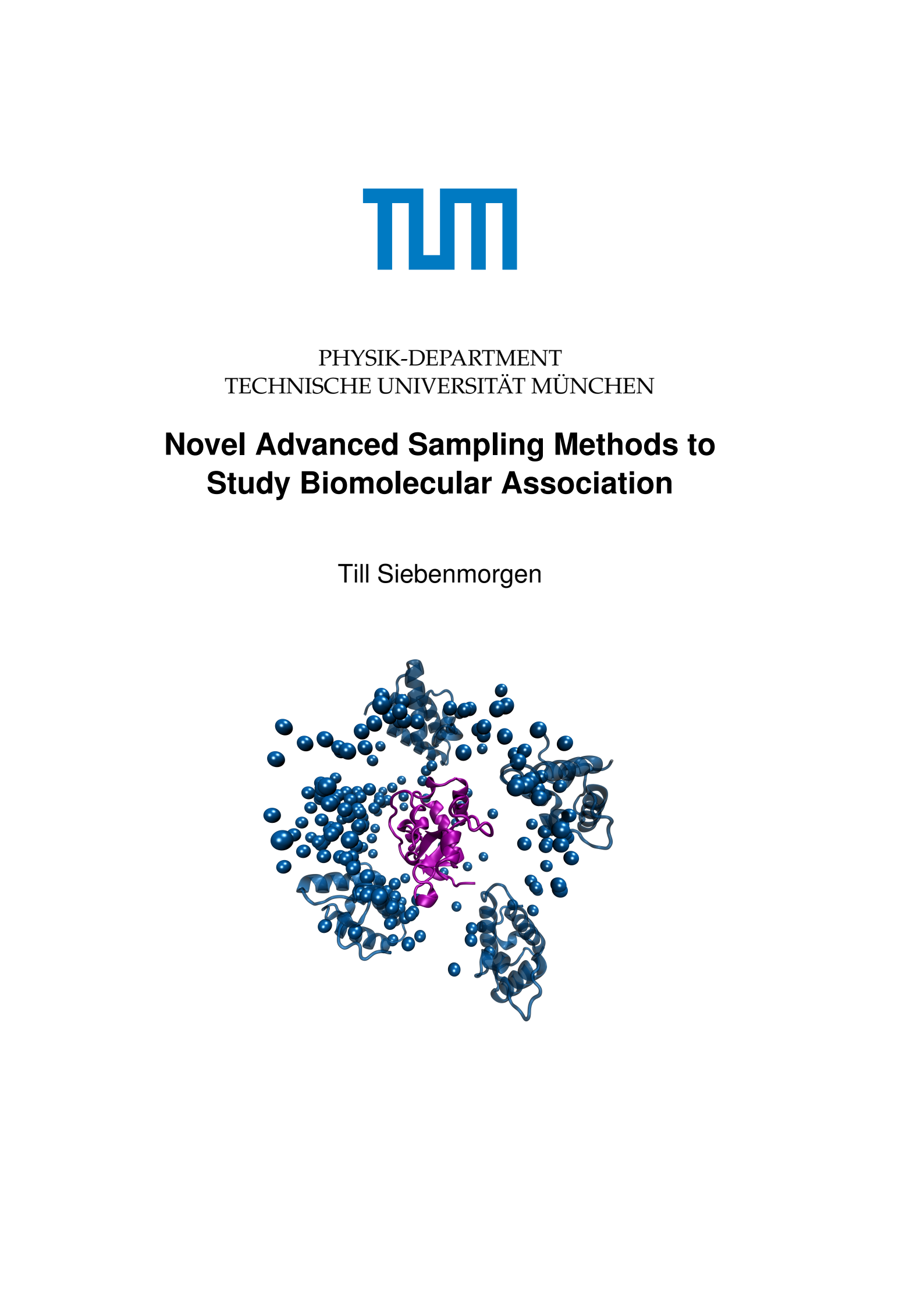


PHYSIK-DEPARTMENT
TECHNISCHE UNIVERSITÄT MÜNCHEN

Novel Advanced Sampling Methods to Study Biomolecular Association

Till Siebenmorgen

PHYSIK-DEPARTMENT
TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Theoretische Biophysik (T38)

Novel Advanced Sampling Methods to Study Biomolecular Association

Till Siebenmorgen

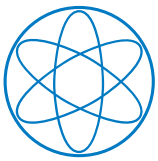
Vollständiger Abdruck der von der Fakultät für Physik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende(r):	Prof. Dr. Friedrich Simmel
Prüfer der Dissertation:	1. Prof. Dr. Martin Zacharias
	2. Prof. Dr. Karen Alim

Die Dissertation wurde am 22.02.2021 bei der Technischen Universität München eingereicht und durch die Fakultät für Physik am 13.04.2021 angenommen.



Für meine Eltern
Dagmar und Ralf

Ich lebe mein Leben in wachsenden Ringen,
die sich über die Dinge ziehn.
Ich werde den letzten vielleicht nicht vollbringen,
aber versuchen will ich ihn.

Rainer Maria Rilke

Abstract

Proteins constitute one of the central molecules of life that are involved in almost all cellular processes. These molecules perform most of their functions interacting through assemblies, like protein-protein complexes or protein complexes with small molecules. Due to huge advances in experimental structure determination (X-ray crystallography, NMR, and cryo-EM), large progress in the study of proteins and their partners could be achieved. In particular, molecular dynamics simulations and docking methods usually rely on such structural data to study proteins in silico. From a computational perspective, biomolecular complexes can be studied using docking approaches, in which one aims to predict the native binding site of the receptor and calculate the associated binding affinity. Compared to traditional docking techniques, molecular dynamics simulations are computationally quite costly, however, they provide higher theoretical rigor. Molecular dynamics simulations account for an atomistic representation of the solute, proper treatment of the aqueous environment, and full flexibility of the partner molecules. These simulations are often assisted by advanced sampling methods to reduce the computational demand. In this thesis, such advanced sampling methods were developed to study biomolecular complexes and tackle the main goals of computational docking.

First, an umbrella sampling approach was applied on a protein-protein benchmark set, calculating the absolute binding free energy. In particular, this approach performed better than the simple energy-based scoring schemes to predict the native binding site for a multitude of docking poses.

Identification of the correct binding site from regular MD simulations alone is a difficult task due to many local energy minima at the receptor surface. To address this issue, the repulsive scaling (RS-REMD) approach is introduced in this thesis. It is shown to speed up MD-based association simulations considerably to capture the native binding placement of protein complexes. The correct binding site was identified for five out of six protein-protein cases and for two proteins bound to small ligands.

The RS-REMD methodology is further extended to estimate the (absolute) binding free energy of protein complexes. A high correlation to experimental affinities is observed for 36 protein-protein cases and for two proteins bound to several small ligands. This binding affinity estimate is shown to be able to discriminate correct ligand binding placements from other poses. Due to their simple implementation, the developed repulsive scaling techniques can be applied to all kinds of molecular complexes, including DNA, RNA, polysaccharides, or small peptides. Such MD-based docking techniques are thought to become increasingly affordable with expected hardware enhancements.

Contents

1 Introduction

Proteins are the building blocks of the cell and mediate most cellular processes. They are constituted of differing amino acid sequences that define the protein's three-dimensional structure and thus code for a remarkably diverse range of functions. Proteins can act as transporters, molecular switches, or catalysts in enzymatic reactions. They aid in the transduction of signal pathways within the cell but also give the cell its structure and provide mechanical stability. Malfunctioning proteins are responsible for several diseases (e.g. Alzheimer's disease) but also carry out specific immune responses through antibodies.

The versatile function of proteins fundamentally relies on the binding to partner molecules. Proteins can bind to nucleic acids, lipids, small molecules, or other proteins and either form stable complexes or transient and reversible assemblies. Understanding the physics behind these binding processes allows the design of small molecular agents [228, 176, 183] in drug discovery projects or completely novel *de novo* proteins [130, 32, 18]. Most effectively these protein interactions are studied in a complementing collaboration of experimental and in silico approaches [281, 176, 183].

In the last decades, significant progress within the field of structural biology has led to a huge number of high-resolution three-dimensional protein models. The data are experimentally obtained using NMR, X-ray crystallography, and lately also cryo-EM. The structural models are collected in the Protein Data Bank (PDB) [27] which currently contains more than 150000 protein entries.

These structural models constitute the starting point of most molecular dynamics (MD) simulations. In MD the details of molecular binding can be revealed with high spatial and temporal resolution. These simulations allow a realistic description of biomolecules in a solvated environment based on an atomistic force field. Most importantly, the absolute free energy of binding, the central feature of molecular association, can be computed with MD.

Simultaneously, a multitude of experimental methods to calculate the binding affinity like surface plasmon resonance and isothermal titration calorimetry are currently available. Such experimental results are united in benchmark sets allowing computational scientists to validate their predictions and develop new approaches [153, 309, 150].

Apart from characterizing the strength of a bimolecular interaction the identification of putative binding sites is of central importance. In the last decades a

vast number of docking algorithms have emerged, e.g. for protein-protein docking [73, 44, 329] as well as protein docking to small ligands [299, 307, 93, 227]. These approaches attempt to predict the complex structures based on the unbound partner molecules by evaluating a very large number of possible placements of one molecule around its partner. Docking algorithms rely on certain scoring functions to rank the predicted placements that can for example be based on statistical- or knowledge-based potentials [329, 217, 260] or might be due to molecular mechanics type force field descriptions [73, 44]. These docking algorithms often have a low computational cost, but can fail to predict the correct binding site in several situations [150]. Common difficulties arise from limited molecular flexibility, coarse pseudo-atomic models, or no proper solvent representation.

In principle, these challenges can be addressed using MD simulations. With unlimited computer power, it would be possible to simulate the reversible association and dissociation of transiently bound partner molecules. Such simulations could potentially predict the native binding site and further allow for the calculation of the free energy of binding. Yet, on an average computing cluster, the required amount of calculations is not practicable in most applications, which increases the need for advanced sampling methods that significantly speed up simulation time. In this thesis such advanced sampling techniques for MD are developed that efficiently allow for the prediction of native protein binding sites, as well as the calculation of absolute binding free energies.

This thesis is organized as follows:

- An introduction to structural and functional aspects of proteins and experimental methods to study the conformation and binding mechanisms of these molecules are given in Chapter 2.

- In Chapter 3, molecular dynamics (MD) simulations are introduced, the computational technique used in this thesis to study biomolecules with an atomistic resolution. The theoretical foundations of MD simulations and advanced sampling methods (Hamiltonian replica exchange molecular dynamics (H-REMD) and umbrella sampling) are explained. Moreover, an introduction to the calculation of binding free energies (e.g. free energy perturbation, Bennett acceptance ratio) is given.

- In silico methods to predict binding affinities with a focus on protein-protein complexes are reviewed in detail next (Chapter 4). Simple scoring functions are discussed but also more advanced methods that rely on the foundations of Chapter 3.

- Chapter 5 applies an umbrella sampling approach to calculate the absolute binding free energy and to refine docking poses for a large protein-protein

benchmark set. The advanced sampling method is compared to simple energy-based scoring functions.

- In Chapter 6, the repulsive scaling (RS-REMD) scheme is designed that relies on an H-REMD approach, introducing repulsive biases between ligand and receptor atoms. The method is shown to predict the native binding site of protein-protein complexes in long association simulations but also in shorter refining simulations.
- After that, the RS-REMD method is extended to yield binding free energies to score native and predicted binding poses (Chapter 7). Different solvent conditions are compared for a protein-protein benchmark set.
- In Chapter 8, RS-REMD is employed to complexes of proteins bound to small ligands. Moreover, a procedure to calculate the native binding site in a blind docking scenario is stated.
- The thesis concludes with a summary and future perspectives of the established methods (Chapter 9).

2 Structure and Function of Proteins

Proteins are one of the fundamental biomolecules that life relies on. In this chapter, I will introduce these molecules in more detail and give an outline of their versatile function. Finally, an overview of experimental strategies is given to elucidate the structure of proteins and to predict the binding affinity to other molecules. Based on such experimental data, computational approaches can be designed and validated.

2.1 From amino acids to folded protein structures

Proteins are formed by amino acid building blocks with their backbone linearly connected by peptide bonds. With little exceptions, the primary structure of proteins that occur in living organisms is built from different arrangements of 20 amino acids. The amino acids have a uniform composition with an amino group, a carboxyl group, a hydrogen atom, and a variable side chain bound to the central C_α atom. The side chain defines the properties of each amino acid, that can for example be grouped into charged and uncharged or hydrophobic, hydrophilic, and amphipathic residues, the latter having both, a polar and a nonpolar character [245].

The primary amino acid arrangement defines the stability and shape of the 3D structure of the folded proteins. According to the thermodynamic hypothesis this native state is given by the free energy minimum of the protein in water [13]. Apart from disulfide bridges, originating from covalently bound cysteine side chains, the folded protein is stabilized to the main degree through a multitude of noncovalent interactions. Oppositely charged amino acids are attracted electrostatically and can form salt bridges within the protein or with a protein partner. Further, hydrophilic amino acids can form hydrogen bonds either with water molecules, the peptide backbone, or with each other. Networks of hydrogen bonded interactions of the backbone lead to secondary structure elements such as alpha helices and β-sheets. Finally, hydrophobic residues are likely to accumulate at the core of the protein (the "hydrophobic effect") so that they are not directly exposed to water molecules. Examined individually, these non-covalent interactions are associated with little increase in energy (e.g. one hydrogen bond in the range of a few kcal/mol) but the accumulation of these contributions can form stable folds.

2.2 How proteins perform their function

A protein's non-bonded interactions not only define its three-dimensional fold but can also lead to the association with other molecules. Intermolecular binding can be viewed as the most fundamental property of proteins, leading to a tremendous functional repertoire including catalysis, switching, and the building of structural elements [245]. The nature of these binding events can be stable in some cases but also short-lived (transient) in many situations.

Protein assemblies can form stable complexes in many designs, including large structural components of the cell or the extracellular matrix. For example, collagen is built from three protein chains that are covalently bound through cross-links and stabilized by interstrand hydrogen bonds in a coiled-coil formation [274]. Other proteins serve as scaffolds that can bind to multiple proteins, aiding in the formation of functional complexes (e.g. signaling complexes) by spatial organization [79].

The improper aggregation of proteins can cause diseases, such as in the case of amyloid fibrils, a protein assembly composed of large stacks of β-sheets that are believed to be connected to Alzheimer's disease. Similarly, pathological assemblies of misfolded prion proteins can lead to a set of diseases, including Creutzfeldt-Jakob in humans, BSE in cattle, and scrapie in sheep [6]. Proteins can assemble as subunits of large structures that form e.g. the spherically shaped capsids of many viruses. On the other hand, recognition of pathogenic viruses by the adaptive immune system can lead to a response of antibodies, which are immune proteins that selectively bind to a specific antigen (*anti*body *gen*erator), ultimately inactivating the virus or mark it as a target for destruction.

In a quite different role, specialized membrane transport proteins move water-soluble proteins and ions across cellular membranes to establish electrochemical gradients, receive nutrients and segregate metabolic waste products. Such proteins can also transfer extracellular molecules that lead to signal activation inside the cell. The signaling pathway can involve a series of protein interactions that ultimately regulate the behavior of the cell and e.g. lead to cell growth, division, or differentiation. Such signaling proteins often act as molecular switches, which are activated through phosphorylation or binding of GTP. With the aid of enzyme proteins, certain stimulatory signals can be amplified by triggering catalytic cascades. These enzymes bind specifically to their substrates, often assisted by cofactors, and can be activated or inhibited by regulatory molecules [6].

The function of proteins is extremely versatile and the study of intermolecular interactions often depends on delicate details. General networks of protein-protein interaction data can in principle be elucidated by high throughput studies that typically rely on yeast two-hybrid screening or affinity purification of complexes followed by mass spectrometry [323, 252, 35]. However, to gain a deeper understanding of protein binding and thus the mechanisms governing life, we must gain

more knowledge of the atomistic constitutions and the energetic contributions that influence intermolecular interactions. This is most effectively undertaken as an interplay of experimental and in silico methods that complement each other in many disciplines.

2.3 Experimental methods for structure prediction

The experimental 3D models of protein structures are archived in the Protein Data Bank (PDB) [27]. Up to date, the PDB contains more than 150000 protein structure files and the number of resolved proteins is growing fast since the early 1970s. The protein structures are mostly obtained via X-ray crystallography (78 %) or solution nuclear magnetic resonance (NMR) (7 %) and increasingly also using single-particle cryo-EM (electron microscopy). The method of choice depends e.g. on the molecular weight of the molecule, the solubility or whether a crystallization is possible.

2.3.1 X-ray crystallography

X-ray crystallography can generate protein structures of high accuracy, for which often a resolution around 2 Å is achieved. The resolution of the huge structure of the 50S ribosomal subunit (nearly 2 MDa) [19] or the 66 MDa bacteriophage PRD1 [1] exemplify that no principal limit in the size of the resolved structures exists that prevents accurate prediction.

In such an experiment a focused X-ray beam is scattered at a single crystal of the protein, leading to a diffraction pattern that is observed. Often the crystal is cooled to cryogenic temperatures to prevent radiation damages during the experiment. Using Fourier transform techniques the electron density of the molecules in the crystal is reconstructed from reciprocal diffraction space to direct space. For this reconstruction two Fourier coefficients are needed, the structure factor amplitude, which is accessible through the measured intensities of the diffraction spots, and the phase angle. The second term can not be accessed directly, but additional phasing experiments have to be conducted, which makes the structure determination quite demanding [257]. Two basic techniques exist to determine the phases, either a structural model of the protein is used as a source for the initial phases, or the protein is labeled by heavy atoms for which the position in the crystal is measured in an additional experiment [245]. Having reconstructed the electron density map, a refinement step that fits the observed data to a structure model is conducted.

2.3.2 Nuclear magnetic resonance

Using nuclear magnetic resonance (NMR) experiments the structures of proteins of small to medium size can be obtained in solution. NMR has a particular advantage to

study partially disordered proteins, proteins that stabilize in multiple conformations or dynamical aspects of certain parts of a protein. A prerequisite is that the proteins do not aggregate in solution.

In NMR spectroscopy the transitions between spin states of nuclei are observed in an external magnetic field. To this end, a second magnetic field is applied that oscillates in the radio-frequency (RF) regime and perturbs the nuclear magnetization. The induced current of the precessing nuclear spins is then measured by the NMR spectrometer. Only the signals of magnetically active atoms can be observed, usually ^{1}H, ^{13}C,^{15}N,^{31}P. In many cases, the sensitivity of the experiment is increased by labeling the molecule with certain isotopes (e.g.^{13}C,^{15}N) due to low natural abundance [43].

Several NMR experiments can be conducted to obtain geometrical constraints on the molecule. For example, the chemical shift is associated with a screening effect of the magnetic field due to the chemical environment of each nucleus. Thus, matching the chemical shifts with specific amino acid residues gives information about the 3D structure of the molecule. Atoms with high proximity can be revealed in 2D NMR observing the cross-peaks of e.g. correlation spectroscopy (COSY) or heteronuclear correlation spectroscopy (HETCOR) experiments [118].

Further, short interproton distances (up to 5 Å) can be observed using the nuclear Overhauser effect (NOE), which leads to constraints on $^{1}\mathrm{H} - {}^{1}\mathrm{H}$ distances for a refinement of the structure e.g. using molecular dynamics simulations. In particular, the 2D NMR (NOESY) experiment reveals residues that are close in space through the observed cross peaks [118, 21].

2.3.3 Cryo-electron microscopy

In the past decade, the number of resolved proteins using single-particle cryo-EM (electron microscopy) increased rapidly, from 320 entries in the PDB until 2010 to over 6000 structures that were resolved until 2020. This huge increment in applicability and resolution of cryo-EM in the last years is due to developments in direct electron detection techniques, the processing steps and in the automation of the setup to obtain high-quality data. Cryo-EM is achieving increasing high resolutions for many structures (under 4 Å for 51 % of the structures in the EMDB in 2019) and has a particular advantage over X-ray crystallography for structures that are difficult to crystallize like e.g. membrane proteins [46].

An electron beam is scattered at a sample of the structure. The scattered electrons can be recorded by film, scintillator based CCD, or nowadays most often with direct detection cameras. The sample is quickly frozen in liquid ethane. Due to the fast cooling rate the water does not crystalize but instead transforms into an amorphous state. This has two main advantages, first amorphous ice is transparent to electrons, second, the sample is retained in its solution environment thus representing a

snapshot of the state before freezing [77]. The ability to take a snapshot of a solution by fast freezing also allows for time-resolved imaging techniques. A protein can be incubated with its substrate and quickly be frozen. The resulting cryo sample will then contain a mixture of different reaction phases which allows studying time-critical processes like substrate processing [91].

The images of a sample represent 2D projections, thus to obtain the 3D volume, these have to be back projected onto this volume. In recent years the step from simple back projection based frameworks toward Bayesian statistics based models, combined with the introduction of direct electron detectors, has led to a significant improvement in the achievable resolution [261].

2.4 Experimental methods for binding affinity prediction

In order to calculate the binding affinity, we are interested in the reversible reaction of molecules of species A and B

$$[\mathrm{A}] + [\mathrm{B}] \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} [\mathrm{AB}], \tag{2.1}$$

with the association rate constant k_{on} and the dissociation rate constant k_{off} and $[\,\cdot\,]$ denoting concentrations. Thus [AB] is the concentration of the product, for example a protein-protein complex. If the reaction is at equilibrium we can define the dissociation constant K_d and the association constant K_a:

$$K_d = \frac{[\mathrm{A}][\mathrm{B}]}{[\mathrm{AB}]} = \frac{k_{off}}{k_{on}} = \frac{1}{K_a}. \tag{2.2}$$

In other words, K_d is the concentration of free A at which half of B is bound to A in equilibrium. The dissociation constant is directly related to the Gibbs free energy of dissociation ΔG_d and can be related to the change in enthalpy, ΔH_d, and entropy, ΔS_d :

$$\Delta G_d = -kT \ln\left(\frac{K_d}{c_0}\right) = \Delta H_d - T\Delta S_d, \tag{2.3}$$

with the Boltzmann constant k, the temperature T, and c_0 the standard state concentration. I will further refer to the binding free energy as the negative of ΔG_d, which one obtains from the association constant K_a, so that a binding event is preferable for negative values of the binding free energy.

In principle, it is possible to deduce the enthalpy and entropy values from equation 2.3 by conducting several experiments to calculate the the dissociation constant K_d

for different temperatures using the van't Hoff equation [173]:

$$\ln\left(\frac{K_d}{c_0}\right) = -\frac{\Delta H_d}{kT} + \frac{\Delta S_d}{k}. \tag{2.4}$$

If H_d and S_d are temperature independent we can obtain the values from a linear fit of the individual experiments. Still, this is not the case for example in reactions where the hydrophobic effect plays a major role, leading to temperature dependent enthalpy changes.

A plethora of experimental methods have been proposed to evaluate the binding affinities of biomolecules. For protein-protein complexes, the most prominent methods are comprised of isothermal titration calorimetry (ITC) and surface plasmon resonance (SPR) (according to the affinity benchmark 1.0 [153]) for binding affinities in the micromolar to nanomolar range. I will discuss these two methods in more detail below and three additional experimental setups, namely, circular dichroism, biolayer interferometry, and fluorescence polarization assays. Although there are still much more methods that can be discussed (see [215, 310]), these give an overview of different physical phenomena to measure the affinities under different experimental conditions, like e.g. immobilization or labeling of one partner molecule, and varying post processing steps that one has to be conscious about to correctly interpret the resulting values. The obtained affinities can vary significantly for the same molecule depending on the accuracy and the experimental method conducted and one should preferably choose an affinity benchmark set that uses the same experimental method that was realized in one laboratory [150, 153].

2.4.1 Isothermal titration calorimetry

Isothermal titration calorimetry gives the opportunity to directly measure thermodynamic parameters of a binding reaction like the change in entropy, enthalpy and the free energy of binding [305]. It is a label-free method that also gives access to the change in heat capacity upon complex formation. However, binding events of very high or very low affinity can be impracticable to study with ITC due to insufficient titration curves [310].

In the experiment, the calorimeter is composed of an adiabatic chamber with two cells that are held at constant temperature: the sample cell that is filled with sample solution and the reference cell. Through an injection syringe, ligand solution is added to the sample cell in multiple steps, and the change in heat is measured associated with the binding reaction between ligand and receptor in the sample cell. This procedure is repeated until the protein in the sample cell is saturated and no heat release upon further ligand addition is measured. From a nonlinear fit of the heat against molar ratio plot, the change in enthalpy and the association constant

can be calculated directly. The change in heat capacity can be obtained by repeating the experiment for different temperatures and gives insights to characteristics of the binding reaction like e.g. the dependence on hydration effects [105, 305].

2.4.2 Surface plasmon resonance

A label-free observation of protein binding to a substrate can be achieved in real-time using surface plasmon resonance (SPR) [231]. Using an optical detection technique the kinetic and equilibrium parameters of the binding events can be extracted.

The ligand is immobilized on a sensor chip surface and an analyte solution flow is passed over it, leading to binding events between ligand and analyte. An incident light beam is refracted at a certain refraction angle at the surface of the gold sensor chip. A change in the mass attached to the sensor chip, due to occurring binding events between analyte and ligand, leads to a change in its refractive index, which can be observed by optical means as the binding response[321].

Different stages in the measurement of the binding response over time are characteristic: First, an increase in binding response due to more and more binding events occurring is observed. Second, an equilibrium phase is reached with no change in the binding response, as binding and dissociation of the analyte are in equilibrium. Finally, the binding response deteriorates again after stopping the injection of analyte solution due to ongoing dissociation of the analyte. Modeling the measured binding response curves with the appropriate binding model determines the binding constants k_{on}, k_{off} and K_{eq}.

2.4.3 Optical biosensors: Biolayer interferometry

Biolayer interferometry, as an example for optical biosensors, allows label-free and real-time measurement of kinetic rates and equilibrium constants. In biolayer experiments the ligand is immobilized on a biosensor surface, surrounded by analyte molecules in solution. Incident white light is directed so that it reflects at the biocompatible surface and at a reference layer. The two reflected beams interfere with each other and the observed spectral pattern changes depending on the thickness of the biolayer surface and thus the number of molecules bound to the ligand. On the sensorgram, a progressively increasing number of bound complexes can be read out until a plateau is reached. Next, the dissociation phase is initialized by dipping the biosensor into a buffer, leading to a decreasing curve on the sensorgram. The kinetic rates and the dissociation constant can be obtained by fitting the binding model to the sensorgram curve [170].

2.4.4 Fluorescence polarization assays

Experiments that rely on fluorescence labeling are prominent, like for example fluorescence polarization assays, in which a bound complex can be distinguished in the fluorescence signal due to an increase in fluorescence polarization. To this end, one of the proteins is labeled with a fluorophore that can either be a small molecule or a variant of GFP. The partner molecule is progressively added in a titration experiment to the solution and the fraction of bound protein can be obtained due to an altered fluorescence signal. The formed complex rotates less rapidly and thus has a higher rotational correlation time (the time it takes the molecule to rotate 68.5°), which leads to a higher fluorescence polarization. The dissociation constant can be obtained from the binding isotherms, i.e. the plot of the fraction of bound protein against the concentration of the free (unlabeled) protein [251].

2.4.5 Circular dichroism

The difference in absorption between left and right circularly polarized light is called circular dichroism (CD) [81]. CD spectroscopy is sensitive to conformational changes of proteins in solution that can be observed during complexation or unfolding. Signals in the far ultraviolet region arise from amides of the protein backbone and give information on the secondary structure of the protein. Signals in the near-ultraviolet and visible regions originate from aromatic amino acid side chains, disulfide bonds, and extrinsic bands from prosthetic groups. The observed CD spectra are fitted to spectral databases to obtain the conformations. Also, the employment of neural networks that are first trained on a set of known proteins is possible [112].

Changes in the CD spectra are directly related to the number of protein complexes formed and can thus be used in titration experiments to determine dissociation constants. Moreover, differences in the thermal stability of unbound and bound proteins can be used in thermal denaturation experiments to calculate the binding constants, assuming that the change in free energy of folding is due to the binding free energy [111].

3 Theory

To rationalize and predict the behavior of biomolecules in silico, the basic concept of the simulation model has to be introduced first. Molecular dynamics (MD) simulations combine approaches from various disciplines in physics. Throughout this chapter, I will give an overview of the theoretical foundations and the general strategies applied in MD.

3.1 Simulation of molecular systems

The theory of quantum mechanics provides the most rigorous method to simulate biomolecules from first principles. Numerical solutions to the time-independent Schrödinger equation can be gained e.g. with Hartree Fock methods [269]. In this context, processes of interest are changes in the electron distribution of a molecule, like bond-breaking or bond-forming [185]. However, if one does not seek to gain insight into such chemical reactions there are less computationally demanding approaches that give access to the dynamics of biomolecules, such as molecular dynamics (MD) simulations.

In MD simulations, a classical mechanics approach is used to describe the atom-atom interactions by solving Newton's second law of motion

$$\boldsymbol{F}_i = m_i \frac{d^2 \boldsymbol{r}_i}{dt^2}. \tag{3.1}$$

The forces acting on each atom $\boldsymbol{F}_i$ are calculated from an empirical potential function termed "force field" (see Section 3.2.1). An atomistic model of the biomolecules is used due to two main assumptions: First, the nuclear motion is considered separate from the electron motion according to the Born-Oppenheimer approximation [222, 2]. Moreover, the nuclei are modeled as point particles surrounded by an average electron density [212]. Thus, the movement of classical atoms is captured by integrating Newton's equations of motion (see Section 3.2.2). In order to simulate these moving macromolecules under physiological conditions, they are solvated (Sections 3.3.3 and 3.3.4) and coupled to a heat bath (Section 3.3.2).

3.2 Force field and dynamics in simulations

3.2.1 Force field

In molecular dynamics simulations, the quantum mechanical interactions of molecules are approximated based on a classical force field. The parameters of these empirical force fields are gained using quantum mechanical calculations in conjunction with experimental measurements. Underlying approximations include, that the parameters calculated for a small set of molecular compounds are also valid for a much higher number of similar molecules (transferability) and that the functional form of the force field has a simple physical interpretation and can be expressed through the sum of several individual potential energy terms (additivity) [222]. The typical bonded and non-bonded force field contributions are expressed as follows [222]:

$$U = \sum_{N_b} \frac{1}{2} k_{b_i} (b_i - b_{i,0})^2 + \sum_{N_\Theta} \frac{1}{2} k_{\Theta_i} (\Theta_i - \Theta_{i,0})^2 + \sum_{N_t} \sum_{n} \frac{V_n}{2} (1 + cos(n\tau - \tau_0)) + \sum_{n_b} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \sum_{n_b} \frac{q_i q_j}{\epsilon r_{ij}} \quad (3.2)$$

The first two terms describe the bond length and angle contributions of covalently bound atoms according to harmonic potentials. The force constants are represented by k_{b_i} and k_{Θ_i}(e.g. originating from experimental vibrational frequencies), the reference position is described by $b_{i,0}$ and the reference angle by $\Theta_{i,0}$. The periodic cosine function characterizes the torsional potential with the torsion angle τ, the phase factor τ_0, the barrier height V_n and the multiplicity n [264].

The van der Waals (vdW) contributions are described by the Lennard-Jones (LJ) potential, with the well depth parameter ϵ_{ij} and the effective (pairwise) van der Waals radius σ_{ij}. The first term contributes to short-range repulsion, known as Pauli's exclusion principle. The second part accounts for attraction between the particles, according to London dispersion interactions, that arise from quantum mechanical dipole fluctuations. Finally, the electrostatic interaction between two atoms with distance d_{ij} is calculated based on Coulomb's law. The particles are assigned partial charges q_i and q_j.

All these parameters are specified for each atom type. In many force fields, multiple atom types are introduced for the same element to account for the chemical environment of the elements. For example, the standard AMBER force field consists of 13 different carbon atom types [60].

In general, deviations from this functional form are possible, as a wide range of different force fields (e.g. AMBER [60], CHARMM [196]) exist for differing purposes.

For example, polarisation effects are included in some force fields which are in theory more rigorous but have the drawback of higher computational costs [139]. Also, coarser force field models lacking an atomistic resolution are possible that replace certain groups of atoms with pseudo particles to lower the computational effort [329, 199]. In the Attract force field for protein-protein docking, each amino acid is represented by up to 4 pseudo atoms. The backbone is represented by 2 pseudo atoms, depending on their size side chains get assigned one or two atoms [329]. The Attract force field is an empirical model consisting of only non-bonded interactions: the sum of a Coulomb type term and a soft LJ type potential [87].

3.2.2 Integrating the equations of motion

In MD simulations, one is interested in calculating the system's dynamics by integrating Newton's equations of motion. The starting positions are often known from experimental data (see Section 2.3), their time evolution corresponds to a multi-body problem that has to be solved numerically with a finite difference method. Typical integration schemes are the Verlet algorithm [308] and the LeapFrog algorithm [78], both being symplectic and time-reversible.

In order to exemplify such an integration scheme, the derivation of the Verlet algorithm starts by expanding the position at times $t + \Delta t$ and $t - \Delta t$ by a Taylor series [300]:

$$\boldsymbol{r}_i(t + \Delta t) = \boldsymbol{r}_i(t) + \Delta t \boldsymbol{v}_i(t) + \frac{\Delta t^2}{2m_i}\boldsymbol{F}_i(t), \tag{3.3}$$

$$\boldsymbol{r}_i(t - \Delta t) = \boldsymbol{r}_i(t) - \Delta t \boldsymbol{v}_i(t) + \frac{\Delta t^2}{2m_i}\boldsymbol{F}_i(t). \tag{3.4}$$

Adding equations 3.3 and 3.4 yields the Verlet algorithm:

$$\boldsymbol{r}_i(t + \Delta t) = 2\boldsymbol{r}_i(t) - \boldsymbol{r}_i(t - \Delta t) + \frac{\Delta t^2}{m_i}\boldsymbol{F}_i(t). \tag{3.5}$$

This equation can be solved by calculating the force that acts on each particle $\boldsymbol{F}_i$ by taking the derivative of the potential-energy function U, known from the force field (Equation 3.2):

$$\boldsymbol{F}_i = -\frac{\partial U}{\partial \boldsymbol{r}_i}. \tag{3.6}$$

A long time step Δt is advantageous to cover the highest amount of phase space with the lowest number of iterations and thus the lowest computational effort. Unfortunately, an overly raised time step can lead to instabilities in the integration scheme because of high energy overlaps between atoms. Thus, the integration step is restrained into the regime below the fastest occurring motions of the system,

which is the bond stretching vibrations of small hydrogen atoms [2]. Constraining the bond lengths to hydrogen atoms by algorithms like SHAKE [258] a time step Δt of 2 fs is accessible. A further increase of the time step by a factor of 2 is possible through repartitioning the heavy atoms' masses into the bonded hydrogen atoms [126].

3.3 Simulating physiological conditions

The MD system until now describes a microcanonical ensemble as a consequence of the conservation of the total energy (E) with a constant number of particles (N) and volume (V). However, general biological experiments are performed under different conditions, for instance fixing the temperature rather than the energy. Thus, to simulate under the more realistic conditions of a canonical ensemble (NVT) - all MD simulations in this thesis are performed in this ensemble - thermodynamic properties that rely on statistical mechanics have to be introduced.

3.3.1 Basic statistical mechanics concepts

The canonical partition function of a system consisting of N identical particles that interact through a potential $U(\boldsymbol{r}_1, ..., \boldsymbol{r}_N)$ in a volume V and at temperature T can be expressed, integrating over the spatial domain $D(V)$ [300]:

$$Q(N,V,T) = \frac{1}{N!h^{3N}} \int d^N\boldsymbol{p} \int\limits_{D(V)} d^N\boldsymbol{r}\, exp\left\{-\beta\left[\sum_{i=1}^{N}\frac{\boldsymbol{p}_i^2}{2m} + U(\boldsymbol{r}_1, ..., \boldsymbol{r}_N)\right]\right\}, \quad (3.7)$$

with the factor $1/N!$ to avoid overcounting due to the identical nature of the N particles, Planck's constant h and $\beta = 1/kT$. The canonical partition function is a measure for the total number of accessible microstates. As the individual properties of the system are defined by the potential $U(\boldsymbol{r}_1, ..., \boldsymbol{r}_N)$ it is convenient to introduce the configurational partition function

$$Z(N,V,T) = \int\limits_{D(V)} d\boldsymbol{r}_1 ... d\boldsymbol{r}_N\, e^{-\beta U(\boldsymbol{r}_1,...,\boldsymbol{r}_N)}. \quad (3.8)$$

The ensemble average of a function $a(\boldsymbol{r}_1, ..., \boldsymbol{r}_N)$ is given by,

$$\langle a \rangle = \frac{1}{Z} \int\limits_{D(V)} d\boldsymbol{r}_1 ... d\boldsymbol{r}_N\, a(\boldsymbol{r}_1, ..., \boldsymbol{r}_N) e^{-\beta U(\boldsymbol{r}_1,...,\boldsymbol{r}_N)}. \quad (3.9)$$

The thermodynamic potential that corresponds to the canonical ensemble is the Helmholtz free energy G [265]

$$G = -kT\ln(Q). \tag{3.10}$$

The NVT ensemble is completely described by the partition function or the free energy. In MD simulations, the partition function is accessed by sampling trajectories in phase space over discrete time steps (see Section 3.2.2). In principle, with an infinite amount of time, the whole phase space would be sampled. The ergodic hypothesis states that the ensemble average can be related to the discrete time-average of a property a [300].

$$< a >= \lim_{T\to\infty} \frac{1}{T}\int_0^T dt\, a(x_t) = \frac{1}{M}\sum_{n=1}^{M} a(x_{n\Delta t}) = \overline{a}. \tag{3.11}$$

3.3.2 Statistical ensembles in MD

To generate a canonical ensemble, the system is coupled to a much bigger thermal reservoir that exchanges energy with the system, so that it is kept in thermal equilibrium. In MD simulations a plethora of thermostats have been introduced that lead to constant temperature in the system [11, 233, 25, 241]. For instance, the Berendsen thermostat rescales the velocities according to the following weak coupling scheme [25]:

$$v_i^{new} = v_i^{old} \cdot \sqrt{1 + \frac{\Delta t}{\tau}\left(\frac{T_{bath}}{T(t)} - 1\right)} \tag{3.12}$$

The old velocity is rescaled according to the fraction of target temperature T_{bath} and instantaneous temperature $T(t)$. The strength of the coupling can be modified by the relaxation parameter τ and Δt is the time step. This scaling method is relatively simple but has the disadvantage of not generating rigorous canonical averages which can lead to artifacts [178].

A more robust thermostatting method that generates a canonical system employs the Langevin equation of motion [241, 242]:

$$m\frac{d^2x(t)}{dt^2} = \boldsymbol{F}\{x(t)\} - \gamma\frac{dx(t)}{dt}m + \boldsymbol{R}(t). \tag{3.13}$$

The force on each particle is considered to have three sources: inter-particle interactions, frictional forces mediated by the collision frequency γ and a random-force vector $\boldsymbol{R}$. The random force is usually assumed to be a Gaussian distribution with

the following statistical properties:

$$\langle \boldsymbol{R}(t) \rangle = 0 \quad \text{and} \quad \langle \boldsymbol{R}(t)\boldsymbol{R}(t') \rangle = 2\gamma kTm\delta(t-t'), \tag{3.14}$$

relating γ and $\boldsymbol{R}$ through the fluctuation-dissipation theorem, with the Dirac symbol δ and the temperature T. Note that, in the limit of $\gamma \to 0$ of equation 3.13 we obtain Newton's equation as the coupling to the heat bath disappears. In the simulations, the Langevin equation has to be integrated numerically, for example using a generalized Verlet algorithm [36], which is valid in the low friction regime ($\gamma \Delta t \ll 1$).

Furthermore, it is also possible to simulate the system in a constant pressure environment, generating an isobaric-isothermal ensemble. For example, the Berendsen barostat controls pressure analogous to equation 3.12: the volume (instead of the velocity) of the system is updated by coupling the positions of the particles to a scaling factor [25].

3.3.3 Explicit solvent and non-bonded interactions

Biological systems are characterized as an integral part by the solvent surrounding the solute molecules. An aqueous environment can either be modeled explicitly or using a continuum solvent approach (see Section 3.3.4). Having an explicit water model hydrogen bonds can be modeled accurately, which potentially stabilize the molecule or mediate between different sites. In implicit solvent models, this feature of aqueous solutions is not considered.

Various approaches were introduced to explicitly describe water molecules that differ in their number of sites, the amount of flexibility of the atoms, and whether they account for polarization effect [143, 144, 26, 135, 38, 315]. A popular and simple choice is the rigid TIP3P water model, with one interaction point for each atom, which is used throughout this work.

To mimic the simulation of a large number of solvent molecules around the solute and to reduce artifacts due to hard boundaries, a simulation box (e.g. of cubic or truncated octahedron shape) with periodic boundary conditions is usually used [92]. The original box has infinitely many identical images in space. In practice, only the particles in the original box are propagated. As soon as a molecule leaves the box, it is replaced by an image reentering from the opposite side, conserving the total number of particles. Due to the minimum image convention, only the closest interaction for each atom pair is considered [2].

Electrostatic interactions decay slowly with r_{ij}^{-1} and a simple evaluation would introduce a considerable computational effort that scales quadratically with the number of particles [2]. Instead, these long-range contributions are typically evaluated using an Ewald sum that splits the total Coulomb energy into several terms.

First, a short-range term represents the original point charges that are neutralized by introducing Gaussian charge distributions of the same magnitude but opposite sign. A second long-range term is evaluated in Fourier space and exactly counteracts the first neutralizing contribution. Using a grid-based charge distribution the fast Fourier transform (FFT) can be applied to evaluate the reciprocal long-range term yielding a scaling of $\mathcal{O}(N \log N)$ in the particle mesh Ewald method [63, 297]. Usually, a cutoff is introduced limiting the space in which van der Waals interactions and real-space Coulomb interactions are calculated.

3.3.4 Implicit solvent models

A high fraction of the total computation time in most explicit solvent simulation setups is spent on the solvent molecules. To save computer time, but still incorporate a solvated environment for the investigated molecules as a dielectric continuum, implicit solvent models can be applied.

In principle, the solvation free energy ΔG_{Solv} is defined as the energy needed to transfer a molecule from vacuum into the solvent [235]. It is usually split into a nonpolar and a polar contribution:

$$\Delta G_{Solv} = \Delta G_{Nonpolar} + \Delta G_{Polar}. \tag{3.15}$$

The hydrophobic contribution ($\Delta G_{Nonpolar}$) consists of a favorable van der Waals interaction between solute and solvent and the unfavorable effect of breaking the molecular structure of the solvent molecules around the solute [235]. This term is approximated via the solvent-accessible surface area (SASA) for which a linear dependence on the surface tension parameter γ (obtained from experimental solvation free energies of alkanes) was found,

$$\Delta G_{Nonpolar} = \gamma \cdot SASA. \tag{3.16}$$

The SASA can be calculated by rolling a spherical probe over the surface of the molecule [275].

The polarization term (ΔG_{Polar}) is the most time-consuming part of the solvation free energy computation and accounts for the electrostatic interactions through the solvent. The electrostatic potential $\Phi(\boldsymbol{r})$ of a molecule with charge density $\phi(\boldsymbol{r})$ in an ionic dielectric continuum (position dependent dielectric constant $\epsilon(\boldsymbol{r})$) is given by the Poisson-Boltzmann (PB) equation, which can be written in a linearized form [178]:

$$\nabla \cdot [\epsilon(\boldsymbol{r}) \nabla \Phi(\boldsymbol{r})] - \kappa^2 \epsilon(\boldsymbol{r}) \Phi(\boldsymbol{r}) = -4\pi\phi(\boldsymbol{r}). \tag{3.17}$$

The Debye-Hückel parameter κ accounts for screening effects through the salt concentration. This differential equation can be solved numerically for example

using finite difference methods. The electrostatic contribution to the solvation free energy can now be calculated performing two independent calculations, the first with the exterior dielectric of the medium $\epsilon_{out,1}$ and the second in vacuum ($\epsilon_{out,2} = 1$) [178]:

$$\Delta G_{Polar} = \frac{1}{2}\sum_i q_i(\Phi^{\epsilon_{out,1}}(\boldsymbol{r}_i) - \Phi^{\epsilon_{out,2}}(\boldsymbol{r}_i)), \tag{3.18}$$

with the summation for all charges i in the solute. A high number of other applications are possible, like for example the calculation of the association free energy of two molecules using a thermodynamic cycle [107, 48].

A computationally efficient approximation of the PB equation is given by the generalized Born (GB) model [289, 4]

$$\Delta G_{Polar} = -\frac{1}{2}\sum_{ij}\frac{q_i q_j}{f_{GB}}\left(\frac{1}{\epsilon_{in}} - \frac{1}{\epsilon_{out}}\right) \tag{3.19}$$

$$f_{GB} = \sqrt{(r_{ij}^2 + R_i R_j e^{-D_{ij}})};\ D_{ij} = r_{ij}^2/4R_iR_j. \tag{3.20}$$

The summation goes over all atom pairs i,j, with the pairwise distance r_{ij}, the effective Born radius R_i and charge q_i of atom i. The dielectric constant of the molecule (ϵ_{in}) is usually low, while the solvent dielectric constant (ϵ_{out}) is high. The design of the function f_{GB} can be understood looking at two scenarios. For $i = j$ we have $f_{GB} = R$ and the Born equation is obtained, which describes the electrostatic component of the free energy of solvation for a single ion. Assuming a large separation of the two charges ($r_{ij} \gg R_i, R_j$) yields approximately the sum of the Coulomb interaction with the Born expression.

The amount of descreening of each atom i is measured through the effective Born radius R_i that is the degree of burial of the atom inside the solute. More precisely it is defined through the self-energy of the atom inside a molecule, which is the polar solvation free energy of a molecule, where all charges except the atom's charge are turned off. The effective radius is the radius of a corresponding spherical ion that has the equivalent polar solvation free energy as the self-energy of the atom [221]. The descreening effect of the surrounding molecule typically leads to a bigger effective radius than the intrinsic radius of the atom.

The effective Born radius can be computed using a Coulomb field integral [221, 146]:

$$R_i^{-1} = \rho_i^{-1} - \frac{1}{4\pi}\int_{|\boldsymbol{r}|>\rho_i}^{Solute} |\boldsymbol{r}|^{-4} d\boldsymbol{V}. \tag{3.21}$$

The integration is performed over the volume inside the solute and outside of atom i, with corresponding radius ρ_i (the origin of the integration is shifted to the center

of atom i). This approach introduces a bias in ΔG_{Solv}, it is only exact in case of a spherical solute, for a point charge located at its center [117, 221].

3.4 Free energy calculation with advanced sampling methods

We have treated the basic concepts that enable us to simulate biomolecules under physiological conditions that resemble experimental setups. Yet, these free MD simulations are in most cases too expensive to reach the timescales between microseconds and several seconds, in which most biophysical mechanisms occur, like protein folding or ligand-receptor binding [335]. In this section, some methods are presented that enhance the sampling of simulations. These methods make it possible to predict the free energy of processes that are out of the scope of free simulations.

3.4.1 Perturbation methods to access free energy differences

Let us consider the transformation of a thermodynamic state 1 to another state 2, with the potential energy functions $U_1(\boldsymbol{r}_1, ..., \boldsymbol{r}_N)$ and $U_2(\boldsymbol{r}_1, ..., \boldsymbol{r}_N)$. The Helmholtz free energy difference between the two states results from equation 3.10,

$$\Delta G = G_2 - G_1 = -kT \ \ln\left(\frac{Q_2}{Q_1}\right) = -kT \ \ln\left(\frac{Z_2}{Z_1}\right). \tag{3.22}$$

In the last step we could replace the canonical partition functions by the configurational partition functions as the momentum integrations cancel out of the ratio.

If one considers the change in the potential energy as a small perturbation

$$U_2(\boldsymbol{r}_1, ..., \boldsymbol{r}_N) = U_1(\boldsymbol{r}_1, ..., \boldsymbol{r}_N) + \Delta U(\boldsymbol{r}_1, ..., \boldsymbol{r}_N) \tag{3.23}$$

the division of both partition functions yields the ensemble average taken with respect to state 1.

$$\begin{aligned}
\frac{Z_2}{Z_1} &= \frac{\int d^N\boldsymbol{r}\, e^{-[U_1(\boldsymbol{r}_1,...,\boldsymbol{r}_N)+\Delta U(\boldsymbol{r}_1,...,\boldsymbol{r}_N)]/kT}}{\int d^N\boldsymbol{r}\, e^{-U_1(\boldsymbol{r}_1,...,\boldsymbol{r}_N)/kT}} \\
&= \frac{\int d^N\boldsymbol{r}\, e^{-U_1(\boldsymbol{r}_1,...,\boldsymbol{r}_N)/kT} \cdot e^{-\Delta U(\boldsymbol{r}_1,...,\boldsymbol{r}_N)/kT}}{\int d^N\boldsymbol{r}\, e^{-U_1(\boldsymbol{r}_1,...,\boldsymbol{r}_N)/kT}} \\
&= \left\langle e^{-\Delta U(\boldsymbol{r}_1,...,\boldsymbol{r}_N)/kT} \right\rangle_1 .
\end{aligned} \tag{3.24}$$

Substitution of equation 3.24 into equation 3.22 yields the equation of Zwanzig [338].

$$\Delta G = -kT \ln\left(\frac{Z_2}{Z_1}\right) = -kT \ln\left\langle e^{-[U_2(\boldsymbol{r}_1,\ldots,\boldsymbol{r}_N)-U_1(\boldsymbol{r}_1,\ldots,\boldsymbol{r}_N)]/kT}\right\rangle_1 . \tag{3.25}$$

It is crucial that the perturbation ΔU is small, which means that the configuration spaces of states 1 and 2 need to have enough overlap [300]. If states 1 and 2 differ significantly, it is possible to create a set of intermediate states in λ steps, for example using a linear interpolation:

$$U(\lambda) = (1-\lambda)\cdot U_1 + \lambda \cdot U_2 \tag{3.26}$$

The free energy difference is the sum of all individual contributions from equation 3.25

$$\Delta G = \sum_{i=1}^{N-1} \Delta G_{\lambda_i,\lambda_{i+1}} \tag{3.27}$$

$$= -kT \sum_{i=1}^{N-1} \ln\left\langle e^{-[U(\lambda_{i+1})-U(\lambda_i)]/kT}\right\rangle_i . \tag{3.28}$$

It is also possible to derive the free energy difference between state 1 and state 2 of equation 3.22 following the Bennett acceptance ratio (BAR) method [24]. We start, by exploring the following identity:

$$\frac{Z_2}{Z_1} = \frac{Z_2}{Z_1}\frac{\int d^N\boldsymbol{r}\, w(\boldsymbol{r}_1,\ldots,\boldsymbol{r}_N)\, e^{-[U_2(\boldsymbol{r}_1,\ldots,\boldsymbol{r}_N)+U_1(\boldsymbol{r}_1,\ldots,\boldsymbol{r}_N)]/kT}}{\int d^N\boldsymbol{r}\, w(\boldsymbol{r}_1,\ldots,\boldsymbol{r}_N)\, e^{-[U_2(\boldsymbol{r}_1,\ldots,\boldsymbol{r}_N)+U_1(\boldsymbol{r}_1,\ldots,\boldsymbol{r}_N)]/kT}}$$

$$= \frac{\left\langle w\, e^{-U_2/kT}\right\rangle_1}{\left\langle w\, e^{-U_1/kT}\right\rangle_2}. \tag{3.29}$$

The weighting function $w(\boldsymbol{r}_1,\ldots,\boldsymbol{r}_N)$ is arbitrary and can be expressed so that the statistical estimate of the free energy has the highest accuracy. Using Lagrange multipliers the Fermi Dirac function $f(x) = 1/(1+\exp(x/kT))$ results as the optimal choice and we obtain the Bennett acceptance ratio,

$$\frac{Z_2}{Z_1} = \frac{\langle f(U_2-U_1+C)\rangle_1}{\langle f(U_1-U_2-C)\rangle_2}\, e^{C/kT}. \tag{3.30}$$

This equation is valid for any value of the constant C. In practice, C is determined so that the following condition is fulfilled, when summing over all configurations

m_1 (m_2) of state 1 (2):

$$\sum_{m_1} f(U_2 - U_1 + C) = \sum_{m_2} f(U_1 - U_2 - C). \tag{3.31}$$

The BAR method, in comparison to equation 3.25, makes use of the sampled data in state 1 and state 2, which can lead to statistically more precise results [272]. It is further possible to extend the BAR approach to the multistate case (MBAR), which is the estimator with the lowest variance in case of multiple states [271]. Instead of considering only adjacent states of a λ coordinate as in equation 3.27, an estimator for the free energy differences of all states is produced:

$$\Delta G_{i,j} = -kT \ln \frac{Z_j}{Z_i} \tag{3.32}$$

The approach seeks weighting functions $\alpha_{i,j}$ according to the identity that was already used in the BAR approach (equation 3.29):

$$Z_i \left\langle \alpha_{i,j} \ \exp(-U_j/kT) \right\rangle_i = Z_j \left\langle \alpha_{i,j} \ \exp(-U_i/kT) \right\rangle_j \tag{3.33}$$

We now sum over the index j and yield K estimating equations, which are known as extended bridge sampling estimators:

$$\sum_{j=1}^{K} \frac{\hat{Z}_i}{N_i} \sum_{n=1}^{N_i} \alpha_{i,j} \ \exp(-U_j(\boldsymbol{r}_{in})/kT) = \sum_{j=1}^{K} \frac{\hat{Z}_j}{N_j} \sum_{n=1}^{N_j} \alpha_{i,j} \ \exp(-U_i(\boldsymbol{r}_{jn})/kT), \tag{3.34}$$

where we substituted $N_i^{-1} \sum_{n=1}^{N_i} g(\boldsymbol{r}_{i,n})$ for the expectation values of $\langle g \rangle_i$. For this class of estimating equations the choice of $\alpha_{i,j}$ with the lowest variance is known from the statistics literature [294]. A self-consistent solution for the free energy estimates $\hat{G}_i$ can be computed iteratively, obtained by combining the choice of $\alpha_{i,j}$ and equation 3.34:

$$\hat{G}_i = -kT \ln \sum_{j=1}^{K} \sum_{n=1}^{N_j} \frac{\exp\left[-U_i(\boldsymbol{r}_{jn})/kT\right]}{\sum_{k=1}^{K} N_k \exp\left[(\hat{G}_k - U_k(\boldsymbol{r}_{jn}))/kT\right]}. \tag{3.35}$$

Choosing $K = 2$ states the BAR equation can be reproduced after some rearrangements [271].

