116] M. Totrov and R. Abagyan. Flexible protein-ligand docking by global energy optimization in internal coordinates. *Proteins*, Suppl 1:215–220, 1997.

[117] J. Fernández-Recio, M. Totrov, and R. Abagyan. ICM-DISCO docking by global energy optimization with fully flexible side-chains. *Proteins*, 52(1):113–117, 2003.

[118] W. Sherman, T. Day, M. P. Jacobson, R. A. Friesner, and R. Farid. Novel procedure for modeling ligand/receptor induced fit effects. *J Med Chem*, 49(2):534–553, 2006.

[119] K. Bastard, A. Thureau, R. Lavery, and C. Prevost. Docking macromolecules with flexible segments. *J Comput Chem*, 24(15):1910–1920, 2003.

[120] K. Bastard, C. Prevost, and M. Zacharias. Accounting for loop flexibility during protein-protein docking. *Proteins*, 62(4):956–969, 2006.

[121] A. Shehu, C. Clementi, and L. E. Kavraki. Modeling protein conformational ensembles: from missing loops to equilibrium fluctuations. *Proteins*, 65(1):164–179, 2006.

[122] D. Seeliger, J. Haas, and B. L. de Groot. Geometry-based sampling of conformational transitions in proteins. *Structure*, 15(11):1482–1492, 2007.

[123] S. Eyrisch and V. Helms. What induces pocket openings on protein surface patches involved in protein-protein interactions? *J Comput Aided Mol Des*, 23(2):73–86, 2009.

[124] D. Seeliger and B. L. de Groot. Conformational transitions upon ligand binding: holo-structure prediction from apo conformations. *PLoS Comput Biol*, 6(1):e1000634, 2010.

[125] A. Benedix, C. M. Becker, B. L. de Groot, A. Caflisch, and R. A. Böckmann. Predicting free energy changes using structural ensembles. *Nat Methods*, 6(1):3–4, 2009.

[126] M. L. Teodoro, G. N. Phillips, and L. E. Kavraki. Understanding protein flexibility through dimensionality reduction. *J Comput Biol*, 10(3-4):617–634, 2003.

[127] D. Mustard and D. W. Ritchie. Docking essential dynamics eigenstructures. *Proteins*, 60(2):269–274, 2005.

[128] C. Cavasotto, J. A. Kovacs, and R. Abagyan. Representing receptor flexibility in ligand docking through relevant normal modes. *J Am Chem Soc*, 127(26):9632–9640, 2005.

[129] Y. P. Pang and A. P. Kozikowski. Prediction of the binding sites of huperzine A in acetylcholinesterase by docking studies. *J Comput Aided Mol Des*, 8(6):669–681, 1994.

[130] X. Barril and S. D. Morley. Unveiling the full potential of flexible receptor docking using multiple crystallographic structures. *J Med Chem*, 48(13):4432–4443, 2005.

[131] E. Moreno and K. León. Geometric and chemical patterns of interaction in protein–ligand complexes and their application in docking. *Proteins*, 47(1):1–13, 2002.

[132] J.-H. Lin, A. L. Perryman, J. R. Schames, and J. A. McCammon. Computational drug design accommodating receptor flexibility: the relaxed complex scheme. *J Am Chem Soc*, 124(20):5632–5633, 2002.

[133] H. A. Carlson, K. M. Masukawa, and J. A. McCammon. Method for including the dynamic fluctuations of a protein in computer-aided drug design. *The Journal of Physical Chemistry A*, 103(49):10213–10219, 1999.

[134] F. Österberg, G. M. Morris, M. F. Sanner, A. J. Olson, and D. S. Goodsell. Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in autodock. *Proteins*, 46(1):34–40, 2002.

[135] H. Claussen, C. Buning, M. Rarey, and T. Lengauer. FlexE: efficient molecular docking considering protein structure variations. *J Mol Biol*, 308(2):377–395, 2001.