3.4.2 Umbrella sampling

To study biomolecular systems it is often desirable to shrink the available phase space to a subspace of particular interest. The design of adequate reaction coor-

dinates is strongly dependent on the process that one seeks to understand, for example, the dissociation of two proteins could be parametrized by a center of mass (COM) distance coordinate between ligand and receptor proteins (see Chapter 5). The corresponding free energy profile is associated with the probability that the generalized coordinate $q = f(\boldsymbol{r}_1, ..., \boldsymbol{r}_N)$ has the predefined value s [300]:

$$G(s) = -kT \ln P(s) \tag{3.36}$$

$$= -kT \ln \frac{C_N}{Q(N,V,T)} \int d^N\boldsymbol{p}\, d^N\boldsymbol{r}\, e^{-\beta\mathcal{H}(\boldsymbol{r},\boldsymbol{p})}\, \delta(f(\boldsymbol{r}_1, ..., \boldsymbol{r}_N) - s), \tag{3.37}$$

with the constant $C_N = 1/N!h^{3N}$, the total Hamiltonian $\mathcal{H}(\boldsymbol{r},\boldsymbol{p})$ and the Dirac function that ensures to take only states with appropriate s into account while integrating over the phase space. Often, the free energy profile along a reaction coordinate is called the potential of mean force (PMF). In order to enhance the sampling along a defined path, umbrella potentials w can be applied, that restrain the reaction coordinate $q = f(\boldsymbol{r}_1, ..., \boldsymbol{r}_N)$ in several intermediate steps $k = 1, ..., n$ using harmonic biasing potentials (force constant κ):

$$w(f(\boldsymbol{r}_1, ..., \boldsymbol{r}_N), s^{(k)}) = \frac{\kappa}{2}(f(\boldsymbol{r}_1, ..., \boldsymbol{r}_N) - s^{(k)})^2. \tag{3.38}$$

This umbrella potential is added to the potential U($\boldsymbol{r}$) so that a biased probability distribution on the predefined path is sampled. Here, substantial overlap between adjacent umbrella windows is important, to ensure adequate sampling of the whole coordinate range of interest. In order to obtain the corresponding PMF, the histograms of the simulated states have to be combined and the underlying unbiased probability distribution can be reconstructed e.g. by using the weighted histogram analysis method (WHAM) [116]. The WHAM equations have to be solved iteratively until they reach self consistency [300]:

$$P(q) = \frac{\sum_{k=1}^{n} n_k P_k(q)}{\sum_{k=1}^{n} n_k e^{\beta(G_K - G_0)} e^{-\beta w(q,s^{(k)})}}, \tag{3.39}$$

$$e^{-\beta(G_K - G_0)} = \int dq\, P(q) e^{-\beta w(q,s^{(k)})}, \tag{3.40}$$

with the full unbiased probability distribution $P(q)$. Interestingly, the WHAM equations can be understood as an approximation to MBAR using histogram kernel density estimators and an equivalence of both methods was shown for histograms of zero bin widths [271].

3.4.3 Hamiltonian replica exchange molecular dynamics

Many biological systems are described by rough potential energy landscapes with multiple local minima that are separated by high barriers. Hamiltonian replica exchange molecular dynamics (H-REMD) combines MD simulations with a Monte Carlo method, to further improve the sampling in phase space [95]. It enables the system if it is for example trapped by a potential well, to surpass it by simply switching coordinates instead of drifting over the barriers, which would take a lot of simulation time.

In this scheme a set of n regular simulations (replicas), that differ in their Hamiltonian, is started in parallel. After a certain amount of simulation steps, the exchange probability between neighboring replicas, i and j, according to a metropolis criterion, is evaluated [214]

$$P_{acc} = \min\left(1, \frac{e^{-\beta[U_i(\boldsymbol{r}_j)+U_j(\boldsymbol{r}_i)]}}{e^{-\beta[U_i(\boldsymbol{r}_i)+U_j(\boldsymbol{r}_j)]}}\right), \tag{3.41}$$

which yields correct probability distributions in the sampled thermodynamic ensemble. Basically, the acceptance criterion is one if the Boltzmann weighted sum of the replicas potential energies before the attempt (denominator), is smaller than after the attempt (numerator), resulting in a guaranteed exchange. Otherwise, if the exchange is energetically unfavourable, the coordinates of the systems are exchanged depending on the calculated probability.

Such an H-REMD technique can be combined with free energy perturbation methods (Section 3.4.1) to improve the sampling of intermediate states to calculate free energy differences for alchemical transformations [280, 239, 192]. In combination with umbrella sampling approaches (see Section 3.4.2), H-REMD is used on a regular basis to improve the convergence in the simulations [331, 194, 157].

The repulsive scaling (RS-) REMD technique can be used to study bimolecular complexes. It is a H-REMD based method that increases the effective pairwise van der Waals radii along the replica ladder. Thus, a physical dissociation of the ligand is introduced while the interactions within the molecules and with the solvent are not altered. The resulting phase space sampling of the higher replicas is improved by avoiding local energy minima on the receptor surface, which can efficiently drive the system into the global energy minimum (see Chapter 6) [276]. In addition to that, a free energy difference along the physical dissociation path can be computed using perturbation approaches like FEP, BAR or MBAR (see Chapter 7) [278].

4 Computational Prediction of Binding Affinities

In order to evaluate the functional relevance of putative protein complexes (see Chapter 2), realistic binding affinity prediction is of increasing importance. Several computational tools ranging from simple force field or knowledge-based scoring of single complexes to ensemble-based approaches and rigorous binding free energy simulations are available to predict relative and absolute binding affinities. In the present chapter, an overview of such in silico methods will be given, with most of these techniques relying on the theoretical foundations of Chapter 3. I will focus on protein-protein complexes, although the most important approaches are also applicable to other ligands. [1]

4.1 Introduction

The interaction of proteins is of fundamental importance for basically all processes in living systems. Most functions in a cell are mediated by the assembly of proteins to form transient dimers or oligomers to act as enzymes, transporters, or to stabilize the shape of the cell [232, 200]. Numerous interactions between proteins in a cell are in principle possible but only a fraction of putative complexes and assemblies are indeed formed and of functional relevance [141]. The associated binding free energy of protein-protein interactions determine the stability of association and the conditions for complex formation. Hence, a full understanding of cellular processes requires not only knowledge of all possible protein-protein interactions but also a quantitative insight into the structure and stability of the formed complexes [200, 224]. This also includes the effect of mutations in proteins that can modulate or even disrupt the binding to partner proteins. Protein-protein interactions are often mediated by reoccurring subdomains [273, 133]. Within different protein families, these domains can differ in sequence resulting in different binding specificities and different combinations of possible interactions. Predicting and understanding the significance of putative domain-domain interactions requires a quantitative understanding of protein-protein binding affinity.

[1] The contents of this chapter have been published in a similar form in [277].

In recent years, the possibility to design new synthetic protein-protein complexes with the desired function has gained significant momentum [166, 159]. One goal is to modify existing natural proteins in such a way that the geometry and affinity of a known protein-protein interaction may change or the interaction with a different protein surface on another protein partner becomes possible. In the longer run, it is desired to create completely new protein partners with programmed surface properties to allow for new interactions with selected candidate partners [159, 130]. Such efforts may form the basis for synthetic complexes that can act as designed molecular machines with a desired new function. A prerequisite for the successful rational design of such interactions and new complexes is the detailed understanding of protein-protein interaction and affinity.

The driving force for a protein binding process corresponds to the associated change in free energy that can be related to the structural and physicochemical properties of the protein binding partners. Initially proposed by Fischer [167] the "lock and key" concept of binding emphasizes the importance of optimal complementarity of binding partners at the interface as the decisive element for high binding affinity and specificity. Nevertheless, proteins and other biological macromolecules (e.g., RNA and DNA) can undergo various types of conformational fluctuations at physiological temperatures and, hence, are not rigid objects (illustrated in Figure 4.1). Often significant conformational changes of the binding partners upon association have been observed leading to the induced-fit binding concept of partner proteins [167, 61]. During binding, proteins appear to induce conformational changes in the partners that are a prerequisite for specific complex formation. In principle, all protein binding processes require some conformational adaptation but these changes can in certain cases be less than 1-2 Å for the root mean square deviation (RMSD) between bound and unbound conformations [181]. Besides of the induced-fit concept, the idea of a pre-existing ensemble of several interconvertible conformational states of proteins at equilibrium has been postulated [61, 319]. Within this ensemble, there are structures close to the bound and unbound forms and the process of binding to partner molecules shifts the ensemble toward the bound form. The mechanism of conformational selection versus induced-fit has been systematically studied for many known protein-protein interactions [288]. However, every conformation is, in principle, accessible even in the unbound state albeit with a potentially very low statistical weight. Hence, the original induced-fit concept is a special case of conformational ensemble selection where only the presence of a ligand gives rise to an appreciable concentration of the bound partner structure. For an accurate calculation of protein-protein binding free energies, it is desirable to account for conformational changes and also for changes in the conformational freedom (conformational entropy) upon binding [12, 326].

A great variety of experimental techniques is available to determine the structure of protein-protein complexes and assemblies (see Section 2.3). Protein X-ray

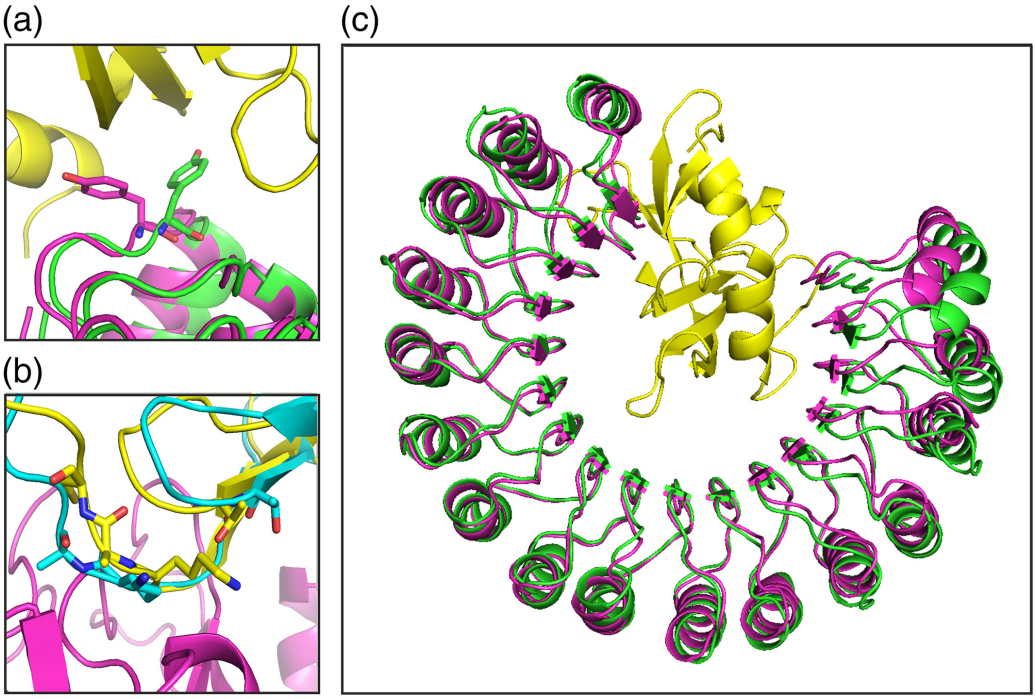


Figure 4.1: protein-protein association can induce different types of conformational changes. Conformational changes can involve side chain flips (a) indicated for a tyrosine side chain flip in the complex of RNAseA (yellow) inhibitor (green: unbound structure, pink: bound structure) complex, pdb1DFJ). For a protease-inhibitor complex (pdb1GL1), a loop refolding is observed (b, blue: unbound inhibitor; yellow: bound inhibitor). Besides local changes also global adaptations can accompany binding (c) demonstrated for the bound (pink cartoon) and unbound (green cartoon) of an RNAseA inhibitor (yellow cartoon: RNAseA).

crystallography is still the most common high-resolution technique to determine the structure of protein complexes. However, it requires the formation of sufficiently well-ordered crystals that can be difficult or impossible to obtain especially for low-affinity transient complexes and for large assemblies containing multiple partners. For the latter cases, recent improvements of cryo-EM (electron microscopy under cryogenic conditions) resulted in solving the structure of many new large and transient biomolecular assemblies achieving often an atomic or near-atomic resolution [17, 80]. The cryo-EM method is limited to complexes above a certain size limit ($\sim$ 100 kD) to achieve sufficient contrast of the image relative to a noisy background but no crystals of the complex are necessary. It requires only a sufficiently large number of object images from different viewpoints that can be combined to

solve the three-dimensional (3D) structure [17]. Nuclear-magnetic resonance (NMR) spectroscopy in solution can also be used for structure determination of mostly small dimeric protein-protein complexes [121, 57]. Several NMR techniques can be used to quantify protein-protein affinity and also help in modeling complexes if the structure of partner proteins is known [337].

A significant fraction of protein interactions are mediated by protein domains and for many domain-domain pairs, 3D structures are available and also data on the range of domain-domain affinities. In such cases, the prediction of domain-domain interaction affinities is often sufficient to estimate the affinity of whole complexes. Databases of protein domain-domain interactions such as 3DiD [225], SCOPPI [322], or PIBASE [65] are available that allow identification of interfaces and are helpful to predict the range of binding affinities for domain-mediated protein complexes.

Despite the great progress in recent years, the experimental determination of protein-protein complexes remains a challenging task and it will be impossible to determine experimentally all putative and transient protein-protein complexes of a cell in the foreseeable future [133, 285]. An additional difficulty arises because many protein-protein interactions involve conformational changes or even the coupled folding or refolding of disordered segments. In particular, transient protein-protein interactions in cells frequently include disordered protein segments. Such cases complicate both the structure prediction of complexes but also the prediction of binding affinities.

Computational prediction of protein-protein binding affinity typically requires the three-dimensional (3D) structure of the complex or at least a model of the complex structure. As the first part of the present review, we will first briefly outline the computational methods to generate structural models of protein-protein complexes (see also Figure 4.2).

A second prerequisite for evaluating methods to calculate and predict binding free energies are accurate experimental binding affinity data. Several different methods can be used to measure experimental binding equilibria (see Section 2.4) and one can distinguish between methods that require the separation of the bound complex from the isolated partners such as gel filtration, electrophoretic separation, or ultra-centrifugation approaches or methods that directly measure the concentration of bound and unbound proteins [94]. Methods based on the separation of complex and unbound partners require long lifetimes of the complex (small dissociation rate constants) beyond the time scale of the experiment. In direct methods, the complex formation can be detected by changes in heat capacity of the solution upon varying partner concentration (e.g., in isothermal titration calorimetry [175]) or optically using for example an associated change in absorbance or fluorescence [94, 22]. In the surface plasmon resonance technique, one partner is immobilized on a sensor surface. Addition of a partner protein that binds to the sensor surface results in a change of the refractive index that can be optically detected [320]. With the method,

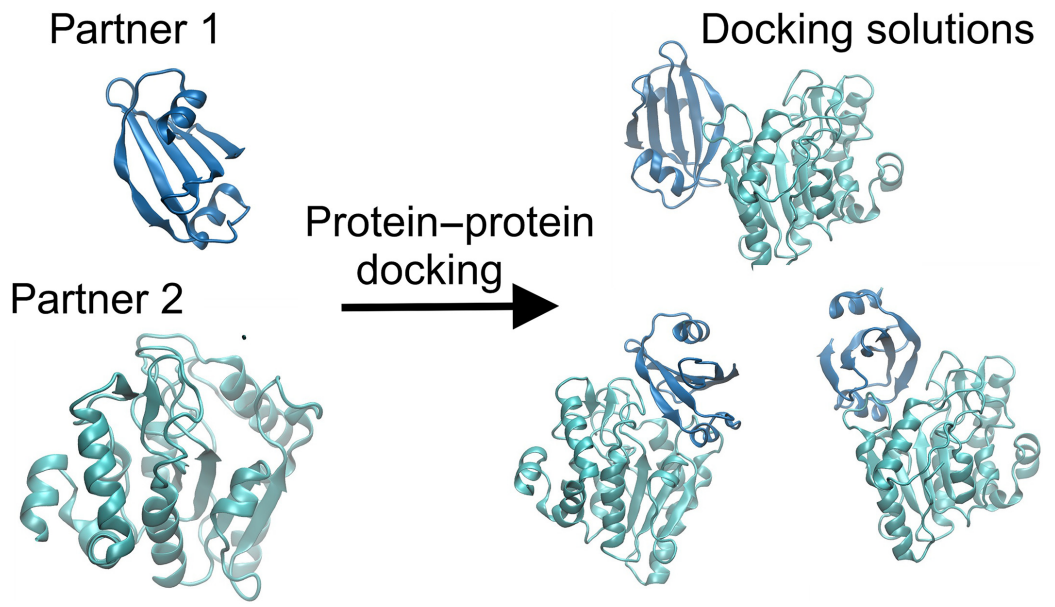


Figure 4.2: Computational protein-protein docking starting from separate unbound partners typically results in several putative sterically possible complex geometries. Task of a scoring step is to identify the most realistic geometry and to estimate the relative binding affinity of putative docked geometries.

it is also possible to analyze the time dependence of binding and to determine kinetic constants of protein-protein association and dissociation. For a critical comparison of methods to predict protein-protein binding affinities, the accuracy and consistency of the experimental reference data are of critical importance. It has been found that experimental protein-protein binding affinities can depend significantly on the experimental method [150, 153]. In addition, binding data for protein complexes are obtained under different experimental conditions (variation in ionic strength, pH, and temperature) whereas computational approaches typically assume the same conditions for all evaluated complexes. However, curated protein-protein binding affinity benchmark sets are available for which also structures of the complexes and the unbound protein partners have been determined [150, 153, 115]. These sets can serve as references for evaluating computational efforts; however, one should always be aware of experimental conditions and the method for binding affinity determination compared to the computational setup.

In the present review on calculating and predicting protein-protein binding affinity, we will first give a brief overview on modeling and predicting the structure of protein-protein complexes. These methods can also be used to predict which proteins in a cellular system may interact [318]. In the second and third sections,

we will focus on rapid approaches to rank predicted complexes and to estimate binding affinities based on single complex structures or ensembles of structures with a focus on force field-based methods. In the last sections, we introduce the application of ensemble-based methods and rigorous free energy simulations either for calculating relative binding free energies, for example, for predicting the effect of mutations on binding affinity or to calculate the absolute binding free energy of protein-complexes. As will be discussed below, such methods can also include the contribution of conformational changes to binding. The last sections will also include the application of different advanced sampling approaches and a discussion of the possibility to use such approaches for systematic applications of evaluating docked protein-protein complex geometries and possible future developments.

4.2 Predicting the structure of protein-protein complexes

The prediction of the binding affinity of putative protein-protein interactions is related to the computational modeling of the structure of protein-protein complexes and the ranking of possible predicted solutions [152, 150]. Besides experimentally determined complexes, one can distinguish two main computational techniques to provide structures of putative protein-protein complexes. First, computational protein-protein docking methods can be considered as "ab initio" approaches for generating complex geometries although in practice experimental (or other bioinformatics) data can be included to restrict the search for possible solutions [12, 326, 285, 29]. In addition, the de novo design of protein-protein interactions requires docking or modeling of new protein-protein interfaces [166]. Second, for the majority of natural stable protein-protein interactions one often finds homologous complex geometries in the data-base of experimentally determined complexes [172]. Hence, many interactions especially those mediated by reoccurring protein domains can be modeled based on similarity to an already known interaction [115]. In the following, we briefly introduce existing protein-protein docking approaches and also discuss template-based protein-protein complex prediction.

4.2.1 Protein-protein docking

The aim of computational protein-protein docking is the prediction of the structure of protein-protein complexes based on the structure of the isolated protein partners (Figure 4.2). The great majority of protein-protein docking algorithms use optimal complementarity as the main target criteria for predicting interactions [303]. A variety of computational methods exist to efficiently generate a large number of putative binding geometries typically with an initial systematic docking search keeping partner structures rigid. Subsequently, one or more refinement and scoring

steps of a set of preselected rigid docking solutions are added to achieve a closer agreement with the native geometry and to recognize near-native docking solutions. Among the most common are geometric hashing methods to rapidly match geometric surface descriptors of proteins [206, 134] and fast Fourier transform (FFT) correlation techniques [326, 303, 155, 40] to efficiently locate overlaps between complementary protein surfaces. Alternatively, molecular dynamics (MD), Brownian dynamics, Monte Carlo, or multi start energy minimization can be used to generate locally optimized protein-protein docking complexes [326]. To enhance the speed, often coarse-grained (CG) instead of atomistic protein models are used. These methods have in principle the capacity to introduce conformational flexibility of binding partners already at the initial search step but are, nevertheless, slower than FFT-based correlation methods or geometric hashing. Some approaches included conformational changes during docking already during the initial search including soft collective normal mode (NM) directions as additional variables during docking by energy minimization [327, 329] or in using swarm optimization [219]. If experimental data or data based on bioinformatics analysis are available, this information can be included either directly during the docking search, for example, in the HADDOCK, [73] ATTRACT [329], or RosettaDock [44, 203] approaches, or can be used to screen and rerank the docking solutions obtained from rapid FFT-based methods.37 The final stage of a protein-protein docking protocol consists of a structural refinement at atomic resolution of the binding partners and often rescoring of the various docking solutions. Often the final scoring step employs potentials that are intended to also provide an estimate of the binding affinity of the predicted complex (discussed below). protein-protein docking performance is regularly evaluated in the community-wide blind docking prediction challenge Critical Assessment of PRedicted Interactions (CAPRI). [181, 180, 182]

4.2.2 Prediction of protein-protein complexes based on homology to known structures

Due to the growing number of solved protein-protein complex structures, it is often possible to generate a structural model of a complex by using a known complex structure as a template [171]. In fact, it has been found that the majority of stable natural protein-protein interactions can in principle be modeled by a template-based approach [172]. The main task is then to find an appropriate and realistic alignment of the target protein sequence to the template sequence. In combination with an accurate prediction of the binding free energy for a template-based complex model, one could then predict if the putative protein-protein interaction is likely to occur.

In general, to enhance the impact of protein-protein docking in structural biology, it is highly desirable to be able to use partner protein structures obtained by comparative (homology) modeling. The accuracy of such comparative models depends on

the correct alignment of target and template sequence. Even in cases of significant average target-template similarity, the quality of the alignment is often not uniform along the whole protein sequence, for example, due to insertions or deletions in the aligned sequences which can result in structural inaccuracies. Overlap of such inaccurate structural segments with the protein region in contact with binding partners may interfere with the possibility to produce near-native complexes using template-based modeling or rigid docking methods. This is also reflected in the fact that docking cases that involve homology modeled protein partners belong to the most difficult cases in the CAPRI docking challenge. [181, 182]

4.3 Force field and knowledge-based scoring methods for ranking and to predict binding affinities

The comparison of known native protein-protein interfaces indicates typically a well-packed interaction region with high shape complementarity between protein partners and few cavities some of which might be occupied by water molecules [53, 137, 16, 15]. Polar side chain or backbone groups buried at the interface are forming hydrogen bonds or other polar contacts. Assuming that both nonpolar, as well as polar contacts, contribute on average favorably to protein-protein interactions one early simple idea is to relate binding affinity to the size of the buried solvent accessible surface area (buried SASA: BSA) upon complex formation. The BSA is obtained by subtracting the SASA of the complex from the SASA of the two protein partners. Surprisingly, the BSA without considering the detailed nature of the interface correlates already quite well with the experimentally measured binding affinity ($R \sim -.55$), if one excludes structures that undergo large changes upon complex formation. [153, 15] More sophisticated physics-based or knowledge-based approaches often result in correlation coefficients that are not much larger. However, interestingly, in systematic docking searches, one frequently obtains incorrect solutions with similar BSA as the near-native docked complexes. Often for pairs of proteins, one can generate complex geometries that have a similar or larger BSA than the native geometry. It indicates that the assumption that polar and nonpolar interface regions all contribute on average favorably to binding might be reasonable for the native interface but not for alternative non-native interfaces. Hence, these incorrect interfaces are either not well packed or contain other unfavorable contacts that prevent favorable association. Recently, it was also found that residues outside of the interface (defined by the BSA) can contribute to binding [154]. In an extension of the BSA approach for binding prediction, Vangone and Bonvin [304] suggested a contact-based scoring with optimal weights on various types of contacts at an interface (e.g., polar-polar, polar-nonpolar, etc.) that showed improved correlation with experimental binding affinity data on a benchmark set ($R \sim -.73$).

More sophisticated statistical- or knowledge-based potentials can be designed that either are based on the statistics of residue or atom contacts at interfaces or can even include the distance and orientation of residues (atoms) around interfaces. As the term "knowledge-based" indicates such potentials are extracted from known protein-protein complexes. The underlying concept is to relate the observed frequency of atom-atom or chemical group-group contacts (or distances) to the corresponding expected frequency assuming a random distribution. Overrepresentation or underrepresentation of a given pair of atoms or residues relates to favorable or unfavorable interactions. In most cases, the inverse-Boltzmann statistics with an appropriate reference state is used to derive an effective potential in terms of group distances and possibly also group-group orientation. The idea to extract effective interaction energies between groups or atoms based on contact frequencies in known protein structures dates back to Tanaka and Scheraga [295] and was further pioneered by Miyazawa and Jernigan [218] as well as Sippl and Weitckus [283]. For evaluating protein-protein complexes, a variety of knowledge-based statistical potentials have been designed in recent years [90, 217, 334, 54, 34]. The potentials differ in the resolution of the interface description. Several potentials are based on interatomic distances and possibly also orientation [334, 190, 138, 186, 191] or representing the protein on the level of whole residues or chemical groups [332, 47, 74, 260]. The potentials also vary in the number of atom or pseudo atom types or the reference state from which the expected contact or distance probability for atom-pairs are derived. Although mostly used for scoring and ranking docked protein-protein complexes, statistical potentials can also be optimized for predicting binding affinities. In this latter case, one should keep in mind that many contributions to binding ranging from solvent effects, restriction of conformational, rotational, and translational mobility of partners (that can all be different for individual complexes) are all averaged over many (training) complexes and merged into atom (or group) pair potentials to estimate binding affinities of predicted complexes. In additon, typically a knowledge-based scoring includes only interaction terms between partners and does not account for possible internal energy changes of each partner. Given these approximations, it is surprising that often quite reasonable correlations to experimental data are observed [304, 334].

Besides using knowledge-based statistical potentials for ranking of docked complexes, it can also be based on a molecular mechanics (MM) type force field description of binding partners [73, 44, 263]. Similar to the statistical potentials discussed above, these scoring potentials are mostly used to provide a relative ranking of single predicted complex structures but can also be optimized to predict the binding affinity of protein-protein complexes [115, 182]. For ranking, various terms of the force field are weighted to give an optimal correlation to experimental data on a training set of complexes [9, 67]. The majority of force field scoring potentials neglect any intramolecular energy changes of the binding partners and therefore do not

include changes in intramolecular interactions or changes in conformational entropy (e.g., restriction of conformational fluctuations upon binding).

For a binding process, one can, however, distinguish several energetic and entropic contributions. Binding results in an interaction between binding partners that can involve electrostatic and van der Waals intermolecular interactions. The partners will also change their average conformation (see Figure 4.1) that is accompanied by a change in intramolecular interactions which is typically an unfavorable contribution. Additionally, the interaction of each partner with the solvent (and surrounding ions) will change upon complex formation. For the change in solvation, one typically distinguishes between a nonpolar contribution (related to hydrophobic effect) and a polar (electrostatic solvation or reaction field) contribution [114, 66, 89, 311]. For rapid evaluation of single complex structures, usually explicit solvent molecules are not included. Hence, solvation contributions are calculated using an implicit solvent model (see Figure 4.3).

It is more realistic to represent a complex geometry not by a single structure but by an ensemble of relevant conformations and to estimate binding affinities from the evaluation of the ensemble of conformations. In the linear interaction energy (LIE) method, this is achieved by taking appropriately weighted averages of interactions between partners from MD simulations [14, 123]. More frequently, approaches are used that involve a reevaluation of explicit solvent MD trajectories after replacement of the surrounding environment by an implicit solvent model (see next paragraph).

4.4 MM-Poisson-Boltzmann/surface area and MM-generalized Born/surface area ensemble-based "endpoint" free energy methods

In recent years, full atomic resolution MM approaches for the binding partners combined with an implicit continuum solvent model have been applied to evaluate protein-protein complexes. In contrast to the scoring of single complexes, an ensemble of complex conformations is evaluated in MM Poisson-Boltzmann/surface area (MM-PBSA) or MM-GBSA (using the generalized Born method instead of the Poisson-Boltzmann) approaches. In most cases, the ensemble is generated using MD simulations with an explicit solvent representation. To limit the computational demand and to also keep a narrow distribution of conformations near an initial state, usually, simulations of a few nanoseconds are performed [311]. Due to the large energy fluctuations of the explicit solvent molecules, the reanalysis of the trajectory (ensemble) is performed after removing the explicit water molecules (sometimes interface waters are retained) employing an implicit solvent model.

For each evaluated complex structure, the mean partner interaction energy can be

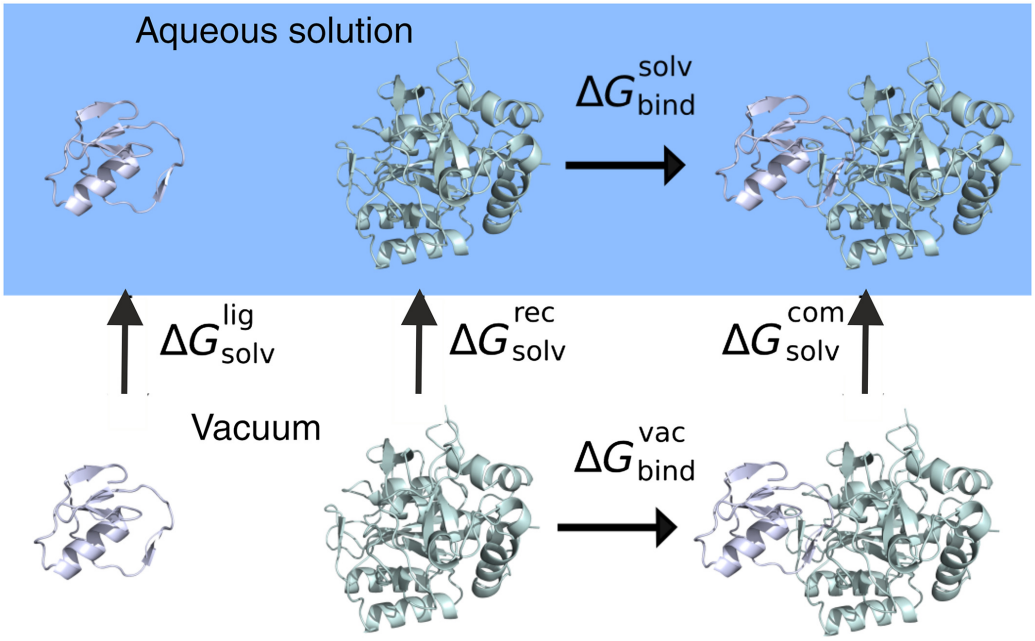


Figure 4.3: Evaluation of protein-protein complexes based on a continuum solvent model during MM-PBSA or MM-GBSA calculations. The binding process consists of an interaction contribution indicated in the lower panel (interaction energy is calculated as the difference in the vacuum energies of the complex and the separate partners). The transfer of the partners and the complex into the aqueous environment (upper panel) adds a solvation contribution (also calculated as the difference between complex and partner contributions). The solvation part consists typically of separate cavity terms and van der Waals interaction with the solvent plus an electrostatic reaction field (solvation) contribution either based on the generalized Born (GB) method or based on solving the finite-difference Poisson-Boltzmann (FDPB) equation numerically.

calculated. In the most basic single trajectory approach, this is achieved by taking the ensemble average energies of the complex and subtracting the corresponding energies of the partners from the same trajectory. This approximation implies that ligand and receptor do not undergo significant conformational changes upon binding and changes in intra-molecular energies can be neglected. The mean interaction energy consists of pairwise electrostatic Coulomb interactions and electrostatic (polar) solvation contributions obtained using the finite-difference Poisson-Boltzmann or generalized Born (GB) equations. Nonpolar solvation (or desolvation) is usually calculated from the BSA surface upon complex formation using an empirical surface tension parameter that represents both cavity creation and van der Waals interaction

of the protein with the solvent. Alternatively, the nonpolar part can also be split further into a surface area dependent cavity or hydrophobic term and a change in van der Waals interaction between proteins and solvent. The latter contribution can be estimated from a solvent grid representation around the complex [311] or a surface integral approach [328]. To include changes in intramolecular contributions, it is possible to run MD simulations not only for the complex but also for separate (unbound) partners and evaluate each ensemble separately. However, the single trajectory approach is used much more frequently and gives typically better convergence of the mean energies due to the cancellation of the intramolecular contributions. Nevertheless, interaction energies obtained from MM-PBSA or MM-GBSA can show significant statistical errors due to numerically small interaction energies that need to be calculated from the subtraction of numerically large and slowly converging mean energies (of the complex and the individual partners).

In addition to interaction energies, changes in the conformational entropy of binding partners can be estimated from an NM analysis of the complex and the isolated partners. This term is often neglected due to its large computational costs or replaced using alternative approaches based on the energy fluctuations within the ensemble or a quasi-harmonic (QH) analysis of the trajectories. Methods have also been developed to just obtain the change in translational and orientational (external) entropy of one partner to the other [52, 23].

MM-GBSA and MM-PBSA have been used both for the calculation of absolute protein-protein binding affinities and to evaluate docked protein-protein complex structures (reviewed in Reference [311]). For example, Gohlke et al [104, 103]. investigated the Ras-Raf and Ras-RalGDS complex and reported binding free energy in good agreement with the experiment, however, depending on how conformational entropy was estimated and with errors of about several kcal/mol. MM-GBSA and MM-PBSA were successfully used in several studies for reranking protein-protein docking solutions [148, 75, 45, 197, 187]. A very systematic study to predict the free energy of binding and to score docked complexes was recently performed by Chen et al [45]. The authors compared various force fields, protocols for performing MD simulations and using the Poisson-Boltzmann or the GB solvation model (not including conformational entropy effects) for 46 protein-protein complexes. The highest correlation between the predicted binding affinities and the experimental data was -0.64 using MM-GBSA, a low interior dielectric constant of 1 and the AMBER ff02 force field. This correlation was better than using MM-PBSA for which the highest correlation was -0.523.

Often water molecules form specific water-mediated contacts at protein-protein interfaces that may not be accurately represented in an implicit solvent model [51]. It is straightforward to include interface water molecules in the calculations and only treat the bulk water as a dielectric continuum [197]. The inclusion of an explicit water model has been proven to give good results in various protein-protein

binding affinity predictions for single complexes (e.g., Ulucan et al [301].). For example, the correlation of MM-GBSA results to experimental binding affinities of 20 native protein-protein complexes was shown to increase significantly (up to 30%) by the inclusion of 30 explicit water molecules at the binding interface [197]. On a smaller test set of four proteins, crystal water molecules were added in the evERdock approach that improved water-mediated contacts so that the identification of near-native binding decoys could be improved [293].

The changes in conformational entropy upon protein-protein complex formation are usually neglected in the MM-PBSA evaluation due to the large computational costs to perform NM analysis on protein-protein complexes and the isolated partners. However, alternative methods to estimate the conformational entropy contribution have recently been introduced and tested also for evaluating predicted protein-protein complexes. In the interaction entropy (IE) approach, the protein-ligand or protein-protein interaction energy fluctuations during the MD trajectories are used to estimate the conformational entropy [76]. This method does not allow to calculate absolute entropy values but is applicable to calculate relative entropy changes, for example, after protein-ligand binding. As no extra computational effort is needed, it has been concluded that for receptor-ligand binding affinity prediction using IE is superior to the standard NM analysis for estimating entropy effects. IE has been successfully applied in studies in combination with MM-PBSA and MM-GBSA calculations of protein-protein [292, 291] or protein-ligand [76, 291] binding affinities. Interestingly, quite substantial differences in the resulting free energies of binding using NM and IE were encountered in several studies [76, 291, 163]. In a recent study of Sun et al [291]., entropy effects on the performance of endpoint methods of over 1,500 protein-ligand systems were assessed. The best correlation to the experimental binding affinities was gained with IE, whereas the absolute binding free energy values had the highest correspondence to experimental values using NM calculations. Recently, however, Kohut et al [163]. proposed that the reproducibility of IE is less robust than that of NM or QH, especially for flexible systems. The calculated entropy value is mainly determined by the highest spikes of interaction energy and it is argued that the calculated entropies are difficult to converge as the simulations are prolonged. Aldeghi et al [8]. also found a higher sensitivity of the IE term to the simulated ensemble than the other MM-PBSA terms for three sets of bromodomain-inhibitor pairs. Hence, further testing could be useful to check the robustness of the IE approach. In a study of 20 protein-protein systems using IE with MM/GBSA, the mean absolute error to experimental binding affinities could be substantially reduced by optimizing the residue type-specific dielectric constants, the errors were especially lower than with NM analysis using a standard dielectric constant of 1 [187].

Formally, the solute entropy change during association can be split into an external entropic contribution due to the reduction of motion in external degrees

of freedom (relative position and orientation of ligand and receptor) and internal entropy (conformational) upon complex formation. Although a full decoupling of external and internal entropy is not in general possible, one can still compute the lowest upper bound of the external entropy [23]. The number of configurations needed to obtain converged results is, however, quite high using the approach with simulations of over 1 μs required for a Barnase-Barstar complex. Furthermore, an external entropy correction alone has been shown to not necessarily improve the correlation to experimental binding affinities for protein-ligand systems [213].

4.4.1 Mutations influencing protein-protein binding affinities

Mutagenesis of residues at protein-protein interfaces has demonstrated that the contributions to binding affinity are not uniformly distributed but can often be attributed to a small number of residues called hot spots [156, 249]. For protein engineering, it is of significant interest to predict changes in binding affinity of protein-protein complexes due to mutations and to identify important residues for the interaction (hotspots). Several approaches based on just single complex structures are available to estimate the effect of interface mutations (recently reviewed in [100]). The ensembles-based MM/PBSA or MM/GBSA approaches can also be used to identify hot spots and to calculate the change in the binding free energy upon mutation of interface residues [282]. The hot spots of 15 protein-protein complexes were calculated recently with MM/PBSA using residue type-specific dielectric constants (11 for charged residues and 7 for nonpolar and polar residues).77 In this study, a mean SE of 1.1 kcal/mol was achieved in 210 mutations after geometry optimization and subsequent MD simulations in explicit solvent. Using an extension of the MM/PBSA method with residue-specific dielectric constants, Petukh et al [246]. achieved a high correlation (correlation coefficient of -0.62) with experiment for a set of 1,300 mutations in 43 protein-protein complexes (several other applications are reviewed in Reference [311]).

Besides estimating binding free energy changes due to mutations using single complex conformations or changes in mean interaction energies obtained from the MM/PBSA or MM/GBSA methods, it is also possible to perform alchemical transformations to mutate residues in silico. In alchemical free energy simulations, one represents the selected amino acid side chain by two force fields, one representing the wild type (State A) and the other the mutated residue (State B). During a series of MD simulations, the force field for State A is switched off (decoupled from the interaction with other parts of the system) whereas the force field of State B is switched on. The changes in free energy can be calculated by integrating the generalized force along the switching pathway (thermodynamic integration [TI] [160]), free energy perturbation (FEP) [338] or using alternative methods such as Bennett acceptance ratio (BAR) method (see Section 3.4.1 for additional details on these methods) [24].

To obtain the effect of a mutation on binding affinity, the transformation needs to be performed in the complex and for the unbound solvated partner. The advantage of the approach is that all energetic as well as entropic contributions that may influence the change in binding affinity are accurately included (within the limits of a molecular mechanic force field description of the system). The disadvantage is the typically higher computational cost compared to the above d-scribed endpoint methods. Due to methodological progress and increased computational power, alchemical free energy methods are increasingly being used to study the effect of mutations on protein stability and protein-protein binding [98, 254, 161, 99, 216, 193]. Although typically performed in explicit solvent, it is also possible to perform alchemical transformations in implicit solvent [239] with computational costs similar to MM/PBSA but avoiding the endpoint approximations inherent to endpoint ensemble approaches. A recent systematic assessment of more than 100 mutations (including charge changing mutations) in four protein-protein complexes resulted in better performance and higher correlation of the alchemical FEP approach (root-mean-square error [RMSE] = 1.2 kcal/mol) than MM/GBSA (RMSE = 1.5 kcal/mol) in reproducing experimental affinities [56].

4.5 Rigorous free energy approaches to calculate absolute binding free energies

4.5.1 Analyzing protein-protein binding by multiple simulations and advanced sampling approaches

In principle, MD simulations allow studying protein-protein association at full atomic detail, including full flexibility of binding partners and explicit inclusion of surrounding water molecules and ions [148, 75]. The most straight forward approach is then to start from separate protein molecules and follow binding during sufficiently long MD simulations. Indeed, in early applications starting from separate components (in bound conformation) of the Barnase-Barstar complex relatively short MD simulations ($<$ 100 ns) were sufficient to observe complex formation to form a complex in close agreement with the native geometry [5]. However, such association simulations do not allow direct extraction of the associated binding free energy. To obtain the free energy of binding from unrestrained simulations requires both association to the native binding site but also sampling of dissociation from the bound complex. The dissociation constants of typical protein-protein complexes are in the nanomolar range and the dissociation rate can reach minutes or even larger times much beyond the timescale of current MD simulations. Hence, it seems that direct extraction of binding free energies from counting association and dissociation events during sufficiently long MD simulations is in principle only possible for

weakly bound complexes with fast association and dissociation characteristics. However, several approaches have been developed to overcome this sampling dilemma in recent years. Using a large number of simulations starting from various initial placements of the Barnase-Barstar system, the Noe group observed multiple binding events [248]. With a further adaptive selection of new starting configurations and addition of a biasing potential to promote also dissociation of the bound complex in Hamiltonian replica exchange (H-REMD) simulations [95] it was possible to sample sufficient association and dissociation events to generate a Markov model for the binding process [248]. Such Markov model allows the extraction of kinetic rates for transitions between various transient states of the system and also of the associated thermodynamic quantities [31, 50]. It includes not only the sampling of the native binding but allows also the characterization of alternative and intermediate encounter binding states. For the Barstar-Barnase system, the kinetic rates and the free energy of binding could be extracted in good agreement with experimental data. As an alternative to multiple unrestrained simulations to obtain a Markov model for the binding process, it is possible to use weighted ensemble (WE) methods to study binding processes along a preset reaction coordinate (RC) [131, 336]. Briefly, in this technique, the space along one or more RCs is discretized in intervals, and simulations are distributed along the coordinates such that each interval is populated by a fixed number of simulations. Each simulation is assigned a statistical weight that can be transported to neighboring intervals and new simulations are then started or eliminated such that the total statistical weight and the number of simulations per interval remain constant. The WE technique allows eventually the extraction of both kinetic and thermodynamic data along the binding RC [336]. Although so far only applied to study small-ligand binding to proteins, [336, 339] in principle, it could also be used to study protein-protein binding.

Finally, the idea of destabilizing the bound state with an added biasing potential during otherwise unrestraint MD simulations has been utilized in recent simulated tempering simulations applied to a set of six different protein-protein complexes.125 During the simulated tempering, different levels of the biasing potential that weakens the protein-protein interaction are applied. Still, extremely long MD simulations in explicit solvent were required (> 100 μs) to allow reversible association and dissociation of the protein-protein complexes [240]. It is also possible to extract kinetics and binding free energies from these simulations. Despite the rapid increase in computational resources, it is, however, unlikely that such methods will be used as routine methods to predict protein-protein binding affinities in the near future. One should keep in mind that the costs for the electricity to run such simulations alone exceed by far any experimental effort to determine the corresponding protein-protein binding affinities. An approach using a restraining potential to avoid trapping of protein-protein complexes in nonspecific transient states but at the same time restraining the partners not to separate has been introduced to refine docked

complexes [238]. Different levels of the biasing potential are applied in an H-REMD simulation with one replica running under the control of the original force field allowing rapid identification of putative complex geometries and possibly also an estimate of the binding free energy.

4.5.2 Binding free energies from advanced sampling including geometrical restraints

The complex formation of two partner molecules leads to the restriction of the translational and rotational degrees of freedom of the partner molecules relative to each other (Figure 4.4). In addition, the conformation and conformational freedom of the partners can change and the interaction with the surrounding solvent and ions is affected. Finally, van der Waals and electrostatic interactions may stabilize or destabilize a bound state. All these contributions influence the affinity or binding free energy of the complex. In binding free energy simulations, one ultimately aims at calculating free energies of binding including accurately all the above contributions.

Using Monte Carlo or MD simulations, it is indeed possible to calculate these contributions rigorously and to obtain absolute binding free energies of protein-protein complex formations. In principle, an alchemical transformation pathway to annihilate one partner protein in the complex and the unbound state is possible (as described in the previous section). The difference in free energy changes for these two calculations (and inclusion of standard state conditions) gives the absolute binding free energy of the two proteins. However, the annihilation of a complete protein force field in an explicit solvent simulation box may suffer from insufficient convergence. The resulting binding free energy is a small number obtained from the subtraction of large free energies of annihilating (or creating) the protein force field in the bound and unbound states.

Instead of an alchemical pathway, it is also possible to use a spatial coordinate to dissociate (or associate) a complex during an MD or MC simulation and record the associated free energy change. Since dissociation of a high-affinity complex typically does not occur during unrestrained MD simulations, a key element is to induce such dissociation (and/or association) along the preselected RC. In the majority of cases, this RC is a distance, for example, between the center of mass of binding partners or involving other subsets of atoms near or around the binding sites on the partners.

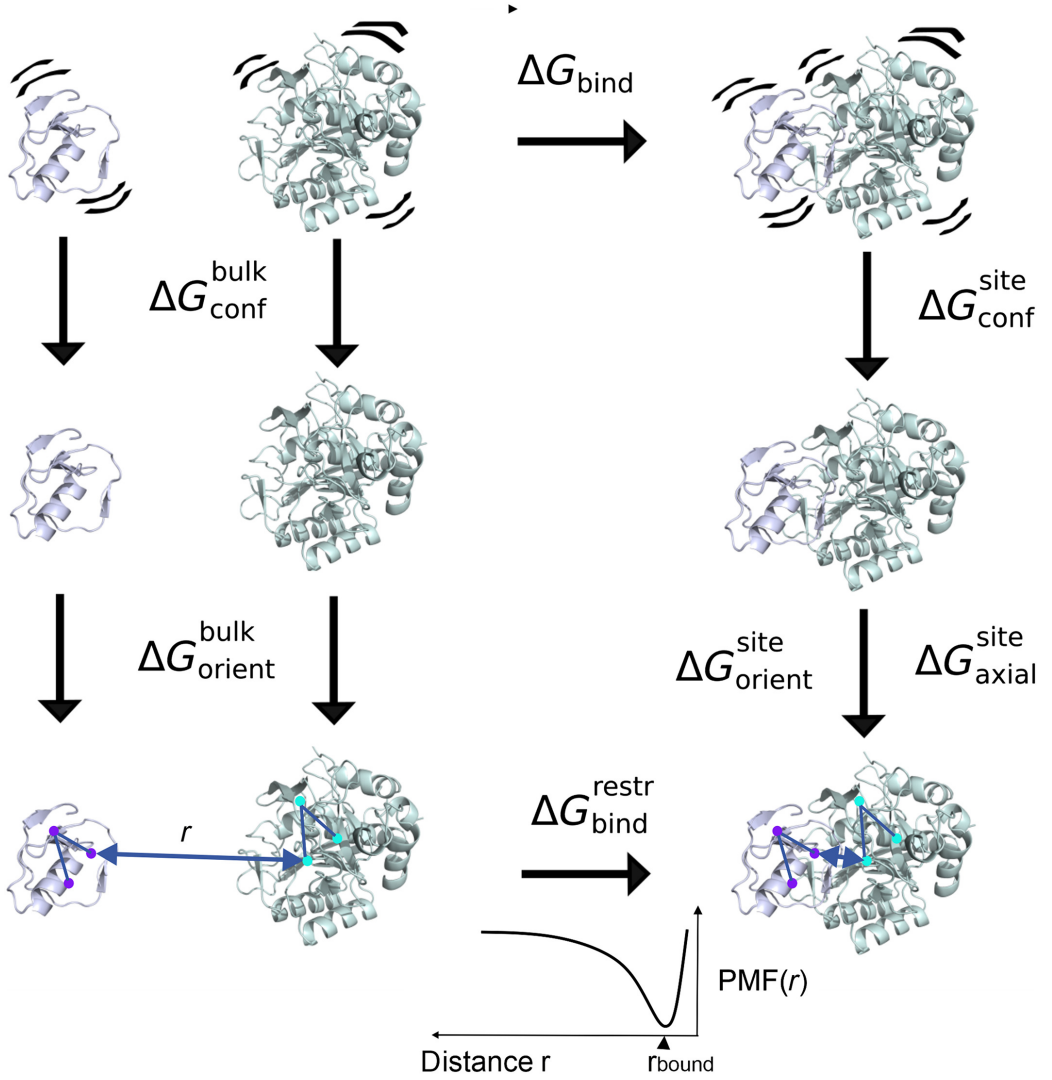


Figure 4.4: Binding free energy calculations including geometrical restraints. The conformational flexibility as well as the relative rotational/axial degrees of freedom of the binding partners are restrained during the calculation of the potential of mean force (PMF) along a separation coordinate (typically a distance r). This limits the necessary sampling of relevant states during the PMF calculation (third row). The contributions of the restricted mobility can be calculated at the endpoints (bound and fully separated states) of the PMF simulation using either analytical or free energy perturbation (FEP) methods. The absolute binding free energy between the unrestrained proteins ΔG_{bind} (top row) is calculated by accounting for several free energy contributions through the illustrated thermodynamic cycle. It requires the separate calculation of ΔG_{conf} (site,bulk) (second row), orientation ΔG_{orient} (site,bulk), and direction of separation ΔG_{axial_site} (third row).

By adding a biasing potential (umbrella potential) along this RC during the MD simulations, it is possible to induce dissociation or guide association of the complex. In most applications, this is achieved in discrete steps by adding a quadratic umbrella potential with reference values for the RC change in steps of 0.5-2 Å in the case of a distance RC. The weighted-histogram analysis method [169, 286] or related algorithms are then used to calculate a potential of mean force (PMF) along the RC. This PMF represents the free energy change associated with the dissociation or association of the partners along the RC. Sufficient sampling of relevant states within each umbrella window and overlap of sampled states between neighboring windows is of critical importance for convergence of the calculated free energy profile [161, 194, 49]. Significant improvements can often be achieved by coupling umbrella sampling with H-REMD allowing exchanges of sampled states between separate US windows [194, 164, 62]. Another possibility is to calculate first an approximate PMF and add this to bias the simulations along the RC to smooth the effective interaction for calculating a precise PMF in a second step or iteratively in several steps [331]. Alternatively, it is possible to use umbrella integration, [149] metadynamics, [59, 64, 177] or to apply an adaptive force (ABF) [59, 64] along the RC to offset the effective interaction between partners. In the case of metadynamics, this is achieved by adding a series of biasing potentials in the form of Gaussian functions along the RC [177]. The process is continued until a uniform diffusive sampling along the RC is observed. The added sum of Gaussian biasing potentials represents the free energy change along the RC. In the ABF approach, the added biasing force along the RC can be integrated to obtain an associated free energy change. Another recent approach to efficiently obtain free energy changes for reversible dissociation and association of a protein-protein complex is the perturbed distance restraints approach [286]. In this case, the bound and unbound states are characterized by different sets of distances that are included as distance restraints, and a coupling parameter lambda is employed to transform between the sets of restraints. The free energy along the coupling parameter can be extracted by TI or BAR methods and H-REMD can be used to improve the convergence of the binding free energy calculation [244].