[136] G. R. Smith, M. J. E. Sternberg, and P. A. Bates. The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes with an application to protein-protein docking. *J Mol Biol*, 347(5):1077–1101, 2005.

[137] M. Totrov and R. Abagyan. Flexible ligand docking to multiple receptor conformations: a practical alternative. *Curr Opin Struct Biol*, 18(2):178–184, 2008.

[138] S.-Y. Huang and X. Zou. Ensemble docking of multiple protein structures: considering protein structural variations in molecular docking. *Proteins*, 66(2):399–421, 2007.

[139] E. M. Novoa, L. R. d. Pouplana, X. Barril, and M. Orozco. Ensemble docking from homology models. *Journal of Chemical Theory and Computation*, 6(8):2547–2557, 2010.

[140] G. Bottegoni, I. Kufareva, M. Totrov, and R. Abagyan. Four-dimensional docking: a fast and accurate account of discrete receptor flexibility in ligand docking. *J Med Chem*, 52(2):397–406, 2009.

[141] I. R. Craig, J. W. Essex, and K. Spiegel. Ensemble docking into multiple crystallographically derived protein structures: an evaluation based on the statistical analysis of enrichments. *J Chem Inf Model*, 50(4):511–524, 2010.

[142] M. Rueda, G. Bottegoni, and R. Abagyan. Consistent improvement of cross-docking results using binding site ensembles generated with elastic network normal modes. *J Chem Inf Model*, 49(3):716–725, 2009.

[143] S.-J. Park, I. Kufareva, and R. Abagyan. Improved docking, screening and selectivity prediction for small molecule nuclear receptor modulators using conformational ensembles. *J Comput Aided Mol Des*, 24(5):459–471, 2010.

[144] A. Amadei, A. B. Linssen, and H. J. Berendsen. Essential dynamics of proteins. *Proteins*, 17(4):412–425, 1993.

[145] M. Zacharias and H. Sklenar. Harmonic modes as variables to approximately account for receptor flexibility in ligand–receptor docking simulations: Application to DNA minor groove ligand complex. *Journal of Computational Chemistry*, 20(3):287–300, 1999.

[146] G. M. Keserû and I. Kolossváry. Fully flexible low-mode docking: application to induced fit in HIV integrase. *J Am Chem Soc*, 123(50):12708–12709, 2001.

[147] R. Tatsumi, Y. Fukunishi, and H. Nakamura. A hybrid method of molecular dynamics and harmonic dynamics for docking of flexible ligand to flexible receptor. *J Comput Chem*, 25(16):1995–2005, 2004.

[148] M. Zacharias. Rapid protein–ligand docking using soft modes from molecular dynamics simulations to account for protein deformability: Binding of FK506 to FKBP. *Proteins: Structure, Function, and Bioinformatics*, 54(4):759–767, 2004.

[149] M. Shatsky, R. Nussinov, and H. J. Wolfson. Flexprot: alignment of flexible protein structures without a predefinition of hinge regions. *J Comput Biol*, 11(1):83–106, 2004.

[150] B. Sandak, H. J. Wolfson, and R. Nussinov. Flexible docking allowing induced fit in proteins: insights from an open to closed conformational isomers. *Proteins*, 32(2):159–174, 1998.

[151] B. Sandak, R. Nussinov, and H. J. Wolfson. A method for biomolecular structural recognition and docking allowing conformational flexibility. *J Comput Biol*, 5(4):631–654, 1998.

[152] D. Tobi and I. Bahar. Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. *Proc Natl Acad Sci U S A*, 102(52):18908–18913, 2005.

[153] A. May and M. Zacharias. Protein-ligand docking accounting for receptor side chain and global flexibility in normal modes: evaluation on kinase inhibitor cross docking. *J Med Chem*, 51(12):3499–3506, 2008.

[154] A. May and M. Zacharias. Accounting for global protein deformability during protein-protein and protein-ligand docking. *Biochim Biophys Acta*, 1754(1-2):225–231, 2005.

[155] S. Kazemi, D. M. Krüger, F. Sirockin, and H. Gohlke. Elastic potential grids: accurate and efficient representation of intermolecular interactions for fully flexible docking. *ChemMedChem*, 4(8):1264–1268, 2009.

[156] X. Barril and X. Fradera. Incorporating protein flexibility into docking and structure-based drug design. *Expert Opinion on Drug Discovery*, 1(4):335–349, 2006.

[157] M. J. Osborne, J. Schnell, S. J. Benkovic, H. J. Dyson, and P. E. Wright. Backbone dynamics in dihydrofolate reductase complexes: role of loop flexibility in the catalytic mechanism. *Biochemistry*, 40(33):9846–9859, 2001.

[158] A. L. Bowman, M. G. Lerner, and H. A. Carlson. Protein flexibility and species specificity in structure-based drug discovery: dihydrofolate reductase as a test system. *J Am Chem Soc*, 129(12):3634–3640, 2007.

[159] T. A. Fritz, D. Tondi, J. S. Finer-Moore, M. P. Costi, and R. M. Stroud. Predicting and harnessing protein flexibility in the design of species-specific inhibitors of thymidylate synthase. *Chem Biol*, 8(10):981–995, 2001.

[160] H. Heaslet, R. Rosenfeld, M. Giffin, Y. C. Lin, K. Tam, B. E. Torbett, J. H. Elder, D. E. McRee, and C. D. Stout. Conformational flexibility in the flap domains of ligand-free HIV protease. *Acta Crystallogr D Biol Crystallogr*, 63(Pt 8):866–875, 2007.

[161] V. Hornak and C. Simmerling. Targeting structural flexibility in HIV-1 protease inhibitor binding. *Drug Discov Today*, 12(3-4):132–138, 2007.

[162] I. Bahar, B. Erman, R. L. Jernigan, A. R. Atilgan, and D. G. Covell. Collective motions in HIV-1 reverse transcriptase: examination of flexibility and enzyme function. *J Mol Biol*, 285(3):1023–1037, 1999.

[163] R. A. Engh and D. Bossemeyer. Structural aspects of protein kinase control-role of conformational flexibility. *Pharmacol Ther*, 93(2-3):99–111, 2002.

[164] M. L. Teodoro and L. E. Kavraki. Conformational flexibility models for the receptor in structure based drug design. *Curr Pharm Des*, 9(20):1635–1648, 2003.

[165] S. J. Teague. Implications of protein flexibility for drug discovery. *Nat Rev Drug Discov*, 2(7):527–541, 2003.

[166] C. Beier and M. Zacharias. Tackling the challenges posed by target flexibility in drug design. *Expert Opinion on Drug Discovery*, 5(4):347–359, 2010.

[167] J. D. Durrant and J. A. McCammon. Computer-aided drug-discovery techniques that account for receptor flexibility. *Curr Opin Pharmacol*, 10(6):770–774, 2010.

[168] J. Desmet, I. A. Wilson, M. Joniau, M. D. Maeyer, and I. Lasters. Computation of the binding of fully flexible peptides to proteins with flexible side chains. *FASEB J*, 11(2):164–172, 1997.

[169] N. Guex and M. C. Peitsch. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, 18(15):2714–2723, 1997.

[170] L. Prade, R. A. Engh, A. Girod, V. Kinzel, R. Huber, and D. Bossemeyer. Staurosporine-induced conformational changes of camp-dependent protein kinase catalytic subunit explain inhibitory potential. *Structure*, 5(12):1627–1637, 1997.

[171] R. A. Engh, A. Girod, V. Kinzel, R. Huber, and D. Bossemeyer. Crystal structures of catalytic subunit of cAMP-dependent protein kinase in complex with isoquinolinesulfonyl protein kinase inhibitors H7, H8, and H89. structural implications for selectivity. *J Biol Chem*, 271(42):26157–26164, 1996.