The calculation of the PMF along an RC for dissociating a complex without any other restraints up to a state where the interaction of the partners can be neglected gives directly the free energy of binding (after accounting for standard state conditions of the binding partners). However, complete freedom of the conformation and relative orientation of the binding partners means also that to achieve converged free energy results sufficient sampling of relevant states at every position along the RC is required. This includes all possible orientations, placements, and conformations within each window in case of umbrella sampling. The convergence of calculating the free energy of dissociation/association along an RC can be enhanced by restricting the relative orientation and conformational freedom of the binding partners.

Woo and Roux [324] devised a method by using a series of simulations including restraints on the spatial arrangement and conformation of the binding partners to further reduce the necessary sampling at every step of the PMF simulation (Figures 4.4 and 4.5).

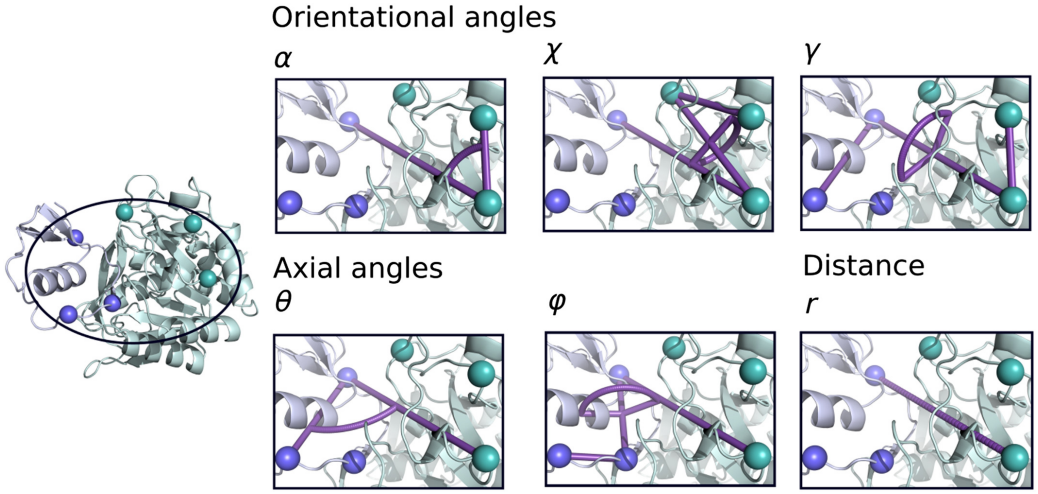


Figure 4.5: Illustration of the rotational and axial angles that are typically used to restrain the rotation of one partner and to restrain the axial placement (relative to the second partner) in PMF-based free energy simulations. For the definition, three centers in each partner need to be selected (indicated as blue and green spheres). These can be centers-of-mass of groups of atoms. The three Euler angles α, χ, and γ are used to restrain the rotation of one partner, and the two axial angles θ and ϕ restrict the direction of r to separate the protein partners. Additionally, the conformation of both partner proteins is restrained, preferably via root mean square deviation (RMSD) restraints to a reference structure. The distance r is typically used as a reaction coordinate for the PMF calculation to induce dissociation (or guide association) of the complex.

The contributions due to restricting orientation and placement as well as conformation need to be calculated only at the endpoints of the dissociation/association process (bound state vs. noninteracting separated state). Hence, much fewer states need to be sampled within each US window resulting in enhanced convergence of the PMF calculation. The restriction of orientation and axial placement in the separated state can be calculated analytically whereas commonly the FEP approach is employed to obtain the corresponding contributions in the bound complex (see Appendix A.1 for further details).

The contribution due to restriction of the conformation of the partners typically using a restraining term based on the RMSD of partners from a reference conforma-

tion is more difficult to estimate. Frequently, a single-step perturbation approach is used with an unrestrained simulation of the isolated partners and the complex as a reference and treating the conformational restraining potential as a (single-step) perturbation. The treatment assumes that an unrestrained simulation samples all relevant conformations in case of the complex but also for the unbound partners. If significant conformational changes are associated with complex formation (e.g., refolding processes), this may not be sufficient to estimate the restriction of conformational space due to binding. More accurate but also more demanding is the possibility to perform free energy simulations along an RMSD coordinate to a reference conformation [324].

Free energy calculations based on PMFs along a spatial (not alchemical) coordinate have been widely used to calculate absolute binding free energies [49]. However, most of these applications focused on drug binding, ion binding, or other small organic molecules binding to biological macromolecules. Typically, such PMF simulations are performed in explicit solvent that results in large system sizes and also requires long simulation times to achieve reasonable convergence. Only recently, applications to calculate protein-protein affinities have been published [244, 119, 120].

These examples indicate that the methodology, in principle, allows accurate calculation of protein-protein binding free energies if the bound complex structure is known. In practice, it is, however, desirable to use accurate free energy calculations to predict if a given docking geometry or designed interaction is realistic and to predict the affinity of the predicted interaction. Since PMF simulations in explicit solvent can take several days for a single binding mode depending on the system size an application for systematic evaluation of many docked protein-protein complexes is out of reach with current resources. The representation of both binding partners by a CG model is one possibility to drastically reduce the computational demand [208]. This was indeed successfully employed on T-cell receptor interactions with MHC-peptide complexes [208]. However, one should keep in mind that a CG model misses many details of intermolecular interactions such as hydrogen bonds and several conformational degrees of freedom are merged into effective interaction potentials. The agreement with experiment will in general dependent on how it is parameterized with respect to experimental binding data or effective interactions between CG centers.

Instead of using an explicit solvent representation, it is also possible to use an implicit, for example, GB, solvent model during US-based binding free energy simulations at atomic resolution. On modern graphical processing units (GPUs), GB simulations can be performed very efficiently. Due to the instantaneous solvent response (at every time point in equilibrium with the solute structure) and the possibility to use a low viscosity, it also allows for faster convergence than explicit solvent simulations (in each US window). In a recent study, it has been demonstrated that it

is possible to directly employ such binding free energy simulations to score a reasonable set of 50 decoy complexes for 20 test complexes within about a day on a GPU cluster (see Chapter 5) [279]. The calculated binding free energies for the near-native complex geometries were in reasonable agreement with experiment. Additionally, an improved ranking of near-native docking solutions compared to simple single complex structure evaluation after energy minimization or short MD simulations was observed [279]. The study demonstrated the feasibility for systematic evaluation of predicted complexes based on calculated binding free energies instead of single point interaction, knowledge-based scores, or mean interaction energies. Further developments of the approach that preferably employ an explicit solvent representation are desirable for evaluating predicted complexes and allowing a clear judgment if a given geometry is thermodynamically stable or may represent only a transient state or represent an unfavorable binding geometry.

4.6 Conclusion

In recent years, the number of experimentally determined protein-protein complex structures and structures of large multiprotein assemblies has grown rapidly. In addition, bioinformatics data, data on residue conversation at putative interfaces, low-resolution experimental data, and the increasing number of protein-protein template structures allow creation of structural models of many putative complexes and interface geometries. Hence, judging if a predicted complex geometry or putative domain-domain interaction is stable and of functional relevance is of increasing importance. This is also of importance for the design of new protein-protein interactions preferably with a controlled affinity. For protein-protein interactions in the cell, one should keep in mind that the effective interaction is influenced by the distribution of proteins in the different compartments of a cell and the crowded environment. Nevertheless, an accurate estimate of the affinity of a putative complex structure is already valuable for the isolated complex. Rigorous binding free energy calculations employing a spatial coordinate for protein-protein binding and unbinding on a so-far limited number of cases show promise and may offer a route to obtain the desired accuracy to offer reliable predictions (that do not need to be controlled by experiment). However, if the computational effort is far larger than an experimental affinity measurement the practical value is limited. The combination of fast scoring-type approaches either based on single structures or ensembles to preselect likely putative natural or designed interfaces and limit the application of the most time-consuming and accurate methods to a small subset might be a reasonable route for practical applications. It is important to note, however, that rigorous binding free energy calculations even on known complexes with known binding affinity allow one to distinguish and characterize different energetic and

entropic contributions to binding often difficult to obtain experimentally. Hence, these studies give valuable insights into the mechanism of specific protein-protein recognition. It is important to note that in vivo not all interactions reach equilibrium and hence not only the binding affinity but also the kinetics of interactions are of increasing relevance [250, 39]. Finally, with increasing computational resources and the development of improved advanced sampling methods, the option to directly follow protein-protein association and dissociation offers the opportunity for an in-depth understanding of transient nonspecific and specific binding.

5 Evaluation of Predicted Protein-Protein Complexes by Binding Free Energy Simulations

As discussed in Chapter 4 the binding free energy of a predicted protein-protein complex can be calculated using umbrella sampling (US) (see Section 3.4.2) along a predefined dissociation/association coordinate of a complex. In the current Chapter, such atomistic US-molecular dynamics simulations are employed including appropriate conformational and axial restraints and an implicit generalized Born solvent model to calculate binding free energies of a large set of docked decoys for 20 different complexes. In principle, the approach includes all energetic and entropic contributions to the binding process. Although time-consuming it may open up a new route for realistic ranking of predicted geometries based on the calculated free energy of binding.[1]

5.1 Introduction

Protein-Protein interactions play a key role in basically all cellular processes, ranging from signal transduction to enzymatic transformations and transport phenomena. Understanding the three-dimensional (3D) structure of protein-protein complexes is of major importance for understanding the molecular mechanism of many diseases and to identify potential targets for drug design [128, 226, 145]. Structure determination of protein-protein complexes at high resolution and binding affinity measurements are challenging experimentally, and for transient interactions not always possible [287, 224]. The theoretical prediction of the structure of protein-protein complexes is not less challenging due to the complicated shape of proteins, the variety of interactions, and possible conformational changes associated with binding [29, 326, 330].

A large variety of docking methods are available to generate putative complexes based on the unbound protein or modeled protein conformations [326, 129]. Simple scoring functions are typically used during the systematic search in the six degrees of freedom (three translational and three rotational) of possible relative placements

[1] The contents of this chapter have been published in a similar form in [279].

to rank predicted solutions [302, 129, 184, 333, 150].
Large efforts have been made to introduce flexibility of the proteins in docking schemes since a major limiting factor of most rigid docking methods are induced-fit conformational changes [174, 296, 326]. This can for instance be targeted using soft contacts, i.e. reducing the energy potentials at sterical contacts [83, 192], employing soft coarse-grained force fields [329, 327, 69] or incorporating site chain flexibility [44, 211] or possible global motions based on soft collective modes [209, 210, 219]. Frequently, a refinement step is employed on a subset of solutions obtained in the first systematic docking approach to obtain higher quality solutions. Experimental data also can help to limit the selected subset of putative solutions. Several refinement methods based on Monte Carlo (MC) or molecular dynamics (MD) simulations have been developed in recent years [67, 204, 263, 306, 69]. The final identification of the most realistic predicted complex involves a scoring function that ideally represents directly the binding affinity of the bound protein partners. Common scoring functions either use force field-based approaches, i.e. scoring functions that are based on a physical model for the potential energy functions [168, 67, 44, 302] or empirical scoring approaches, that incorporate experimental data in a statistical scoring [325, 327, 333, 113].

Typically, a scoring function involves the evaluation of one complex structure and therefore neglects any entropic contributions to the restricted conformational flexibility and translational/rotational mobility of the partner molecules. Furthermore, most scoring functions consist of pairwise additive terms and often neglect or only approximately account for solvent effects.
In recent years, approaches with a full atomic resolution of the binding partners and a continuum model for the surrounding solvent have been applied to evaluate docked complexes [103, 45, 223]. Instead of scoring only one complex an ensemble of complex geometries is evaluated in Molecular Mechanics Poisson-Boltzmann/SurfaceArea (MM-PBSA) or MM-GBSA (using the generalized Born method instead of the Poisson-Boltzmann) approaches [165]. For each complex structure, the partner interaction energy is calculated and electrostatics are obtained using the finite-difference Poisson-Boltzmann or generalized Born equations. Hydrophobic contributions are calculated from the buried solvent accessible surface upon complex formation. A mean over the ensemble results in a calculated score. MM-GBSA and MM-PBSA have been used successfully to improve the prediction of near-native protein-protein complex geometries [45]. However, MM-GBSA or MMPBSA are end-point methods that only give a mean interaction energy between partners without taking into account the energetic changes within each partner upon binding and also usually do not account for changes in the conformational entropy upon binding.

In addition, interaction energies obtained from MM-PBSA can show large statistical errors due to numerically small interaction energies that need to be calculated

from the subtraction of numerically large and slowly converging mean energies (of the complex and the individual partners). For example, Gohlke et al. [103] investigated the Ras-Raf and Ras-RalGDS complex and reported errors of calculated binding free energies of about 6 kcal/mol corresponding to about one third of the calculated total binding free energy [103].

However, using MD simulations there are rigorous approaches of calculating relative or even absolute binding free energies based for example on umbrella sampling (US) along a dissociation (or association) coordinate [102, 324, 71, 119, 331]. The convergence of calculating the free energy of dissociation/association along a typical distance coordinate can be enhanced by adding restraints to the relative orientation and conformational freedom of the binding partners [324]. By adding contributions due to restraining degrees of freedom to the potential of mean force along the reaction coordinate it is in principle possible to extract an absolute binding free energy. In contrast to MM-PBSA or MM-GBSA such an approach includes all energetic and entropic contributions to binding, hence, is ideal for ranking of binding modes. Typically and preferably, US free energy simulations are performed on ligand-receptor or protein-protein complexes including explicit solvent. However, only a few but quite successful applications to protein-protein or peptide-protein binding have so far been published [120, 119] i.e. due to the large computational demand of explicit solvent MD simulations. It should be emphasized since such simulations can take several days for a single binding mode an application for systematic evaluation of many docked protein-protein complexes is out of reach with current resources.

However, instead of using an explicit solvent representation it is also possible to perform US-based binding free energy simulations at atomic resolution in an implicit generalized Born (GB) solvent [230, 236]. On modern graphical processing units (GPUs) GB simulations can be performed very efficiently [108] and also allow for faster convergence than explicit solvent simulations (in each US window) because the solvent response is instantaneous for each sampled structure. We demonstrate that it is possible to directly employ such binding free energy simulations to score a reasonable set of 50 preselected complexes for a series of 20 test cases within a day on a GPU cluster.

This demonstrates a basis for systematic evaluation of predicted complexes based on calculated binding free energies instead of single point interaction or knowledge-based scores or mean interaction energies. We compared the docking ranking based on calculated binding free energies with simple one-point energy evaluation of energy-minimized complexes, and after MD-based refinement of docked complexes and compare the results to experimental binding free energies.

5.2 Material and methods

5.2.1 Protein-protein docking using ATTRACT

Since umbrella sampling simulations for evaluation of docked complexes are computationally highly demanding it was necessary to limit the number of systems to a set of 20 complexes from the protein docking benchmark 5 [132] (Table 5.1) with small partner structures. For each system, the bound and unbound conformations of both partners are known. The docking was performed using the unbound partner structures with the program ATTRACT [329, 69]. A standard docking protocol with rigid partner proteins and grid-acceleration for fast energy evaluation was used [68]. After the final docking scoring step (based on the ATTRACT scoring potential), the 300 top-ranked complexes were considered. We distinguish between receptor and ligand-protein according to the assignment in the benchmark 5 [132] (typically the large protein partner is the receptor and the smaller is the ligand partner). For evaluation of the deviation of a decoy complex from the native geometry, we use the RMSD (root mean square deviation of the complex after best superposition of the complex to the native complex structure, only the C_α atoms considered). The 50 models with the lowest RMSD to the native complex structure were used for further evaluation based on atomistic simulations.

5.2.2 Refinement of docking solutions using molecular dynamics simulations

Atomistic refinement of docked complexes was performed with the Amber16 molecular dynamics (MD) Package [42] using the pmemd.cuda module [108] in combination with the ff14SB [198] force field for proteins following a standard protocol developed previously [263]. Energy minimization and MD simulations employed the generalized Born implicit solvent model [230] (igb=8 in the Amber input section) and an infinite cutoff. Energy minimization consisted of 2500 minimization steps (400 steps of steepest descent, 2100 steps of conjugate gradient). Scoring after minimization was performed by subtracting the potential energy of the partners from the energy of the complex. For further MD-refinement the systems were heated in three steps (each 5 ps) to 300 K using a Langevin thermostat for temperature scaling. For the production run (mass-weighted) RMSD restraints of the C_α atoms of each individual protein were applied (force constant 10000 kcal/mol $\cdot$ Å^2 for each protein), together with a small distance restraint (force constant 0.5 kcal/mol $\cdot$ Å^2) between the COM of ligand and receptor that prevents the protein partners from dissociating. Finally, weak distance restraints between the C_α atoms at the interface of the proteins were applied to gently push the binding partners towards each other (force constant 0.25 kcal/mol $\cdot$ Å^2). The purpose of the restraints during the MD

simulation is to avoid dissociation due to possible initial sterical overlap and to improve the sterical complementarity at the interface. In total, the refined structures were simulated for 30 ps and evaluated after a final minimization for another 2500 steps. On average the refinement of 50 models took approximately 1 hour on a PC with a GeForce GTX 1080 for the different proteins.

Table 5.1: Table of the 20 protein complexes analyzed in this study with according PDB-id and difficulty (as defined in the benchmark [132]).

PDB	Difficulty	Protein1	Protein2
7cei	Rigid Body	Colicin E7 nuclease	Im7 immunity protein
1ak4	Rigid Body	Cyclophilin	HIV capsid
1ay7	Rigid Body	Rnase SA	Barstar
1ppe	Rigid Body	Trypsin	CMTI-1 squash inhibitor
1r0r	Rigid Body	Subtilisin carlsberg	OMTKY
2i25	Rigid Body	Shark single domain antigen receptor	Lysozyme
1j2j	Rigid Body	Arf1 GTPase.GNP-RanBD1	GAT domain of GGA1
1z0k	Rigid Body	RAB4 binding domain of Rabenosyn	Rab4A GTPase
1qa9	Rigid Body	CD2	CD58
1gcq	Rigid Body	GRB2 C-ter SH3 domain	Vav N-ter SH3 domain
2oob	Rigid Body	Ubiquitin ligase	Ubiquitin
1ffw	Rigid Body	Chemotaxis protein CheY	Chemotaxis protein CheA
1zhi	Rigid Body	BAH domain of Orc1	Sir Orc-interaction domain
3a4s	Rigid Body	SUMO-conjugating enzyme UBC9	NFATC2-interacting protein
1fle	Rigid Body	Elastase	Elafin
2sni	Rigid Body	Subtilisin	Chymotrypsin inhibitor 2
3sgq	Rigid Body	Ovomucoid inhibitor third domain	Streptogrisin B
1syx	Medium	Spliceosomal U5 15 kDa protein	CD2 receptor binding protein 2
2cfh	Medium	BET3	TPC6
1z5y	Rigid Body	N-term of DsbD	E.coli CCMG protein

5.2.3 Restraint umbrella sampling

The complexes obtained after the MD-based refinement served as the starting structures for the umbrella sampling (US) simulations (details on the calculations of absolute binding free energies are given in Appendix A.1). The same temperature and solvent conditions as for the refinement were used. The center of mass distance between ligand's and receptor's C_α atoms served as the reaction coordinate. A harmonic potential $U(\xi) = \frac{k_\xi}{2}(\xi_0 - \xi_i)^2$ was applied, with the force constant k_ξ ranging from 4 $kcal/mol \cdot \AA^2$ to 6 $kcal/mol \cdot \AA^2$. The reference distance ξ_0 was modified in each US window.

In addition to the distance restraint along the reaction coordinate 5 angular restraints (2 axial restraints and 3 orientational restraints) were applied ac-

cording to a harmonic potential $U(\alpha) = \frac{k_\alpha}{2}(\alpha_0 - \alpha_i)^2$ with a force constant of k_α=10 kcal/mol $\cdot$ rad^2. The angular restraints were based on 3 additional centers-of-masses of subsets of 20 atoms in each partner protein. The separation of the new COMs was taken as large as possible. Taking the angles between three or four points the axial or orientational restraints were defined (see Appendix Figure A.1). We gained the reference positions α_0 by evaluating the starting structures. To allow limited backbone and full side-chain flexibility of partner proteins the same harmonic intra-molecular RMSD restraints of the C_α atoms were applied as during MD refinement (force constant 0.05 kcal/mol $\cdot$ Å^2 per atom).

First, an initial series of MD-simulations for 20-60 ps (per window) was performed with a separation along the reaction coordinate of 0.15 to 0.5 Å between the windows. In each run, the coordinates and velocities of the previous window were used as starting conditions. The data gathering in each US window was extended to 1ns that allowed in general good overlap of sampled distance distributions and rapid convergence of the calculated PMF (see Appendix Figures A.3,A.4). A great advantage of the US technique is that it is possible to add additional windows after a fist inspection of the distribution overlap between neighboring US windows to improve the sampling in regions with too little sampled distance overlap between neighboring windows. Occasional new US windows were added to improve the convergence. Furthermore, due to the differences in the interaction strength of the protein partners in the different complexes the force constants and spacing were adjusted and slightly vary for each case due to variation in the slope of the free energy curves (steep or weakly varying PMF) (Table 5.2). Finally, the potentials of mean force (free energy change) along the COM distance (PMFs) were evaluated using the WHAM algorithm (in Section 3.4.2 the WHAM equations are given) [116]. Due to the small US windows, the conformational, axial and orientational restraints and the use of an implicit solvent reasonable convergence could be achieved with a simulation length of 1ns per window (Appendix Figures A.3,A.4,A.5)

Additionally, to obtain the absolute binding free energy, three free simulations of the complex, the ligand, and the receptor were performed for 4 ns, with no restraints applied. As starting structure the umbrella sampling window was used that corresponded to the minimum position of the PMF. The equilibrium binding constant can be calculated from the work of bringing the binding partner from the bulk to the native binding state (see details in Appendix Section A.1). [324]

$$K_{eq} = 4\pi \cdot \xi_{bulk}^2 \cdot \int_{site} d\xi \; e^{-\beta(A(\xi) - A(\xi_{bulk}))}. \tag{5.1}$$

The integration is performed over the phase space region representing the bound state. ξ_{bulk} indicates a distance in the bulk region where ligand and receptor are not bound.

Furthermore, the standard binding free energy can be expressed as follows [102, 324]

$$\Delta G_{bind} = -kT \ln(K_{eq} C^{\circ}), \tag{5.2}$$

with the standard concentration C°=1/1661 Å^3. The incorporation of this concentration is crucial, in order to compare the free energy values to experimental results. The advantage of this method is that the whole dissociation/association process is considered. Hence, also entropic effects are included and the whole free energy profile of binding is obtained.

A typical PMF $A(\xi)$ has the following characteristics: It lowers with increasing COM distance of the C_α atoms between ligand and receptor (x-axis) until the optimum distance is reached at the minimum PMF. After that, the PMF increases until the ligand and the receptor cease to influence each other and the PMF reaches a plateau (bulk region).

The region representing the bound state was defined as the states around the minimum position where the PMF does not exceed 2 kcal/mol. Note that the definition of the binding site is not prone to errors, as high PMF values are exponentially suppressed. Hence, by considering a much larger bound region, deviations of less than 0.01 kcal/mol were observed for the calculated free energy. Thus, for PMFs where the bulk state was below 2 kcal/mol, the barrier height was diminished to approximately 1 kcal/mol. Like this, we were able to obtain reasonable results also for models with a low affinity. Nevertheless, nearly all cases included some decoy complexes that were so weakly bound that no proper PMF was obtained and thus no free energy of binding could be calculated. In particular, for the protein 1qa9 only 21 models could be resolved followed by 3a4s with 38 results. As we are primarily interested in the ranking of the complexes with high affinity, the lack of on average 4 weakly bound decoy models does not deteriorate the results.

Additionally, the RMSD from the native structure was recalculated by evaluating the deviation at the minimum position of the PMF. Moreover, the uncertainties of the RMSD (standard deviation of the mean RMSD at the minimum PMF) and the free energy of binding (mean of the standard deviation was $\pm$1.0 kcal/mol) are presented. In this context, the simulations were split into 6 parts, where each one was treated separately, and the mean free energy of binding with its standard deviation was calculated.

The mean difference in ΔG between evaluating the parts of the simulations and evaluating the whole trajectory is 0.03 kcal/mol. It lies at least one order of magnitude under the range of the uncertainty and is clearly smaller than the scale in which changes in the specificity occur. Compared to the MD refinement the US simulations were much more expensive and took 5.0 days on average for all decoy complexes for one protein-protein complex (see Table 5.2) using one GeForce GTX 1080 or GeForce GTX 1080 Ti GPU card.

Table 5.2: Umbrella Sampling simulation setups for each protein-protein complex.

PDB	Number of windows	Simulation time days	Separation Å	Force constant kcal/mol · Å^2
7cei	30	4.1	0.30	5.0
1ak4	30	6.1	0.30	5.0
1ay7	25	3.1	0.25	6.0
1ppe	24	3.8	0.40	4.0
1r0r	30	5.7	0.30	5.0
2i25	40	5.8	0.30	5.0
1j2j	40	5.3	0.25	6.0
1z0k	24	3.4	0.30	5.0
1qa9	24	3.0	0.30	5.0
1gcq	24	1.5	0.40	4.0
2oob	20	0.6	0.50	4.0
1ffw	55	6.0	0.15	5.0
1zhi	30	6.9	0.30	5.0
3a4s	55	7.6	0.15	5.0
1fle	40	7.4	0.30	5.0
2sni	30	7.3	0.30	5.0
3sgq	25	3.2	0.30	6.0
1syx	40	4.0	0.25	6.0
2cfh	30	6.8	0.30	5.0
1z5y	40	6.7	0.30	5.0

Each complex is indicated by PDB-id. The number of windows corresponds to the number of separate umbrella sampling simulations. The distance between each US interval, force constant for the quadratic distance restraint and total simulation time on a single GPU (for 50 complexes) are also indicated.

5.3 Results and discussion

Realistic ranking of predicted protein-protein complex structures is one of the key challenges in the field of protein-protein docking and decisive for using such techniques in any useful application. So far the ranking of complexes is typically based on scoring functions based on force field or knowledge-based evaluation of single complex structures or on mean interactions energies between partners. However, protein-protein binding is determined by the associated free energy change that includes solvent effects and entropic as well as energetic contributions. Hence,

the binding free energy associated with complex formation can be considered as the ideal ranking score for evaluating docked complexes. The aim of our study is to evaluate the possibility to use a calculated binding free energy obtained from restraint US simulations as docking scoring function and to compare it with a score based on single-point energy minimization and after a short MD-based refinement (Figure 5.1).

Figure 5.1: The different scoring methods that are used in this study. The binding energy was calculated after energy minimization (EM) and MD based refinement. Restraint umbrella sampling (US) yields the binding free energy and the absolute binding free energy was accessed by accounting for the restraints via the Woo and Roux scheme [324].

Preferably, one would use an explicit solvent representation but due to the many water molecules (large box sizes are necessary for allowing to separate the protein partners to an appropriate unbound distance) such approach is computationally too demanding for application to many decoy complexes of a given protein-protein complex system. Besides, since it is possible to use a reduced effective viscosity in the implicit solvent case (see Methods) faster convergence of translational and rotational (diffusive) sampling of the partners is possible. The use of an implicit solvent has the additional advantage that the same force field and implicit solvent model is employed for scoring based just on energy minimization, or following short MD-based refinement or US-based free energy calculation.

Initial systematic docking on 20 benchmark test cases with unbound partner structures using ATTRACT [329, 69, 68] resulted in 50 docked decoy complexes with an RMSD less than 15 Å from the native geometry (see Table 5.2). The scoring based on the ATTRACT force field is indicated in the Supporting Information (Figure S8). The RMSD indicates the root-mean-square deviation of the complex relative to the native structure after best superposition onto the native complex. The docking result corresponds to a frequently encountered prediction case in which the approximate binding regions on partner molecules are known and only a limited set of solutions need to be considered.

For a first docking evaluation we use the same force field model as used in the US simulations based on the Amber ff14SB [198] in combination with an advanced GB model specifically designed for proteins [234, 136] (igb=8 option in Amber) after extensive energy minimization of each docking solution (Figure 5.2). The binding interaction energy ΔE_{Bind} was calculated according to equation 5.3, by subtracting the complex energy from the sum of the corresponding ligand and receptor energies,

$$\Delta E_{Bind} = \Delta E_{Coul} + \Delta E_{LJ} + \Delta G_{Solv} \tag{5.3}$$

with contributions due to Coulomb interaction (ΔE_{Coul}), Lennard-Jones interaction (ΔE_{LJ}) and difference in solvation free energy (ΔG_{Solv}). The last term consists of a polar component (ΔG_{GB}) and a nonpolar surface tension component that is proportional to the loss of solvent accessible surface area (SASA) upon complex formation. However, it turns out that the SASA surface tension contribution differed only little between the sampled conformations in the bound complexes (in comparison to the other interaction energy contributions) and therefore it was only calculated for the final structures and not included during energy minimization or MD simulations.

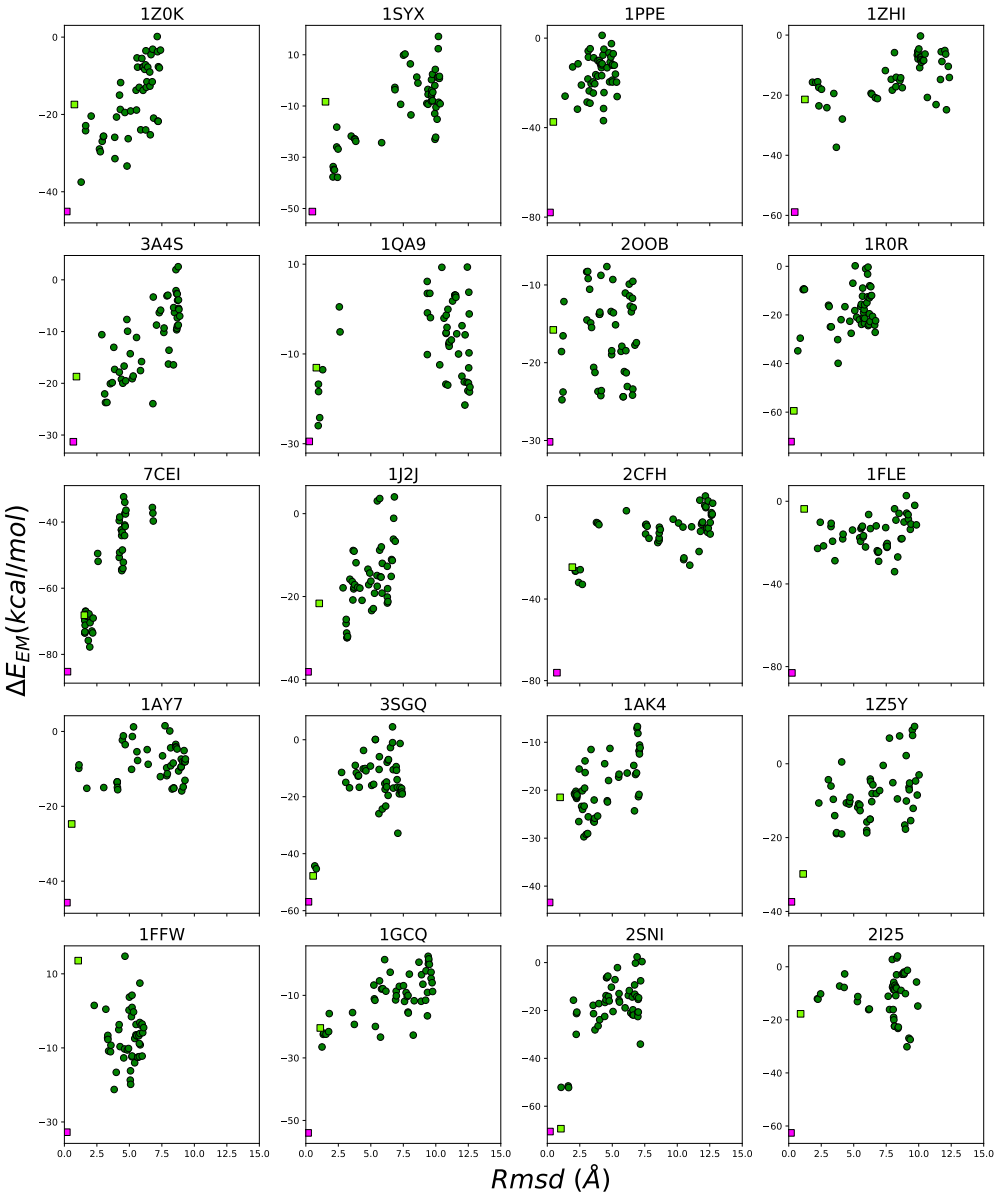


Figure 5.2: Force field scoring after energy minimization of all 50 decoy protein-protein complexes for each test case (indicated as pdb-id). The calculated interaction energy (green circles) between both protein partners is plotted vs. the RMSD from the native complex. The green squares display the unbound partners after superposition on the bound native complex and subsequent energy minimization (best possible placement for unbound rigid docking), the magenta square depicts the interaction energy of the native bound complex (after energy minimization).

Docking models with a very small RMSD of around 1 Å from the native complex, were obtained for the structures 3sgq, 1gcq, 2oob, 1qa9, 1ay7, 2sni and 1r0r. In the case of 1j2j and 3a4s the predicted structures closest to the native show an RMSD of 2-3 Å. While, in most cases, the RMSD was relatively evenly distributed in the range from 1-12 Å, for the complex 1qa9 only solutions with RMSD lower than 3 Å or higher than 9 Åwere obtained. The calculated scoring interaction energy of the best-scored models was highest for the structures 7cei (-78 kcal/mol), 2sni (-52 kcal/mol), and 3sgq (-45 kcal/mol), respectively. The lowest scoring value of around -15 kcal/mol was obtained for complex 1ay7, followed by -19 kcal/mol for 1z5y. Evaluation of the bound complex structure with the same energy minimization and force field protocol resulted in the best score for the bound complex in all test cases (Figure 5.2) supporting the quality of the continuum solvent force field score. In the case of the native complex with unbound partners (by fitting the C_α atoms of the partners to the native complex, corresponding to the best possible rigid docking solution for unbound docking) the resulting complex often scored considerably worth than the bound solution and other solutions obtained from the systematic search (Figure 5.2). In contrast, the RMSD of the energy minimized unbound form relative to the native complex was around 1 Å for nearly all cases. Only in the case of 7cei, 1syx and 2cfh the deviations were higher, up to 2 Å. These can be due to variations of the backbone orientation (7cei) or a conformational twist that involves whole α-helices (2cfh). The unfavorable scoring of the energy minimized unbound complex (in near-native starting geometry) observed for several cases is likely due to trapping of the incorrect side-chain and possibly backbone states at the protein-protein interface compared to the bound docking case. For four structures docking models were found that had nearly the same RMSD as the unbound form (3sgq, 1gcq, 1r0r and 2cfh). The correlation of binding energy to RMSD value was reasonable for 1syx, 3sgq, 7cei, 2cfh and 2sni. Here, models with a small deviation to the native complex were scored more favorable compared to high RMSD solutions. Nevertheless, for most proteins, the docking run produced several solutions with high RMSD that scored well (1ffw, 1qa9, 3a4s, 1ay7, 1gcq, 1ppe, 2oob, 1r0r, 1fle, 1z5y and 2i25). For some structures, e.g. 3a4s, 1j2j, 1ak4, 1r0r and 1qa9, low RMSD solutions with predicted high affinity were found, but at the same time also models with a similar RMSD but calculated low score.

The atomistic scoring function with implicit GB solvent model identifies the bound complex in most but not all cases as the best scoring solution. In some cases, solutions very close to the bound complex can significantly vary in scoring because of side chain or small backbone differences at the interface region. Hence, an MD-based refinement might be necessary to improve the specificity of the scoring.

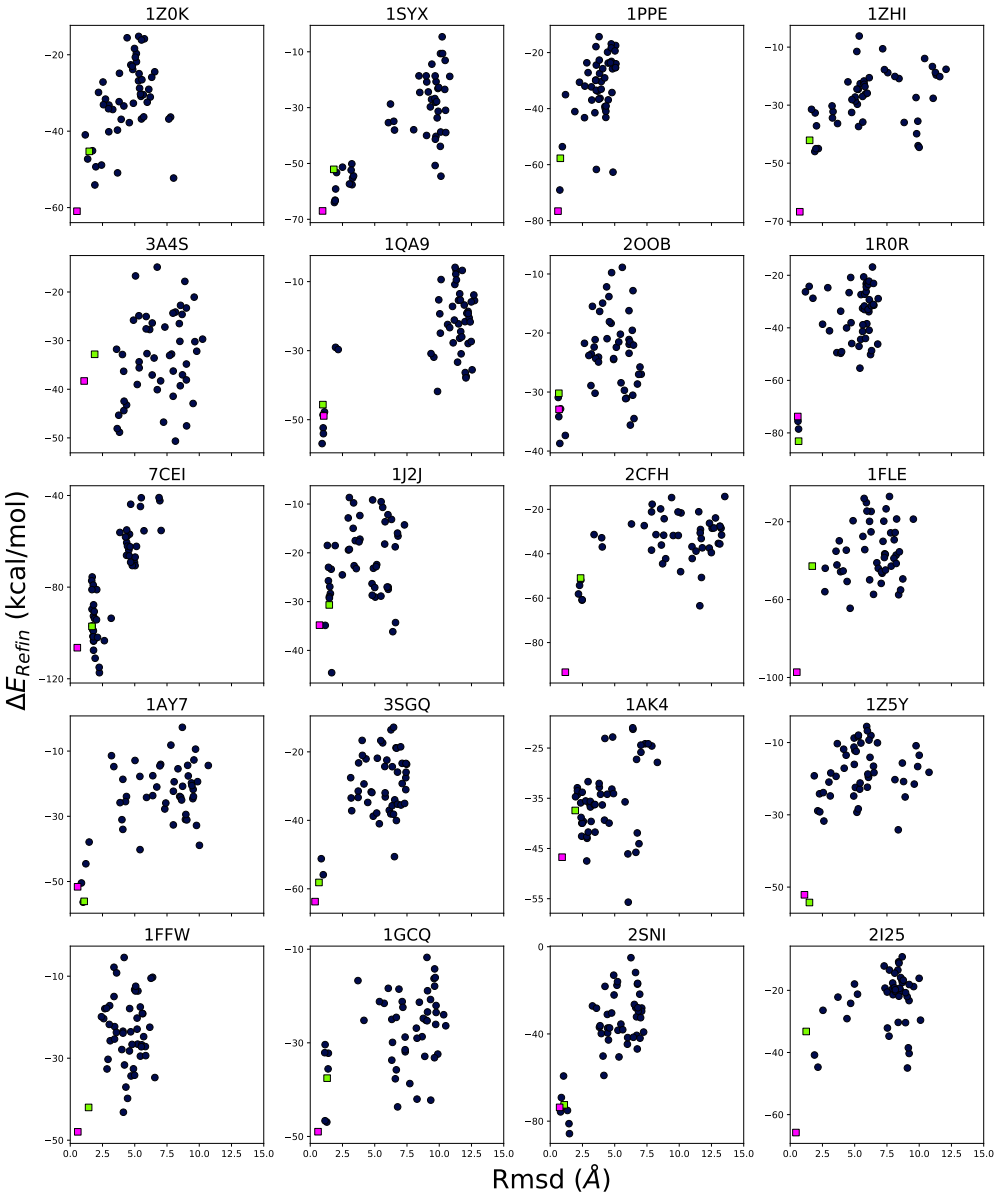


Figure 5.3: Force field scoring after molecular dynamics refinement of all 50 decoy protein-protein complexes for each test case (indicated as pdb-id).The calculated protein-protein interaction energy (blue circles) is plotted vs. the RMSD from the native complex. The interaction energy after refinement of the unbound complexes is indicated as green squares. The magenta squares depict the interaction energy of the native bound complex (after refinement).

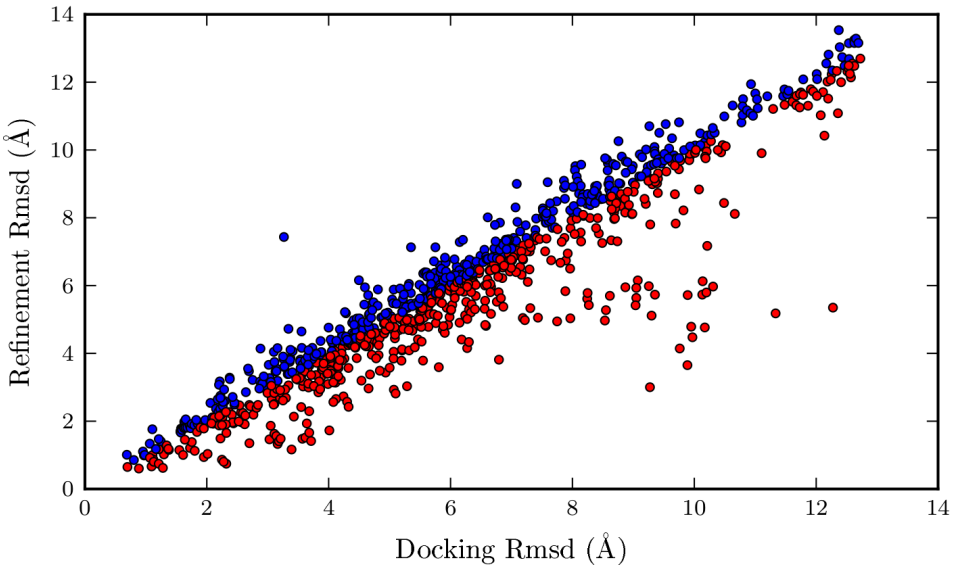


Figure 5.4: RMSD of all refined complexes from the native complex vs. the initial RMSD after the protein-protein docking. Red circles (522) mark models, where the RMSD decreased upon refinement and blue circles (459) identify models for which the RMSD increased after the refinement.

5.3.1 Molecular dynamics refinement of docked complexes

A standard MD-based refinement procedure was applied to all energy minimized complexes following closely a standard procedure that is used as the final refinement step in the ATTRACT docking protocol [263]. It consists of a short MD simulation including restraints to keep the backbone structure of each partner reasonably close to the start conformation but allowing full side-chain flexibility and some adjustment of relative partner translation and orientation followed by a final energy minimization (see Methods section for details). During the refinement, the same force field was used as in the energy minimization (EM) protocol. It is important to emphasize the MD-refinement is not intended to serve as a new search procedure to identify new relevant docking geometries (this is done by the initial systematic docking run, see above). The purpose of the short MD-refinement is to more effectively reach a nearby local minimum (compared to just energy minimization). It relates to procedures widely used in the protein-protein docking field to rapidly locally optimize thousands of docking solutions obtained from a systematic docking run using various docking methods (e.g. FFT-dock [168], ATTRACT [263], HADDOCK [73]).

On average the refinement procedure resulted in a better agreement of predicted complexes compared to the native complex for many cases (Figure 5.3). The lowest sampled RMSD decreased for 14 out of the 20 structures after the refinement. In particular, for cases with initial RMSD of less than 5 Å frequently further improvement was observed. The largest impact was found for 1j2j, where the RMSD improved by 1.7 Å. The calculated interaction energies are lower for all complexes compared to the start structures. The improvement ranged from 9 kcal/mol (1zhi) to 40 kcal/mol (1ay7) for the best-scored model. Also, the scoring of native-like complexes with unbound partner structures improved in most cases with a reduced deviation from the scoring of the bound complex indicating sterical adjustments at the partner interface. For 13 cases (1z0k, 1j2j, 2sni, 3sgq, 1zhi, 7cei, 1gcq, 1ay7, 1ppe, 1r0r, 2oob, 1syx and 1qa9) a favorable scoring of low RMSD solutions was observed. Moreover, compared to the EM of docked complexes for 9 cases an improved ranking of structures close to the native complex compared to high RMSD solutions were found (1ay7, 1ffw, 1qa9, 1ppe, 2oob, 2sni, 1zhi, 1r0r and 2i25). The scoring got worse compared to EM for eleven docked complexes. However, only in the case of 1ak4 the funnel plot changed from a higher specificity to a lower specificity after MD refinement, while the specificity improved for four structures (1ay7, 1ppe, 1zhi and 1r0r).

Overall, the refinement step enhanced the quality of the results. In many cases, the shape of the resulting funnel plots improved. Further refinement runs under different temperature conditions (100K, 200K, 300K) as well as longer simulation times were tested but did not improve the relative scoring (not shown). A final re-evaluation after the refinement results with a more sophisticated implicit solvent treatment using both, a R-6 integration scheme introduced by Aguilar et al. [4] or solving the linearized finite-difference Poisson-Boltzmann (FDPB) equation instead of a GB model also did not noticeable improved scoring specificity (Appendix Figures A.6,A.7).

5.3.2 Binding free energy calculation using umbrella sampling

The present force field scoring based on energy minimization or after MD-based refinement identified in several cases complex structures close to the native geometry as complexes with the lowest interaction energy. However, still, for several cases, non-native complex geometries achieved better scores than the native structures and even in successful cases the separation between the best scoring near-native solutions and several incorrect complexes was small relative to the range of scores.

To obtain a more realistic free energy based scoring, we sample the binding dissociation/association process along a center of mass (COM) distance coordinate using umbrella sampling (US). Such simulations are performed by applying a set of harmonic potentials that restrain the relative position of the partners to an

interval along a distance reaction coordinate in several simulations. The potential of mean force (PMF) is then calculated by evaluating the distance histograms of the simulations by employing the WHAM algorithm [116] (see Methods section for details). To achieve rapid convergence, both the conformational flexibility and the relative rotational degrees of freedom of the protein partners were restricted (three torsional angles between receptor and ligand: α, χ and γ and two axial angles: θ and ϕ, see Appendix Figure A.1). The use of a continuum solvent model combined with the conformational and orientational restraints resulted in reasonable convergence of the calculated PMFs within 1 ns sampling per US window (using 20-55 US windows, see Appendix Figures A.2,A.3,A.4).

The orientational and conformational restraining contributions can be calculated separately at the two end-points of the US simulations (bound state vs. bulk state) as described in the methods section (see also Appendix Section A.1). However, in a first attempt we compare the calculated free energies without accounting for the orientational and conformational restraining contributions (Figure 5.6), hence assuming that the calculated PMFs dominate the binding free energy.

The calculated PMF binding free energy ranges from -6 ± 1 kcal/mol (3a4s) to -23 ± 1 kcal/mol (2sni) for the best-bound model much more positive than the interaction energies obtained with the same force field before or after refinement (see above). Encouragingly, the range of calculated free energies is similar to experimental binding free energies of protein-protein complexes [150, 153] (see also next section). Using the native bound complexes as start structure resulted in the most favorable calculated PMF for 5 complexes and in 12 cases the bound structure ranks among the best-ranked complexes. Thus, only for 3 proteins (3a4s, 2oob and 1j2j) it failed to score the native binding mode among the top-ranked complexes. The difference in the calculated binding PMF of unbound to bound form was high (above 10 kcal/mol) for the complexes 1fle, 2sni and 2i25, indicating that differences in backbone and side-chain conformations still exists and shifts the results to lower calculated binding PMFs for the unbound complex. Nevertheless, in many cases, other models close to the native complex scored very favorably.
The selectivity of the funnel plot (for the definition of selectivity see the paragraph on selectivity) was favorable for 16 structures (1z0k, 7cei, 1ffw, 1syx, 1qa9, 1j2j, 3sgq, 1gcq, 1ppe, 2cfh, 1ak4, 2sni, 1zhi, 1r0r, 1z5y and 2i25), with a clear preference of the models having a low RMSD. In addition to that, an unfavorable selectivity was obtained for four models. Here, the highest scored pose at the binding site has a lower ranking than the highest scored pose not at the binding site (2oob, 1ay7, 3a4s and 1fle).

Interestingly, for two cases a large gap between the calculated binding free energy in case of starting from the bound complex vs. any other decoy including the unbound complex was found (1fle and 2i25). Inspection of these cases identified indeed an unfavorable unbound backbone conformation at the interface that interferes with

the correct interface structure (see Figure 5.5). For example, the protein 2i25 has a loop at the interface that directs towards the receptor in the native bound conformation but not in the unbound conformation. The switching of the loop during the simulations towards the bound form is not observed during the US simulations (presumably because of the RMSD constraints to limit the conformational flexibility during the US simulations).

Compared to the single-point scoring after refinement a clear improvement of the results could be achieved for 15 of the 20 structures (1z0k, 3a4s, 7cei, 1ffw, 1syx, 1qa9, 3sgq, 1gcq, 1ppe, 2cfh, 1ak4, 2sni, 1zhi, 1z5y and 2i25). Most importantly, in six cases a low selectivity during refinement changed to a high selectivity (1ffw, 2cfh, 1ak4, 1z5y and 2i25) after the restraint US simulations. The ranking of the models did not change considerably for five structures (3sgq, 7cei, 1fle, 1syx and 1qa9). Importantly, the employment of restraint umbrella sampling did only yield two funnel plots that were considered as less selective than after the simple refinement (1ay7 and 2oob).

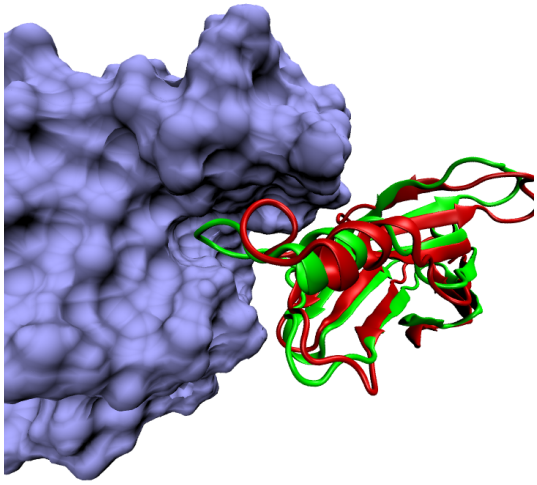


Figure 5.5: Superposition of the decoy complex (ligand protein in red cartoon, receptor as blue surface representation) of the pdb2i25 case with the smallest deviation from the native complex (ligand protein as green cartoon). The loop of the native bound ligand conformation at the interface (green cartoon) fits well into a receptor pocket, whereas the loop of the decoy complex (partners in unbound conformation) deviates significantly from the placement in the native interface.