[172] N. Narayana, T. C. Diller, K. Koide, M. E. Bunnage, K. C. Nicolaou, L. L. Brunton, N. H. Xuong, L. F. T. Eyck, and S. S. Taylor. Crystal structure of the potent natural product inhibitor balanol in complex with the catalytic subunit of cAMP-dependent protein kinase. *Biochemistry*, 38(8):2367–2376, 1999.

[173] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin. UCSF Chimera – a visualization system for exploratory research and analysis. *J Comput Chem*, 25(13):1605–1612, 2004.

[174] *MarvinSketch, ChemAxon*. http://www.chemaxon.com/download/marvin/.

[175] O. Sperandio, L. Mouawad, E. Pinto, B. O. Villoutreix, D. Perahia, and M. A. Miteva. How to choose relevant multiple receptor conformations for virtual screening: a test case of CDK2 and normal mode analysis. *Eur Biophys J*, 39(9):1365–1372, 2010.

[176] M. A. Miteva, F. Guyon, and P. Tufféry. Frog2: Efficient 3D conformation ensemble generator for small compounds. *Nucleic Acids Research*, 38(suppl 2):W622–W627, 2010.

[177] E. Jenwitheesuk and R. Samudrala. Improved prediction of HIV-1 protease-inhibitor binding energies by molecular dynamics simulations. *BMC Struct Biol*, 3:2, 2003.

[178] S.-Y. Huang and X. Zou. Efficient molecular docking of NMR structures: application to HIV-1 protease. *Protein Sci*, 16(1):43–51, 2007.

[179] D. J. Osguthorpe, W. Sherman, and A. T. Hagler. Exploring protein flexibility: incorporating structural ensembles from crystal structures and simulation into virtual screening protocols. *J Phys Chem B*, 116(23):6952–6959, 2012.

[180] J. Fanfrlík, A. K. Bronowska, J. Rezác, O. Prenosil, J. Konvalinka, and P. Hobza. A reliable docking/scoring scheme based on the semiempirical quantum mechanical PM6-DH2 method accurately covering dispersion and H-bonding: HIV-1 protease with 22 ligands. *J Phys Chem B*, 114(39):12666–12678, 2010.

[181] A. W. Schüttelkopf and D. M. F. van Aalten. PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallogr D Biol Crystallogr*, 60(Pt 8):1355–1363, 2004.

[182] T. Yamazaki, A. P. Hinck, Y. X. Wang, L. K. Nicholson, D. A. Torchia, P. Wingfield, S. J. Stahl, J. D. Kaufman, C. H. Chang, P. J. Domaille, and P. Y. Lam. Three-dimensional solution structure of the HIV-1 protease complexed with DMP323, a novel cyclic urea-type inhibitor, determined by nuclear magnetic resonance spectroscopy. *Protein Sci*, 5(3):495–506, 1996.

[183] J. Smith. http://structbio.vanderbilt.edu/jsmith/suppose/.

[184] J. Janin and S. J. Wodak. Protein modules and protein-protein interaction. *Adv Protein Chem*, 61:1–8, 2002.

[185] D. Reichmann, O. Rahat, M. Cohen, H. Neuvirth, and G. Schreiber. The molecular architecture of protein-protein binding sites. *Curr Opin Struct Biol*, 17(1):67–76, 2007.

[186] A. M. J. J. Bonvin. Flexible protein-protein docking. *Curr Opin Struct Biol*, 16(2):194–200, 2006.

[187] J. An, M. Totrov, and R. Abagyan. Comprehensive identification of "druggable" protein ligand binding sites. *Genome Inform*, 15(2):31–41, 2004.

[188] J. Janin and C. Chothia. The structure of protein-protein recognition sites. *J Biol Chem*, 265(27):16027–16030, 1990.

[189] S. Jones and J. M. Thornton. Principles of protein-protein interactions. *Proc Natl Acad Sci U S A*, 93(1):13–20, 1996.