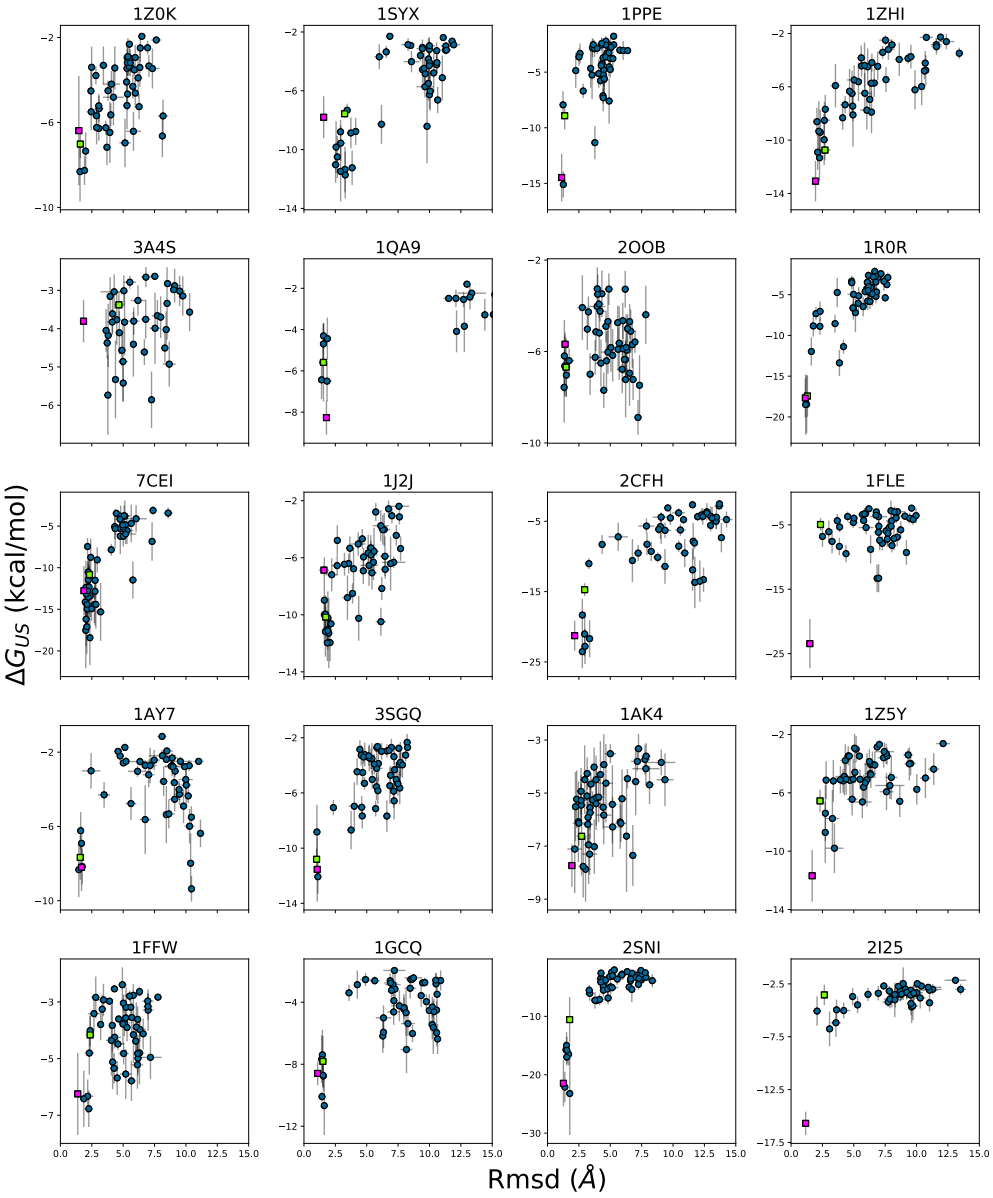


Figure 5.6: Scoring of decoy complexes based on the potential of mean force free energy obtained from restrained umbrella sampling simulations vs. deviation from the native complex geometry. The green squares display the results for starting from the native complex but using unbound partner structures whereas the magenta squares depict results obtained starting from the native bound complex.

5.3.3 Absolute binding free energy calculation

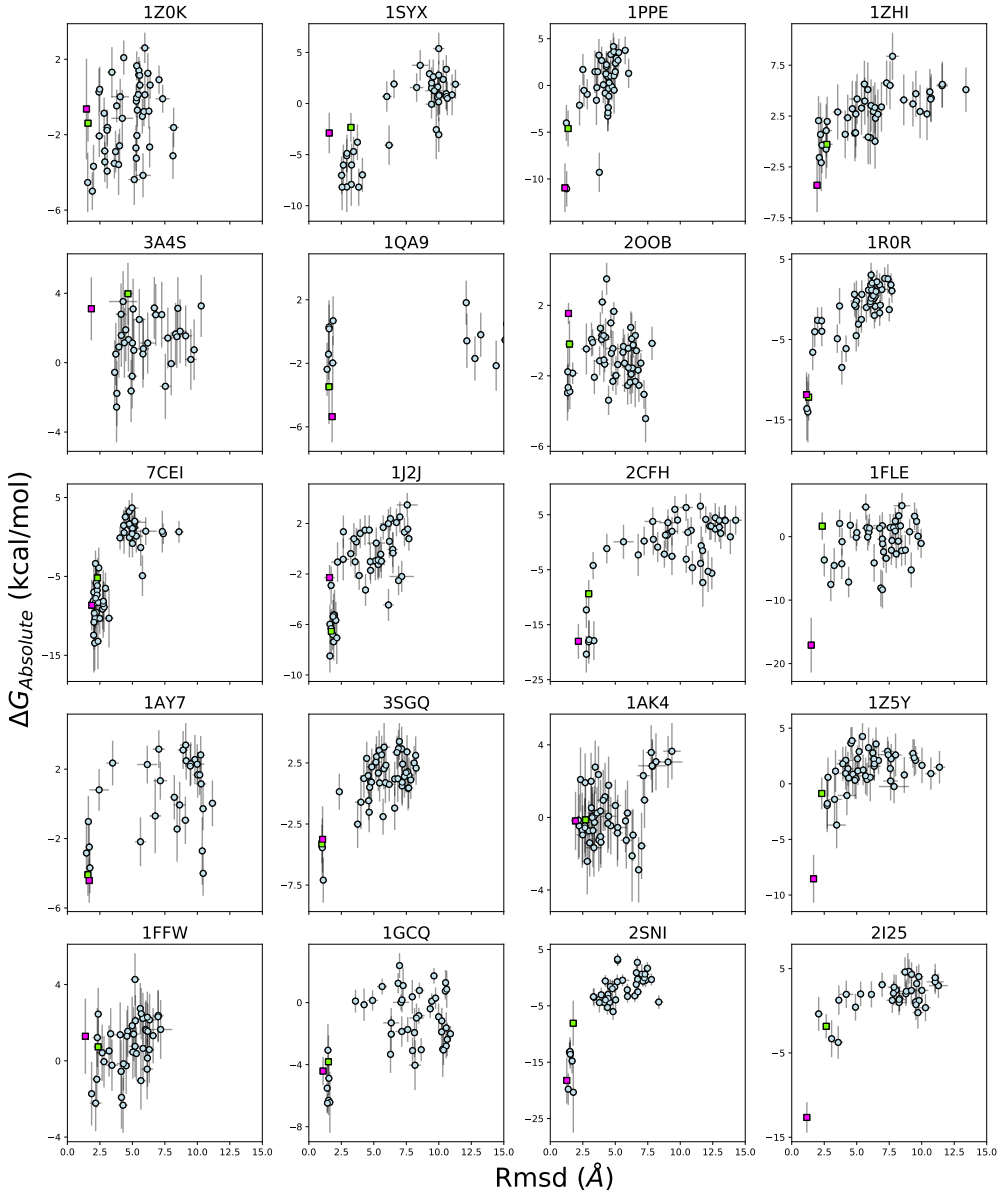


Figure 5.7: Same as Figure 5 but plotting the calculated absolute binding free energies vs. deviation from the native complex. The green squares display the results for starting from the native complex but using unbound partner structures whereas the magenta squares depict results obtained starting from the native bound complex.

In the previous section, the ranking of decoy complexes was based on free energies calculated including restraints on orientation and conformation of the binding partners (PMF obtained along the reaction coordinate). It is possible to account for the axial, orientational (orient) and conformational (conf) contributions following an approach introduced by Woo and Roux [324] that yields the absolute binding free energy (details are given in the Methods section and Appendix Section A.1):

$$\Delta G_{bind} = -kT \ln \left[C^{\circ} e^{-\beta[\Delta G_{bind}^{restr,axial} + (\Delta G_{orient}^{bulk} + \Delta G_{conf}^{bulk}) - (\Delta G_{orient}^{site} + \Delta G_{axial}^{site} + \Delta G_{conf}^{site})]} \right]$$

$$\Delta G_{Absolute} = \Delta G_{bind} + \Delta G_{SASA}$$

In the case of the bound state, the calculation of the conformational and axial and orientational site (bound) terms require additional simulations of the complex (starting from the window of the minimum PMF) without applying restraints. Based on free energy perturbation (FEP) it is possible to evaluate the corresponding free energy contributions (see Methods and Appendix section A.1).

The conformational restraining contribution can also be estimated through an additional simulation of the individual ligand and receptor proteins without restraining potentials and using FEP to estimate the associated free energy contribution. The orientational- and axial bulk (unbound) contributions can be calculated without additional simulations (see Appendix Section A.1 for a more detailed description). Finally, since we are now interested in absolute binding free energies the nonpolar solvation contributions were accounted for by adding a cavity solvent tension term, calculated from the buried SASA, to the binding free energy.

In Figure 5.7, the funnel plots (absolute binding free energy vs. RMSD from the native complex) are shown for the 20 proteins. Accounting for the additional contributions resulted in less favorable binding free energies for all complexes (mean difference 5 kcal/mol), so that for most of the decoys (55%) the associated binding free energy is positive predicting unfavorable binding. Hence, in contrast to the simple scoring schemes described above the absolute binding free energy calculations predict that the majority of decoy complexes do not form at all. However, the binding free energy was still attractive for the best-scored model of all structures and for all cases when using the bound partner conformations in the starting complex except for three cases: 1ffw, 2oob, and 3a4s (Figure 5.7).

Apart from the PMF contribution, the term that varied the most between the different systems was the conformational bulk contribution (the orientational, and axial contributions did not vary strongly between the systems due to the relatively small force constants). The shift towards more unfavorable binding affinity was generally due to a higher bulk contribution of axial, orientational and conformational restraints in comparison to the same terms for the associated state (see Figure 5.8). It reflects a free energy cost related to a loss of freedom due to a restricted

axial, orientational and conformational mobility in the associated state. For loosely bound complexes the free energy cost due to restriction of axial, orientational and conformational mobility is expected to be smaller than for binding to the high-affinity sites, and indeed considering absolute binding free energy gives a ranking that favors the near-native complexes less than the binding free energy based only on the PMF calculation (previous paragraph). Thus, the preference improved compared to the US evaluation for eight structures (3a4s, 7cei, 1ay7, 1j2j, 3sgq, 2cfh, 1r0r and 1fle) and deteriorated for twelve structures (1z0k, 1ffw, 1syx, 1qa9, 1gcq, 1ppe, 2oob, 1ak4, 2sni, 1zhi, 1z5y and 2i25). Overall, the selectivity changed from a high preference to a low preference for three structures (1ffw, 1ak4 and 2i25) and vice versa for one structure (3a4s). Thus, also for the proteins where a change in the specificity was observed, this was mainly due to an alteration in the relative scoring of only a few models.

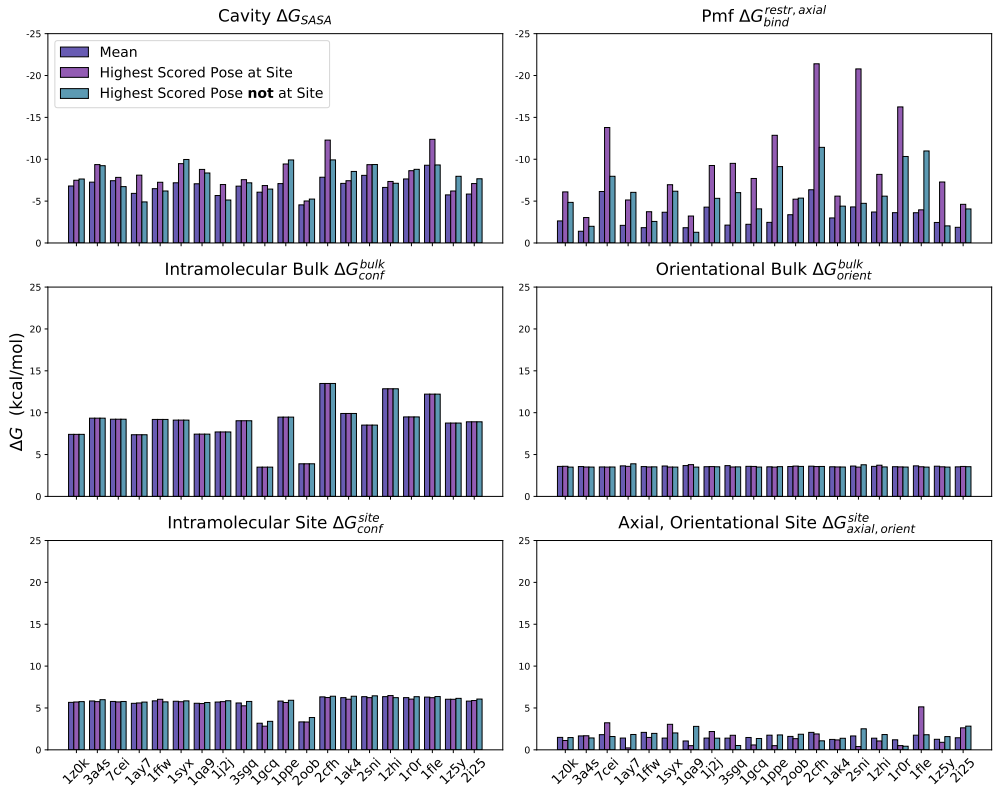


Figure 5.8: Contributions of individual free energy terms (from the absolute binding free energy calculations) to the mean scoring, scoring of the near-native or scoring of the most favorable non-native decoy.

5.3.4 Evaluation of the binding selectivity of the scoring methods

To be able to compare the different scoring methods, a measurement for selectivity has to be defined, reflecting how well the poses at the binding site score compared to all other decoys. The overall shape of the funnel plots of each structure is pretty similar (besides a shift in the scoring axis) for the different scoring methods, it's the scoring of just a few poses per structure that leads to high differences in the selectivity. Thus, it is possible to scale the different scorings to relative values and consequently be able to compare the selectivity by identifying the key poses.

First, comparable scores were obtained by shifting each score S_i by the mean value of all scores $\overline{S}$ and dividing it by the minimum scoring value S_{min}:

$$S_i' = \frac{S_i - \overline{S}}{S_{min}} \tag{5.4}$$

Second, a decoy was defined as being at the binding site if its RMSD was smaller than $R_{min} + 1.5$Å, with R_{min} being the smallest RMSD value of all poses. The pose with the highest score (minimum binding energy), while being at the binding site is called the true (T) result. The highest-ranked pose that is not at the binding site is called the false (F) result. The selectivity is measured by taking the difference of these key poses

$$Selectivity = S_T' - S_F'. \tag{5.5}$$

The selectivity lies between 1.0 and -1.0 (as $S_T,\ S_F < \overline{S}$ for all structures and scoring methods). A value of 1.0 indicates a perfectly selective scoring, the best-scored pose at the binding site is separated clearly from the other poses. On the other hand, a value of -1.0 indicates that there is a high selectivity for the poses not being at the binding site. Finally, a selectivity of 0 means that the scoring of false and true decoys was the same. Note that due to the definition of the origin of the scaled scores as the mean scoring value (in contrast to eg. the mean scoring value minus one standard deviation), the selectivity values are only in a few cases higher than 0.5 although the selectivity is good regarding the funnel plots (eg. 1z0k, 1ppe, 1ffw with US). Thus, the defined selectivity should be considered as a relative value to compare the different scoring approaches, not as an absolute value for the selectivity of individual funnel plots (as this value highly depends on the defined origin of the scaled scores).

In Figure 5.9 the calculated selectivity is shown for all structures and all scoring approaches. The scoring with EM had the highest selectivity for two structures (7cei and 1ak4), single-point scoring after MD refinement had the highest selectivity for three structures (1ay7, 2oob and 1r0r), the US approach was the most selective for ten structures (1z0k, 1ffw, 1syx, 1qa9, 1gcq, 1ppe, 2sni, 1zhi, 1z5y and 2i25) and the absolute binding free energy calculation had the highest selectivity for five structures

(3a4s, 1j2j, 3sgq, 2cfh, 1fle). Considering the mean selectivity for all structures US performed best with a value of 0.33, the second highest was the absolute binding free energy calculation (selectivity = 0.28). Less selective were the single-point scoring after refinement (0.14) and single-point scoring of the docked and minimized poses (0.06). This reflects a clear improvement in the overall selectivity using the more sophisticated scoring methods in contrast to single-point scoring.

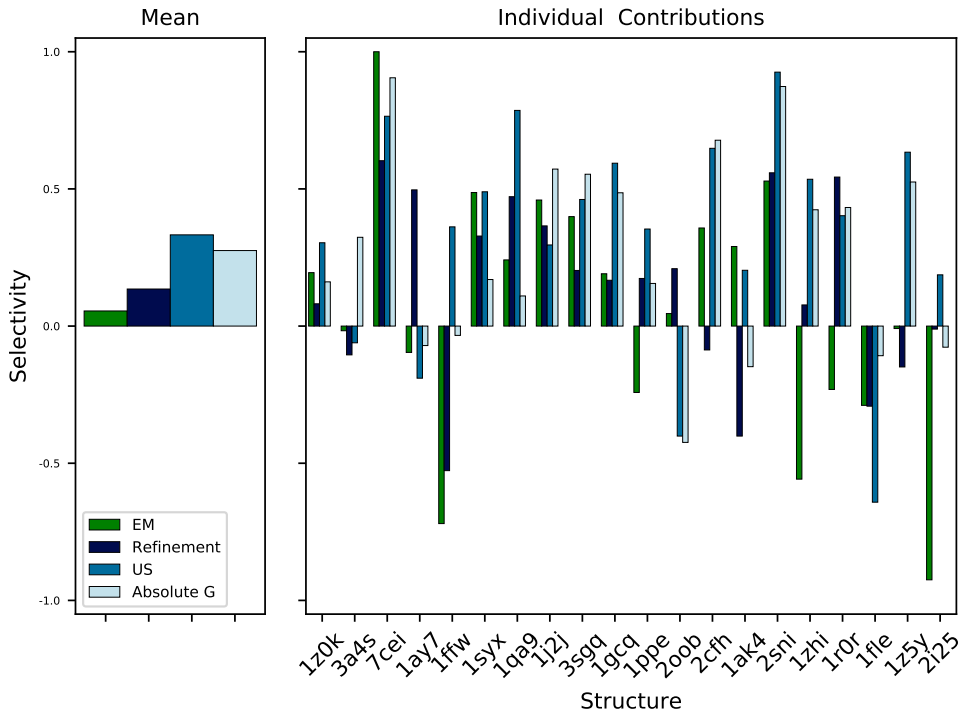


Figure 5.9: Mean selectivity of all structures and the selectivity of each structure for the different scoring methods. The selectivity was calculated as described in the main text (see equation 5.5), considering the highest-ranked poses at the binding site and not at the binding site, respectively.

5.3.5 Comparison of the scores with experimental binding free energies

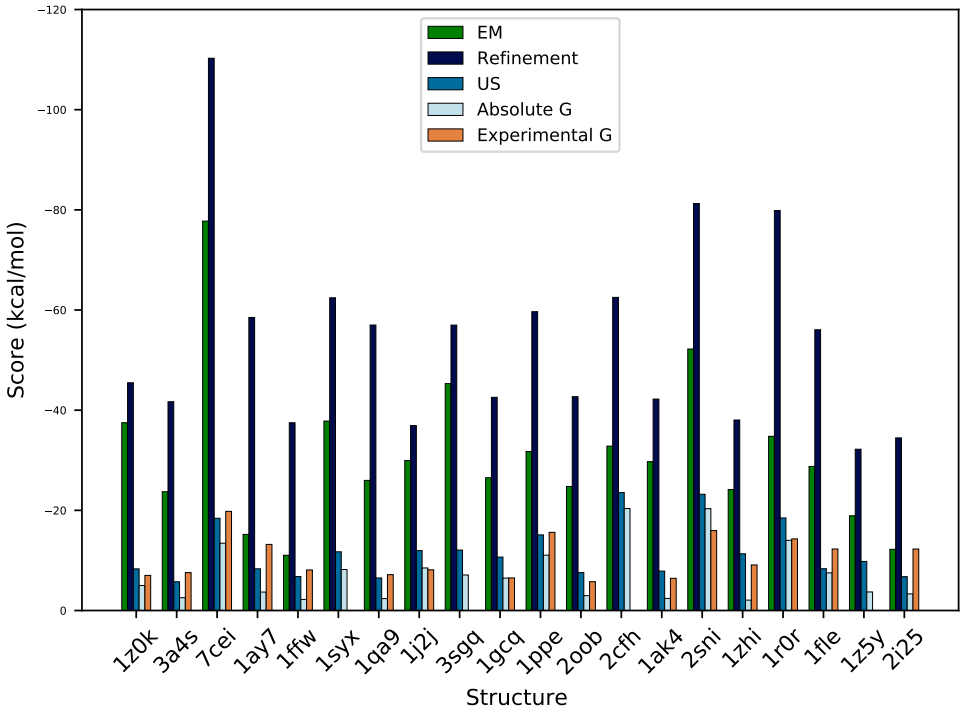


Figure 5.10: Binding affinity of the best-scored pose at the binding site (defined as in the selectivity paragraph) for the different scoring methods. The values for all structures are shown and can be compared to the experimental binding affinities (orange bars) which were available for 16 structures.

For 16 evaluated structures, experimental binding free energies are available (Appendix Table A.1). These values have to be regarded with caution, as not all results were obtained by the same experimental approach. It has been found that experimental binding affinities of protein-protein complexes can depend significantly on the experimental method [150, 153]. For the comparison with experiment, the best scoring complex among the ensemble of native-like complexes (see above) was considered. Interestingly, the restraint US approach (just considering the PMF along the dissociation/association coordinate) showed the best agreement with the experimental data, followed by the absolute binding free energy calculation. Both methods performed remarkably better than simple single-point scoring, due to a shift towards higher absolute energy values (see Figure 5.10). Hence, for 50 % of the structures the binding free energies calculated from the PMF only were within a range of

$\pm$2 kcal/mol of the experimental binding affinities (7cei, 1ppe, 1z0k, 1qa9, 1ak4, 2oob, 1ffw, 3a4s) and a small mean deviation of $\Delta\Delta G_{Mean} = \overline{\Delta G_{exp} - \Delta G_{calc}} = 0.4$ kcal/mol was observed. The highest scored pose at the binding site was considered for each structure, as described in the last paragraph. Accounting for the restraining contributions 25 % of the structures had a scoring in very good agreement ($\pm$2 kcal/mol) with the experimental values (1r0r, 1j2j, 1z0k and 1gcq) with a mean deviation of -3.8 kcal/mol. In case of the single-point scoring, after the EM evaluation two structures (13 %) scored close to the experimental values (2i25 and 1ay7). Note, that these two structures had a negative selectivity (see Figure 5.9), meaning that the pose at the binding site that was considered for the evaluation was not the overall best-scored pose. With MD refinement no structure had a pose at the binding site with a scoring value in agreement to the experimental binding affinity. In both cases, the shift in the scoring towards too high energy values is pointed out by a mean deviation of 19.8 kcal/mol and 47.1 kcal/mol to the experimental binding free energies.

In case of just limiting the analysis to the scoring of the native (bound) poses a slightly lower agreement to the experiment was obtained: 39 % of the structures were in close agreement for the US evaluation (1ppe, 1j2j, 1z0k, 1qa9, 1ak4, 2oob and 1ffw) and 13 % for the absolute binding free energy calculation (1qa9 and 2i25). For the single-point interaction energy evaluations the native models were never in close agreement with the experimental binding free energies.

We also calculated the Pearson correlation coefficient (PCC) between the highest scored binding affinity at the binding site and the experimental data. Note that a coefficient of 1 represents a perfectly correlated system and 0 a completely uncorrelated state. This yielded significant correlations of 0.73 and 0.74, for the PMF-based approach and the absolute binding free energy calculation, respectively. Thus, we find no difference between these methods if the PCC is considered which is in accordance with the finding of the previous section that calculating the absolute binding free energy has only a minor influence on the selectivity. The highest PCC of 0.85 was found for the MD refinement, the scoring with EM had the lowest value of 0.60.

The results have to be regarded with caution, as we are dealing with a limited data size that incorporates much more low energy binder than high energy binder, the latter consequently have a much higher impact on the correlation coefficient. Moreover, calculating the correlation coefficient omits the fact that the US calculations have to be much more precise than the single-point scoring values to achieve the same correlation coefficient, due to the much higher absolute scorings of the single-point evaluation.

5.4 Conclusions

A long-term goal of our efforts is to design a scoring approach that is reliable enough to realistically predict the binding free energy of complexes and is so accurate that new interactions between proteins (not those already discovered experimentally) can be reliably predicted. In the present study, the possibility of evaluating docked protein-protein complexes using US-based free energy simulations was explored. The results were compared to scoring using the same force field approach but evaluating the interaction energy of energy-minimized or MD-refined complexes. In the present study we used the ATTRACT docking program [329, 69] for generating putative docked complexes (typically generating hundred of thousands of solutions within hours of computer time). However, this step can also be replaced by other search techniques to obtain putative protein-protein binding geometries. In terms of selectivity, the free energy simulation approach resulted in improved scoring compared to the simpler approaches based on the interaction energy of single complexes. Encouragingly, the calculated binding free energies are in quite reasonable agreement with experimental data. The remaining differences might be due to inaccuracies of the force field and implicit solvent model but can also be in part due to the experimental methods used for affinity measurements [150].

In most protein-protein docking tasks one ends up with a limited number of possible complex geometries (e.g. often experimental data may limit the regions of interaction on partner proteins). In such cases, the present binding free energy simulation approach might be very useful to further limit the number of putative complexes or to even reliably identify the most realistic complex structure.

Interestingly, scoring based just on the restraint US approach resulted in the most precise results with a high selectivity found for 16 complexes and only in four cases failed to predict the native binding site. For two proteins a substantial difference in scoring of the bound and the unbound native conformation was observed. In these cases, conformational changes at the binding site, leading to a high affinity of the complex, were not captured (due to conformational restraints during the simulations). Nevertheless, for all other complexes, the near-native models showed a high ranking. Interestingly, the selectivity did not change significantly upon calculation of the absolute binding free energy indicating that the individual restraining contributions play a smaller role (and may compensate each other in part) compared to the effective receptor-ligand interactions captured through the US calculations along the distance coordinate.

Furthermore, we found a significant correlation to experimental results of 0.73 and 0.74 for the restraint US free energy values and the absolute binding free energy, respectively. The MD refinement gave an even higher correlation of 0.85, due to the overall higher scorings the PCC is less prone to deviations of some kcal/mol in the interaction energy. In particular, the calculated PCCs were higher than the

correlation of all scoring methods evaluated by Kastritis and Bonvin [150] yielding a maximum correlation coefficient of -0.32 but on a larger benchmark set of 81 complexes.

A better correlation was achieved recently by Chen and coworkers that investigated the performance of MM/PBSA and MM/GBSA methods on 46 proteins of the same protein-protein benchmark as in our study [45]. They achieved a maximum correlation of 0.647 using MM/GBSA, a low interior dielectric ($\epsilon = 1$) and the force field ff02. Interestingly, testing the same force field that was used in our study, ff14SB, a worse correlation of 0.578 was achieved.

The inclusion of explicit solvent represents an obvious possible improvement of the present approach. However, although in principle more accurate than an implicit solvent model, it is significantly more demanding especially for small systems and requires longer simulation times for convergence. Nevertheless, in cases where the number of decoy complexes is even smaller than the 50 complexes considered in the present study inclusion of explicit solvent may represent a realistic option. Another limitation of the present approach are the conformational restraints employed to achieve rapid convergence. Although allowing also limited deformations of the backbone such restraints may prevent a transition to a near-native bound partner structure especially if the conformational difference between unbound and bound conformations is significant. In future studies, we plan to test if optimization of the conformational restraints employed during the US simulations can further improve the results. Secondly, the employment of advanced sampling methods like replica exchange umbrella sampling (REUS or HREUS) along the distance coordinate [189, 62] can also further improve the results.

6 Prediction of Protein-Protein Complexes Using Replica Exchange With Repulsive Scaling

Molecular dynamics (MD) simulations allow one to follow the association process of protein-protein complex structures under realistic conditions (see Chapter 4), including full partner flexibility and surrounding solvent. However, due to the many local binding energy minima at the surface of protein partners, MD simulations are frequently accumulated for long simulations in transient trapped states. A H-REMD (see Section 3.4.3) based scheme is designed in the following chapter, employing different levels of a repulsive biasing between partners in each replica simulation. The bias acts only on intermolecular interactions due to an increase in effective pairwise van der Waals radii (repulsive scaling (RS)-REMD) without affecting interactions within each protein or with the solvent. [1]

6.1 Introduction

Biomolecular binding and in particular protein-protein binding processes to form functional complexes are key elements of almost all biological processes. Knowledge of the three-dimensional (3D) structure of protein-protein complexes is a prerequisite for understanding its function. Experimental structure determination as well as prediction of protein-protein complex structures are also of significant interest for the rational design of drug molecules to influence biological processes. Computationally efficient docking algorithms are frequently applied to identify putative protein-protein binding geometries based on surface complementarity or simple pairwise interaction potentials [110, 255, 326, 115]. Molecular docking, however, often largely neglects or only approximately accounts for the flexibility of the binding partners and interactions with the solvent [296, 326, 312, 72]. It is possible to include a moderate degree of flexibility using for example deformations in soft normal modes at reasonable computational costs [209, 210, 205, 306, 219]. In some approaches a refinement stage with side-chain flexibility is performed, mainly focusing on interfacial rearrangements [83, 67]. In addition, the evaluation of identified

[1] The contents of this chapter have been published in a similar form in [276].

binding geometries is largely based on empirical scoring functions applied to single complex conformations neglecting conformational and orientational entropic contributions to binding. Ideally, molecular association should be simulated including full flexibility of both partners and accounting for surrounding water molecules and ions [277]. Molecular dynamics (MD) simulations are in principle well suited for investigating biomolecular association processes including full atomic flexibility. The methodology has already been used to refine potential binding geometries identified in the docking efforts [253, 293, 270]. In selected cases, it is even possible to skip any initial docking but to use ultra-long atomistic MD simulations and directly mimic the physical binding process [37, 266]. This is, however, computationally very demanding and only possible up to timescales on the order of microseconds to milliseconds for individual examples with current computational resources. The search for putative binding regions on the surface of proteins is associated with a rough energy landscape. Hence, the binding partners often get kinetically trapped in local energy minimum for long time intervals resulting in a waste of computational resources. Several efforts have been undertaken to accelerate the search for binding sites. It is possible to employ temperature replica exchange molecular dynamics (TREMD) with multiple parallel MD simulations and periodic exchanges. It can improve the sampling by exploring the surface of the receptor more rapidly at higher temperatures and extracting relevant states at lower temperatures. However, TREMD does not scale well with the system size and another method, Hamiltonian REMD (H-REMD) might be more suitable [95] because one can specifically scale force field parameters affecting receptor-ligand interactions. One possibility is to linearly scale the Lennard-Jones and electrostatic potential across replicas [314] or reduce the ruggedness of the energy landscape by introducing soft core potentials [192]. The latter method has shown promising results to refine complex geometries close to the native binding mode but do not effectively reduce the problem of trapped binding sites on the receptor surface [192]. Transient binding states in agreement with experiment could be recognized using replica exchange Monte Carlo simulations for three protein-protein complexes using a coarse-grained representation of the molecules [158].

It is also possible to use meta-dynamics methods to reconstruct the free energy surface of association and dissociation of protein-ligand systems by gradually adding biasing potentials that destabilize already sampled protein surface regions [101]. In the latter study the choice of only two collective variables (CVs) was enough to identify the binding site of four protein-ligand systems. In general, a higher number of CVs is necessary to completely describe the relative ligand-receptor position and orientation [30]. A larger set of CVs can be chosen using reconnaissance meta-dynamics that incorporates a self-learning algorithm that gradually pushes linear combinations of the CVs [284]. For a protein-ligand system, this method was able to identify multiple binding sites of the protein.

In a recent study by Pan et al. reversible association and dissociation of five protein-protein complexes was observed using tempered binding MD simulations [240]. However, the binding and unbinding events were captured in still expensive computer simulations (simulation times of several hundred microseconds) on the special purpose machine Anton [267]. In tempered binding, the interaction strength between the two solutes (but not within each protein partner) is used as tempering coordinate (instead of the total temperature as used in standard simulated tempering). As another alternative, it is possible to add an explicit repulsive biasing potential between partners in a series of replicas (BP-REMD) that keeps the ligand and receptor at various distance intervals apart in higher replicas [238]. The higher replicas allow to keep some space between partner molecules and therefore result in fast diffusion. Upon exchange with lower replicas, favorable binding sites can be rapidly sampled also in the reference replica. This method showed a promising performance to specifically accelerate the search process for identifying ligand binding sites on protein surfaces under realistic conditions [238]. However, the method requires calculating an ambiguity distance between all pairs of surface atoms of both partners which is computationally demanding and not well suited to run in parallel on many cores such as graphical processing units (GPU)s.

In the present study, the possibility of increasing the repulsion between ligand and receptor by specifically increasing the pairwise effective van der Waals (vdW) radii and reducing the vdW attraction along the replicas in an H-REMD simulation is explored. It weakens not only the Lennard-Jones contribution to binding interactions but also reduces the number of hydrogen bonds and electrostatic interactions due to an increased average distance between ligand and receptor. Hence, the biasing potential in the replicas allows the partners to rapidly dissociate from possible suboptimal binding sites to effectively search the protein surface. The method is promising for its simplicity of implementation and only requires adjusting parameters and can therefore be used with existing simulation software that runs on GPUs. The approach was tested on several protein-protein complexes of different sizes and types. In contrast to regular MD simulations, it allowed the identification of near-native complexes even when starting far from the native binding region. In addition, we tested the approach for the refinement of complexes starting from geometries in the vicinity of the native binding arrangement. In this case, also a slightly better performance than regular MD simulations at the same computational effort was achieved.

6.2 Materials and methods

For all atomistic simulations the Amber16 or Amber18 software packages [42, 41], were used employing the *pmemd.cuda* module for efficient calculations on Graphical

Processing Units (GPU)s. The ff14SB [198] force field was used together with an implicit water representation using the OBC generalized Born (GB) model [237] (igb=8 option in amber) involving an infinite cutoff.

6.2.1 Simulations of protein-protein complexes starting far from the binding geometry

MD simulations on protein-protein complexes in order to identify the native binding arrangement were started from an initial placement of one partner (termed ligand) at the opposite side of the second (receptor) protein with respect to the native binding site. Six complexes were considered for these simulations (pdb-id of complexes: 2oo9, 2cfh, 7cei, 2sni, 1gcq, and 1syx, see also Supporting Information Table S1). In all cases, the unbound protein structures were used for the simulations. The 6 complexes were selected due to the relatively small size from the docking benchmark 3.0 [132]. We distinguish between receptor and ligand-protein according to the assignment in the benchmark 3.0 [132] (typically the large protein partner is the receptor and the smaller partner is the ligand). The root mean square deviation ($rmsd_{ligand}$: Rmsd of the ligand after best superposition of the receptor with respect to the native complex structure) of the initial relative placement of the partner proteins from the native complex geometry was between 30 Å and 61 Å.

MD simulations were performed using the OBC (Onufriev, Bashford, Case) generalized Born (GB) implicit solvent model [237] (igb=8 option) and using an infinite cutoff radius for both the GB radii and nonbonded interactions. A Langevin thermostat with a collision frequency of $\gamma = 5\,ps^{-1}$ was used to control the temperature. The collision frequency is reduced relative to a more physical value of $50\,ps-1$ to reduce the apparent viscosity of the solvent and speed up sampling [10]. Equilibration of the start geometry was achieved after energy minimization (50 steps steepest descent followed by 1500 steps conjugate gradient) and heating in three steps (each 12 ps) to 300K with positional restraints of 0.05 kcal $mol^{-1}Å^{-2}$ applied on the heavy atoms relative to the starting structure. Since we observed in long MD simulations for some proteins a partial unfolding, positional restraints on the receptor C_α atoms (force constant 0.05 kcal $mol^{-1}Å^{-2}$) were also included during production simulations. Note, that such weak restraints allow still considerable backbone fluctuations and full side-chain flexibility but prevent unfolding or large domain motions in the proteins. To prevent the ligand from diffusing too far away from the receptor, restraints between the center of masses (COM) of the C_α atoms of the proteins were employed. The restraining energy was zero for COM distances below a certain threshold and increased quadratically beyond the threshold (force constant 1.0 kcal $mol^{-1}Å^{-2}$) so that it prevents large receptor-ligand separation but still allows the ligand to dissociate from the receptors up to a certain distance. The

COM distance threshold ranged from 27 Å to 50 Å for the different protein-protein complexes and was slightly larger than the sum of the largest center to surface distances for the two partner proteins (see Table B.1). The mean difference of the applied COM distance restraints and the native COM distances was 10.8 Å. For avoiding unfolding of the ligand-protein additional intra-molecular pairwise distance restraints between the C_α atoms of the ligand-protein (only distances between 5-10 Å) were applied, that prevented the ligand backbone from unfolding (force constant 2.0 kcal $mol^{-1}Å^{-2}$) but allow full side-chain flexibility.

In order to perform Hamiltonian replica exchange simulations (H-REMD), 16 replicas for each protein were generated with different Lennard-Jones (LJ) parameters for atom pairs involving atoms from different protein molecules (all intra-molecular nonbonded parameters within were preserved). The intermolecular LJ potentials were scaled by a parameter d that adjusts the effective van der Waals radius and a factor e that changes the potential well depths (see next section for a detailed description). The following parameter set for d and e, with a smaller step size between the parameters close to the reference replica that increases in the higher replicas gave the best results for protein-protein test simulations (see Table 6.2). For each replica, a short equilibration was performed for 32 ps with no exchange attempts. In the production run every 1000 MD steps an exchange between neighboring replicas was attempted, yielding a total simulation time (per replica) ranging from 340 ns to 845 ns (see Table B.1).

Finally, starting from the same equilibration runs, 16 regular MD simulations with no H-REMD but different initial velocities (using the same restraints) were performed for comparable timescales as the H-REMD simulations (see Table B.1).

6.2.2 Refinement of individual protein-protein docking poses in implicit solvent

In addition to simulations starting far away from the native binding geometry, H-REMD, and regular MD simulations were also performed for arrangements in the vicinity of the native complex structure obtained by an initial protein-protein docking run using the program ATTRACT [329, 69]. The same set of structures and docking procedure as used in a previous study [279] were employed (see Supporting Information Table S2). Since the H-REMD method for refinement of docked complexes is computationally demanding the number of test complexes was limited to 20 complexes from the docking benchmark 3.0 [132]. The docking was performed using a standard docking protocol on the unbound partner structures with the program ATTRACT [329, 69]. The 300 top-ranked complexes were considered. Out of this set the 50 models with the lowest RMSD to the native complex structure were used for further refinement using the RS-REMD or regular MD simulations. In order to refine the docking solutions atomistic replica exchange simulations in

Table 6.1: Simulation setups for each complex indicated by the PDB-id for the repulsive scaling H-REMD (RS-REMD) approach and the regular MD simulations.

PDB	Simulation time		COM d[a]
	RS-REMD ns/replica	regular MD ns/simulation	Å
7cei	772	400	40
2oo9	730	684	27
2cfh	845	899	50
1syx	438	380	35
2sni	340	308	35
1gcq	640	640	30

[a] Distance was chosen slightly larger than the sum of the largest center to surface distances for the two partner proteins.

implicit solvent were performed (OBC model [237], using the same conditions as described above) starting from the 50 docking poses of 20 protein-protein complexes. Energy minimization consisted of 2500 minimization steps (400 steps of steepest descent, 2100 steps of conjugate gradient). The systems were heated gradually in three steps of 15 ps to 300 K using a Langevin thermostat for temperature scaling. For each equilibrated pose 8 replicas were generated with increasing bias for higher replica numbers of the intermolecular LJ parameters. As described above for the simulations starting from the opposite side of the receptor protein a parameter d adjusting the effective van der Waals radius and a factor e that changes the potential well depths were varied between the 8 replicas (see Table 6.2).
Each replica was simulated for 0.5 ns with an exchange attempt every 125 steps amounting to 4 ns simulation time per pose. Intra-molecular pairwise distance restraints between the C_α atoms of each individual protein were applied (force constant 0.5 kcal $mol^{-1}Å^{-2}$) together with a COM distance restraint of interfacial C_α atoms between the ligand and receptor (atoms with distances between 10 and 15 Å were considered) with a half parabolic shape (force constant 1.5 kcal $mol^{-1}Å^{-2}$) that prevents full dissociation in the high replicas and shrinks the possible sampling space for these short simulations. The same simulation conditions and restraints were applied for regular MD simulations of each pose (no replica exchange and bias involved) of the same simulation time (4 ns) following a standard refinement protocol developed previously [263, 279]. For evaluating the interaction energy a short MD simulation (30 ps) was applied on the reference replica of the REMD simulations followed by a minimization (500 steps of steepest descent, 2000 steps of

Table 6.2: Lennard-Jones scaling parameters for the different RS-REMD simulation setups. For the repulsive scaling simulations starting far from the binding geometry and for the refinement of a docking ensemble the 16 replica scheme was used (column 2 and 3). In the refinement simulations of individual docking poses the 8 replica setup was used (column 4 and 5).

Replica Number	16 replicas		8 replicas	
	d(Å)	e	d(Å)	e
1	0.0	1.0	0.0	1.0
2	0.01	0.99	0.015	0.99
3	0.02	0.98	0.03	0.985
4	0.04	0.97	0.045	0.98
5	0.08	0.96	0.06	0.97
6	0.12	0.94	0.075	0.96
7	0.16	0.92	0.09	0.95
8	0.2	0.9	0.12	0.935
9	0.24	0.88		
10	0.28	0.86		
11	0.32	0.84		
12	0.38	0.82		
13	0.44	0.8		
14	0.5	0.78		
15	0.58	0.76		
16	0.68	0.74		

conjugate gradient), which was also applied to evaluate the final structures from regular MD simulations. Finally, the minimized structures were scored by subtracting the potential energy of the partners from the energy of the complex [279]. To access the deviation of the refined structures from the native binding site the $\text{rmsd}_{\text{ligand}}$ was calculated, the root mean square deviation of the ligand to the native ligand after superpositioning the receptor on the native receptor (only heavy atoms were considered).

6.2.3 Refinement of a protein-protein docking ensemble in implicit solvent

Multiple docking poses were considered in a single REMD run to perform refinement simulations for each of the 20 protein-protein complexes. Only docking poses with a $\text{rmsd}_{\text{ligand}}$ above 10 Å (for the complex 7cei 8 Å was chosen, due to a lack of poses

with large RMSD) were considered as starting structures for the subsequent RS-REMD refinement. Each of these poses was first scored based on the potential energy difference of the complex to the individual partners. The 16 highest ranked poses were considered for the subsequent REMD simulations. The poses formed the start structures in the 16 replicas and were distributed based on the ranking (best-ranked pose in replica 1, second-best in replica 2, etc.). Thus, the (initially) best-ranked pose started in the reference replica. Each replica was simulated for 30 ns amounting to a total simulation time of 480 ns per RS-REMD run with an exchange attempt every 250 steps between neighboring replicas. Finally, for comparison, regular MD simulations (no biases and no replica exchange involved) were performed starting from the same poses and simulating the same time as in the RS-REMD case.

6.3 Lennard-Jones parameter scaling between partner molecules

The Lennard-Jones interaction consists of an attractive part proportional to $1/r^6$ and a repulsive contribution typically modeled by a term proportional to $1/r^{12}$. The parameters ϵ_{ij} and R_{ij} in the Lennard-Jones potential determine the magnitude of attractive interaction and the effective (pairwise) van der Waals radius of the interaction between a pair of atoms of type i and j.
Typically, only the parameters between atoms of the same type, ϵ_{ii} and R_{ii} are used and one obtains parameters for pairs of different atom types using the Lorentz-Berthelot rules [188]:

$$R_{ij} = \frac{R_{ii} + R_{jj}}{2} \tag{6.1}$$

$$\epsilon_{ij} = \sqrt{\epsilon_{ii}\epsilon_{jj}}\,. \tag{6.2}$$

By defining new atom types it is possible to specifically modify the Lennard-Jones potential for interactions between the ligand and receptor without affecting the Lennard-Jones interaction within one partner molecule or with the solvent. We used this possibility by adding an adjustable parameter d to an effective pair-wise van der Waals interactions between pairs of ligand and receptor atoms,

$$R'_{ij} = R_{ij} + d\,. \tag{6.3}$$

One might also scale R_{ij} by multiplying with a factor, but this would increase the effective radius of pairs of atoms by different amounts and may strongly distort an

interaction interface. A change in the potential well depth ϵ_{ii} is described by a factor e:

$$\epsilon'_{ij} = e \cdot \epsilon_{ij} \,. \tag{6.4}$$

As this factor enters multiplicative instead of additive the same relative scaling of attractive interactions of different pairs of atom types is possible. One subtle problem with increasing d, however, is that the number of atoms that can interact increases (illustrated in Figure 6.1).

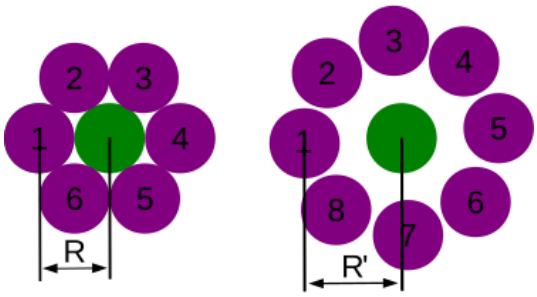


Figure 6.1: Effect of increasing the van der Waals radius on the number of possible interactions.

This increases the total binding strength, even though any individual interaction might be weaker due to an $e < 1$. ϵ_{ij} has to be decreased further to compensate for that effect. The number of atoms that can exactly fit into the energy minimum around one atom is proportional to the surface area of a sphere with van der Waals radius, which would suggest the following quadratic correction:

$$\epsilon''_{ij} = (\frac{R}{R'})^2 \epsilon'_{ij} = (\frac{R}{R+d})^2 \epsilon'_{ij} \,. \tag{6.5}$$

The Lennard-Jones potential minimum also gets wider linearly as R_{ij} increases and more atoms can fit into the minimum along the radial direction, leading to a cubic correction:

$$\epsilon''_{ij} = (\frac{R}{R+d})^3 \epsilon'_{ij} \,. \tag{6.6}$$

In the cubic case, the binding energy stays approximately constant so that the correction with equation (6.6) compensates well for the additional possible interactions and a lowering of $e < 1$ indeed weakens the attractive LJ interaction between the partner molecules. Hence, in all cases, a cubic correction of ϵ_{ij} was used.

6.4 Results and discussion

6.4.1 Simulations of near-native protein-protein complex formation

For 6 protein-protein complexes the regular MD simulations and the RS-REMD (repulsive scaling replica exchange molecular dynamics) method was compared to identify the native complex geometry after starting from distant initial locations of partner proteins. In the starting arrangement the ligand-protein was located on the opposite side of the receptor partner with respect to the native binding site (worst-case scenario of the initial guess). In each case 16 MD simulations with different initial velocities were performed using an implicit generalized Born (GB) solvent model (see Methods for details). The use of an implicit solvent model reduces the computational demand and allows for faster free diffusion of the proteins due to appropriate reduction of the viscosity compared to an explicit solvent model. The simulations started from the unbound ligand and receptor conformations. Only in 2 of the test cases (2oo9, 1syx) individual regular MD simulations reached locations near to the native binding site and sampled it for longer than a few ns with the $16 \times (300 \text{ to } 900)\ ns$ (see Table B.1). In the smallest test case (2oo9) the native binding site ($\text{rmsd}_{\text{ligand}} < 10$ Å) (root mean square deviation of the ligand to the native ligand after best superposition of native and simulated receptors) was identified in 10 runs after an average simulation time of 186 ns (relative occupancy of binding site 47 % in the second half of the simulations, see Supporting Information Table S1). For 1syx one simulation reached placements near the binding site ($\text{rmsd}_{\text{ligand}} < 10$ Å) after a long simulation time of 236 ns where it stayed for a short time span (approximately 44 ns) until the $\text{rmsd}_{\text{ligand}}$ grew again beyond 10 Å. In the other simulations including all 16 regular MD simulations of the other 4 protein cases (2cfh, 2sni, 1gcq, 7cei) trapping at locally stable sites but no approach of the native binding site was observed (see Figure 6.2).

Next, we employed the repulsive scaling (RS)-REMD technique using 16 replicas and starting from the same initial placement as the regular MD simulations. In all but one case sampling of near-native arrangements was observed in the reference replica after $\sim$ 20-400 ns (see Figure 6.2 and Table 6.3). For the 2oo9 system, it took 55 ns and thus a bit longer than in the case of the regular MD simulations (28 ns). Here, the interacting proteins are very small (contain fewer than 70 residues) with an apparently small number of alternative locally stable binding geometries. However, using RS-REMD the ligand of 2oo9 reached the native site on average faster than in the case of regular MD simulations (by $\sim$ 100 ns, see Table 6.3) and the relative occupancy of the binding site was 78 % in the reference replica, higher than the 47 % observed when combining all regular MD results (see Supporting Information Table S1). The RS-REMD simulation on 1syx explored near-native binding arrangements ($\text{rmsd}_{\text{ligand}} < 10$ Å) after 68 ns and thus more than 150 ns faster than the free

simulations. For the other three proteins, the RS-REMD simulations captured the native binding site after 21 ns (2sni), 258 ns (7cei), and 405 ns (2cfh). For these complexes, one can observe that the ligand continuously approaches the native binding site through several intermediate states (see Figure 6.2). It illustrates the advantage of RS-REMD compared to regular MD simulations: while regular MD simulations can get easily trapped in intermediate binding states for significant simulation times the RS-REMD allows the system to more rapidly dissociate from such states and reach near-native geometries. The process of approaching the binding site is illustrated for the 7cei case in Figure 6.3. The initial population of the centers of mass of the ligand-protein is located on the opposing side of the receptor in the first third of the simulation (in the reference replica). The sampled distribution eventually shifts towards the binding site on the receptor protein and there increases continuously until the ligand is mostly populated at the binding site or in the vicinity of the binding site in the reference replica. A similar representation for the highest replica shows a quite uniform spherical population of the ligand around the receptor (see Supporting Information Figure S1).

In all the cases for which the binding site was identified RS-REMD lead occasionally to a very close agreement to the native structure with a lowest $\text{rmsd}_{\text{ligand}}$ of ~ 3 Å (see Table 6.3). In particular, the lowest $\text{rmsd}_{\text{ligand}}$ using RS-REMD was closer than for regular MD also in those cases were both methods identified the correct binding site. Thus, in these cases RS-REMD not only performs better than regular MD in the global searching process for the binding site but also for local rearrangements at the binding site.

In only one case, 1gcq, the near-native binding geometry was not detected after the upper limit of 640 ns of simulation time in the RS-REMD and also not during any of the 16 regular MD simulations (Figure 6.2). In the 1gcq case, the correct position of the ligand at the protein-interface site of the receptor was captured but the orientation of the ligand was incorrect (see Supporting Information Figure S2). It is possible that the force field and implicit solvent representation favor in this case the non-native binding geometry. Stabilization of alternative (non-native) binding geometries (in the current force field setup) is also observed for some of the other test cases. For example, in the 1syx case, complexes with an $\text{rmsd}_{\text{ligand}} < 5$ Å are occasionally visited in the reference replica but alternative states with larger $\text{rmsd}_{\text{ligand}} \sim 8$ Å are more frequently sampled. Besides force field artifacts, such deviation can also be due to the conformational restraining with respect to the unbound (backbone) protein conformation that we include during all simulations. Indeed, the protein association in the case of 1syx involves some backbone changes towards the bound structure at the protein interface (1syx corresponds to a target of medium difficulty, Supporting Information Table S2).

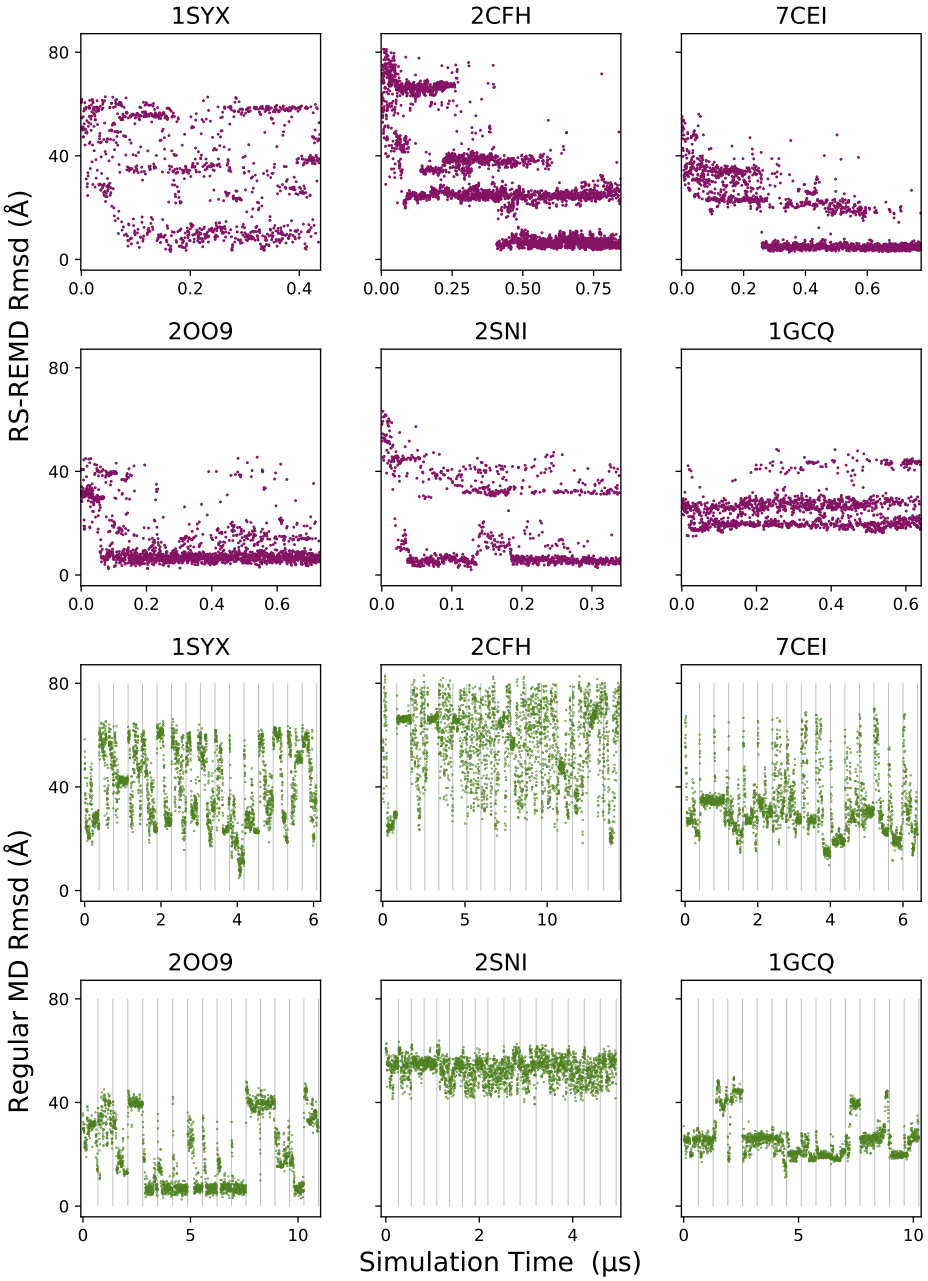


Figure 6.2: $\text{Rmsd}_{\text{ligand}}$ from the native structure for the reference replica of the RS-REMD simulations (magenta dots; first and second row) and for the regular MD simulations (green dots; third and fourth row) of the six protein–protein test cases. The results of the 16 individual simulations (separated by vertical lines) were concatenated in the regular MD cases.

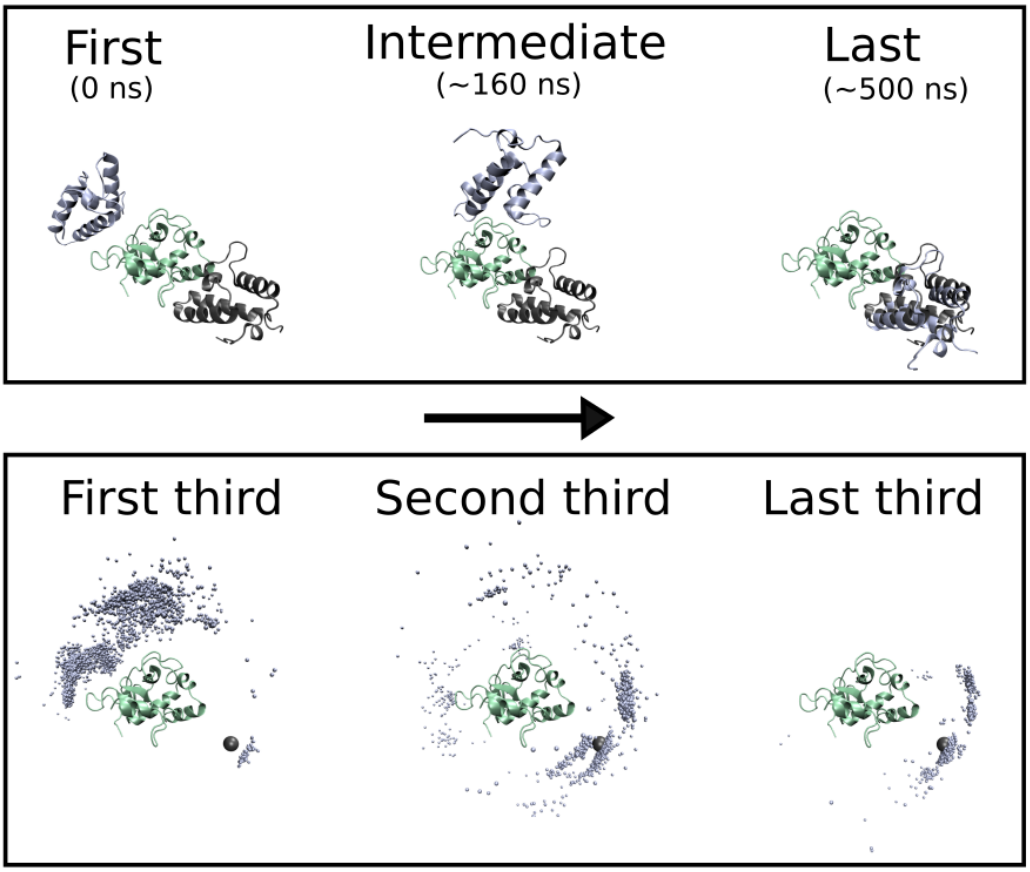


Figure 6.3: (Upper panel) Three snapshots from the reference replica trajectory of the RS-REMD of the 7cei complex example (green cartoon: receptor protein, blue cartoon: ligand-protein, black cartoon: native ligand protein placement). (Lower panel) The population of the sampled ligand (center of mass) placements during RS-REMD is indicated as blue spheres around the receptor (green cartoon). The ligand-protein placement in the native complex is shown by an enlarged black sphere.

In particular, a loop conformation at the interface of the receptor protein differed in the sampled near-native complexes from the structure in the bound form (see Supporting Information Figure S3). Also, states with larger $\text{rmsd}_{\text{ligand}}$ are still populated in the reference replica in the final stage of the RS-REMD simulation (Figure 6.2). In the 2cfh case a near-native geometry ($\text{rmsd}_{\text{ligand}} < 5$ Å) forms the dominant sampled state in the final simulation stage but an alternative binding

Table 6.3: Simulation details for each complex indicated by the PDB-id.

PDB	Time to encounter native site		lowest rmsd	
	RS-REMD	regular MD	RS-REMD	regular MD
	ns/replica	ns/simulation	Å	Å
7cei	258	363[a]	2.7	7.6[a]
2oo9	55	186[b]	2.4	2.4
2cfh	405		3.0	14.9
1syx	68	236	3.1	4.5
2sni	21		2.1	37.4
1gcq			11.6	10.2

[a] The ligand was not stable at the binding site and stayed only for 1.4 ns.
[b] The mean value of all encounter times was taken.

geometry with $\text{rmsd}_{\text{ligand}} \sim 25$ Å remains also highly populated. The result indicates that the force field setup stabilizes in many cases not only exactly the native binding geometry but also alternative states in the vicinity of the native structure but also some binding modes quite far from the experimentally observed complex structure. All protein simulations for which the binding site was captured were extended for more than 300 ns after having encountered the native binding site. The relative population in the reference replica of near-native states at the binding site grows in several cases with ongoing simulation time (reaching > 50%) (see Supporting Information Table S1). This is not the case for 1syx with a population of the near-native complex of $\sim$ 25% (still forming the largest populated cluster; Supporting Information Figure S4) but some alternative binding modes reaching a similar population indicating similar binding affinity. The population of ligand placements at the native binding site is highest in the reference replica and decreases for the higher replicas (see Supporting Information Figure S5 for the example case 7cei) due to the higher repulsive bias. Hence, an advantage of the RS-REMD technique relative to regular MD is that the near-native binding site can be identified by just looking at the population in the different replicas.

6.4.2 Refinement of individual protein-protein docking poses in implicit solvent

Significant computational demand and simulation times are still necessary to reach near-native binding geometries using RS-REMD from distant initial placements. However, this corresponds to a worst-case scenario. Instead, it is also possible to first perform a rapid protein-protein docking (not including solvent or partner flexibility) in order to first identify potential binding sites possibly not too far from the native

binding geometry. In a second step, short RS-REMD simulations are used as a refinement procedure to further improve the docking results. We first performed a docking run on a subset of 20 protein–protein complexes of the protein-protein benchmark 3.0 [132] using the docking program ATTRACT and obtained 50 top-ranked poses with different ligand deviations from the bound complex ($rmsd_{ligand} <$ 25 Å) around the receptor protein (same as used in a recent study on protein-protein docking scoring [279]). For each of these poses, a short RS-REMD refinement (4 ns) was performed to refine the docking results. In order to limit the computational demand for this procedure, an RS-REMD with 8 replicas was performed.

Overall, upon RS-REMD refinement of all 50 decoys for the 20 test cases a slightly larger number of models (65%) with higher $rmsd_{ligand}$ was observed compared to the starting structures (35 % of poses had higher $rmsd_{ligand}$ before refinement) (Figure 6.4). Likely, because of the short simulation time, no $rmsd_{ligand}$ improvements better than 13 Å were sampled. More important than the $rmsd_{ligand}$ improvement of poses with a high deviation from the native binding mode is the refinement performance of the near-native models. For the docking model closest to the native binding site a smaller $rmsd_{ligand}$ after RS-REMD refinement was observed for 13 complexes. Also, the improvement in $rmsd_{ligand}$ was higher, so an overall improvement of 0.49 Å was found considering the mean difference in $rmsd_{ligand}$ before and after RS-REMD refinement of the closest to native pose (see Figure 6.4, left panel).

The RS-REMD scaling results are compared to a well established atomistic refinement procedure [263, 279] using regular MD simulations with 8 times longer simulation time per decoy and final energy minimization (same force field setup and positional and distance restraints as in the RS-REMD, see Methods). In Figure 6.4 (right panel) the $rmsd_{ligand}$ after RS-REMD refinement is plotted vs. $rmsd_{ligand}$ after regular MD refinement. A slightly higher amount of structures had a lower rmsd with RS-REMD refinement (57 %, magenta dots) than with regular MD refinement (43 %, green dots). RS-REMD refinement also performed better than the regular MD refinement considering the model with the lowest $rmsd_{ligand}$ for each protein-protein complex. For 13 complexes the near-native pose was closer for RS-REMD refinement than regular MD refinement and the mean $rmsd_{ligand}$ of the near-native poses of all structures was slightly lower (0.33 Å) for RS-REMD refinement.

Finally, the refined poses were scored based on the interaction energy of ligand and receptor, the total energy of the complex was subtracted from the total energy of the individual ligand and receptor. The selectivity of the resulting funnel plots ($rmsd_{ligand}$ versus scoring) was compared (see Figure 6.5 and Supporting Information Figures S6 and S7), measuring the ability of the refinement procedures to distinguish near-native from other decoys. The selectivity was calculated based on an approach introduced recently (see Chapter 5) [279], by calculating the normalized difference in binding energy of the highest scored pose at the binding site

($\text{rmsd}_{\text{ligand}} < 10$ Å from the pose of minimal $\text{rmsd}_{\text{ligand}}$) S'_T and not at the binding site S'_F:

$$Selectivity = S'_T - S'_F. \tag{6.7}$$

The two key poses were shifted by the mean scoring value of all poses and divided by the minimum scoring value in order to obtain comparable results for each protein.

$$S'_i = \frac{S_i - \overline{S}}{S_{min}}. \tag{6.8}$$

A selectivity of 1 means a perfectly selective funnel plot for the best-bound pose at the binding site and -1 means that the funnel plot is very selective for the highest scored decoy not at the binding site. A value of 0 means that the highest scored near-native pose and the highest scored pose not at the binding site have the same binding affinity.

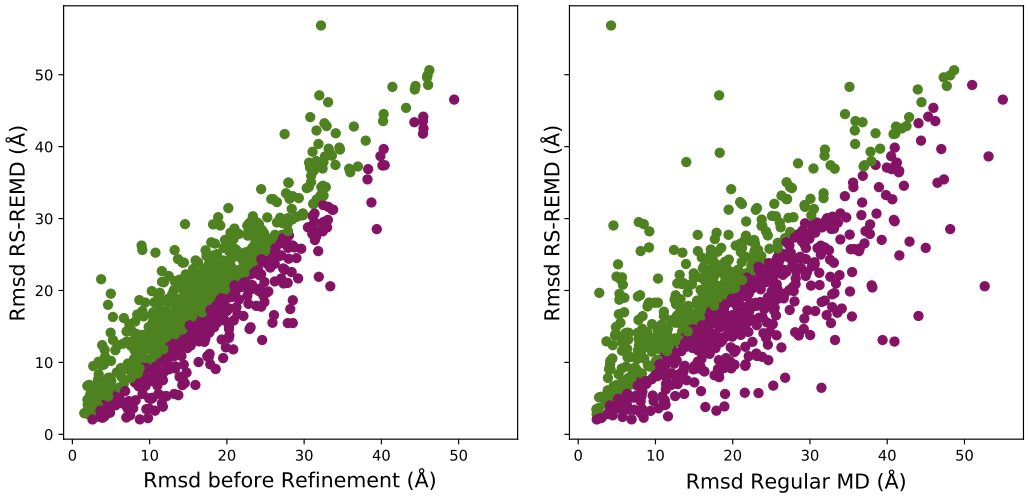


Figure 6.4: The $\text{rmsd}_{\text{ligand}}$ of all refined complexes after RS-REMD refinement is plotted against the $\text{rmsd}_{\text{ligand}}$ before refinement (left panel) and after regular MD refinement (right panel). Magenta dots mark poses where the $\text{rmsd}_{\text{ligand}}$ decreased due to RS-REMD refinement (35 % for the left panel and 56 % for the right panel) and green dots depict the poses for which the $\text{rmsd}_{\text{ligand}}$ increased after RS-REMD refinement.

In 16 cases RS-REMD was able to identify the near-native binding placement (positive selectivity) and only in four cases the refinement approach resulted in a clearly negative selectivity (1ffw, 2oob, 1ak4, 2i25), identifying an incorrect binding site.

The selectivity was slightly higher for 14 structures in RS-REMD (1z0k, 3a4s, 7cei, 1ay7, 1ffw, 1qa9, 3sgq, 1gcq, 2oob, 2cfh, 1ak4, 1fle, 1z5y, 2i25) compared to using regular MD refinement. The mean selectivity of all structures was also higher after the RS-REMD refinement procedure (0.12) in contrast to regular MD refinement (0.06).

In summary, the RS-REMD refinement was able to improve in many cases the placements of the near-native poses. It overall performed slightly better than an established regular MD refinement procedure (at the same computational costs) in terms of the selectivity in identifying the native binding site.

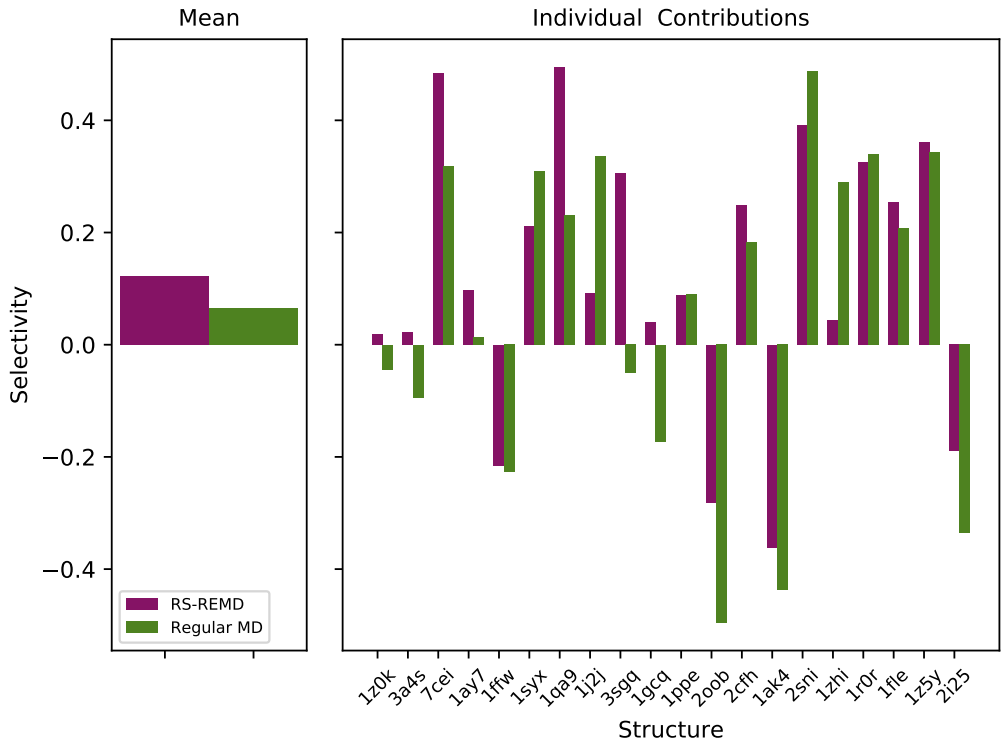


Figure 6.5: Mean selectivity of all structures and the selectivity of each structure for the different refinement procedures. The selectivity was calculated as described in the main text (see equations 6.7, 6.8), considering the highest ranked poses at the binding site and not at the binding site, respectively. The corresponding funnel plots (scoring vs $\text{rmsd}_{\text{ligand}}$) are shown in the Supporting Information Figures S1 and S2.

6.4.3 Refinement of a protein-protein docking ensemble in one RS-REMD

Instead of refining every docked pose separately, it is also possible to start from a different docking decoy in each replica leading to a higher diversity in the starting conditions such that multiple possible binding sites are represented in different replicas. In the case of a sufficient number of replicas it is then possible to perform only one RS-REMD simulation per complex in contrast to 50 separate simulations for individual refinement of decoys (see above). To increase the challenge, the refinement was initialized exclusively from starting placements that were not located at the binding site. Only poses with an $\text{rmsd}_{\text{ligand}}$ above 10 Å (for the complex 7cei 8 Å, due to a lack of poses with large RMSD) were considered as starting structures. An increasing replica number was linked to a lower ranking after docking for the selected starting poses. The results of the RS-REMD (with 16 replicas) are again compared to 16 regular MD simulations of the same length starting from the same initial placements.

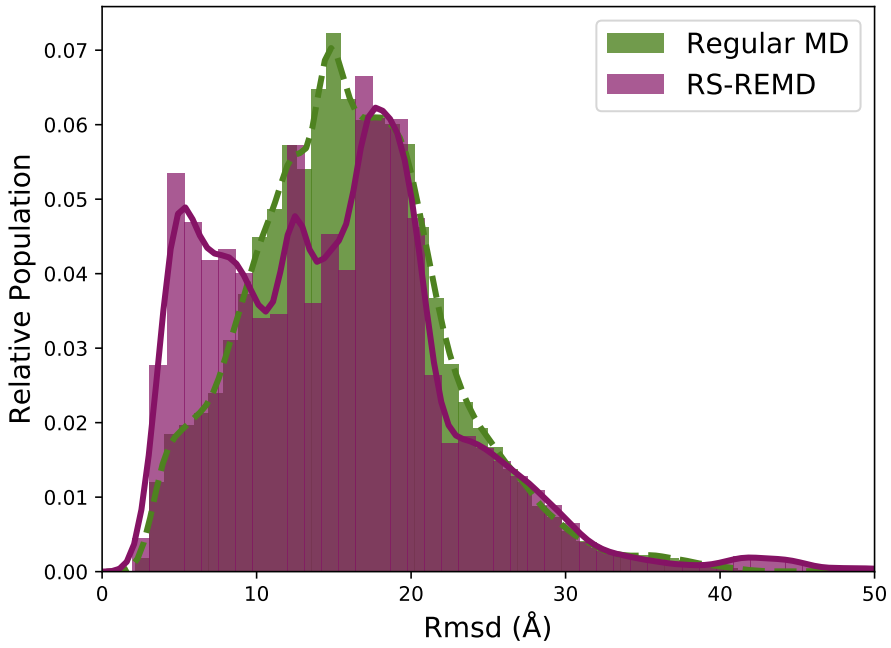


Figure 6.6: Histograms of the $\text{rmsd}_{\text{ligand}}$ from the native structure for the reference replica of the RS-REMD refinement (magenta) (for all 20 protein–protein test cases) is compared to the $\text{rmsd}_{\text{ligand}}$ histograms of the regular MD simulations (green). The refinement was performed starting from initial placements not at the binding site.

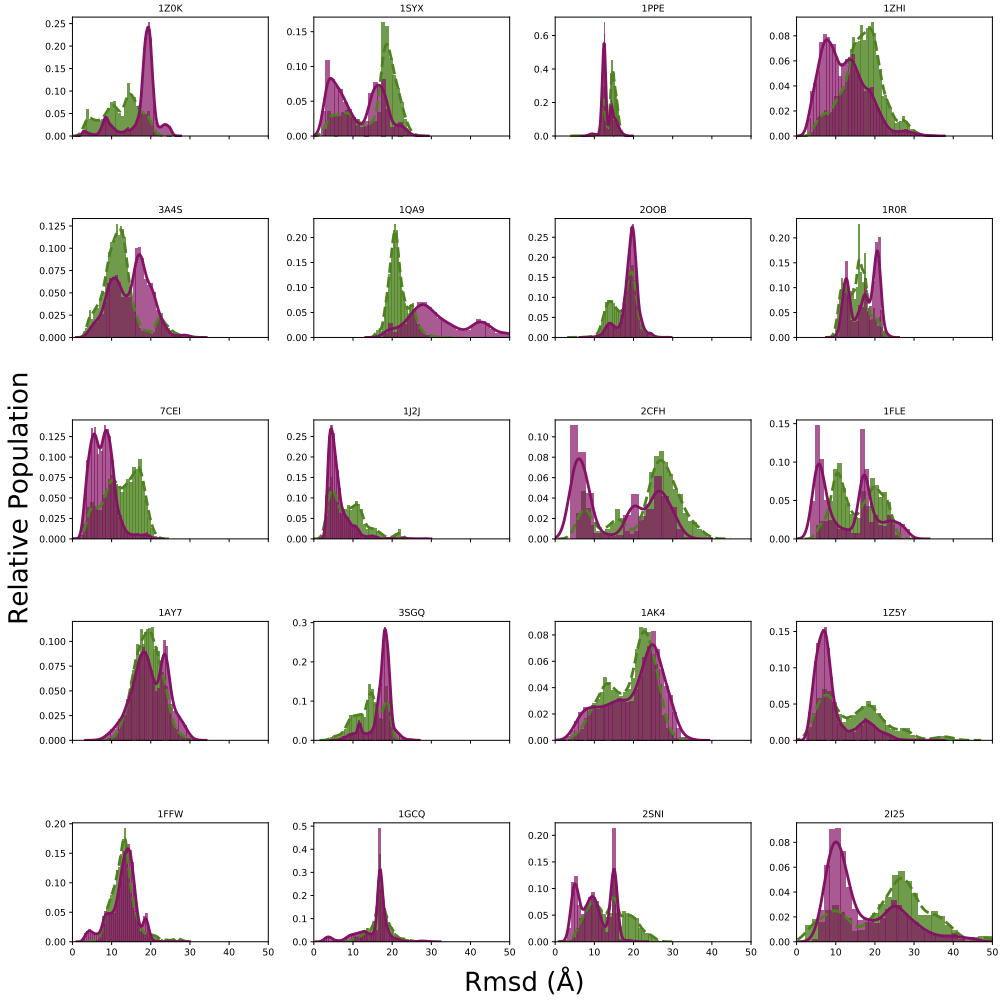


Figure 6.7: Histograms of the $\text{rmsd}_{\text{ligand}}$ from the native structure for the reference replica of the RS-REMD refinement (magenta) (for all 20 protein–protein test cases in an individual figure) is compared to the $\text{rmsd}_{\text{ligand}}$ histograms of the regular MD simulations (green). Both refinement procedures were initialized from poses that were not at the binding site.

The resulting total population of sampled near-native states close to the binding site ($\text{rmsd}_{\text{ligand}} < 10$ Å) increased significantly (30 %) in comparison to regular MD refinement (17 %) (see Figure 6.6). For 14 of the 20 structures, RS-REMD was able to capture the binding site, in some cases, the population was highest at the binding site (see Figure 6.7). It points out, that even relatively short RS-REMD simulations

can capture near-native binding geometries that were not already found in the initial docking search.

Performing only one refinement simulation starting from an ensemble of promising docking solutions and not refining every single pose individually can significantly reduce the computational demand.

Comparing the $rmsd_{ligand}$ histograms of the refinement procedure to the histograms of the pure RS-REMD simulations (see Figure 6.2 and Supporting Information Figure S4) the dominant states are consistent in most cases, especially for $rmsd_{ligand}$ values under 10 Å(see 2cfh, 7cei, 2sni, 1syx). In the case of 1gcq, the binding site was not captured in the long repulsive scaling simulation, still the two populated spikes around 20 Å are also present in the refinement simulations.

6.5 Conclusion and outlook

A new H-REMD scheme is presented that includes a repulsive scaling potential (RS-REMD) between different protein molecules based on a modification of the intermolecular LJ parameters. The bias requires a modification of the simulation parameter file but no changes in the underlying MD program are involved and full GPU support is possible. The replica exchange scheme was applied and tested on three tasks that seek to identify the native binding geometry of protein-protein complexes using an implicit solvent model. First, RS-REMD allowed sampling near-native binding placements in 5 out of 6 example complexes, starting from a random placement far away from the native binding site. In contrast to multiple regular MD simulations, which were stuck mostly at locally stable sticky sites, these sticky sites were overcome through several intermediate steps in the RS-REMD. While the higher replicas sampled the whole receptor surface, the reference replica sampled locally favorable sites quickly until the native binding site was captured but depending on the case alternative binding modes were also still sampled. Although much less demanding than regular continuous (c)MD simulations still quite extensive sampling is needed for this approach that may limit its applicability.

In addition to starting from a worst-case scenario, we also used the approach for refining pre-docked poses. By applying a short RS-REMD run for each of the 50 poses of a benchmark set of 20 protein-protein complexes, it was possible to decrease the mean deviation from the native binding site. Moreover, the mean selectivity of identifying the native binding site according to a simple scoring function (based on the interaction energy) was increased in comparison to a regular MD refinement.

The simulation effort could be further reduced using a refinement scheme that associates each replica in the RS-REMD run with a different docking pose as starting structure. In contrast to the first refinement procedure, only one refining simulation had to be performed for each protein-protein complex. The population of the

ligand-protein partners near the binding site was clearly increased with RS-REMD beyond the result achieved by regular MD simulations. The benchmark set also contained difficult test cases (see Chapter 5 and also reported in [279]), where the identification of the native binding site was not possible in both refinement procedures. Possible reasons are inaccuracies of the implicit solvent model that may not always favor correct complex structure relative to alternative arrangements. Explicit solvent simulations may help to solve this issue and will be tested in future studies. However, the probably slower diffusion and increase in the number of particles will likely demand higher computational efforts. Another limitation of our setup is the inclusion of conformational restraints of the partner molecules with respect to the unbound conformations. This avoids any large-scale conformational change or unfolding of partners but in some cases may prevent conformational adaptations necessary for productive protein-protein complex formation. More global restraining methods like the inclusion of backbone Rmsd restraints can help to overcome this issue in future efforts.

In principle, the RS-REMD biasing scheme can also be helpful to study folding/unfolding events or dissociation/association of parts of a protein structure. In such a case only the interactions of the selected part of the protein with other protein segments are scaled in the replica simulations.

7 Efficient Refinement and Free Energy Scoring of Predicted Protein-Protein Complexes Using Replica Exchange With Repulsive Scaling

Typically, putative protein-protein complexes are predicted based on docking methods and simple force field or knowledge-based scoring functions are applied to evaluate single complex structures (Chapter 4). In the next chapter, we will extend the repulsive scaling RS-REMD scheme of Chapter 6 to simultaneously refine and calculate a free energy score of protein-protein complexes. Originally introduced for implicit solvent, we will employ a more realistic explicit water environment. The approach is computational demanding but may offer a route for refinement and realistic ranking of predicted protein-protein docking geometries.[1]

7.1 Introduction

Protein-protein interactions (PPI) play an important role in nearly all cellular processes. Despite a growing number of experimentally determined protein-protein complex structures, there is still a demand for accurate prediction of complexes especially in combination with limited or low-resolution experimental data. On the computational side, two aspects are of utmost importance to elucidate the details of the PPI, the identification of the native binding site of two protein partners (sampling step) and the correct prediction of the corresponding binding affinity (scoring step). A variety of computational docking and scoring methods have emerged (reviewed in [277, 326, 115, 150, 303, 29, 109]). Current methods can usually generate rapidly numerous putative complex geometries but the generation of near-native geometries and the realistic scoring of the complexes is limited. Hence, a near-native complex structure is not necessarily detected as the most favorable binding geometry. Secondly, during a systematic search step, the partners are typically represented as rigid bodies or at limited resolution using a coarse-grained model. Depending on possible conformational changes associated with complex formation or in the case

[1] The contents of this chapter have been published in a similar form in [278].

of homology-modeled partner structures none of the generated complexes is close to a native geometry [115, 326, 296]. Even with a realistic scoring the prediction fails in this case because of the lack of near-native solutions in the generated pool.

In principle, an atomistic description of the proteins including full flexibility with an explicit water model in a molecular dynamics (MD) simulation should be ideally suited to predict realistic interaction geometries [277, 240, 101]. Indeed, in recent years near-native protein-protein complex structures have been recovered from multiple MD simulations [248]. However, such approaches are computationally very demanding and in recent years promising advanced sampling methods have been developed to identify near-native protein-protein interaction structures [101, 219, 244, 240].

Recently, encouraging strategies to tackle the identification of the native binding site have been conducted on sets of protein-protein complexes using molecular dynamics simulations in combination with advanced sampling methods. One approach used tempered binding to explore several binding events of five protein-protein cases on the special purpose computer ANTON [240, 267]. In tempered binding, the interaction strength between the protein partners is scaled in regular steps to enhance the simulation of rare events. Still, the computational demand of the approach is large. Another more efficient approach employs perturbed distance restraints to allow reversible protein-protein association during MD simulations [244].

A very popular advanced sampling method is replica exchange, in which several copies of the system (replicas) are simulated in parallel, performing exchange attempts between the different replicas in regular time intervals [194]. One possibility is to employ different temperatures along the replica coordinate (T-REMD) [290]. The system is able to escape local minima in the replicas of higher temperature and introduce the newly sampled phase space to replicas of lower temperature. The efficiency of the T-REMD method is, however, strongly limited by the size of the simulation system. It is also possible to exchange the Hamiltonian between the copies (H-REMD) with the temperature held constant [95, 314, 192, 3, 127, 147].

Improvement of efficiency compared to T-REMD can be achieved by specific scaling of relevant interaction parameters. In a recent method, various levels of a biasing potential are introduced in each replica that keep the partner at different distances from the receptor surface and thus accelerates the searching process for the correct binding site [238]. The efficiency can be further enhanced by a specific repulsive intermolecular bias between the ligand and receptor, based on modified Lennard-Jones parameters in the replicas (RS-REMD) (see Chapter 6) [276]. With this method for five protein-protein complexes (out of six) it was possible to identify the native binding site, even when starting from ligand placements on the other side of the receptor. Moreover, short RS-REMD simulations ($<$ 10 ns) allowed frequent

identification of near-native complex structures when starting from an ensemble of pre-docked placements in the vicinity of the binding site.

Atomistic MD-simulations can also be used to rigorously calculate free energy changes of a molecular system including full flexibility and solvent effects. A variety of methods were established in the past decades to calculate the free energy of binding [277]. Some methods to calculate binding free energies are based on multi-step alchemical decoupling (DDM) of the ligand in the complex with a receptor protein vs. free solution [7, 70]. Such alchemical transformations become more expensive with increasing atom number of the conformers, making the use of a DDM inefficient in the context of protein-protein complexes.
For protein-protein complexes usually physical pathway methods are used to predict the binding affinity, introducing a coordinate for physical separation of the partners and measuring the dissociation work [119, 244]. Good agreement with experiment was found for selected cases. Siebenmorgen and Zacharias applied an umbrella sampling (US) approach in all-atom implicit solvent MD simulations to calculate the absolute binding free energy of (50) pre-docked poses for 20 protein-protein complexes (see Chapter 5) [279]. The US free energy method showed improved performance to selectively discriminate native binders compared to simply evaluating the interaction energy [279]. Perthold and Oostenbrink evaluated recently 18 protein-protein docking targets from the CAPRI (Critical Assessment of PRedcited Interactions) docking challenge using short nonequilibrium explicit solvent simulations (GroScore) with very promising results compared to the best CAPRI scorer performance [243].

In the present study, we extend the RS-REMD method to perform simulations in explicit solvent and in addition to also allow extracting a free energy score of binding. For a benchmark set of 36 complexes, the RS-free energy score (in explicit solvent) gives a quite good correlation to experimental binding data with modest computational demand. An application to 50 docked decoys based on unbound partner structures for 20 protein-protein docking cases gave on average a significant improvement of the prediction geometry (deviation from the native complex) and for each decoy a predicted RS-score for binding. In many but not all cases the near-native solutions could be detected as those with the lowest free energy score. The failure in some cases and the differences in the performance for explicit solvent and implicit solvent treatment are investigated and discussed.

7.2 Materials and methods

7.2.1 Explicit solvent simulations of native protein-protein complexes

A subset of the protein-protein affinity benchmark [150] was evaluated with RS-REMD in explicit solvent, containing 36 native protein-protein complexes. The benchmark contained 16 structures that we already evaluated in an earlier study on binding free energy simulations [279] and 20 additional structures for which the experimental affinity measurements were stated as reliable [151]. In order to reduce the computational effort only dimeric complexes with no more than 700 residues were taken into account. Missing residues were added using the program Modeller [201]. All MD simulations of this study were conducted with the pmemd.cuda module of the Amber18 software package [41]. The protein force field ff14sb [198] was used for all protein-protein complexes. The proteins were solvated using the tip3p [144] water model in an octahedron box with periodic boundary conditions (minimum distance between protein and box edge were 15 Å). The charges were neutralized with Na+ and Cl- ions. The systems were minimized (1000 steps of steepest descent), heated in 3 steps to 300 K using a Langevin thermostat, and equilibrated in 16 ps. Next, the repulsive scaling (RS) REMD simulations were performed (5 ns per replica, exchange attempts every picosecond), i.e. H-REMD simulations with 16 replicas and increasing bias between the ligand and receptor Lennard-Jones parameters (see Table 7.1). The trajectories were analyzed using pytraj and the associated relative binding free energies were calculated with MBAR (see Section 3.4.1 for details on the MBAR method) [271].

7.2.2 Implicit solvent simulations of native protein-protein complexes

The same 36 native protein-protein complexes and the same force field as for the explicit solvent simulations were used for the implicit solvent MD simulations. The OBC (Onufriev, Bashford, Case) generalized Born (GB) implicit solvent model [237] (igb=8 option) was used with an infinite cutoff radius for the GB radii and non-bonded interactions. For temperature scaling a Langevin thermostat with a collision frequency of $\gamma = 2\,\mathrm{ps}^{-1}$ was used. For geometry optimization of the starting structures minimization was performed consisting of 400 steps of steepest descent followed by 2100 steps of conjugate gradient. During 45 ps of simulation time, the systems were heated in 3 steps to 300K and equilibrated. RS-REMD simulations were performed for 2.5 ns per replica using the same LJ parameter scaling scheme as in the explicit solvent case (see Table 7.1). Between the C_α atoms of each individual protein harmonic pairwise distance restraints were applied (force constant 0.5 kcal $\mathrm{mol}^{-1}\mathrm{Å}^{-2}$) (with respect to the unbound partner structures!). Comparable restraints were also used in previous implicit solvent simulations of these structures.

[279] Moreover, dissociation far from the receptor was limited using half-parabolic COM distance restraints between the C_α atoms of ligand and receptor (force constant 2.0 kcal $mol^{-1}Å^{-2}$ for distances 15 Å above the COM distance of the equilibrated structure).

7.2.3 Scoring of protein-protein docking poses in explicit solvent

A set of 50 docking poses for 20 protein-protein complexes was generated using ATTRACT [329]. As in previous studies, the 50 models with the lowest root-mean-square deviation (RMSD) to the native structure out of the 300 top-ranked poses were considered [279]. The simulations were prepared in the same way as the native explicit solvent simulations (force field ff14sb with tip3p water representation). Each docking model was minimized, heated to 300 K, and equilibrated as in the case of the native explicit solvent simulations.

For each docking pose a RS-REMD simulation was conducted with 16 replicas and the same parameter scaling as in the preceding simulations (see Table 7.1). The simulation time between each docking pose was constant and varied slightly between different protein-protein cases (5 ns - 5.5 ns per replica).

7.2.4 Scoring of protein-protein docking poses in implicit solvent

The same 50 ligand placements in 20 protein-protein complexes as for the explicit solvent RS scoring of pre-docked poses was used for implicit water RS-REMD simulations. The same force field parameters as for the native implicit solvent simulations were used (force field ff14sb and igb=8 option for the implicit solvent model). Also, the minimization, heating to 300 K and equilibration was conducted for each docking pose as the native implicit solvent simulations with the same restraints applied. A RS-REMD simulation of 2.5 ns (per replica) with 16 replicas (same parameter set as before, see Table 7.1) was performed for every docking pose.

7.2.5 Free energy calculation along RS-REMD replicas

In the RS-REMD simulations, N = 16 replicas were simulated in parallel with an increasing bias between ligand and receptor atoms with higher replica number. The introduced bias is associated with a lower (attractive) potential well depths ϵ and higher (repulsive) effective pair-wise van der Waals radius in the Lennard-Jones parameters, which leads to a dissociation of the ligand in the higher replicas (no contact with the partner). Thus, the reference replica (no bias between ligand and receptor) ideally represents the bound state of the complex ($\alpha = 1$) which is progressively transferred to the unbound state in the higher replicas. This leads to

Table 7.1: Lennard-Jones scaling parameters for the RS-REMD simulations in implicit and explicit solvent with 16 replicas.

Replica Number	16 replicas	
	d	e
1	0.0	1.0
2	0.01	0.99
3	0.02	0.98
4	0.04	0.97
5	0.08	0.96
6	0.12	0.94
7	0.16	0.92
8	0.2	0.9
9	0.24	0.88
10	0.28	0.86
11	0.32	0.84
12	0.38	0.82
13	0.44	0.8
14	0.5	0.78
15	0.58	0.76
16	0.68	0.74

distributions with very good overlap along the distance between partner proteins which is also reflected in the high REMD acceptance rate of usually close to 0.5 (see Figure 7.1). A binding free energy score between the bound and unbound state of the complexes can now be calculated with a perturbation approach along the biasing (α-) coordinate associated with a higher replica number (see Table 7.1).

The introduced bias in each replica is calculated using trajectory re-evaluation. The trajectory of replica $\alpha + 1$ is evaluated with the LJ parameter set of replica α and the LJ parameter set of replica $\alpha + 1$. The associated free energy difference can be obtained using the BAR (Bennet-acceptance ratio) method [24]. This approach can be extended from only evaluating adjacent biases along the α path $\Delta U_{\alpha,\alpha+1}$, to evaluating all pairs of biases $\Delta U_{i,j}$, with $i, j = 1, ..., N$. The corresponding free energy difference is calculated using MBAR [271]. For calculating the relative binding free energies only the second half of the simulation time was incorporated and the first half was considered as equilibration time. The statistical errors were calculated by splitting the simulations into 5 parts and calculating the mean and standard deviation.

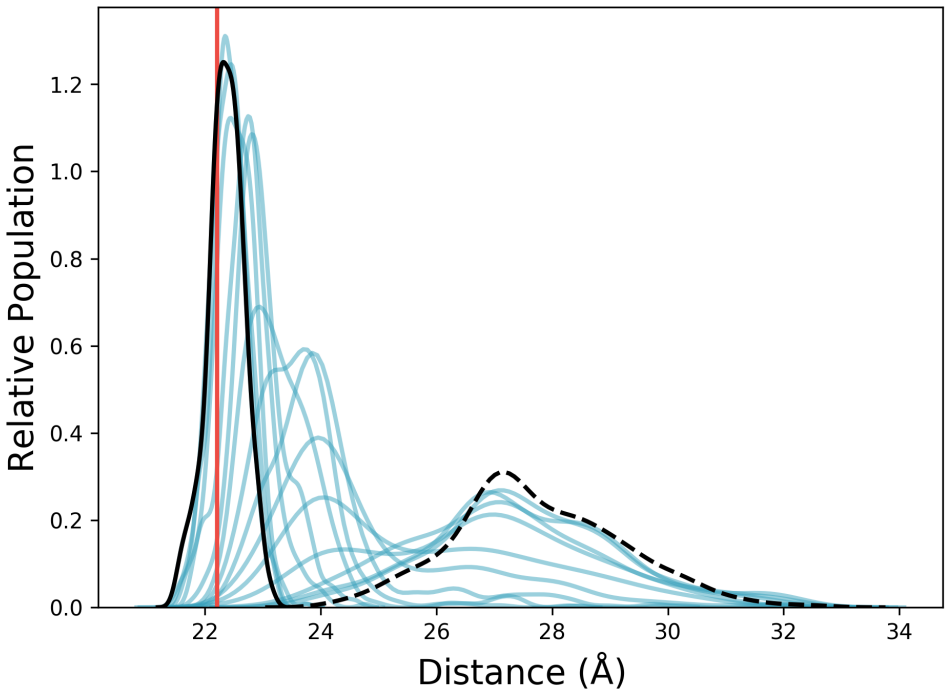


Figure 7.1: Sampled center of mass (COM) distance histograms of all 16 replicas for a typical RS-REMD run in explicit solvent (pdb1z0k case). The solid black line represents the reference replica distribution (associated state) while the dashed black line represents the distribution in the highest replica (sampling mostly the dissociated state). The intermediate replicas (blue histograms) indicate substantial overlap in the transition from the reference to the highest replica distribution. The distance observed in the experimental complex structure is depicted with a red vertical line.

7.3 Results and discussion

7.3.1 Evaluation of the RS-REMD free energy scoring on native protein-protein complexes

The RS-REMD technique employs a series of (16) parallel running replicas with increasing repulsive biasing between protein partners based on a scaling of the effective pairwise van der Waals radius and attractive Lennard-Jones parameters (Table 7.1). Only the intermolecular interactions are modified without affecting intramolecular interactions or interaction with the solvent [276]. In the reference

replica the original unmodified force field is used. Starting from complexed states including explicit solvent the technique allows rapid generation of extensively overlapping partial or fully dissociated distributions of one protein relative to the partner (Figure 7.1) along the replicas. The technique was first applied to a benchmark set of 36 native protein-protein complexes of relatively small to medium size and for which binding free energies are available. In the RS-REMD simulations the bound state has a higher tendency to remain in the lower replicas while the dissociated state is predominantly detected in the highest replicas (see Appendix, Tables C.1, C.2). The histograms of the COM distances of the associated state (replica with minimum $\text{rmsd}_{\text{ligand}}$ calculated in each frame) and the dissociated state (replica with highest COM distance of each frame) evaluated for the second half of the simulation are shown in the Appendix, Figure C.1. Only narrow fluctuations around the native distance (vertical black line) are observed in the bound states (blue histograms). A clear separation between associated and dissociated states (red histograms in Appendix, Figure C.1) is visible. The mean COM distance difference for all complexes between the associated and the dissociated states is 8.4 Å (see Appendix, Table C.1).

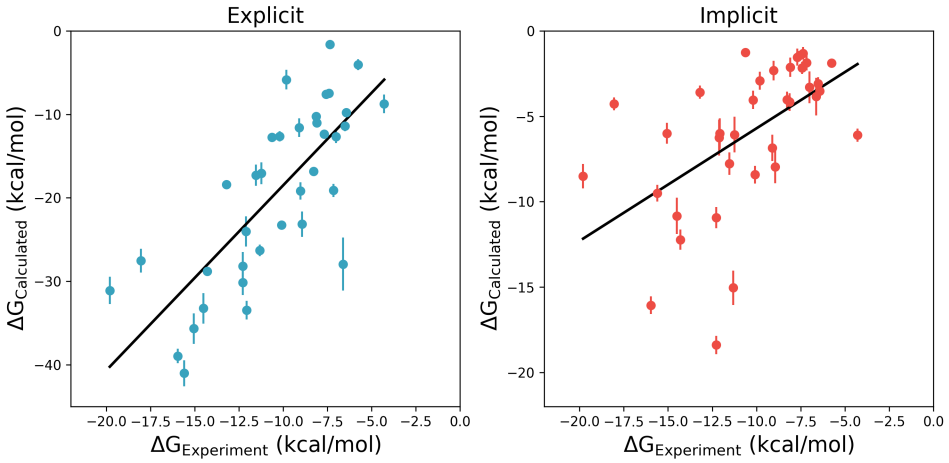


Figure 7.2: Calculated relative binding free energy vs. experimental binding affinity for the explicit water RS-REMD simulations (blue dots in the left panel) and the implicit water simulations (red dots in the right panel) of all 36 native protein-protein complex cases. A linear regression is fitted to the data points with PCCs of 0.77 (explicit solvent) and 0.55 (implicit solvent), respectively.

In order to calculate the free energy difference between bound and unbound states, we applied the MBAR approach along the RS-REMD replicas (see Methods). It is important to note that the calculated free energy change does not represent an

absolute binding free energy of the complex. For example, only a fraction of the full translational and rotational freedom of the partners in the unbound state will be sampled. It is, nevertheless, of interest to compare the calculated free energy change for the set of complexes with corresponding experimental binding free energies (Figure 7.2). The Pearson correlation coefficient (PCC) indicates a quite good correlation of 0.77 (a PCC=1 indicates a perfect linear correlation, 0 = no correlation). It is higher than correlations of a similar benchmark set evaluated with the MMGBSA/MMPBSA (molecular mechanics generalized Born/Poisson-Boltzmann surface area) approaches, that achieved a maximum correlation of 0.65 and 0.52, respectively [45]. A subset of the present benchmark set was recently evaluated by Siebenmorgen and Zacharias (2019) using an umbrella sampling approach in implicit solvent to calculate the absolute binding free energy with a correlation of 0.74 (see Chapter 5) [279]. Taking only those complexes into account that were used in the latter study an even higher correlation of 0.87 can be extracted from the present results.

7.3.2 RS-REMD free energy scoring of native protein-protein complexes in implicit solvent

RS-REMD simulations on the same native protein-protein complexes as used for the explicit water case were also performed with an implicit generalized Born (GB) solvent representation. Also, in case of the implicit GB solvent simulations the RS-REMD allowed to effectively dissociate the native complexes along the replica simulations (see Appendix, Figure C.2 and Table C.2). A mean difference in COM distance between the dissociated and the associated states around 11.9 Å was observed, slightly higher than in the explicit water case. Importantly, the native binding site is stable in the reference replica in all cases, with only small deviations in the COM distances from the distance in the experimental complex structure. However, the average RMSD with respect to the native complex was on average higher compared to the explicit solvent simulations.

The free energy differences are calculated using the same approach as for the explicit water case. A reasonable correlation of the calculated free energy differences and the experimental affinities was obtained (Figure 7.2) with a lower associated PCC of 0.55 compared to the explicit solvent results. The implicit water RS-score has a higher tendency to underestimate the calculated binding affinity, as in 75 % of the cases the achieved ΔG_{Calc} was more than 2.5 kcal/mol too low. Overall, the calculated free energy values are in slightly better agreement with the experimental affinities with a mean difference of $\Delta G_{Diff} = \overline{\Delta G_{Calc} - \Delta G_{Exp}}$ = -4.5 kcal/mol. In 7 cases the difference in affinity was under 2.5 kcal/mol (2sni, 1zhi, 1r0r, 2i25, 1s1q, 2hle, 1b6c).

7.3.3 RS-REMD refinement and free energy scoring of protein-protein docking poses in explicit solvent

RS-REMD simulations were performed on a benchmark set of 50 pre-docked poses of 20 protein-protein complexes in explicit solvent. The complexes were obtained from rigid docking of unbound partner structures (see Methods and [276]). The RS-REMD scoring technique was applied separately to every docking pose to calculate an associated free energy change for the dissociation and also to check for a possible improvement of the complex geometry. The simulations of $50 \times 16 \times 1$ ns for each case (5 ns were simulated overall for each case) took on average 4.9 days (2.3 hours per pose) on a single GPU (see Appendix, Table C.3).

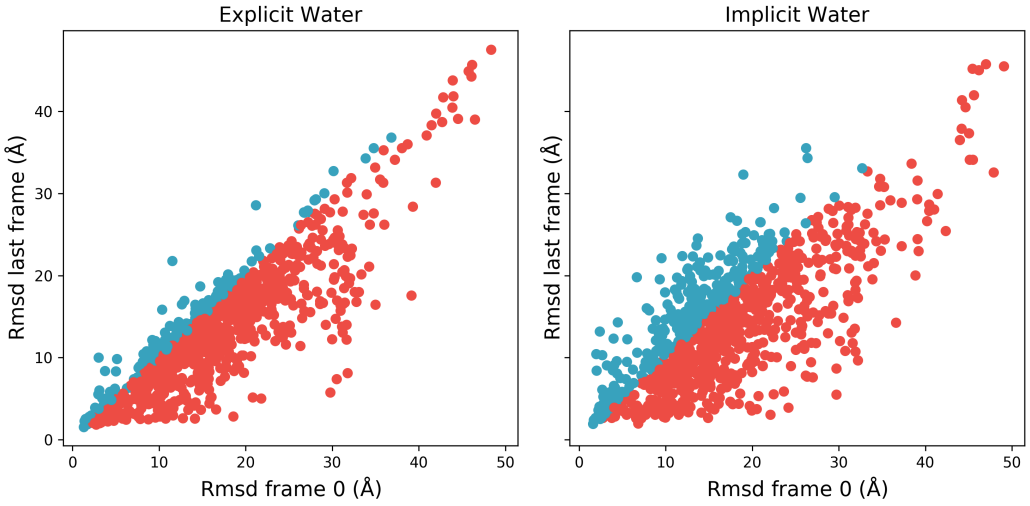


Figure 7.3: The $\text{rmsd}_{\text{ligand}}$ of the first and the last frame of the RS-REMD reference replica simulations for the explicit water representation (left panel) and the implicit water representation (right panel). The minimum rmsd of all replicas was chosen in frame 0 and in the last frame to account for occasional exchanges of the reference structure to higher replicas. An improvement of rmsd (red dots) was observed in 81 % (explicit) and 67 % (implicit) of the cases.

We reported recently, that the RS-REMD technique is able to significantly improve the sampling of near-native binding geometries, especially for start structures in the vicinity of the native binding site (see Chapter 6) [276]. Comparing the $\text{rmsd}_{\text{ligand}}$ from the native binding site of the initial structures and the final sampled geometries of the RS-REMD simulations indicates an improvement in $\text{rmsd}_{\text{ligand}}$ in 81 % of the poses (see Figure 7.3, left panel). Again, especially for structures in the vicinity of the binding site ($\text{rmsd}_{\text{ligand}} < 20$ Å) in many cases quite dramatic improvements were

observed despite relatively short simulation times. It indicates that the repulsive biasing in the replicas is effective in disrupting trapped states and therefore allows rapid diffusion towards low free energy configurations on the protein surfaces. Note that for evaluating the improvement in complex geometry the minimum $\text{rmsd}_{\text{ligand}}$ of all replicas at the end of the simulation was considered, to account for the occasional exchanges from the reference replica to higher replicas. The average minimum $\text{rmsd}_{\text{ligand}}$ of all poses vs. simulation time decays rapidly in the first half of the simulations (Figure 7.4 (blue line)).

The RS-REMD free energy scoring results vs. $\text{rmsd}_{\text{ligand}}$ from the native complex (using again the minimum in the replicas of the last frame of the RS-REMD simulations) are shown in Figure 7.5. The free energy difference was calculated with the same procedure described for the native complexes (from the last 2.5 ns of each RS-REMD simulation, errors were estimated after splitting the data set into 5 parts to calculate mean and standard errors). For most of the structures, clear discrimination between poses close to the native binding site and alternative placements are observed. Comparison to a plot of the same scores vs. initial $\text{rmsd}_{\text{ligand}}$ of the docked start structures also demonstrates the improvement in predicted complex structure (see Appendix, Figure C.3).

Remarkably, in several cases a complex structure is reached (starting from a non-native docked complex based on unbound partner structures!) with an $\text{rmsd}_{\text{ligand}}$ basically identical to the $\text{rmsd}_{\text{ligand}}$ when starting from the native structure and a score very close to the score obtained for the native complex (black dots in Figure 7.5). This indicates both reasonable convergence and sampling power of the approach. However, there are few cases that failed to reach a complex very close to the native complex or resulted in complexes with free energy scores lower than starting from the native complex (e.g. 3a4s, 2oob, and 1ak4).

In order to quantify the ability of the RS-REMD free energy score to differentiate between close to native poses from other poses, a selectivity of each funnel plot was calculated following an approach we introduced previously (see Chapters 5 and 6) [279, 276]. Basically, the selectivity measures the difference in scoring between the highest-ranked ligand pose at the binding site G'_{Site} ($\text{rmsd}_{\text{ligand}} < 8$ Å from the pose of minimal $\text{rmsd}_{\text{ligand}}$) from the highest-ranked ligand pose not at the binding site $G'_{NotSite}$. To this end, each score was first shifted by the mean value of all scores and divided by the minimum scoring value to achieve comparable results for the different protein-protein complex cases:

$$G'_i = \frac{G_i - \overline{G}}{G_{min}}. \tag{7.1}$$

Next, the selectivity is given by the normalized difference of the key poses:

$$Selectivity = G'_{Site} - G'_{NotSite}. \tag{7.2}$$

The resulting selectivity is a value between 1 (perfectly selective for the pose at binding site) and -1 (perfectly selective for pose not at the binding site). A value of 0 means that the scoring of the highest scored pose at the binding site was the same as the scoring of the highest scored pose not at the binding site.