[190] C. J. Tsai, S. L. Lin, H. J. Wolfson, and R. Nussinov. Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Sci*, 6(1):53–64, 1997.

[191] L. L. Conte, C. Chothia, and J. Janin. The atomic structure of protein-protein recognition sites. *J Mol Biol*, 285(5):2177–2198, 1999.

[192] F. Glaser, D. M. Steinberg, I. A. Vakser, and N. Ben-Tal. Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins*, 43(2):89–102, 2001.

[193] J. Janin and B. Séraphin. Genome-wide studies of protein-protein interaction. *Curr Opin Struct Biol*, 13(3):383–388, 2003.

[194] R. P. Bahadur, P. Chakrabarti, F. Rodier, and J. Janin. A dissection of specific and non-specific protein-protein interfaces. *J Mol Biol*, 336(4):943–955, 2004.

[195] R. P. Bahadur and M. Zacharias. The interface of protein-protein complexes: analysis of contacts and prediction of interactions. *Cell Mol Life Sci*, 65(7-8):1059–1072, 2008.

[196] J. Liang, H. Edelsbrunner, and C. Woodward. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci*, 7(9):1884–1897, 1998.

[197] R. A. Laskowski, N. M. Luscombe, M. B. Swindells, and J. M. Thornton. Protein clefts in molecular recognition and function. *Protein Sci*, 5(12):2438–2452, 1996.

[198] C. Mattos and D. Ringe. Locating and characterizing binding sites on proteins. *Nat Biotechnol*, 14(5):595–599, 1996.

[199] D. W. Miller and K. A. Dill. Ligand binding to proteins: the binding landscape model. *Protein Sci*, 6(10):2166–2179, 1997.

[200] S. J. Campbell, N. D. Gold, R. M. Jackson, and D. R. Westhead. Ligand binding: functional site location, similarity and docking. *Curr Opin Struct Biol*, 13(3):389–395, 2003.

[201] S. Vajda and F. Guarnieri. Characterization of protein-ligand interaction sites using experimental and computational methods. *Curr Opin Drug Discov Devel*, 9(3):354–362, 2006.

[202] F. Chiti, M. Stefani, N. Taddei, G. Ramponi, and C. M. Dobson. Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature*, 424(6950):805–808, 2003.

[203] S. Pechmann, E. D. Levy, G. G. Tartaglia, and M. Vendruscolo. Physicochemical principles that regulate the competition between functional and dysfunctional association of proteins. *Proc Natl Acad Sci U S A*, 106(25):10159–10164, 2009.

[204] A.-M. Fernandez-Escamilla, F. Rousseau, J. Schymkowitz, and L. Serrano. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol*, 22(10):1302–1306, 2004.

[205] J.-L. Chung, W. Wang, and P. E. Bourne. High-throughput identification of interacting protein-protein binding sites. *BMC Bioinformatics*, 8:223, 2007.

[206] R. P. Bahadur, P. Chakrabarti, F. Rodier, and J. Janin. Dissecting subunit interfaces in homodimeric proteins. *Proteins*, 53(3):708–719, 2003.

[207] J. Vondrásek, P. E. Mason, J. Heyda, K. D. Collins, and P. Jungwirth. The molecular origin of like-charge arginine-arginine pairing in water. *J Phys Chem B*, 113(27):9041–9045, 2009.

[208] T. Clackson and J. A. Wells. A hot spot of binding energy in a hormone-receptor interface. *Science*, 267(5196):383–386, 1995.

[209] K. S. Thorn and A. A. Bogan. ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*, 17(3):284–285, 2001.

[210] T. Kortemme and D. Baker. A simple physical model for binding energy hot spots in protein-protein complexes. *Proc Natl Acad Sci U S A*, 99(22):14116–14121, 2002.

[211] A. A. Bogan and K. S. Thorn. Anatomy of hot spots in protein interfaces. *J Mol Biol*, 280(1):1–9, 1998.