The selectivity for each structure is given in Figure 7.6. RS score was able to identify the closest to native poses in 11 cases (selectivity higher than 0.1) and only in 4 cases the poses not at the binding site were scored considerably higher (selectivity lower than -0.1). For 5 structures the highest-ranked poses at the binding site and not at the binding site scored quite equally (selectivity between 0.1 and -0.1). A mean selectivity of 0.22 was measured for all structures, which is comparable to the selectivities achieved in a previous study using umbrella sampling for absolute binding free energy calculations in implicit solvent on the same benchmark set (0.28) [279]. Please note that a slightly different rmsd metric and conformational restraining methods were used in the previous study.

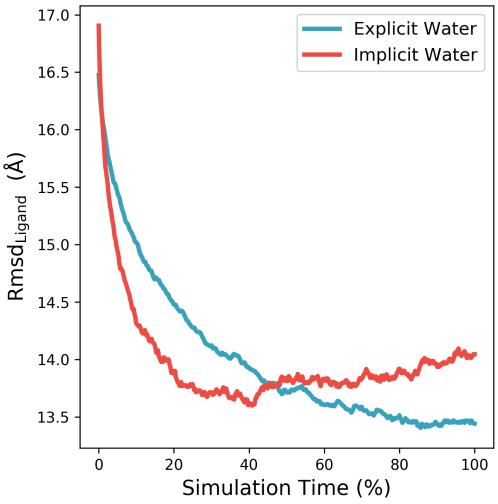


Figure 7.4: The rmsd$_{ligand}$ following the replica of minimal rmsd$_{ligand}$ against percentage of simulation time for the explicit (blue) and implicit (red) water simulations. The mean value of all simulated poses was considered. Due to differing sampling steps we interpolated linearly between the sampled values.

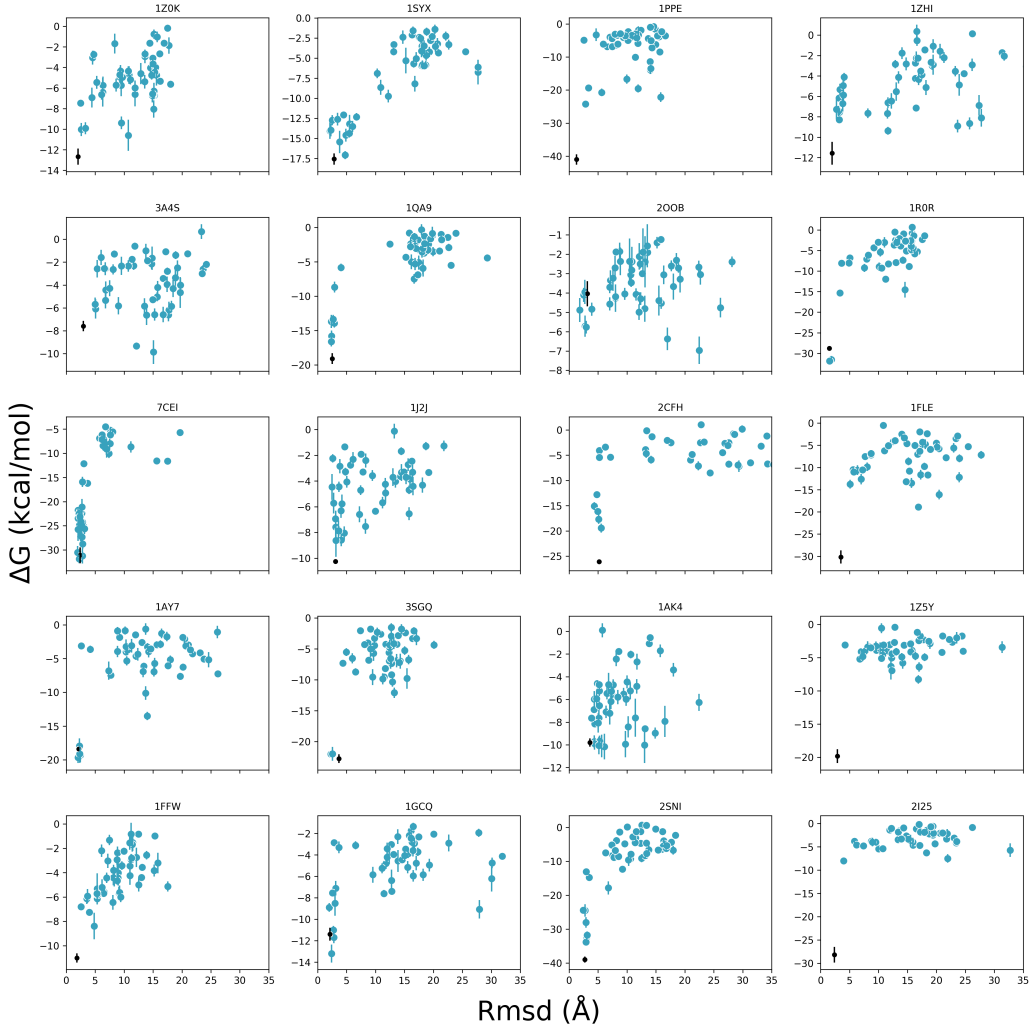


Figure 7.5: RS-REMD scoring of all 50 poses for 20 protein-protein cases against $\text{rmsd}_{\text{ligand}}$ from the native complex. The results of the explicit solvent RS-REMD simulations of the different poses (blue dots) and the native structures (black dots) are given with uncertainties.

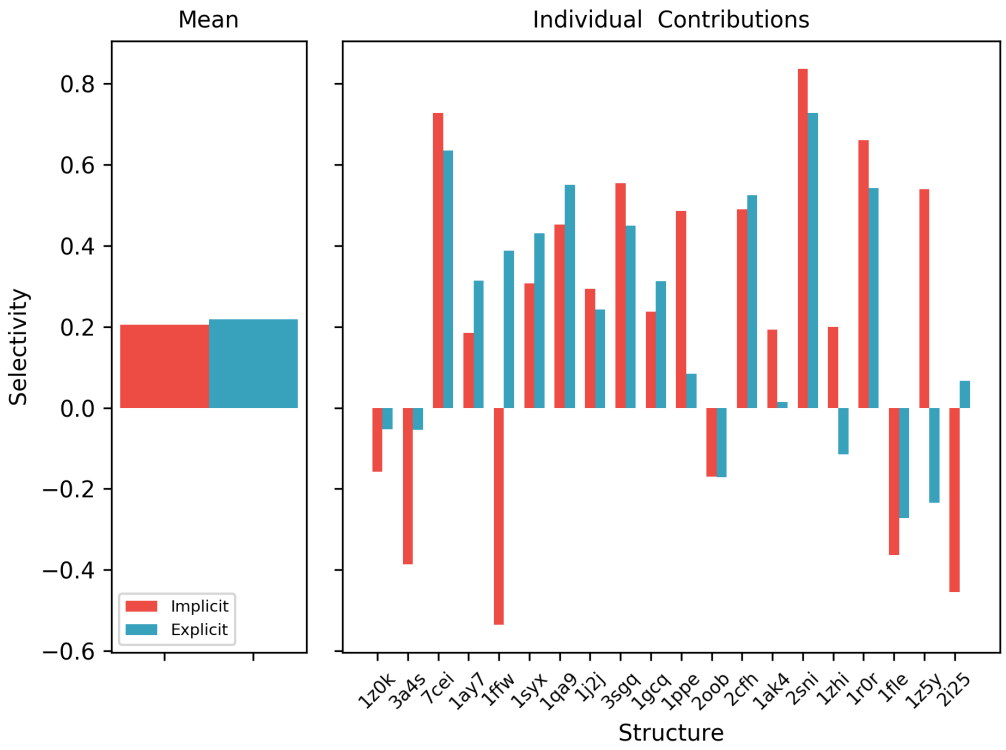


Figure 7.6: Selectivities of the explicit solvent (cyan) and implicit solvent (red) RS-REMD free energy scoring for all 20 protein-protein complex cases (right panel) calculated as described in the main text using equations 7.1 and 7.2. The mean selectivity of all cases reaches 0.20 (left panel) for the implicit solvent simulations and 0.22 in case of the RS-REMD in explicit solvent.

7.3.4 RS-REMD free energy scoring of protein-protein docking poses in implicit solvent

The ability of RS score to differentiate poses at the binding site from other poses was also studied in implicit solvent. The same benchmark set containing 50 poses for 20 protein-protein cases was tested as in the explicit water case. These simulations took approximately half of the time required for the explicit water simulations on one GPU (2.4 days per complex for every 1 ns simulation time of 16 replicas and 50 poses, see Appendix, Table C.3).

An improvement in $\text{rmsd}_{\text{ligand}}$ during the RS-REMD simulations for 67 % of the poses was observed (see Figure 7.3, right panel and Appendix, Figure C.4), a slightly worse performance than the explicit water case. Again, the minimum $\text{rmsd}_{\text{ligand}}$ of

all replicas was chosen in order to account for exchanges between the replicas. The development of the minimum $rmsd_{ligand}$ (mean over all simulated poses) along the simulation time shows the same minimum as in the explicit water case with a high decrease in the first 0.25 ns (see Figure 7.4, red line). Interestingly, the curve reaches the minimum already after 40 % of the simulation time (1 ns), much faster than in the explicit water case (4 ns).

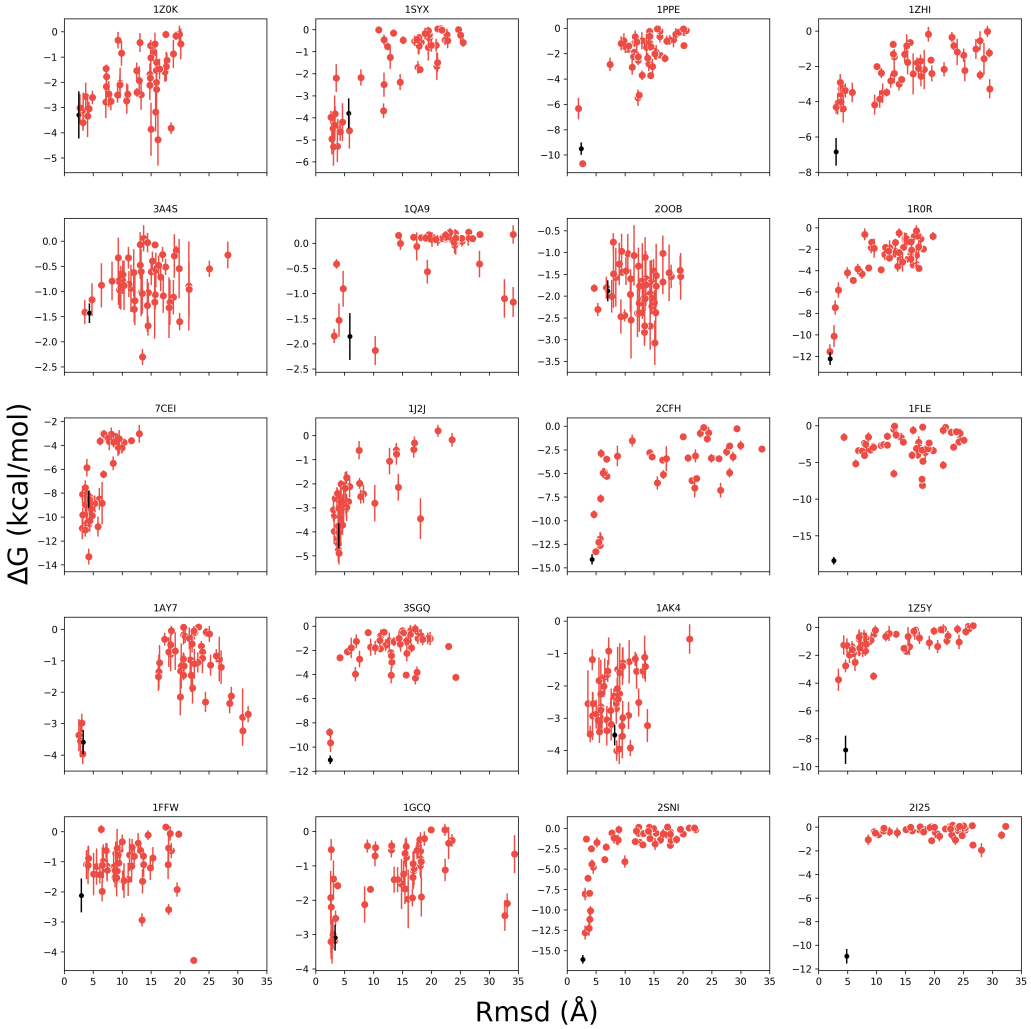


Figure 7.7: RS score of all 50 poses for 20 protein-protein complexes vs. $rmsd_{ligand}$ (replica of minimal $rmsd_{ligand}$ in the last frame) from the native complex. The results of the implicit solvent RS-REMD simulations of the different poses (red dots) and the native structures (black dots) are given with uncertainties.

A funnel-like shape of the RS score versus $rmsd_{ligand}$ from the native complex (minimal $rmsd_{ligand}$ of all replicas in the last frame taken as reference) can be observed for most of the structures (see Figure 7.7). The selectivity was calculated as described for the explicit water scoring. RS score resulted in a high preference for the highest scored pose at the binding site ($rmsd_{ligand} < 8$ Å from the pose of minimal $rmsd_{ligand}$) in contrast to the highest scored pose not at the binding site in 14 cases (selectivity higher than 0.1). A clear preference for poses not at the binding site was found in 6 cases (selectivity lower than -0.1). Overall, the mean selectivity is 0.20 and thus slightly worse than in the explicit solvent case.

7.3.5 Structural details leading to different selectivities

The identification of the native binding site was possible using the RS-REMD free energy scoring for several but not all complexes. Still, we also identified some cases with a medium or low selectivity for the native binding geometry, even in the most rigorous explicit solvent simulation case. The 3 cases 1fle, 1z5y, and 2i25 resulted in very favorable scores when starting from the native bound complexes but gave for all 50 decoy start structures a low-affinity score in the RS-REMD simulations (explicit and implicit solvent), although some decoys reached relatively small final $rmsd_{ligand}$. For these cases, loop rearrangements or misplacements of single side chains at the interface were found to cause the unfavorable scoring (see Appendix Figure C.5 for details). Since our REMD scheme does not drive loop or side chain transitions explicitly along the replicas (but only dissociation of transiently bound states to allow rearrangements) it might be difficult to sample such necessary interface rearrangements within the limited simulation time.

Besides these specific cases, it is also of interest to identify structural and interaction details that could be responsible for the low selectivity in several cases. The cases were split into three groups according to the selectivity (high selectivity: selectivity higher than 0.1; medium selectivity: selectivity between -0.1 and 0.1; low selectivity: selectivity lower than -0.1). The stability of salt bridge contacts is given by a delicate balance between large compensating solvation and Coulomb interaction contributions and are especially sensitive to force field parameterization. In case of low or medium selectivity, the model with the lowest $rmsd_{ligand}$ from the native binding site was selected (this model was not scored realistically). The model with the overall best scoring was considered in case of high selectivity of the structure (and thus a placement close to the native binding structure). An intermolecular salt bridge was identified if the distance between the carboxylate group (centered at the carbon atom) from an aspartic acid or a glutamic acid residue and the ammonium (centered at the nitrogen atom) of either a lysine or an arginine residue was lower than 6 Å. The mean counts and standard deviation were calculated for the second half of the simulation following the replica with lowest $rmsd_{ligand}$ for the explicit

and implicit water cases (see Figure 7.8). Moreover, the number of salt bridges for the native bound complexes were counted.

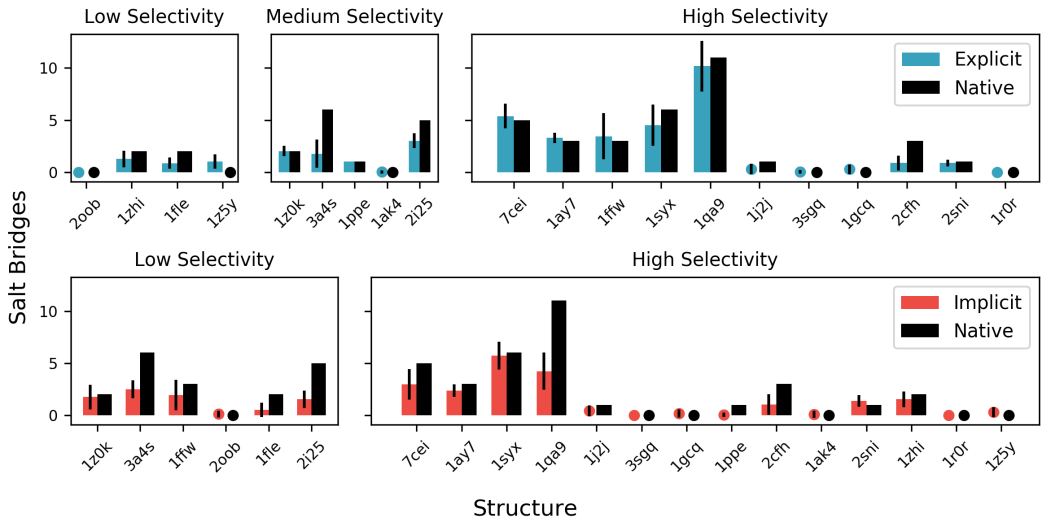


Figure 7.8: The counts of simulated salt bridges (distances lower than 6 Å were defined as contacts for the salt bridges) at the binding site of ligand and receptor proteins for the explicit water simulations (blue; upper row) and implicit water simulations (red; upper row) are compared to the counts of salt bridges in the native complex for each structure (indicated as pdb-id). The mean counts over the last half of the simulations (following the replica with minimum rmsd) with corresponding standard deviation were evaluated. The counts for the structures are separated into three subfigures according to their selectivity (high selectivity: selectivity higher than 0.1; medium selectivity: selectivity between -0.1 and 0.1; low selectivity: selectivity lower than -0.1). In case of low or medium selectivity, the pose with minimal $\text{rmsd}_{\text{ligand}}$ was shown, in case of high selectivity the counts for the best-scored pose of each structure were displayed. A circle was assigned if the number of counts was below 0.5.

For 9 of the 20 structures no or only one salt bridge was found at the native interface but for several structures, a higher number of salt bridges up to 11 were counted (see black bars in Figure 7.8). The structures of low selectivity in the implicit water case seem to have a higher number of native salt bridges (mean count of 3.0) while the low selectivity explicit water cases are dominated by structures with a lower number of native salt bridges (mean count of 1.0). Apparently, the implicit water model has more difficulties in scoring the native binding site correctly if a higher number of charged interactions are involved in stabilizing the binding

geometry. On the other hand, if the number of salt bridges at the interface was below 2 in the implicit case the native binding site was identified correctly, except for the structure 2oob. Interestingly, also some structures with a higher number of salt bridges were scored with high selectivity. Thus, it is not the number of native intermolecular salt bridges alone that leads to scoring of high or low selectivity. Three of the implicit solvent structures with low selectivity and a high number of salt bridges are assigned a medium selectivity in the explicit water case (1z0k, 3a4s, 2i25). These structures are scored better with the explicit water model but still are difficult cases to predict.

Looking at the difference between simulated and native number of salt bridges the explicit water case has a higher tendency to stabilize the correct number of salt bridges ($-1 \leq \text{Count}_{\text{Explicit}} - \text{Count}_{\text{Native}} \leq 1$ in 75 % of the cases) while the implicit water representation is more likely to underestimate the number of salt bridges at the interface ($\text{Count}_{\text{Implicit}} - \text{Count}_{\text{Native}} < -1$ in 35 % cases).

Overall, the analysis of these limited cases indicates that the better performance of the RS-score with an explicit solvent model in predicting the native binding site could in part be due to better representation of the salt bridges at the native binding site. However, it requires further analysis of a larger set of structures.

7.4 Conclusion and outlook

The realistic in silico prediction of protein-protein binding structures and affinities is of importance to better understand many cellular processes. Based on our earlier work in which we introduced repulsive scaling (RS-)REMD simulations (Chapter 6) [279], we have extended the RS-REMD method to improve docked binding geometries in explicit solvent and to estimate binding affinities. The ability of RS-REMD free energy scoring to predict native binding affinities was explored for 36 protein-protein cases in explicit and implicit solvent. The explicit water simulations gave a quite high correlation to experimental binding free energies (0.77) with better performance than the implicit water case (0.55). Note that in the latter case conformational restraints (with respect to the unbound partner structures) were included to avoid unfolding or large conformational changes of protein partners during the simulations. This may in part influence the scoring results (no restraints were used in explicit solvent). The majority of current scoring schemes involve energetic evaluation of single structures or ensembles of docked structures (reviewed in Chapter 4 and in [277]). The current scheme is computationally demanding but results in a free energy like score that includes contributions due to flexibility of the partners as well as due to the surrounding solvent. Only a few recent methods explore related free energy scoring schemes [244, 279, 243]

An advantage of our present RS-REMD scoring approach is that it allows simulta-

neous improvement of the structure of a docked decoy (improvement of the RMSD with respect to the native complex) and free energy type scoring. For a benchmark set of 50 pre-docked poses for 20 protein-protein complexes overall a significant improvement of the docking geometries relative to the native complex was obtained especially for initial docking decoys in the vicinity of the native complex structure. Although for the majority of cases (14 out of 20 in explicit solvent) a near-native structure was found as the best scoring refined decoy for some cases it failed to best score a near-native structure. This can be attributed to force field inaccuracies but in several cases, it is likely due to significant conformational differences between unbound and bound structures at the interface that were too large to be sampled during our RS-REMD simulations. In future efforts, it might be possible to include additional biases to promote conformational transitions at interface regions to also enhance the sampling of such conformational changes. The approach could also be useful to investigate weak transient protein-protein interactions that are often difficult to solve experimentally or to more realistically evaluate docked structures based on homology to a similar protein-protein complex.

8 Accurate Refinement and Calculation of the Absolute Binding Free Energy of Small Ligand Molecules to Proteins Using Replica Exchange With Repulsive Scaling

The repulsive scaling scheme introduced in Chapter 6 for protein-protein complexes has a simple implementation that leads to a straightforward application of the method to other biomolecular assemblies. In the current chapter, we will evaluate RS-REMD for protein complexes with small ligands. We will further extend the ΔG score of Chapter 7 to yield absolute binding free energies and we will demonstrate how the method can be applied in the context of blind docking studies. In the future, RS-REMD may be applicable in drug design campaigns that could prove to be useful to the pharmaceutical industry as well.

8.1 Introduction

Successful characterization of ligands binding to a target protein and identification of high affinity binders, while preserving the general properties of the molecules, is the aim of computer aided drug discovery. The most rigorous in silico approaches to estimate the binding free energy rely on physics based methods using Molecular Dynamics and Monte Carlo approaches. Such approaches are getting increasing attention in the commercial sector [268, 262] due to progress in force field development [313, 124, 298], algorithms [316, 259] and computational hardware [267, 259].

In the last decades a multitude of in silico methods to calculate the binding affinities of bimolecular systems have been developed [277]. In order to access the relative binding free energy of small molecules, free energy perturbation (FEP) approaches are well suited that imply alchemical transformations from one ligand to another [316]. The FEP scheme can be generalized, using additional constraining simulations, to calculate the absolute binding free energy in a double decoupling method (DDM) [102, 30]. Aldeghi and coworkers achieved a high accuracy of such a decoupling

scheme for a Bromodomain target with a mean absolute error of 0.6 kcal/mol [7, 8]. Another method to calculate the binding free energy is by introducing a physical separation of ligand and receptor along a pathway. This can be achieved using the Jarzynski relationship and applying non-equilibrium dynamics or by computing the potential of mean force in umbrella sampling simulations. Using a method introduced by Woo and Roux such umbrella sampling simulations can be extended to yield an absolute binding free energy by constraining the relative orientation and conformation of ligand and receptor [324, 279]. Deng et al. compared the accuracy of DDM with the Woo and Roux scheme on charged ligands and obtained similar accuracy in both methods [70].

Apart from calculating the binding affinity of ligands it is of high importance to predict correct binding modes. Lately, structure based virtual screening predictions were successfully conducted with docking on an ultra-large ligand library resulting in new chemotypes that were experimentally validated [195]. Overall, many docking software packages are available [307, 93, 227, 58], in particular Auto-dock vina showed to be quite successful for native docking experiments [299]. Still, these docking approaches are not completely rigorous and have limitations especially regarding their scoring functions [142]. In principle, full atomistic MD simulations provide the method of highest rigor to dynamically model bimolecular binding, incorporating explicit solvent effects, entropic contributions and partner flexibility [277]. The association process of a ligand molecule to its target was elucidated in unguided MD simulations by the D.E. Shaw lab [266]. Further, to distinguish stable from unstable ligand poses short MD simulations were shown to be effective [179]. In a recent study, Guterres et. al used MD simulations to distinguish active from decoy ligands for a large benchmark set of 560 small ligands by evaluating the ligand stability. They achieved a 22 % improvement of ROC AUC compared to AutoDock Vina results with a moderate refinement of the binding modes [122].

In the field of protein-protein docking the efforts to predict native protein binding sites from MD simulations alone have intensified [240, 276, 238]. A replica-exchange based repulsive scaling (RS-REMD) scheme successfully predicted the native binding site for five protein-protein cases using full partner flexibility and an implicit solvent model (Chapter 6) [276]. Recently, the RS-REMD method was extended to simultaneously refine protein-protein complexes and yield a realistic free energy score in explicit solvent (Chapter 7) [278]. The aim of this study is to apply RS-REMD for the first time on a benchmark set of 24 protein-ligand structures. We will show that complete ligand association is possible with RS-REMD in explicit solvent, starting from a worst case scenario of the original placement. Moreover, we extend the RS-REMD approach to yield absolute binding free energies with a quite good correlation to experimental affinities. We further show that repulsive scaling refinement of ligand placements in the vicinity of the binding site is efficient and a RS-REMD absolute binding free energy score predicts the near native structures

correctly for most cases. Finally, the performance of the repulsive scaling scheme in a fully blind docking context using single point MMGBSA scores is discussed.

8.2 Materials and methods

8.2.1 RS-REMD simulations to estimate the absolute binding free energy of native protein-ligand complexes

As a benchmark set two protein systems were analyzed in this study,the Fk1 domain of FKBP51 (Fkbp) and the first bromodomain of human BRD4 (Br4). In both protein systems the ligand binding site is not deeply buried and the receptor molecules are relatively rigid. For these proteins a large number of ligands with experimental binding affinities and structural models are available in the PDB. In the case of Fkbp 14 small ligands were selected with a broad range of binding affinities. For the bromodomain system 10 ligands were used that were evaluated as a benchmark set in a recent study by Aldeghi and coworkers (incorporating only the ligands for which structural data were available) [7].

The MD simulations throughout this study were conducted using the amber18 [41] software suite with a tip3p explicit water model and the ff14sb [198] protein force field. The ligand parameters were generated using antechamber with AM1-BCC charges and the gaff2 [298, 313] force field. The systems were solvated in an octahedron box with 15 Å minimum distance of the solute to the box edge. After minimization (1000 steps steepest descent) the systems were heated to 300 K and equilibrated for 16 ps. Repulsive scaling replica exchange molecular dynamics (RS-REMD) simulations were performed for 5 ns (per replica) on the equilibrated native structures using 16 replicas with increasing repulsive bias of the Lennard-Jones parameters between ligand and receptor atoms (parameter set given in Table 8.1). Half parabolic COM distance restraints (force constant 15 kcal $mol^{-1}Å^{-2}$) of the C_α atoms between the ligand and the receptor interface were introduced that restricted the accessible volume of the ligand to a sphere of 20 Å radius around the receptor interface. The accessible volume of the ligand was estimated for one snapshot by counting the water molecules in the spherical volume. The second half of the RS-REMD simulation was analysed as the production run, which was split into five parts to give uncertainty estimates. In order to calculate the binding free energy the repulsive biases introduced in each replica were calculated using trajectory reevaluation with pytraj. The total energy (sum of Lennard-Jones and Coulomb interaction) of the ligand and the protein was reevaluated for each replica with each lj parameter set, following the perturbation approach introduced recently (Chapter 7) [278]. The bias between the reference state and the actual state was calculated taking the corresponding difference of the total energy. To reduce the computational

Table 8.1: Lennard-Jones scaling parameters for the different RS-REMD simulation setups. For the repulsive scaling absolute binding free energy calculations the 16 replica scheme was used (column 2 and 3). In the refinement and association simulations the 8 replica setup was used (column 4 and 5).

Replica Number	16 replicas		8 replicas	
	d(Å)	e	d(Å)	e
1	0.0	1.0	0.0	1.0
2	0.01	0.99	0.015	0.99
3	0.02	0.98	0.03	0.985
4	0.04	0.97	0.045	0.98
5	0.08	0.96	0.06	0.97
6	0.12	0.94	0.075	0.96
7	0.16	0.92	0.1	0.94
8	0.2	0.9	0.15	0.9
9	0.24	0.88		
10	0.28	0.86		
11	0.32	0.84		
12	0.38	0.82		
13	0.44	0.8		
14	0.5	0.78		
15	0.58	0.76		
16	0.68	0.74		

effort 50 snapshots per replica were considered. Finally, the statistically optimal estimate of the free energy of binding was calculated using the calculated biases with pymbar (see Section 3.4.1 for details on the MBAR method) [271].

8.2.2 RS-REMD simulation of protein-ligand association starting from an ensemble of incorrect binding poses

For both proteins the performance of longer RS-REMD simulations to predict the native binding site for one ligand (pdb-id 3u5j in case of Br4, pdb-id 3o5r for Fkbp) was tested. Around each protein 16 ligand poses were generated with AutoDock Vina, excluding the ligand placements close to the native binding site. This test case corresponds to a scenario in which only wrong binding sites are predicted by the docking program. The individual placements were prepared, heated to 300 K and equilibrated as described in section 8.2.1. Repulsive scaling REMD simulations were performed using 8 replicas (see Table 8.1 for lj repulsive scaling parameters) for 300ns

(in case of 3u5j) and 600 ns (in case of 3o5r) per replica. The 8 binding poses with the highest MMGBSA score (only last frame evaluated) were used as starting structures for each replica. Additionally, 8 regular MD simulations of the same length and starting from the same placements were performed for both protein-ligand systems.

8.2.3 RS-REMD refinement of ligand poses in the vicinity of the binding site

Ten binding poses in the vicinity of the binding site were generated for both proteins and the respective 13 and 10 ligands. The placements were obtained through short MD simulations with a repulsive COM distance restraint between ligand and protein applied, starting from the different replicas of the RS-REMD scoring simulations of the native structures. Each ligand pose was restricted to be placed in a spherical volume of 20 Å around the receptor interface. Like this, a docking scenario in which the binding site of the protein is approximately known is mimicked, for which still some refinement of the ligand placement is possible. For the 10 poses, refining RS-REMD simulations of 10 ns (per replica) were performed using 8 replicas (see table 8.1 for lj scaling parameter set). Moreover, 8 regular MD simulations starting from the same placement with the same simulation time as the RS-REMD replicas were conducted. Moreover, one regular MD simulation with a simulation time of 80 ns (amounting to the same simulation time as all replicas of the RS-REMD simulations) was executed from the identical starting structure.
Finally, from the resulting structure of the RS-REMD refinement, a RS-REMD absolute binding free energy scoring was performed as described in section 8.2.1. To account for occasional exchanges in the reference replica (no bias between ligand and receptor) different models are evaluated to estimate the replica from which the scoring simulations are started. The best case scenario, the replica of minimal $\text{rmsd}_{\text{ligand}}$, is evaluated against the replica of highest MMGBSA score, the replica of minimal distance between ligand and receptor and the reference replica. These parameters were calculated using pytraj.

8.3 Results and discussion

8.3.1 Evaluation of the RS-REMD absolute binding free energy on native protein-ligand complexes

Repulsive scaling (RS-REMD) simulations were performed to calculate the absolute binding free energy for a benchmark set of 14 ligands for Fkbp and 10 ligands for Br4. To obtain the binding free energy we adopted an approach introduced recently for protein-protein complexes. Basically, we apply repulsive potentials between ligand and receptor for each replica that lead to a dissociation of the ligand from its binding

site. The introduced repulsive bias in each replica is calculated using trajectory reevaluation. From these biases the free energy difference between the associated and the dissociated states is calculated with a perturbation approach using mbar. The ligand was restricted to a spherical shell volume around the receptor binding site, so that a standard state correction could be applied, resulting in an absolute binding free energy.

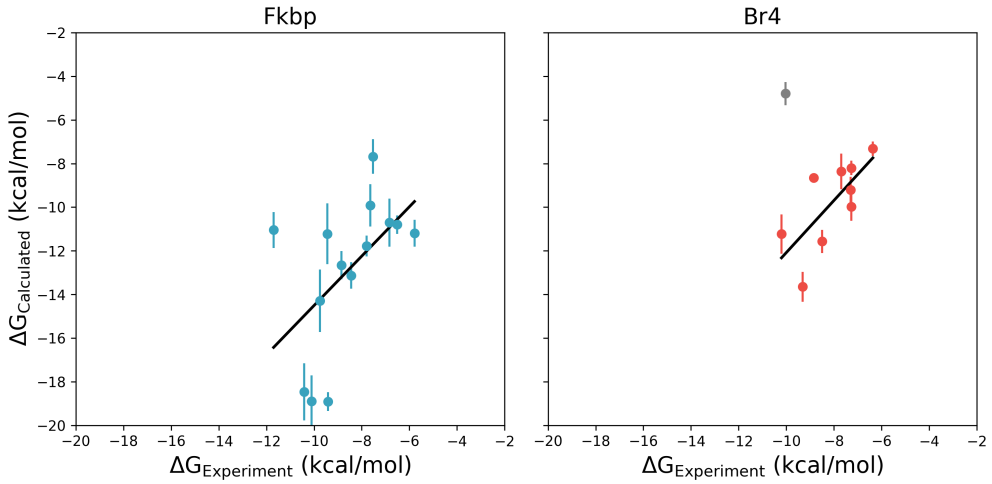


Figure 8.1: RS-REMD calculated absolute binding free energy against experimental binding affinity for the native ligand-protein complexes. On the left panel the 14 complexes of Fkbp, on the right panel the 10 complexes of Br4 are given with the corresponding linear fit. For Fkbp a correlation coefficient of 0.54, in case of Br4 a coefficient of 0.72 was achieved. The result for the structure 3mxf of Br4 was omitted due to its poor performance (given as gray dot) which was probably due to an unfavourable starting position as discussed in the main text.

The introduced repulsive biases resulted in a clear separation of associated and dissociated states as shown in Figures 8.3 and 8.2. The histograms of the COM distance distribution between ligand and receptor heavy atoms of the associated states (replica of lowest $\text{rmsd}_{\text{ligand}}$ in each frame) and dissociated states (replica of highest COM distance in each frame) are given for every native structure.

Table 8.2: Experimental affinities and calculated binding free energies with uncertainties in kcal/mol for the protein Br4 (each structure indicated by pdb-id). In brackets the corresponding references are given. The difference $\Delta G_{Diff} = \Delta G_{Calc} - \Delta G_{Exp}$ is given in the last column.

Structure	ΔG_{Exp}	ΔG_{Calc}	ΔG_{Diff}
4mr4 [247]	-7.3 [256, 20]	-8.2 $\pm$ 0.3	-0.9 $\pm$ 0.3
4mr3 [247]	-7.3 [256, 20]	-10.0 $\pm$ 0.7	-2.7 $\pm$ 0.7
3u5l [85]	-8.5 [85]	-11.6 $\pm$ 0.5	-3.1 $\pm$ 0.5
4ogi [55]	-10.2 [55]	-11.2 $\pm$ 0.9	-1.0 $\pm$ 0.9
3mxf [86]	-10.0 [86]	-4.8 $\pm$ 0.5	5.2 $\pm$ 0.5
3u5j [85]	-7.7 [85]	-8.4 $\pm$ 0.8	-0.7 $\pm$ 0.8
4j0r [125]	-8.8 [125]	-8.7 $\pm$ 0.2	0.2 $\pm$ 0.2
4ogj [55]	-9.3 [229]	-13.6 $\pm$ 0.7	-4.3 $\pm$ 0.7
4hbv [88]	-6.4 [88]	-7.3 $\pm$ 0.3	-1.0 $\pm$ 0.3
3svg [84]	-7.3 [125]	-9.2 $\pm$ 0.6	-1.9 $\pm$ 0.6

Table 8.3: Experimental affinities and calculated binding free energies with uncertainties in kcal/mol for each structure (given as pdb-id) of the protein Fkbp. In brackets the corresponding references are given. The difference $\Delta G_{Diff} = \Delta G_{Calc} - \Delta G_{Exp}$ is given in the last column.

Structure	ΔG_{Exp}	ΔG_{Calc}	ΔG_{Diff}
3o5r [33]	-9.7 [96]	-14.3 $\pm$ 1.4	-4.5 $\pm$ 1.4
4tx0 [28]	-9.4 [28]	-11.2 $\pm$ 1.4	-1.8 $\pm$ 1.4
4jfj [317]	-7.8 [317]	-11.8 $\pm$ 0.5	-4.0 $\pm$ 0.5
4jfm [317]	-7.5 [317]	-7.7 $\pm$ 0.8	-0.2 $\pm$ 0.8
5div [96]	-9.4 [96]	-18.9 $\pm$ 0.4	-9.5 $\pm$ 0.4
4drk [106]	-7.6 [317]	-9.9 $\pm$ 1.0	-2.3 $\pm$ 1.0
4jfl [317]	-6.8 [317]	-10.7 $\pm$ 1.1	-3.9 $\pm$ 1.1
4jfk [317]	-8.8 [317]	-12.7 $\pm$ 0.6	-3.8 $\pm$ 0.6
4tw7 [97]	-10.4 [97]	-18.5 $\pm$ 1.3	-8.1 $\pm$ 1.3
5dit [82]	-8.4 [82]	-13.1 $\pm$ 0.6	-4.7 $\pm$ 0.6
4drh [202]	-11.7 [317]	-11.0 $\pm$ 0.8	0.7 $\pm$ 0.8
4jfi [317]	-6.5 [317]	-10.8 $\pm$ 0.4	-4.3 $\pm$ 0.4
5diu [96]	-10.1 [96]	-18.9 $\pm$ 1.2	-8.8 $\pm$ 1.2
4tw6 [97]	-5.8 [97]	-11.2 $\pm$ 0.6	-5.4 $\pm$ 0.6

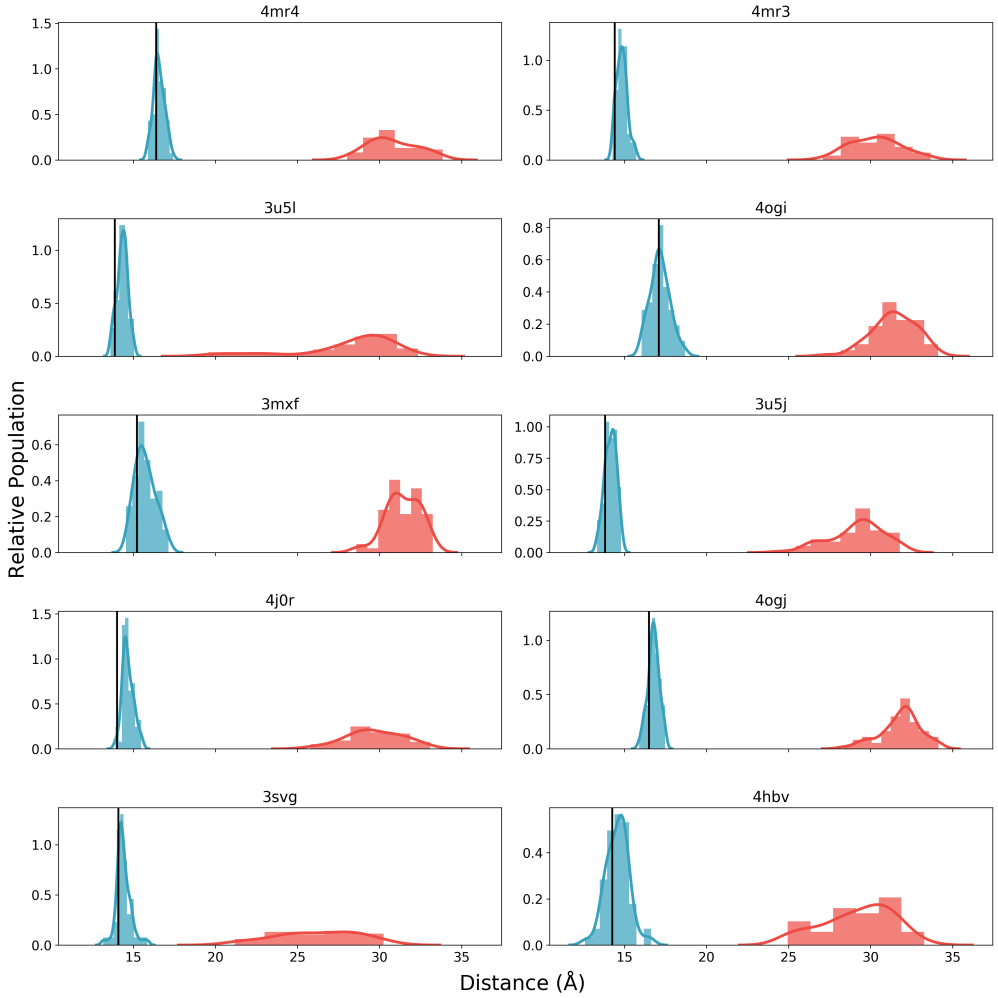


Figure 8.2: Histograms of COM distance between ligand and receptor for the associated state (replica of minimum $\text{rmsd}_{\text{ligand}}$ in blue) and the dissociated state (histogram of the replica of maximum distance in red) for the 10 ligands of Br4 (indicated as pdb-id). Only the second half of the RS-REMD simulation was taken into account. The native COM distance is given by the black vertical line.

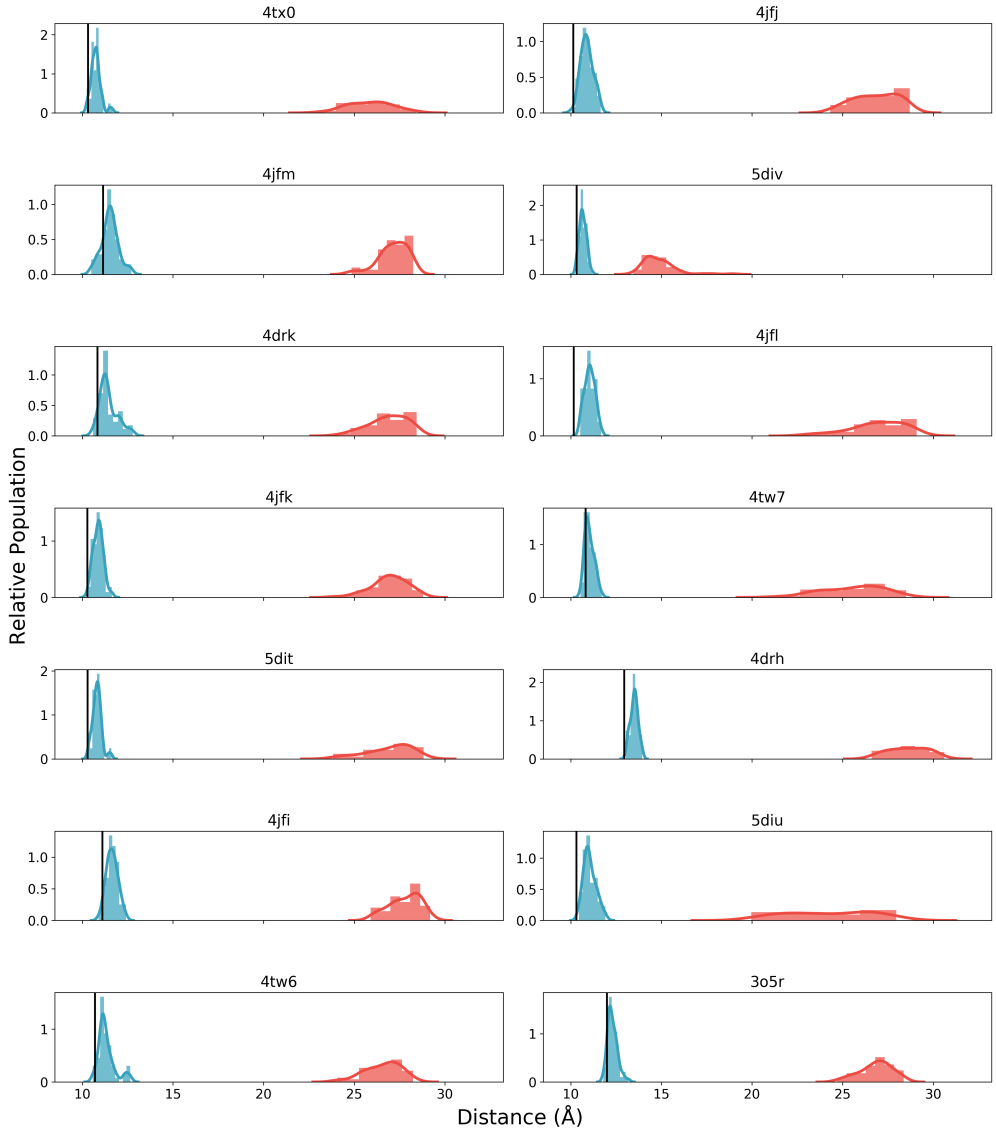


Figure 8.3: Histograms of COM distance between ligand and receptor for the associated state (replica of minimum $\text{rmsd}_{\text{ligand}}$ in blue) and the dissociated state (histogram of the replica of maximum distance in red) for the 14 ligands of Fkbp (indicated as pdb-id). Only the second half of the RS-REMD simulation was taken into account. The native COM distance is given by the black vertical line.

The calculated binding affinities are given Figure 8.1 for both protein systems. Pearson correlation coefficients of 0.72 for Br4 and 0.54 for Fkbp to the experimental

binding free energies were achieved. Due to the small $\Delta\Delta G$ range of the individual ligands ($\Delta\Delta G = \Delta G_{Experiment,max} - \Delta G_{Experiment,min} = 5.9$ kcal/mol for Fkbp and 3.8 for Br4) it is challenging to obtain high correlations for these systems. In the case of Br4 we did not consider the result of 3mxf with a very low affinity of only -4.8 kcal/mol (shown as gray dot). This value differed considerably from the experimental ΔG measurement and is also inconsistent with the binding affinities of the RS-REMD docking results (see section 8.3.3). The poor performance is probably due to an unfavorable ligand placement from which the simulations were started. A slightly varied ligand placement, obtained from a RS-REMD docking simulation, resulted in a two-fold increase in the calculated binding free energy.

The absolute values of the binding affinities were in quite good agreement with the experimental data in case of Br4. For 5 structures the calculated binding affinities had 1.0 kcal/mol difference or lower (see Table 8.2). These results are comparable in accuracy to the binding affinities calculated using a double decoupling scheme (DDM) by Aldeghi and coworkers [7]. In the case of Fkbp, the affinities were overestimated slightly ($\overline{\Delta G_{Diff}} = \overline{\Delta G_{calc} - \Delta G_{Exp}} = -4.3$ kcal/mol) so that only 3 structures had affinities within 2 kcal/mol.

8.3.2 RS-REMD simulation of protein-ligand association starting from an ensemble of incorrect binding poses

Extensive MD simulations of the ligand association were performed for one ligand of each protein (3u5j for Br4 and 3o5r for Fkbp). The ability to sample the native binding site was compared between RS-REMD and regular MD simulations. For both approaches the equal amount of simulation time (between 300ns and 600 ns) was conducted starting from an ensemble of 8 poses that were placed at incorrect binding sites using AutoDock Vina [299]. This test layout characterizes a worst case scenario in which the docking program only generates false positive solutions. Thus the capacity of the MD schemes to overcome the pre-set incorrect binding minima can be investigated. RS-REMD was performed with 8 replicas and also regular MD simulations with differing starting conditions were executed.

The $\text{rmsd}_{\text{ligand}}$ against simulation time can be compared for all individual replicas for 3o5r and 3u5j in Figures 8.5 and 8.4, respectively. In both cases the RS-REMD simulations were able to sample the native binding site with $\text{rmsd}_{\text{ligand}}$ values around 5 Å. In case of the regular MD simulations for 3u5j the binding site was not sampled at all (lowest $\text{rmsd}_{\text{ligand}}$) and for 3o5r the ligand was sampled in a stable position close to the binding site ($\text{rmsd}_{\text{ligand}}$ around 10 Å) in two simulations. The simulations stuck at local minima in many simulations. These observations are in compliance with a recent study of protein-protein association using RS-REMD with an implicit solvent model (see Chapter 5) [276].

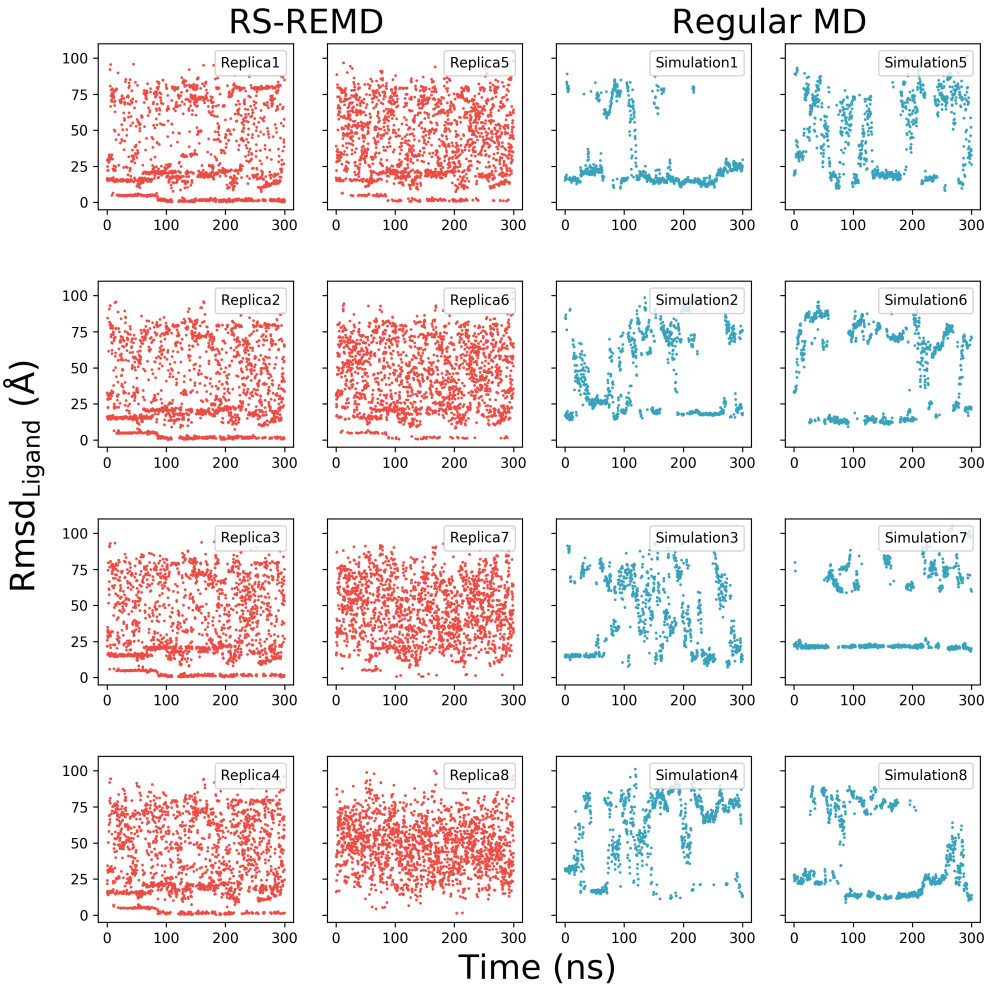


Figure 8.4: Development of the $rmsd_{ligand}$ for the simulation of ligand association to Br4 (pdb-id 3u5j). The results for every replica of RS-REMD (red) and every simulation of the regular MD case (blue) are given.