[212] I. S. Moreira, P. A. Fernandes, and M. J. Ramos. Hot spots – a review of the protein-protein interface determinant amino-acid residues. *Proteins*, 68(4):803–812, 2007.

[213] S. J. Darnell, D. Page, and J. C. Mitchell. An automated decision-tree approach to predicting protein interaction hot spots. *Proteins*, 68(4):813–823, 2007.

[214] J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, and L. Serrano. The FoldX web server: an online force field. *Nucleic Acids Res*, 33(Web Server issue):W382–W388, 2005.

[215] N. J. Burgoyne and R. M. Jackson. Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces. *Bioinformatics*, 22(11):1335–1342, 2006.

[216] A. T. R. Laurie and R. M. Jackson. Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screening. *Curr Protein Pept Sci*, 7(5):395–406, 2006.

[217] S. Leis, S. Schneider, and M. Zacharias. In silico prediction of binding sites on proteins. *Curr Med Chem*, 17(15):1550–1562, 2010.

[218] G. P. Brady and P. F. Stouten. Fast prediction and visualization of protein binding pockets with pass. *J Comput Aided Mol Des*, 14(4):383–401, 2000.

[219] J. Konc and D. Janezic. Protein-protein binding-sites prediction by protein surface structure conservation. *J Chem Inf Model*, 47(3):940–944, 2007.

[220] M. Hendlich, F. Rippmann, and G. Barnickel. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J Mol Graph Model*, 15(6):359–63, 389, 1997.

[221] F. Glaser, T. Pupko, I. Paz, R. E. Bell, D. Bechor-Shental, E. Martz, and N. Ben-Tal. Consurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, 19(1):163–164, 2003.

[222] J. Dundas, Z. Ouyang, J. Tseng, A. Binkowski, Y. Turpaz, and J. Liang. CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res*, 34(Web Server issue):W116–W118, 2006.

[223] A. T. R. Laurie and R. M. Jackson. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics*, 21(9):1908–1916, 2005.

[224] B. Huang and M. Schroeder. LIGSITEcsc: predicting ligand binding sites using the connolly surface and degree of conservation. *BMC Struct Biol*, 6:19, 2006.

[225] M. Nayal and B. Honig. On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins*, 63(4):892–906, 2006.

[226] M. Bryliński, K. Prymula, W. Jurkowski, M. Kochańczyk, E. Stawowczyk, L. Konieczny, and I. Roterman. Prediction of functional sites based on the fuzzy oil drop model. *PLoS Comput Biol*, 3(5):e94, 2007.

[227] N. Eswar, D. Eramian, B. Webb, M.-Y. Shen, and A. Sali. Protein structure modeling with MODELLER. *Methods Mol Biol*, 426:145–159, 2008.

[228] A. Sali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*, 234(3):779–815, 1993.

[229] J. C. Fuller, N. J. Burgoyne, and R. M. Jackson. Predicting druggable binding sites at the protein-protein interface. *Drug Discov Today*, 14(3-4):155–161, 2009.

[230] S. Eyrisch and V. Helms. Transient pockets on protein surfaces involved in protein-protein interaction. *J Med Chem*, 50(15):3457–3464, 2007.

[231] N. Moitessier, P. Englebienne, D. Lee, J. Lawandi, and C. R. Corbeil. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br J Pharmacol*, 153 Suppl 1:S7–26, 2008.

[232] D. Plewczynski, M. Łaźniewski, R. Augustyniak, and K. Ginalski. Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. *J Comput Chem*, 32(4):742–755, 2011.

[233] G. Neudert and G. Klebe. DSX: A knowledge-based scoring function for the assessment of protein–ligand complexes. *Journal of Chemical Information and Modeling*, 51(10):2731–2745, 2011.

[234] T. Tuccinardi, M. Botta, A. Giordano, and A. Martinelli. Protein kinases: docking and homology modeling reliability. *J Chem Inf Model*, 50(8):1432–1441, 2010.

[235] O. Trott and A. J. Olson. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*, 31(2):455–461, 2010.