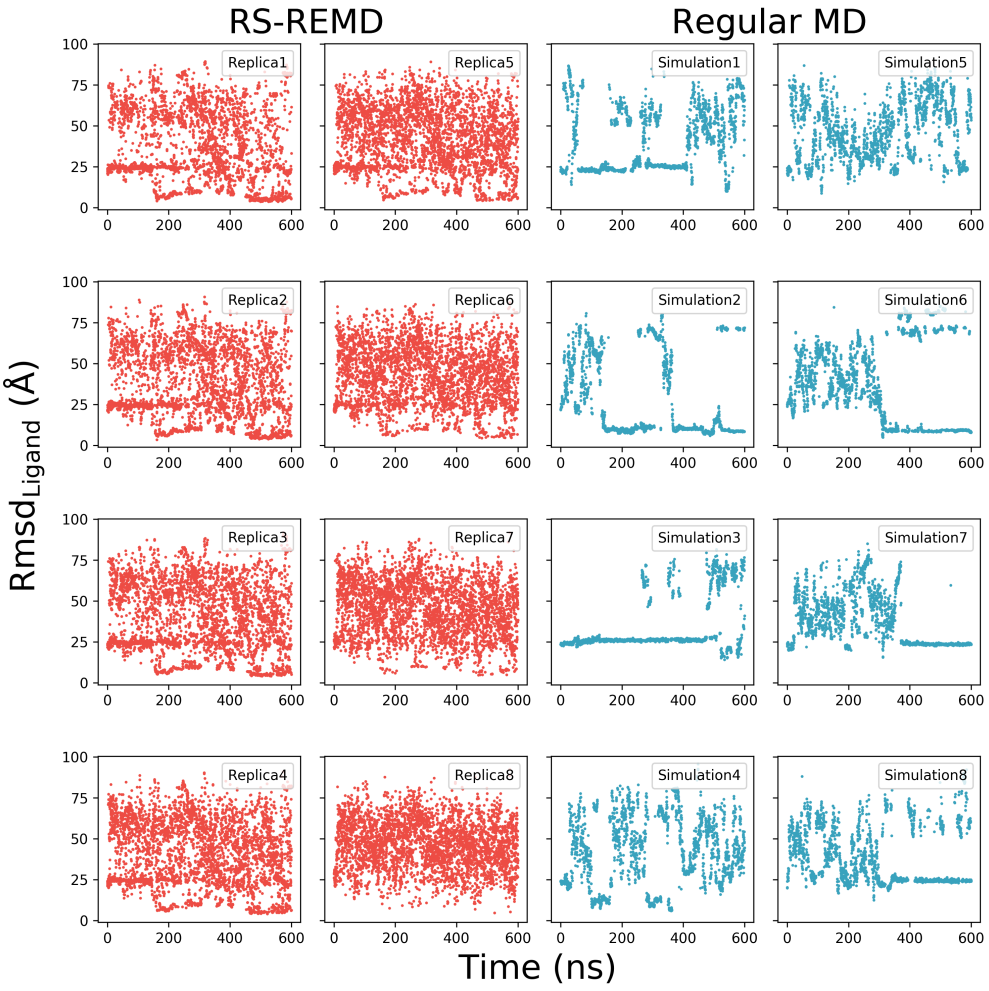


Figure 8.5: Development of the $rmsd_{ligand}$ for the simulation of ligand association to Fkbp (pdb-id 3o5r). The results for every replica of RS-REMD (red) and every simulation of the regular MD case (blue) are given.

The development of the $rmsd_{ligand}$ of the replica with lowest $rmsd_{ligand}$ (accounting for exchanges between the replicas) for both protein cases is given in Figure 8.6. A multi-step process to identify the native binding site is observable, with different residence times for the local minima. For the RS-REMD simulations of 3u5j occasional dissociations of the ligand from the binding site can be witnessed, followed by reassociations. However, the ligand stayed most of the time (over 200 ns) at the binding site ($rmsd_{ligand}$ below 5 Å) with a minimum $rmsd_{ligand}$ of 0.3 Å. Such a small $rmsd_{ligand}$ was possible due to the small size and little flexibility of

the ligand. In case of the regular MD simulations the native binding site was not encountered at all, the smallest $rmsd_{ligand}$ was 7.4 Å that was reached after 275 ns.

For 3o5r, binding placements in closer proximity to the native binding site ($rmsd_{ligand}$ below 10 Å) were encountered slightly faster in the regular MD simulations (after 102 ns) than with RS-REMD (after 152 ns). Still, in the regular MD simulations the ligand is stuck in $rmsd_{ligand}$ regions between 5 and 10 Å, whereas the RS-REMD simulations accumulated over 50 ns at the native binding site ($rmsd_{ligand} <$ 5 Å) with the first encounter after 161 ns. Overall, the lowest measured $rmsd_{ligand}$ in case of RS-REMD was 3.5 Å, for regular MD it was 4.9 Å. The latter was the only snapshot for the regular MD simulations with an $rmsd_{ligand}$ below 5 Å. Interestingly, in the RS-REMD simulation the ligand was not bound firmly after the first encounter of the native binding site, which lastet around 60 ns. After a period of multiple binding and unbinding events (240 ns) a ligand placement around 5 Å $rmsd_{ligand}$ was explored that was stable for over 100 ns.

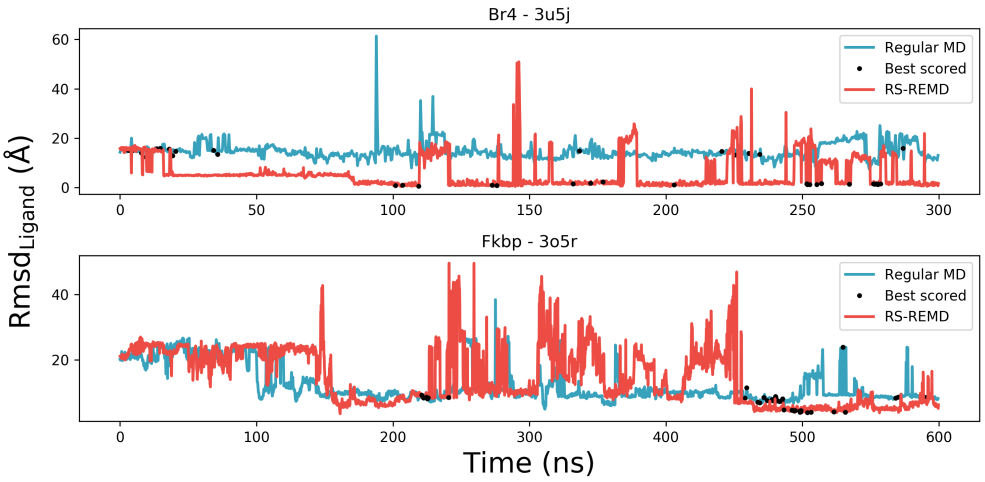


Figure 8.6: $Rmsd_{ligand}$ against simulation time (result for the replica with minimum $rmsd_{ligand}$ in every snapshot) for the RS-REMD (red) and the regular MD simulations (blue). The results for Br4 (pdb-id 3u5j) are given in the upper panel for Fkbp in lower panel (pdb-id 3o5r). The black dots mark the 20 snapshots of highest single point MMGBSA score.

For a blind docking scenario it is of interest how well a successfully captured binding site is predicted using a scoring function. As it turns out, a simple MMGBSA single point scoring of the trajectory is sufficient to predict the snapshots of correct associations. The 20 best scored snapshots are given as black dots in Figure 8.6. For both cases the frames of very low $rmsd_{ligand}$ for the RS-REMD simulations are

selected. In case of 3o5r the best scored snapshot of the RS-REMD simulations has an $rmsd_{ligand}$ of 5.7 Å (9.0 Å for the regular MD) and for 3u5j the best scored $rmsd_{ligand}$ was 1.4 Å (14.0 Å for the regular MD). Interestingly, using such a scoring scheme it is also possible to identify the parts in the simulations of occasional dissociation, which are given by a substantially lower binding score (see Figure 8.7).

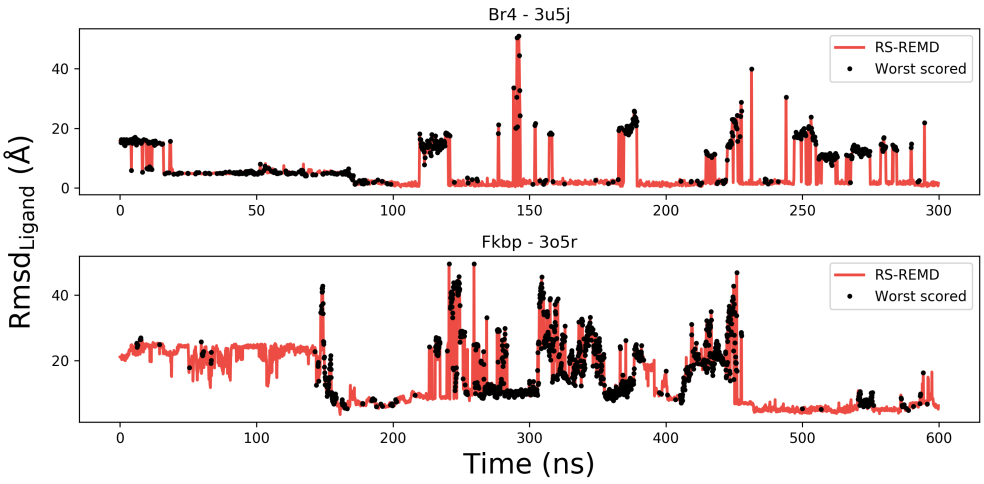


Figure 8.7: $Rmsd_{ligand}$ against simulation time (result for the replica with minimum $rmsd_{ligand}$ in every snapshot) for the RS-REMD (red) simulations. The results for Br4 (pdb-id 3u5j) are given in the upper panel for Fkbp in the lower panel (pdb-id 3o5r). The 30 percent of snapshots with the lowest single point MMGBSA scoring are shown as black dots.

8.3.3 RS-REMD refinement of ligand poses in the vicinity of the binding site

The ability of different MD refinement schemes to predict the native ligand binding pose starting in the vicinity of the binding site (less than 20 ÅCOM distance separation to the interface of the native binding site and $rmsd_{ligand}$ up to around 20 Å) are evaluated. For each of the 10 (for Br4) and 14 (for Fkbp) ligands 10 ligand placements were considered. For the refinement simulations RS-REMD was compared to multiple regular MD simulations (the same number of simulations as replicas for RS-REMD but unaltered parameter set and different starting conditions in each simulation) and a single regular MD simulation of the same total simulation length of 80 ns (starting from the identical ligand placement in each scheme). The $rmsd_{ligand}$ of the first and the last frame of all structures for three MD refinement methods can be compared in Figure 8.8. For the schemes using multiple simulations

(RS-REMD and regular multiple MD) the replica of lowest $rmsd_{ligand}$ was considered for the two evaluated frames. A clear improvement of the $rmsd_{ligand}$ (red dots) for RS-REMD (94 % for Br4, 89 % for fkbp) and regular multiple MD (95 % for Br4, 97 % for fkbp) can be witnessed that was especially higher than for the single regular MD simulation (58 % for Br4, 62 % for fkbp). In contrast to the scenario in which association simulations from local minima of the receptor were carried out (see section 8.3.2), the regular multiple MD simulations performed substantially better in particular the sampling of the native binding site was equally good (for Br4) or slightly better than using RS-REMD. This is probably due to a low number of alternative sticky minima around the vicinity of the binding site that had to be overcome in this test case.

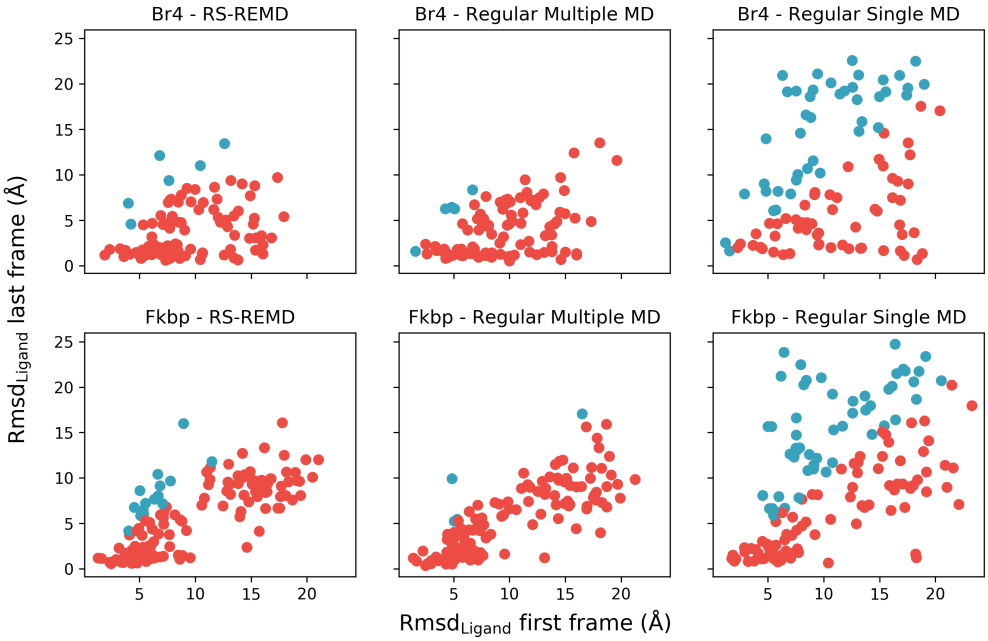


Figure 8.8: $Rmsd_{ligand}$ of the last frame againts $Rmsd_{ligand}$ of the first frame of the different refinement schemes (RS-REMD in first column, regular MD with multiple simulations in second column, regular MD with a single simulation in third column) for the cases Br4 (upper row) and Fkbp (lower row). The replica of lowest $rmsd_{ligand}$ was considered in the schemes using multiple simulations. Red dots indicate an improvement, blue dots a deterioration in $rmsd_{ligand}$ due to the refinement.

In case of a blind docking scenario the replica of lowest $rmsd_{ligand}$ is not known, so that a different measurement to predict the replica of lowest $rmsd_{ligand}$ has to be executed. Calculating the replica of minimal single point MMGBSA score is able to

identify a $rmsd_{ligand}$ enhancing replica for most of the poses (see Figure 8.9) with improvements of 83 % (Br4) and 79 % (fkbp) for RS-REMD and 83 % (Br4) and 86 % (fkbp) for regular multiple MD.

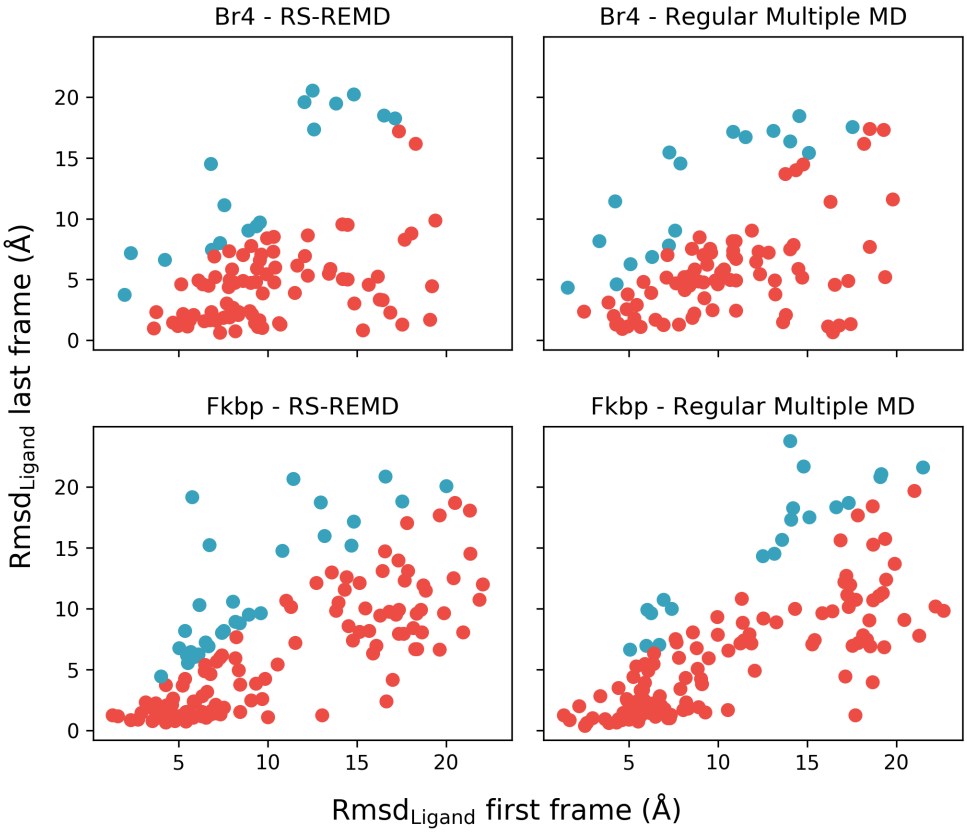


Figure 8.9: $Rmsd_{ligand}$ of the last frame againts $Rmsd_{ligand}$ of the first frame of the different refinement schemes (RS-REMD in first column, regular MD with multiple simulations in second column) for the cases Br4 (upper row) and Fkbp (lower row). The replica of lowest MMGBSA single point score was considered in the schemes, representing the result of a blind refinement scenario. Red dots indicate an improvement, blue dots a deterioration in $rmsd_{ligand}$ due to the refinement.

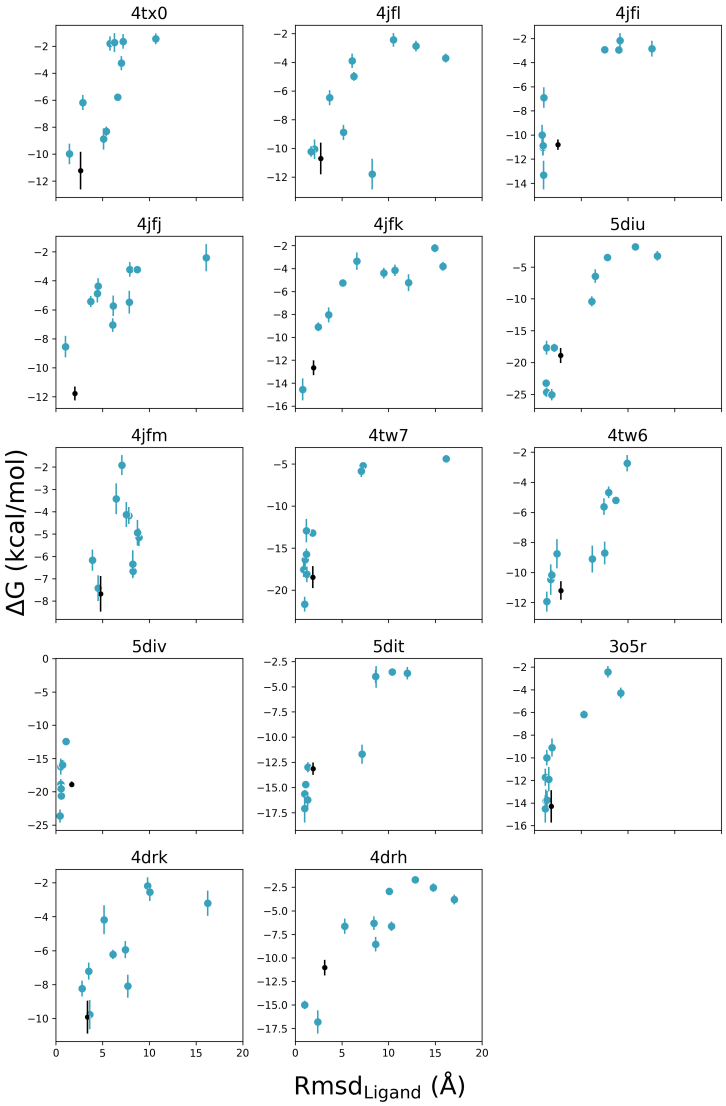


Figure 8.10: Absolute binding free energy against rmsd$_{ligand}$ from the native structure for the 14 ligands (indicated as pdb-id) of Fkbp. For each ligand the results for the 10 refined poses (blue dots) and the results of the native pose (black dots) are given. The uncertainties were calculated as the standard deviation from splitting the simulation into five parts.

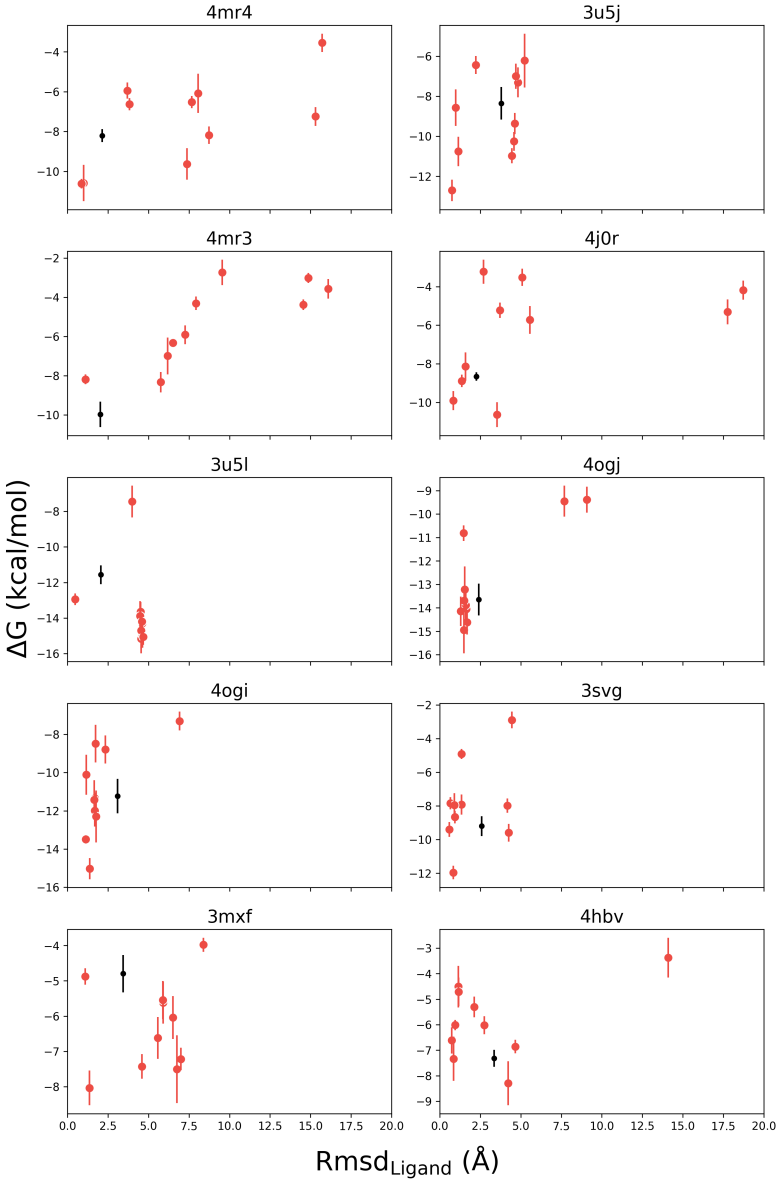


Figure 8.11: Absolute binding free energy against rmsd$_{ligand}$ from the native structure for the 10 ligands (indicated as pdb-id) of Br4. For each ligand the results for the 10 refined poses (red dots) and the results of the native pose (black dots) are given. The uncertainties were calculated as the standard deviation from splitting the simulation into five parts.

For each RS-REMD refined structure (starting from the replica with highest MMG-

BSA score in the last frame) we calculated the absolute binding free energy with RS-REMD just as for the native complexes in section 8.3.1. The resulting ΔG against $\text{rmsd}_{\text{ligand}}$ is given in Figures 8.10 (Fkbp) and 8.11 (Br4) including the native results (black dots). For every ligand structure a binding pose at the binding site was sampled with an $\text{rmsd}_{\text{ligand}}$ similar or better than the native case. The RS-REMD absolute binding free energy predicted the arrangement at the native binding site correctly (assigning it the highest ΔG score) with only one exception for Fkbp (4jfl) and four exceptions for Br4 (4mr3, 4j0r, 3u5l, 4hbv) of which only one favored a pose above 5 Å $\text{rmsd}_{\text{ligand}}$. The structure 3mxf is notable as the score of the native binding pose was the second worst, emphasizing that the adopted ligand arrangement was unfavorable in the native case.

8.4 Conclusion and outlook

The correct in silico prediction of ligand-protein binding geometries as well as having a realistic binding affinity estimate is of increasing importance in drug design projects. In this study, we applied the repulsive scaling (RS-)REMD scheme, that was originally developed using protein-protein benchmark sets, to small ligand protein complexes in explicit solvent. Basically, a repulsive biasing potential between ligand and receptor lj parameters is increased along a replica ladder which leads to dissociation of the ligand from free energy minima at the receptor surface in the higher replicas. This effect can be used in applications to improve the sampling of accurate ligand placements as well as the scoring of these placements in docking efforts. On the one hand, the binding free energy of the ligand can be calculated with a perturbation approach accounting for the effective biasing energy introduced in the system. On the other hand, RS-REMD can be used to effectively explore the receptor surface escaping local minima in the higher replicas to eventually identify the native binding site. These methods can be implemented in blind docking studies using single point MMGBSA scores to identify the correct results.

The absolute binding free energy was calculated with RS-REMD to a benchmark set of two proteins with 10 and 14 ligands. A standard state correction could be included restricting the ligand to a spherical shell volume around the receptor binding site. A good correlation of 0.72 for Br4 (omitting the results of 3mxf) and 0.54 for Fkbp was achieved to the experimental binding free energies with an overall slight overestimation of the affinities.

Moreover, we tested the ability of RS-REMD to identify the native binding site for one ligand of each protein. The initial placements represent a worst case scenario starting from an ensemble of incorrect binding sites based on an AutoDock Vina run. In both cases the RS-REMD simulations were able to escape the local minima and sample the native binding site. The lowest sampled $\text{rmsd}_{\text{ligand}}$ values were 0.3 Å for

3u5j and 3.5 Å for 3o5r. The performance was superior to regular MD simulations that in one case failed to predict the native binding site completely and in the second case stuck at $rmsd_{ligand}$ values above 5 Å. Using a single point MMGBSA score it is possible to effectively classify RS-REMD simulations in a blind docking scenario. The frames of lowest $rmsd_{ligand}$ were indicated by a high score, whereas the regions of occasional dissociations are given by a low MMGBSA affinity.

For less computationally expensive refinement efforts, RS-REMD simulations were conducted starting from ten placements in the vicinity of the native ligand pose. These were compared to the performance of a single regular MD simulation and multiple regular MD simulations with differing starting conditions. Both, the RS-REMD and the regular multiple MD simulations resulted in a drastic improvement of the starting $rmsd_{ligand}$ and performed in particular better than the single regular MD. Interestingly, the results of regular multiple MD were slightly superior to the RS-REMD results, probably due to a low number of competing local minima around the binding site, diminishing the need for repulsive sampling. Identification of a replica of low $rmsd_{ligand}$ for a blind docking study is possible using a single point MMGBSA score. From the identified binding placements we calculated the RS-REMD absolute binding free energy. This scoring scheme assigned the highest binding affinity to the lowest $rmsd_{ligand}$ pose for 19 of the 24 structures.

9 Summary and Outlook

The study of protein interactions is of major importance to ultimately elucidate the human interactome. To classify and understand the versatile complexes that proteins form through computational approaches, atomistic molecular dynamics simulations are best suited. In these simulations a proper treatment of the aqueous environment and inclusion of all energetic and entropic contributions to binding for free energy calculations is possible. Throughout this work, several MD advanced sampling approaches were established to efficiently identify the native binding site of proteins and also predict the binding affinity of different ligand poses.

In Chapter 5 the correct identification of near-native binding poses was possible using umbrella sampling simulations to apply a perturbation scheme that calculates the absolute binding free energy. The approach was validated on a benchmark set of 20 protein-protein complexes with 50 pre-docked poses using an implicit solvent model. The umbrella sampling approach was in particular advantageous comparing it to a single point score using the same force field description. The atomistic refinement procedure of the docked poses could improve the ligand conformation for some placements but was not able to sample the native state unless the simulations started in proximity of the binding site.

This task was addressed with the repulsive scaling scheme introduced in Chapter 6. The RS-REMD method establishes a repulsive bias along a replica ladder between ligand and receptor atoms. The method only affects the intermolecular pairwise van der Waals interactions but no intramolecular or solvent terms. Like this, the ligand is driven out of multiple local energy minima at the receptor surface to eventually encounter and stabilize the native binding site. Starting from worst case scenarios of the original ligand placements the native binding site was identified for five out of six protein-protein complexes in implicit solvent and two protein-ligand complexes in explicit solvent (Chapter 8). In particular, an improved sampling of the binding site in comparison to regular MD simulations was shown. In addition to that, the refinement of pre-docked protein poses (Chapter 7) and small ligand poses (Chapter 8) was possible with repulsive scaling and showed a substantial improvement in the $\mathrm{rmsd}_{\mathrm{ligand}}$ in many cases.

Furthermore, a reliable free energy estimate that can be obtained from repulsive scaling simulations was introduced (Chapter 7). In the method the applied bias in each replica is calculated via trajectory reevaluation. Using a perturbation approach, the free energy difference between the associated and the dissociated states can be

obtained. This free energy score gave a high correlation in explicit solvent to experimental binding affinities and also proved to successfully distinguish placements at the binding site from other poses for several protein-protein cases.

Finally, the RS-REMD score was extended to yield absolute binding free energies, restricting the ligand to a spherical volume around the receptor, for several protein complexes bound to small ligands (Chapter 8). The possibility of using single point MMGBSA scores for blind docking challenges to predict associated and dissociated sampling stages of repulsive scaling simulations was also elucidated.

I envision that, due to its simple implementation based on the force field parameters the RS-REMD method can be used for a much wider range of applications. Studying the association of other biomolecules is possible including structures like DNA, RNA or peptide-protein complexes. For example, the RS-REMD approach was successfully employed on protein-polysaccharide complexes in recent studies [207, 162]. Moreover, the repulsive scaling approach can be applied to investigate the stability of certain regions of a biomolecule. For example, the affinity of a specified alpha helix on a protein could be measured against different mutations and thus give suggestions to experimentally customize the protein.

The usability of the method could be increased by making it available to other software packages like GROMACS, CHARMM or NAMD. It is already possible to convert a parameter set between different types of force fields, but a direct implementation in packages like CHARMM-GUI [140] would highly facilitate the applicability to the MD community.

In future efforts, imposing an additional layer of guidance in the dissociated replicas could be of high benefit. In the current scheme the dissociated replicas still sample the receptor surface completely unguided which can lead to relatively high simulation times. Choosing an ensemble based initial placement of the ligands can already speed up the sampling step of the approach considerably (see Chapter 6). Still, a guidance of the dissociated states inspired by a metadynamics type of approach in which already sampled states are progressively avoided in the highest replicas could decrease the sampling time for the native binding site drastically.

To validate the accuracy of scoring functions the correlation to experimental binding affinities is usually calculated. Still, it is often difficult to assess the reliability of a novel method from the publication alone, as especially the choice of the benchmark can determine to a great part the complexity of obtaining a high correlation. As a perspective, it would be desirable that the community agrees on specific test sets and validation schemes that include verified experimental data [220]. These common evaluation standards should incorporate a correlation to experimental binding affinities but also a measurement of the deviation to native binding free energies in absolute terms.

In the long term, MD-based docking techniques are expected to increasingly gain importance due to probable enhancements in the computation hardware. Even

realistically resolving complete protein interaction networks could come into sight with such techniques, that might eventually shed light on the complete protein interactome.

Appendix A

Evaluation of Predicted Protein-Protein Complexes by Binding Free Energy Simulations

A.1 Absolute binding free energy calculation

Based on an approach introduced by Woo and Roux [324] the sampling of a dissociation/association process can be enhanced by introducing geometrical restraints. Apart from radial distance restraints, additional biasing potentials that reduce the axial (direction of separation), orientational (orient) and conformational (conf) degrees of freedom, are applied. The absolute binding free energy can be accessed by accounting for each contribution in supplementary simulations and calculations (illustrated in Figure A.2). The free energy is split into the following parts (standard concentration C°=1/1661 Å^3)

$$\Delta G_{bind} = -kT \ln \left[C^\circ e^{-\beta[\Delta G^{restr}_{bind} + (\Delta G^{bulk}_{orient} + \Delta G^{bulk}_{conf}) - (\Delta G^{site}_{orient} + \Delta G^{site}_{axial} + \Delta G^{site}_{conf})]} \right]$$

with

$$\Delta G^{restr}_{bind} = -kT \ln(IS) \tag{A.1}$$

being the restraint binding free energy. Superscripts bulk and site represent the unbound and bound conformation, respectively. The separation of ligand and receptor is denoted similar to equation 1 by an integral over the bound state of a distance PMF $A(\xi)$ with ξ_{bulk} being the distance in the unbound state

$$I = \frac{\int_{site} d\xi \; e^{-\beta A(\xi)}}{e^{-\beta A(\xi_{bulk})}}. \tag{A.2}$$

S accounts for the restraining of receptor and ligand to a fixed axial orientation towards each other. A sphere shell volume element at the bulk radius ξ_{bulk} is

calculated via

$$S = \xi_{bulk}^2 \int_0^{\pi} d\theta sin(\theta) \int_0^{2\pi} d\phi e^{-\beta U_{axial}(\theta,\phi)}. \tag{A.3}$$

U_{axial} denotes the harmonic potential of the axial angles, θ and ϕ, that restrict the relative movement of the two bodies to a single axis. The bound contribution that corrects the orientational and axial restraining of ligand and receptor $\Delta G_{orient}^{site} + \Delta G_{axial}^{site}$, is calculated by employing the FEP equation. The harmonic bias potential U_{orient}, that accounts for the rotation of the bodies via the three euler angles α, χ and γ as well as the axial contribution U_{axial}, are treated as a perturbation of an unrestrained simulation:

$$\Delta G_{orient}^{site} + \Delta G_{axial}^{site} \quad = \quad -k_B T \; ln \left\langle e^{-\beta[U_{axial}(\theta,\phi)+U_{orient}(\alpha,\chi,\gamma)]} \right\rangle_U. \tag{A.4}$$

The corresponding contribution $\Delta G_{conf}^{bulk,site}$ for either the bound (site) or unbound case (bulk), is corrected by again using the FEP equation

$$\Delta G_{conf}^{bulk,site} = -kT \ln \left\langle e^{-\beta U_{conf}^{bulk,site}} \right\rangle_U. \tag{A.5}$$

The remaining term to obtain ΔG_{bind} is the orientational contribution in the bulk ΔG_{orient}^{bulk}, that can be solved analytically by integration over the Euler angles

$$\Delta G_{orient}^{bulk} = -kT \ln \left[\frac{1}{8\pi^2} \int_0^{\pi} d\alpha \; sin(\alpha) \int_0^{2\pi} d\chi \int_0^{2\pi} d\gamma \; e^{-\beta U_{orient}(\alpha,\chi,\gamma)} \right]. \tag{A.6}$$

Here, the bulk is assumed to be isotropic and thus does not depend on the relative orientation of the bodies.

A.2 Figures

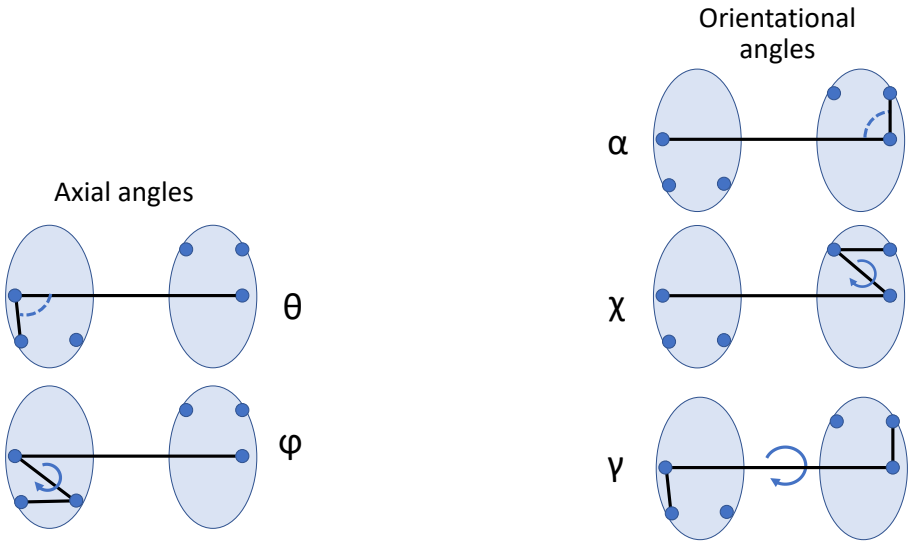


Figure A.1: The three orientational (α, χ and γ) and two axial restraint angles (θ and ϕ) were defined as shown using three COM positions (blue circles) on each partner protein (gray ellipse).

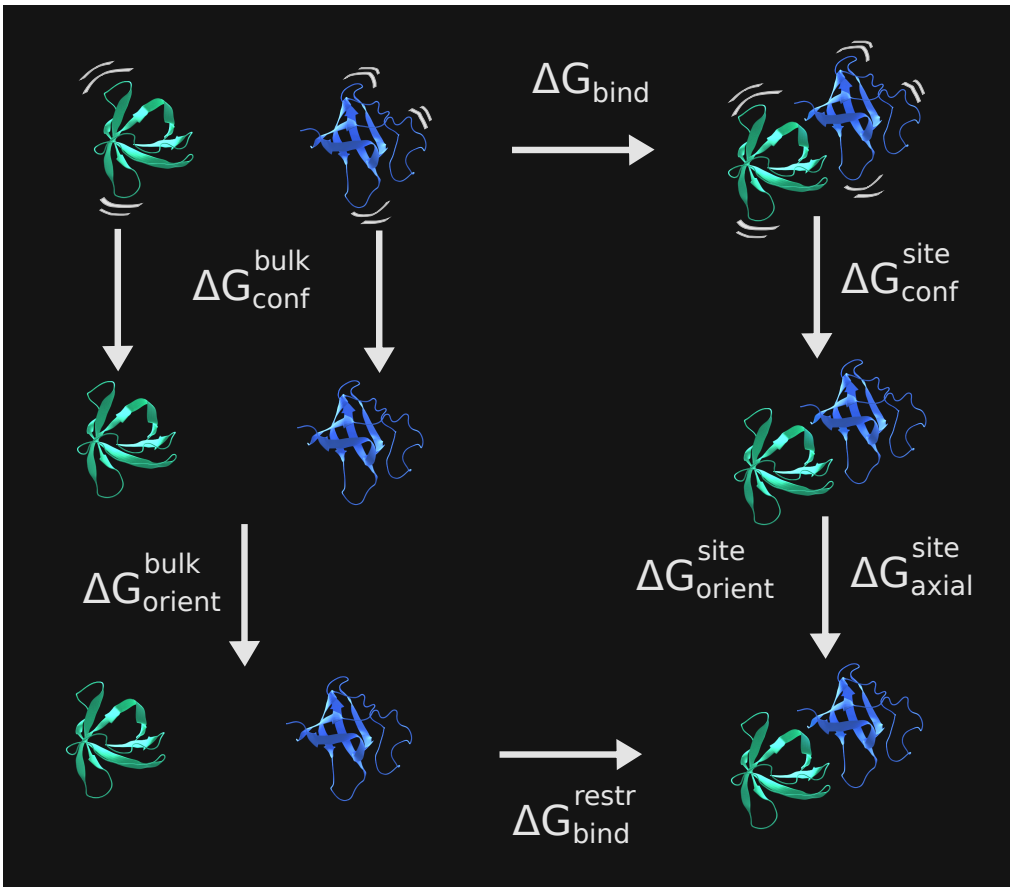


Figure A.2: The thermodynamic circle illustrating the different free energy contributions gained with FEP ($\Delta G_{conf}^{bulk,site}$, ΔG_{orient}^{site} and ΔG_{axial}^{site}), the distance PMF (ΔG_{bind}^{restr}) and integration over the Euler angles (ΔG_{bulk}^{orient}).

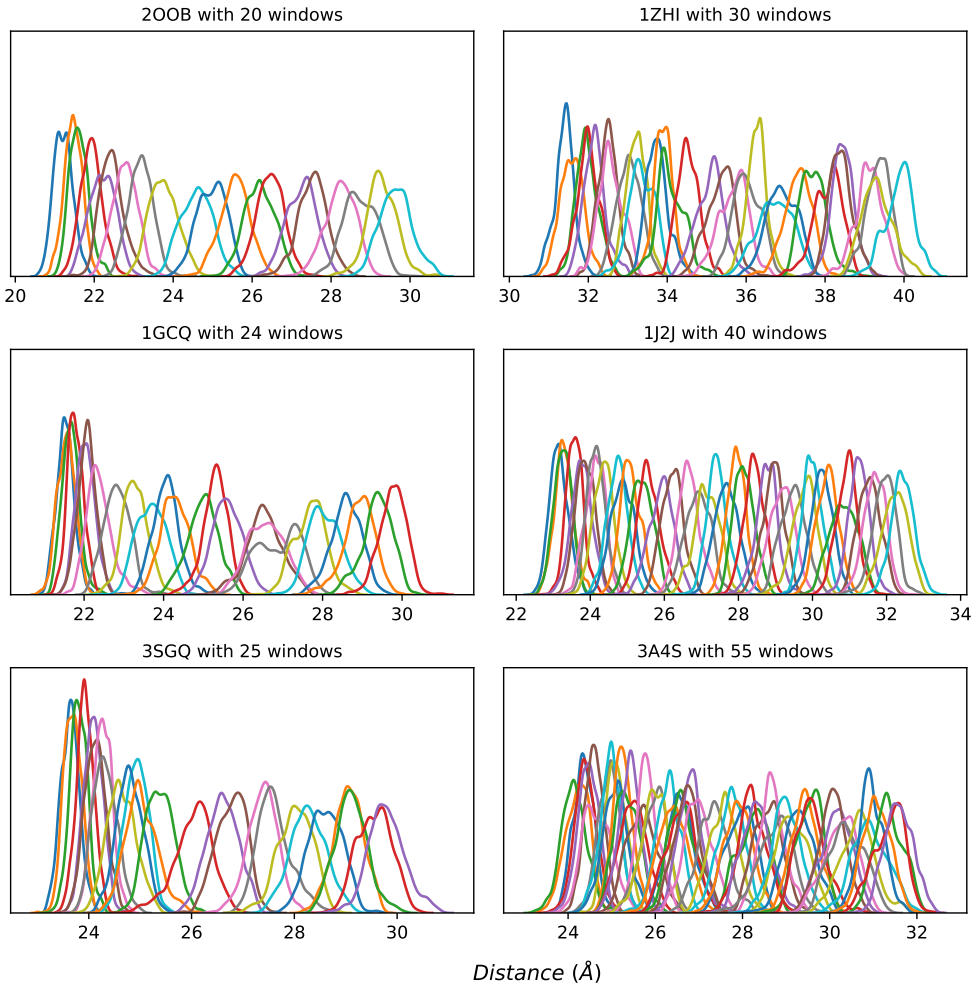


Figure A.3: Typical distance probability (y-axis) histograms of the US windows indicating sufficient overlap for accurate free energy estimation. For six different protein-protein complexes the case closest to the native complex is shown. In Figure A.4 the corresponding PMFs and pdb-entries for each case are given.

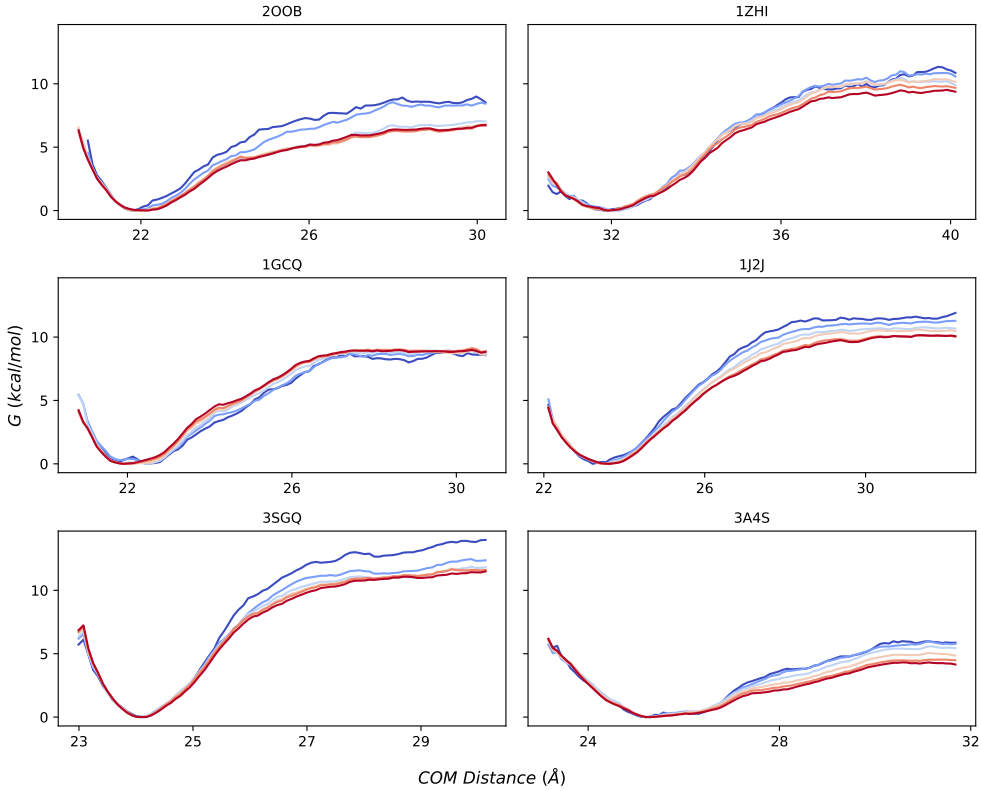


Figure A.4: The PMFs associated with the distance histograms of Figure A.3 are shown. The simulation was split into six parts and the corresponding running PMFs (from blue to red) show good convergence in all cases.

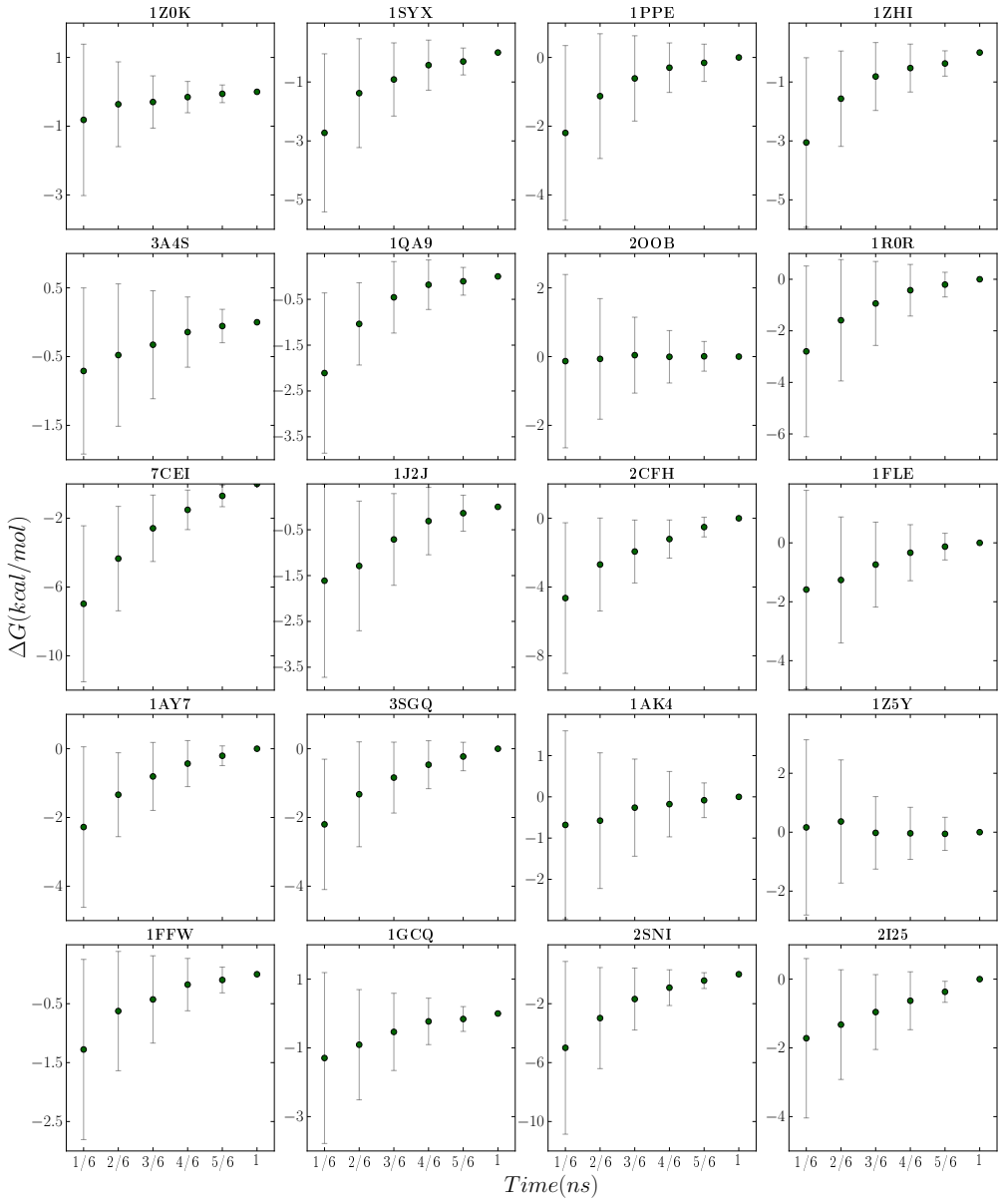


Figure A.5: The PMF free energy for each decoy was calculated after 1/6, 2/6, etc. of the total simulation time and subtracted from the calculated free energy obtained after the total simulation time for each US interval (values are given relative to the final simulation results). The error bars represent the standard variation over the results for all decoys of a given system.

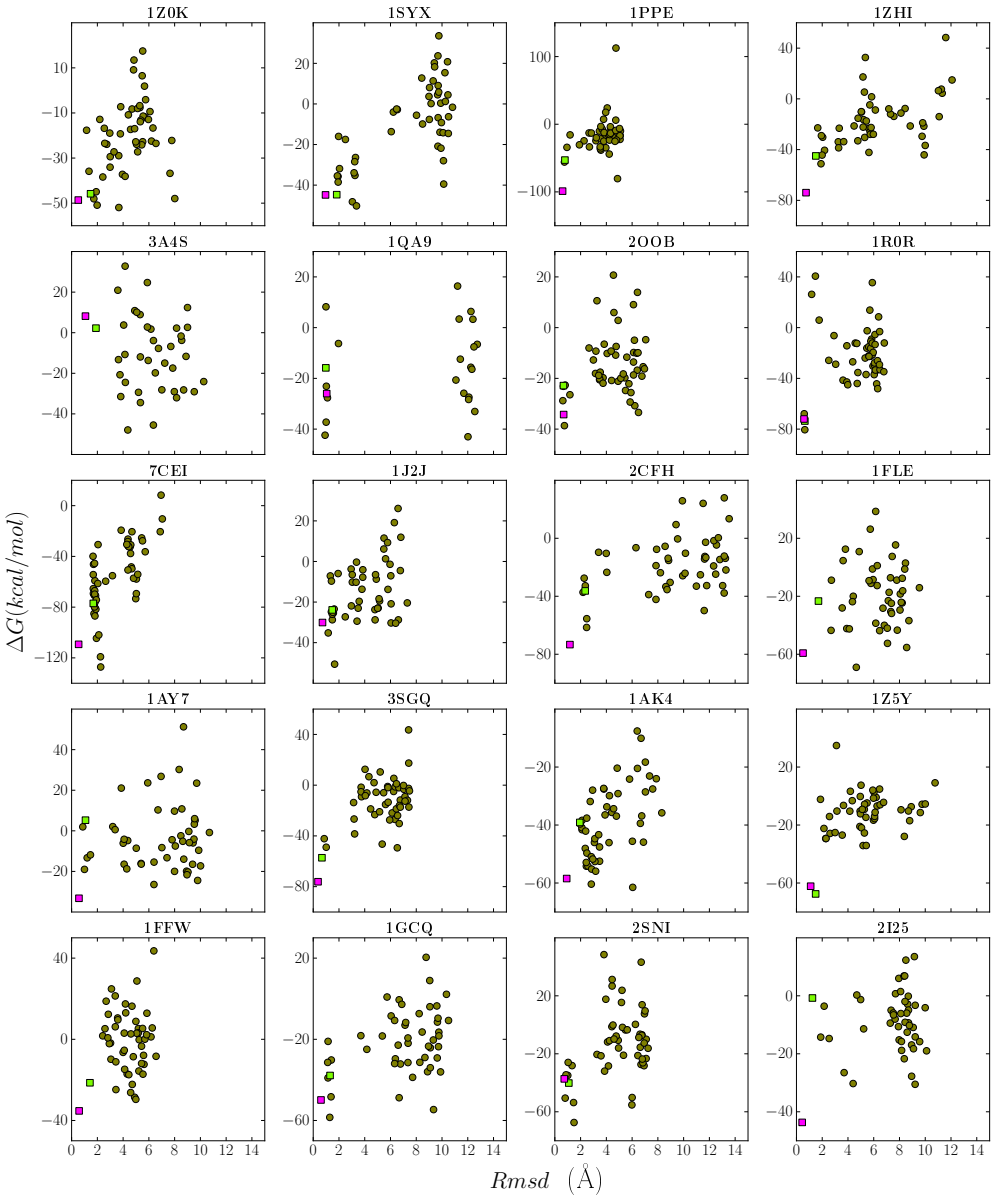


Figure A.6: Re-scoring after MD refinement/energy minimization and subsequent recalculation of the polar solvation free energy by solving the linearized finite-difference Poisson-Boltzmann equation. The pink and green squares indicate results for the bound and unbound starting conformations, respectively. The mean selectivity was 0.04 and thus lower than for MD refinement using the GB-model as described in the main text.

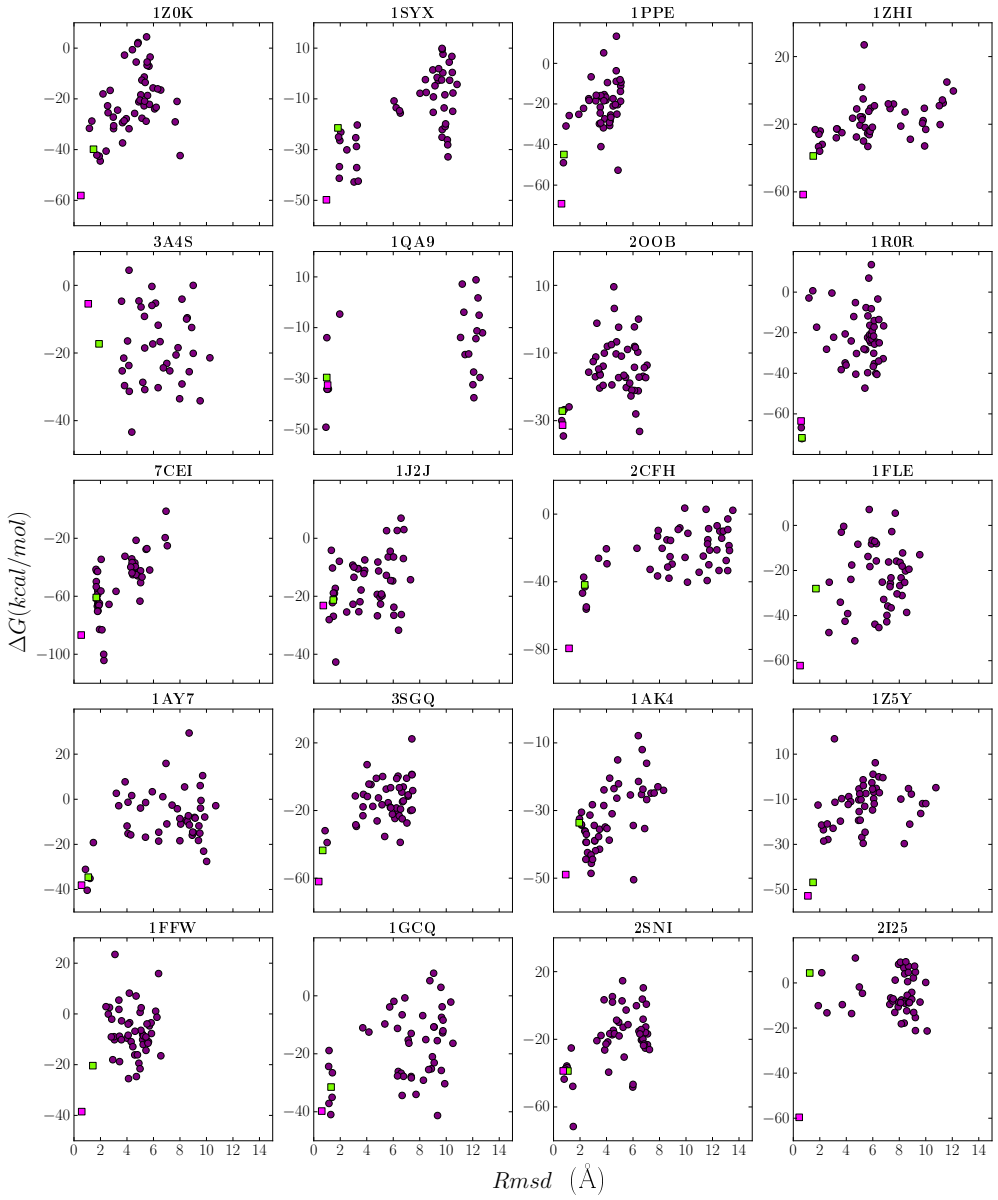


Figure A.7: Same as Figure S3 after re-scoring of the MD refinement results using a R^{-6} integration scheme [4] for calculating effective Born radii in the generalized Born model to estimate the polar solvation free energy. The mean selectivity was slightly higher (0.16) than using the standard GB model as described in the main text (0.14).

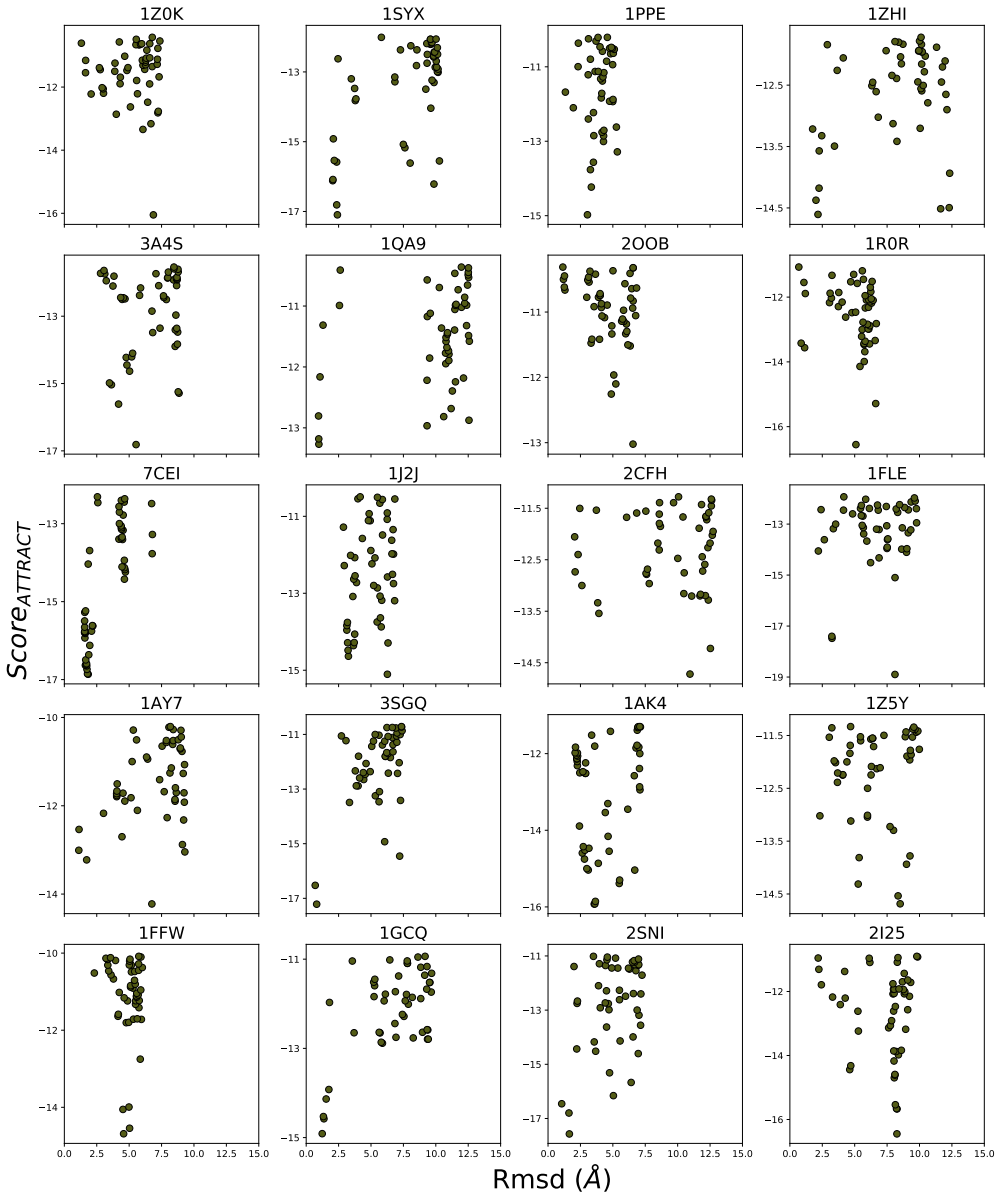


Figure A.8: Original Attract [329] docking score of the 50 models used in this study for 20 protein-protein complexes with rmsd to the native complex. The mean selectivity was -0.29 and thus lower than the selectivity of all subsequent refinement procedures.

Table A.1: Table of the experimental binding affinity ΔG_{Exp} [150, 153] and the binding affinity of the best scored model at the binding site evaluated with US ΔG_{US} and the absolute binding free energy $\Delta G_{Absolute}$ for all structures (only for 16 structures experimental binding affinities are available).

Structure	ΔG_{Exp} [kcal/mol]	ΔG_{US} [kcal/mol]	$\Delta G_{Absolute}$ [kcal/mol]
7cei	-19.8	-18.4 $\pm$ 3.3	-13.4 $\pm$ 3.5
1ay7	-13.2	-8.3 $\pm$ 1.5	-3.7 $\pm$ 1.5
1ppe	-15.6	-15.1 $\pm$ 1.2	-11.1 $\pm$ 1.9
1r0r	-14.3	-18.5 $\pm$ 3.7	-14.0 $\pm$ 3.8
2i25	-12.28	-6.8 $\pm$ 1.6	-3.3 $\pm$ 2.2
1j2j	-8.13	-12.0 $\pm$ 1.3	-8.5 $\pm$ 1.3
1z0k	-7.01	-8.3 $\pm$ 1.4	-5.0 $\pm$ 1.0
1qa9	-7.16	-6.5 $\pm$ 1.0	-2.4 $\pm$ 1.7
1ak4	-6.43	-7.9 $\pm$ 1.2	-2.4 $\pm$ 1.8
1gcq	-6.51	-10.7 $\pm$ 1.9	-6.5 $\pm$ 0.6
2oob	-5.76	-7.6 $\pm$ 1.5	-3.0 $\pm$ 1.6
1ffw	-8.1	-6.8 $\pm$ 0.6	-2.2 $\pm$ 1.5
1zhi	-9.1	-11.3 $\pm$ 1.8	-2.1 $\pm$ 2.8
3a4s	-7.57	-5.7 $\pm$ 1.0	-2.6 $\pm$ 2.0
1fle	-12.28	-8.3 $\pm$ 1.3	-7.5 $\pm$ 2.6
2sni	-15.96	-23.2 $\pm$ 7.1	-20.3 $\pm$ 7.2
1syx		-11.7 $\pm$ 1.6	-8.2 $\pm$ 2.2
3sgq		-12.1 $\pm$ 1.2	-7.1 $\pm$ 1.8
2cfh		-23.5 $\pm$ 2.4	-20.4 $\pm$ 3.3
1z5y		-9.8 $\pm$ 1.7	-3.7 $\pm$ 2.1

Appendix B

Prediction of Protein-Protein Complexes Using Replica Exchange With Repulsive Scaling

B.1 Figures and tables

Table B.1: Population of the binding site ($rmsd_{ligand} < 10$ Å) of the second half of the simulations (see Figure S1 for the whole $rmsd_{ligand}$ populations) for each complex indicated by the pdb-id for the repulsive scaling (RS-REMD) approach and the regular MD simulations.

PDB	RS-REMD	Regular MD
7cei	77 %	0 %
2oo9	78 %	47 %
2cfh	46 %	0 %
1syx	23 %	1 %
2sni	70 %	0 %
1gcq	0 %	0 %

Table B.2: Table of the 20 protein complexes analyzed in this study with according pdb-id and difficulty in terms of the magnitude of deviation between unbound and bound structures (as defined in the benchmark [132]).

PDB	Difficulty	Protein1	Protein2
7cei	Low	Colicin E7 nuclease	Im7 immunity protein
1ak4	Low	Cyclophilin	HIV capsid
1ay7	Low	Rnase SA	Barstar
1ppe	Low	Trypsin	CMTI-1 squash inhibitor
1r0r	Low	Subtilisin carlsberg	OMTKY
2i25	Low	Shark single domain antigen receptor	Lysozyme
1j2j	Low	Arf1 GTPase.GNP-RanBD1	GAT domain of GGA1
1z0k	Low	RAB4 binding domain of Rabenosyn	Rab4A GTPase
1qa9	Low	CD2	CD58
1gcq	Low	GRB2 C-ter SH3 domain	Vav N-ter SH3 domain
2oob	Low	Ubiquitin ligase	Ubiquitin
1ffw	Low	Chemotaxis protein CheY	Chemotaxis protein CheA
1zhi	Low	BAH domain of Orc1	Sir Orc-interaction domain
3a4s	Low	SUMO-conjugating enzyme UBC9	NFATC2-interacting protein
1fle	Low	Elastase	Elafin
2sni	Low	Subtilisin	Chymotrypsin inhibitor 2
3sgq	Low	Ovomucoid inhibitor third domain	Streptogrisin B
1z5y	Low	N-term of DsbD	E.coli CCMG protein
1syx	Medium	Spliceosomal U5 15 kDa protein	CD2 receptor binding protein 2
2cfh	Medium	BET3	TPC6

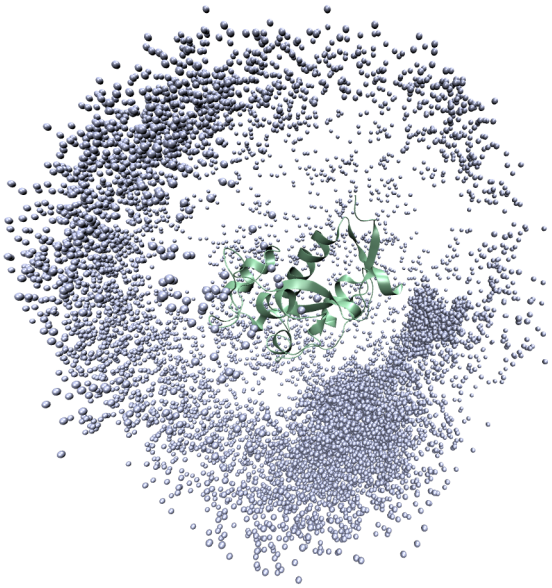


Figure B.1: The population of the ligand (placements shown as blue spheres, each depicts the placement of the same ligand C_α atom) around the receptor (green cartoon) in the highest replica of the extensive RS-REMD simulation of the structure 7cei.

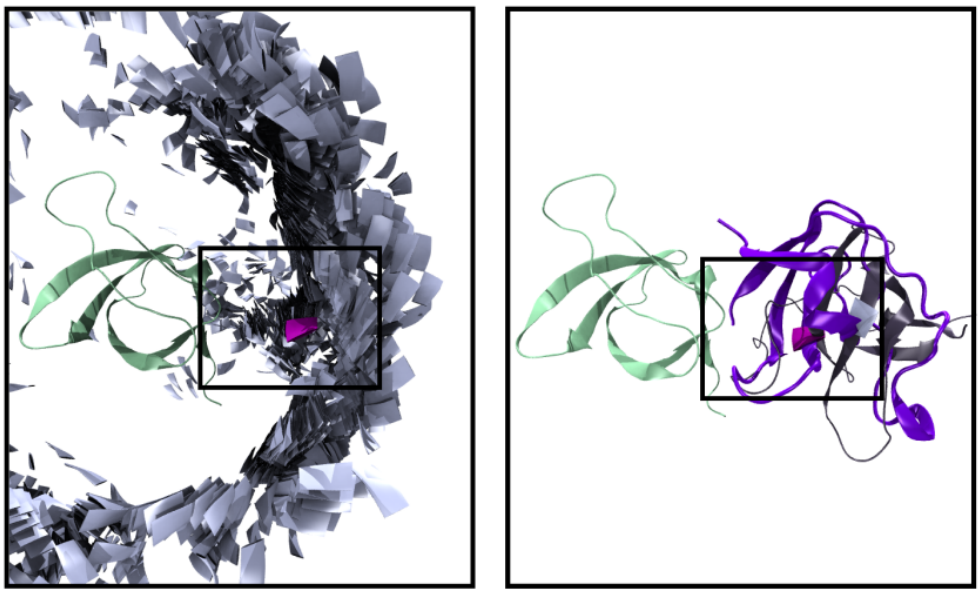


Figure B.2: (Left panel) The population of the ligand (placements shown as blue rectangles, each depicts the orientation of the same ligand C_α atom) around the receptor (green cartoon) of the reference replica in the extensive RS-REMD simulation of the structure 1gcq. The native ligand placement and orientation is shown by a magenta rectangle (same C_α atom as for the blue rectangles chosen, see right panel). The native ligand is oriented differently than all the sampled ligands. (Right panel) The corresponding native (purple cartoon) and simulated (dark grey cartoon) ligand orientations.

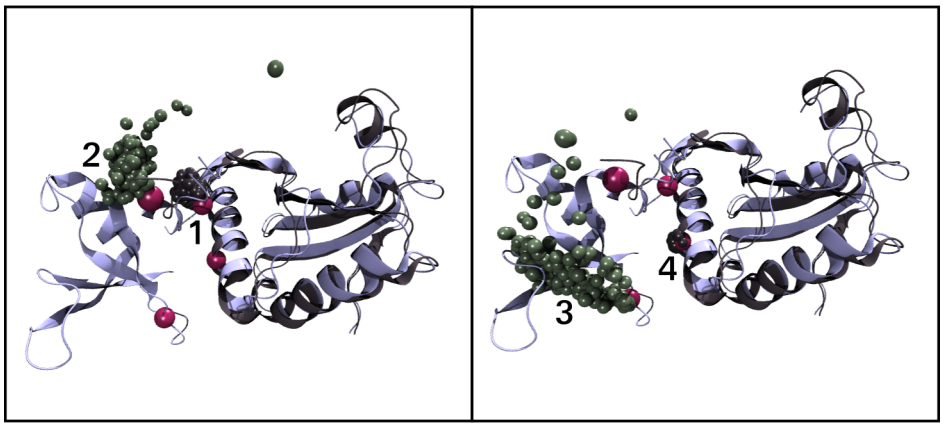


Figure B.3: Snapshots of the reference replica of the extensive RS-REMD simulations of structure 1syx with an $rmsd_{ligand}$ under 18 Å. The native structure is shown in blue cartoon, the simulated receptor protein is aligned on the native receptor protein (black cartoon). Four C_α atoms of the native structure (red spheres) and the corresponding C_α atoms of simulated snapshots of the ligand (green spheres) and the receptor (black sphere) (both restrained on the unbound conformation) are shown. (1) The simulated receptor loop that coordinates the ligand-receptor interaction is shifted apart from the receptor COM compared to the native state (black spheres compared to the closest red sphere on the left panel), presumably due to the restraining on the unbound conformation. (2) Still, the simulated ligand is coordinated by this loop as little deviation is visible in the distribution of the green spheres on the left panel. (3) On the contrary the lower part of the simulated ligand shows high deviations in the same snapshots (green spheres right panel), while the receptor has low fluctuations (4, black spheres right panel). Thus the interaction between the lower part of ligand and receptor is not stable, presumably due to the deviation of the coordinating loop between unbound and native conformation (1).

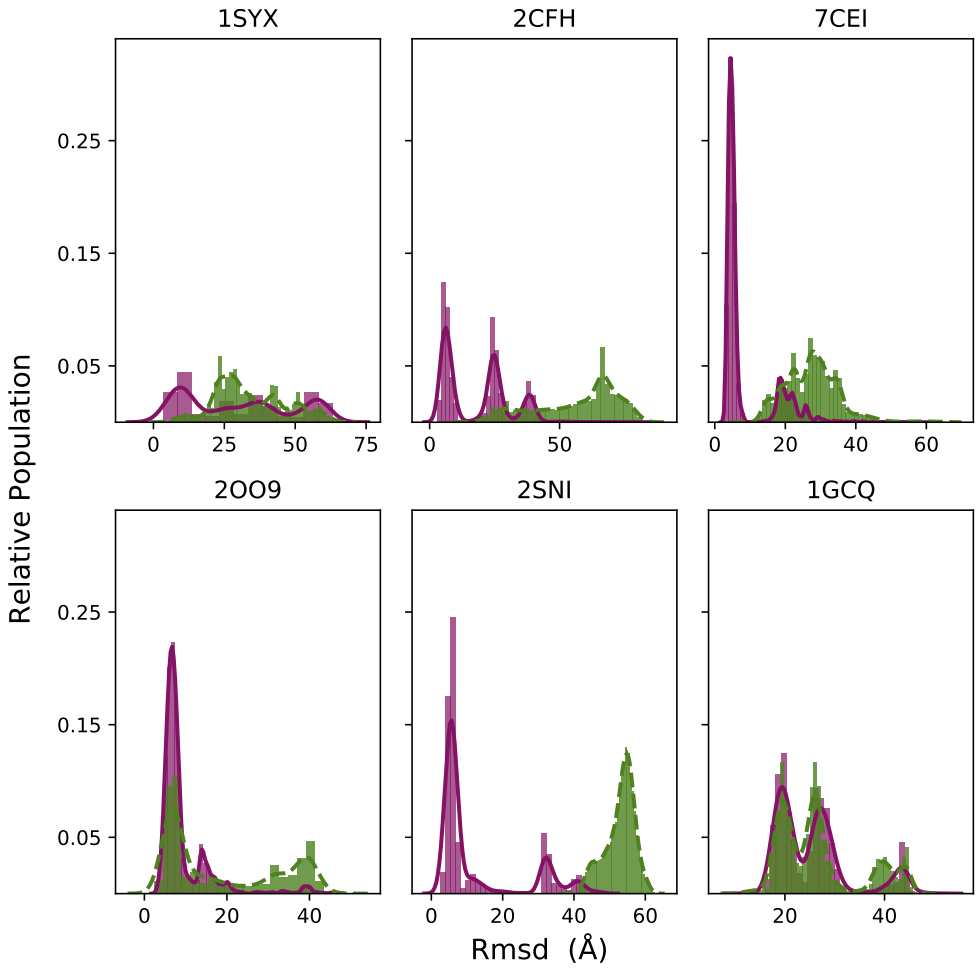


Figure B.4: Histograms of the $\text{rmsd}_{\text{ligand}}$ from the native structure for the reference replica of the RS-REMD simulations (for all 6 protein–protein test cases in an individual figure) is compared to the $\text{rmsd}_{\text{ligand}}$ histograms of the regular MD simulations (green), considering only the last half of the simulation time, respectively. The corresponding population at the binding site is given in Table S1.

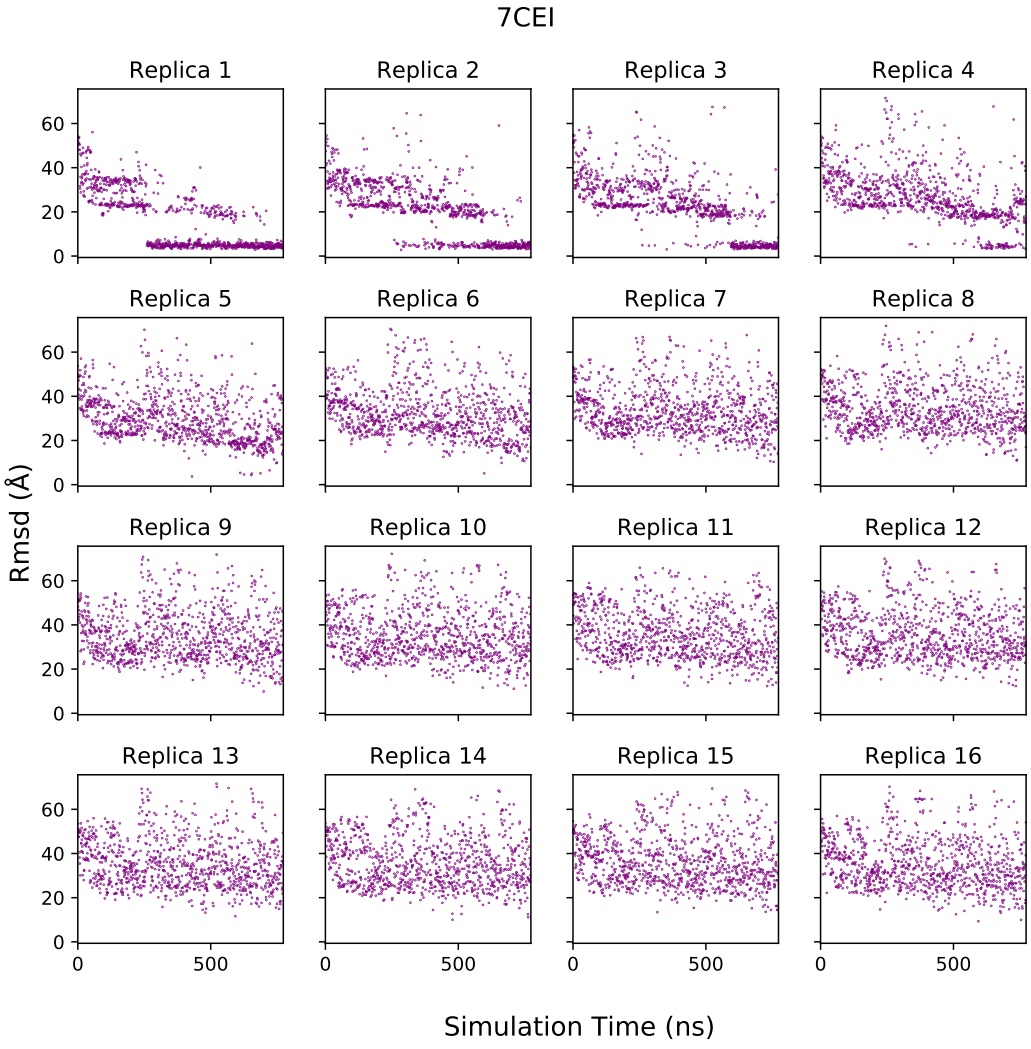


Figure B.5: $\text{Rmsd}_{\text{ligand}}$ from the native structure vs. simulation time for all 16 replicas of the RS-REMD simulation of the 7cei protein-protein complex starting with the ligand partner on the opposite site of the native binding region of the receptor protein.

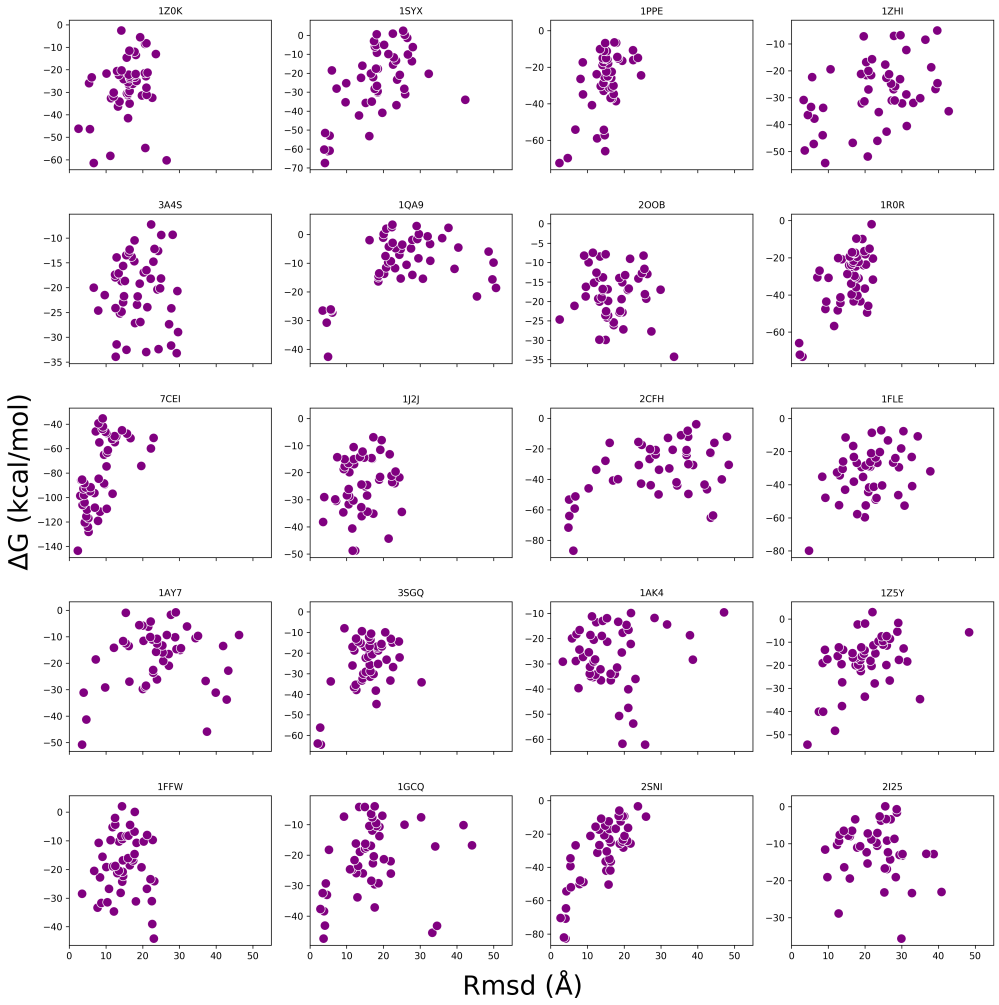


Figure B.6: Force field scoring after RS-REMD refinement of the 50 decoy complexes for each of the 20 protein-protein complex test case (indicated by pdb-id). Calculated final protein-protein interaction energy (violet circles) is plotted vs. the $\text{rmsd}_{\text{ligand}}$ with respect to the native complex.

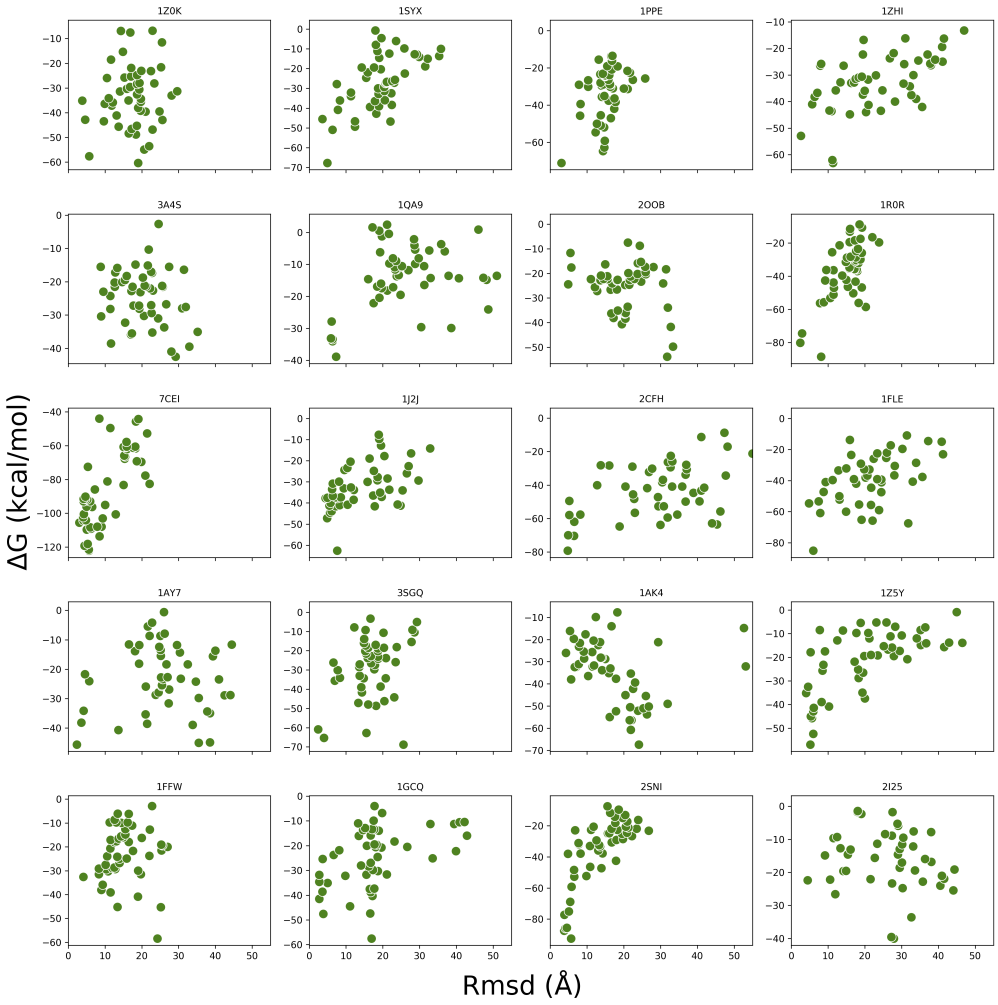


Figure B.7: Force field scoring after regular MD refinement of the 50 decoy complexes for each protein-protein complex test case (indicated by pdb-id). Calculated protein-protein interaction energy (green circles) is plotted vs. the $\text{rmsd}_{\text{ligand}}$ from the native complex.

Appendix C

Efficient Refinement and Free Energy Scoring of Predicted Protein-Protein Complexes Using Replica Exchange With Repulsive Scaling

C.1 Figures and tables

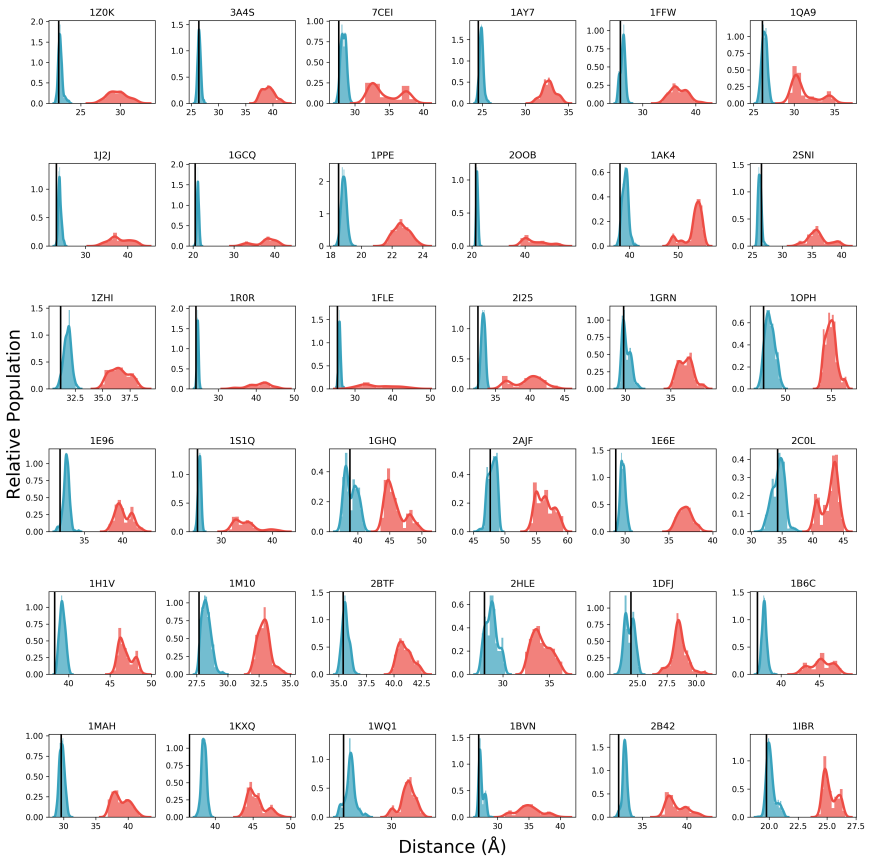


Figure C.1: Histograms of the protein-protein COM distances of the bound states (blue, replica with minimal $\text{rmsd}_{\text{ligand}}$ for each frame in the second half of the simulation) and the dissociated state (red, replica with the highest COM distance in each frame of the second half of the simulation) observed during the explicit water RS-REMD simulations for all native protein-protein cases (indicated by pdb-id). The native distances are given as black vertical lines. The mean values and distance differences are given in Table C.1.

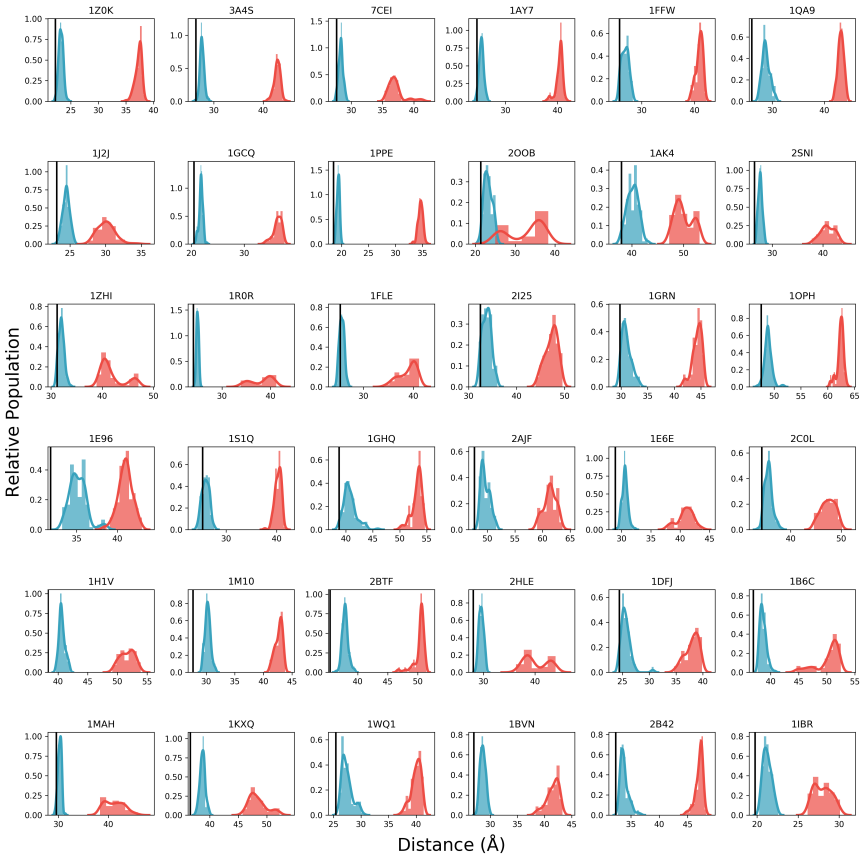


Figure C.2: Histograms of the protein-protein COM distances of the bound states (blue, replica with minimal $\text{rmsd}_{\text{ligand}}$ for each frame in the second half of the simulation) and the dissociated state (red, replica with the highest COM distance in each frame of the second half of the simulation) observed during the implicit water RS-REMD simulations for all native protein-protein cases (indicated by pdb-id). The native distances are given as black vertical lines. The mean values and distance differences are given in Table C.2.

Table C.1: For all protein-protein complexes in explicit solvent the native COM distance (d_{Nat}) is given. The mean COM distance between ligand and receptor in the associated state (d_{Asso}) and the dissociated state (d_{Disso}) and the corresponding difference in distance are shown. Additionally, the mean replica number (counting from 0) of the associated state (Rep_{Asso}) and the dissociated state (Rep_{Disso}) are given.

Structure	d_{Nat}	d_{Asso}	d_{Disso}	$d_{Disso} - d_{Asso}$	Rep_{Asso}	Rep_{Disso}
1z0k	22.2	22.4 ± 0.3	29.7 ± 1.2	7.2 ± 1.3	2.1 ± 1.9	12.9 ± 1.8
3a4s	26.4	26.5 ± 0.3	39.1 ± 1.1	12.7 ± 1.2	1.2 ± 1.1	11.6 ± 2.6
7cei	27.6	28.3 ± 0.4	34.5 ± 2.3	6.1 ± 2.3	4.8 ± 2.0	13.2 ± 1.4
1ay7	24.6	24.9 ± 0.2	32.7 ± 0.8	7.8 ± 0.8	2.4 ± 1.8	12.6 ± 1.8
1ffw	25.8	26.2 ± 0.4	36.8 ± 1.6	10.6 ± 1.7	2.8 ± 1.9	13.6 ± 1.4
1qa9	26.1	26.3 ± 0.3	31.3 ± 1.6	4.9 ± 1.7	3.0 ± 1.9	12.3 ± 2.0
1j2j	23.1	24.0 ± 0.4	38.1 ± 2.5	14.1 ± 2.5	5.6 ± 3.6	14.2 ± 1.1
1gcq	20.5	21.2 ± 0.3	37.6 ± 2.7	16.4 ± 2.7	2.9 ± 2.2	13.3 ± 1.5
1ppe	18.5	18.9 ± 0.2	22.6 ± 0.5	3.8 ± 0.6	2.5 ± 1.9	14.0 ± 0.9
2oob	21.4	22.0 ± 0.3	43.8 ± 4.3	21.8 ± 4.3	2.0 ± 1.7	11.8 ± 2.5
1ak4	38.0	39.1 ± 0.6	53.2 ± 1.9	14.1 ± 2.0	5.5 ± 2.6	13.6 ± 1.2
2sni	26.5	26.2 ± 0.3	36.0 ± 1.8	9.8 ± 1.8	3.0 ± 2.4	14.6 ± 0.5
1zhi	31.2	31.8 ± 0.3	36.6 ± 0.9	4.8 ± 0.9	2.1 ± 2.0	13.3 ± 1.7
1r0r	24.1	24.6 ± 0.2	40.8 ± 2.9	16.2 ± 2.9	2.5 ± 2.0	13.2 ± 1.4
1fle	25.3	25.9 ± 0.3	36.1 ± 4.6	10.2 ± 4.6	6.5 ± 1.9	14.3 ± 0.8
2i25	32.5	33.2 ± 0.3	39.7 ± 2.1	6.5 ± 2.1	3.0 ± 2.0	12.7 ± 1.5
1grn	29.8	30.1 ± 0.5	36.6 ± 0.8	6.5 ± 0.9	2.3 ± 2.3	13.0 ± 1.4
1oph	47.6	48.3 ± 0.5	55.0 ± 0.6	6.7 ± 0.8	2.8 ± 2.2	13.1 ± 1.2
1e96	31.8	32.6 ± 0.4	40.1 ± 1.0	7.5 ± 1.1	1.7 ± 1.4	12.4 ± 2.1
1s1q	25.3	25.7 ± 0.3	34.8 ± 2.4	9.0 ± 2.4	1.8 ± 1.5	12.4 ± 2.1
1ghq	38.8	39.0 ± 1.0	45.8 ± 1.6	6.8 ± 1.8	2.6 ± 2.6	12.7 ± 2.1
2ajf	47.7	48.0 ± 0.7	56.4 ± 1.4	8.4 ± 1.5	5.7 ± 3.3	13.8 ± 1.2
1e6e	28.9	29.8 ± 0.3	36.9 ± 0.8	7.1 ± 0.8	1.7 ± 1.1	12.9 ± 1.7
2c0l	34.3	34.4 ± 1.0	42.8 ± 1.3	8.4 ± 1.7	5.7 ± 4.5	13.0 ± 1.9
1h1v	38.3	39.3 ± 0.3	46.9 ± 0.8	7.6 ± 0.9	2.7 ± 1.8	14.2 ± 1.2
1m10	27.7	28.3 ± 0.4	32.9 ± 0.5	4.6 ± 0.6	1.6 ± 1.4	12.2 ± 2.1
2btf	35.4	35.6 ± 0.3	41.0 ± 0.6	5.4 ± 0.7	1.8 ± 1.4	12.8 ± 1.9
2hle	28.0	28.8 ± 0.6	34.1 ± 1.0	5.3 ± 1.2	2.7 ± 3.0	13.7 ± 1.2
1dfj	24.4	24.3 ± 0.4	28.6 ± 0.7	4.2 ± 0.8	1.2 ± 1.3	12.8 ± 2.0
1b6c	36.9	37.8 ± 0.3	45.2 ± 1.4	7.4 ± 1.5	3.3 ± 2.4	14.4 ± 0.8
1mah	29.6	29.7 ± 0.4	38.9 ± 1.3	9.2 ± 1.4	3.4 ± 2.9	13.7 ± 1.1
1kxq	36.6	38.5 ± 0.3	45.5 ± 1.2	7.0 ± 1.2	4.6 ± 3.2	13.7 ± 1.2
1wq1	25.4	26.1 ± 0.5	31.6 ± 0.7	5.5 ± 0.9	5.2 ± 2.2	13.4 ± 1.4
1bvn	27.0	27.4 ± 0.4	35.0 ± 1.9	7.5 ± 2.0	4.5 ± 2.9	14.1 ± 0.9
2b42	32.3	33.0 ± 0.3	38.9 ± 1.1	5.8 ± 1.2	1.4 ± 1.6	14.1 ± 1.0
1ibr	19.8	20.1 ± 0.4	25.3 ± 0.6	5.1 ± 0.7	1.9 ± 1.8	13.2 ± 1.5

Table C.2: For all protein-protein complexes in implicit solvent the native COM distance (d_{Nat}) is given. The mean COM distance between ligand and receptor in the associated state (d_{Asso}) and the dissociated state (d_{Disso}) and the corresponding difference in distance are shown. Additionally, the mean replica number (counting from 0) of the associated state (Rep_{Asso}) and the dissociated state (Rep_{Disso}) are given.

Structure	d_{Nat}	d_{Asso}	d_{Disso}	$d_{Disso} - d_{Asso}$	Rep_{Asso}	Rep_{Disso}
1z0k	22.2	23.2 ± 0.4	37.3 ± 0.6	14.1 ± 0.7	1.6 ± 1.5	10.4 ± 3.2
3a4s	26.4	27.6 ± 0.4	42.9 ± 0.6	15.2 ± 0.8	0.9 ± 1.1	9.4 ± 3.8
7cei	27.6	28.3 ± 0.4	37.1 ± 1.3	8.8 ± 1.4	1.3 ± 1.2	10.7 ± 2.8
1ay7	24.6	25.4 ± 0.4	40.4 ± 0.7	15.0 ± 0.8	1.4 ± 1.5	10.3 ± 3.3
1ffw	25.8	26.8 ± 0.7	40.9 ± 0.7	14.1 ± 0.9	1.6 ± 1.6	10.2 ± 3.0
1qa9	26.1	28.8 ± 0.7	43.1 ± 0.5	14.4 ± 0.9	1.2 ± 1.0	8.9 ± 4.0
1j2j	23.1	24.5 ± 0.5	30.1 ± 1.2	5.6 ± 1.3	3.8 ± 3.2	12.7 ± 2.0
1gcq	20.5	21.8 ± 0.4	36.3 ± 0.8	14.5 ± 0.9	1.9 ± 1.5	10.3 ± 2.9
1ppe	18.5	19.4 ± 0.3	34.7 ± 0.5	15.3 ± 0.5	2.0 ± 1.5	11.3 ± 2.3
2oob	21.4	23.2 ± 1.0	32.6 ± 4.5	9.4 ± 4.6	5.4 ± 4.2	11.0 ± 3.2
1ak4	38.0	40.2 ± 1.1	50.0 ± 1.7	9.8 ± 2.0	3.6 ± 2.9	11.4 ± 2.7
2sni	26.5	27.6 ± 0.4	40.9 ± 1.4	13.3 ± 1.4	3.1 ± 2.2	11.7 ± 2.0
1zhi	31.2	32.0 ± 0.6	41.9 ± 2.4	9.9 ± 2.5	3.5 ± 3.1	12.2 ± 2.3
1r0r	24.1	25.0 ± 0.2	38.1 ± 2.3	13.2 ± 2.3	2.2 ± 1.7	11.5 ± 2.5
1fle	25.3	25.7 ± 0.5	38.9 ± 1.8	13.2 ± 1.8	3.5 ± 2.5	12.6 ± 1.5
2i25	32.5	33.8 ± 0.9	47.2 ± 1.3	13.5 ± 1.6	5.4 ± 3.2	11.7 ± 2.1
1grn	29.8	31.0 ± 0.9	44.3 ± 0.9	13.3 ± 1.3	1.8 ± 2.6	10.6 ± 3.0
1oph	47.6	48.9 ± 0.7	62.4 ± 0.6	13.5 ± 0.9	3.1 ± 2.2	11.5 ± 2.2
1e96	31.8	35.2 ± 1.0	41.0 ± 0.9	5.8 ± 1.4	7.0 ± 4.5	10.7 ± 3.1
1s1q	25.3	25.9 ± 0.7	40.3 ± 0.6	14.3 ± 1.0	4.6 ± 3.3	11.2 ± 2.4
1ghq	38.8	40.8 ± 1.2	53.2 ± 0.9	12.4 ± 1.5	3.7 ± 3.5	9.9 ± 3.3
2ajf	47.7	49.7 ± 0.8	61.4 ± 1.1	11.7 ± 1.4	1.8 ± 1.9	9.6 ± 3.2
1e6e	28.9	30.6 ± 0.5	40.9 ± 1.4	10.3 ± 1.4	1.5 ± 2.2	11.3 ± 2.5
2c0l	34.3	35.6 ± 0.8	47.4 ± 1.4	11.8 ± 1.6	4.6 ± 2.9	10.7 ± 2.5
1h1v	38.3	40.7 ± 0.5	51.6 ± 1.2	11.0 ± 1.3	1.7 ± 1.6	10.8 ± 2.4
1m10	27.7	30.3 ± 0.5	42.7 ± 0.6	12.4 ± 0.8	2.0 ± 1.6	9.5 ± 3.0
2btf	35.4	37.9 ± 0.5	50.3 ± 0.9	12.4 ± 1.1	1.6 ± 1.6	9.1 ± 3.5
2hle	28.0	29.6 ± 0.4	40.2 ± 2.3	10.6 ± 2.3	1.9 ± 2.1	12.1 ± 2.1
1dfj	24.4	25.6 ± 1.0	38.0 ± 1.3	12.3 ± 1.6	1.5 ± 1.4	10.7 ± 2.8
1b6c	36.9	38.6 ± 0.6	50.3 ± 2.0	11.6 ± 2.1	2.8 ± 3.1	11.8 ± 2.0
1mah	29.6	30.3 ± 0.4	41.2 ± 1.9	10.9 ± 1.9	2.7 ± 2.0	11.2 ± 2.3
1kxq	36.6	38.7 ± 0.5	48.4 ± 1.6	9.7 ± 1.6	2.8 ± 2.8	12.7 ± 1.9
1wq1	25.4	27.5 ± 1.0	40.0 ± 0.9	12.5 ± 1.3	2.6 ± 3.8	10.5 ± 2.8
1bvn	27.0	28.7 ± 0.5	41.5 ± 1.1	12.8 ± 1.3	2.4 ± 1.8	10.9 ± 2.6
2b42	32.3	33.8 ± 0.7	47.0 ± 0.7	13.2 ± 1.0	1.7 ± 2.0	11.1 ± 2.4
1ibr	19.8	21.2 ± 0.5	28.0 ± 1.1	6.8 ± 1.2	1.8 ± 1.9	11.1 ± 2.8

Table C.3: RS-REMD timings for simulating every ns (per replica and pose) for all protein-protein cases (50 poses and 16 replicas in each simulation) using implicit and explicit water representation on one GPU. On average 1 ns (per pose and replica) for one complex took 2.4 days in the implicit case and 4.9 days in the explicit water case.

Structure	Implicit		Explicit	
	GPU	Time (days)	GPU	Time (days)
1z0k	GTX 1080 Ti	2.1	GTX 1080	5.4
3a4s	GTX 1080 Ti	2.1	GTX 1080 Ti	4.4
7cei	GTX 1080 Ti	2.1	GTX 1080	5.7
1ay7	RTX 2080 Ti	1.9	GTX 1080 Ti	4.2
1ffw	GTX 1080	2.2	GTX 1080 Ti	3.6
1syx	GTX 1080 Ti	1.7	GTX 1080 Ti	4.4
1qa9	RTX 2080 Ti	1.7	GTX 1080	7.2
1j2j	RTX 2080 Ti	2.2	GTX 1080 Ti	3.1
3sgq	GTX 1080 Ti	2.1	GTX 1080 Ti	3.9
1gcq	RTX 2080 Ti	1.4	GTX 1080 Ti	3.5
1ppe	RTX 2080 Ti	3.0	GTX 1080 Ti	2.8
2oob	RTX 2080 Ti	1.4	GTX 1080 Ti	2.4
2cfh	RTX 2080 Ti	2.8	GTX 1080 Ti	5.3
1ak4	RTX 2080 Ti	2.5	GTX 1080	8.2
2sni	GTX 1080 Ti	3.5	GTX 1080	5.7
1zhi	RTX 2080 Ti	3.1	GTX 1080 Ti	7.5
1r0r	GTX 1080 Ti	3.3	GTX 1080	5.2
1fle	GTX 1080 Ti	3.3	GTX 1080 Ti	4.6
1z5y	GTX 1080 Ti	2.5	GTX 1080 Ti	5.4
2i25	GTX 1080	3.5	GTX 1080 Ti	6.0

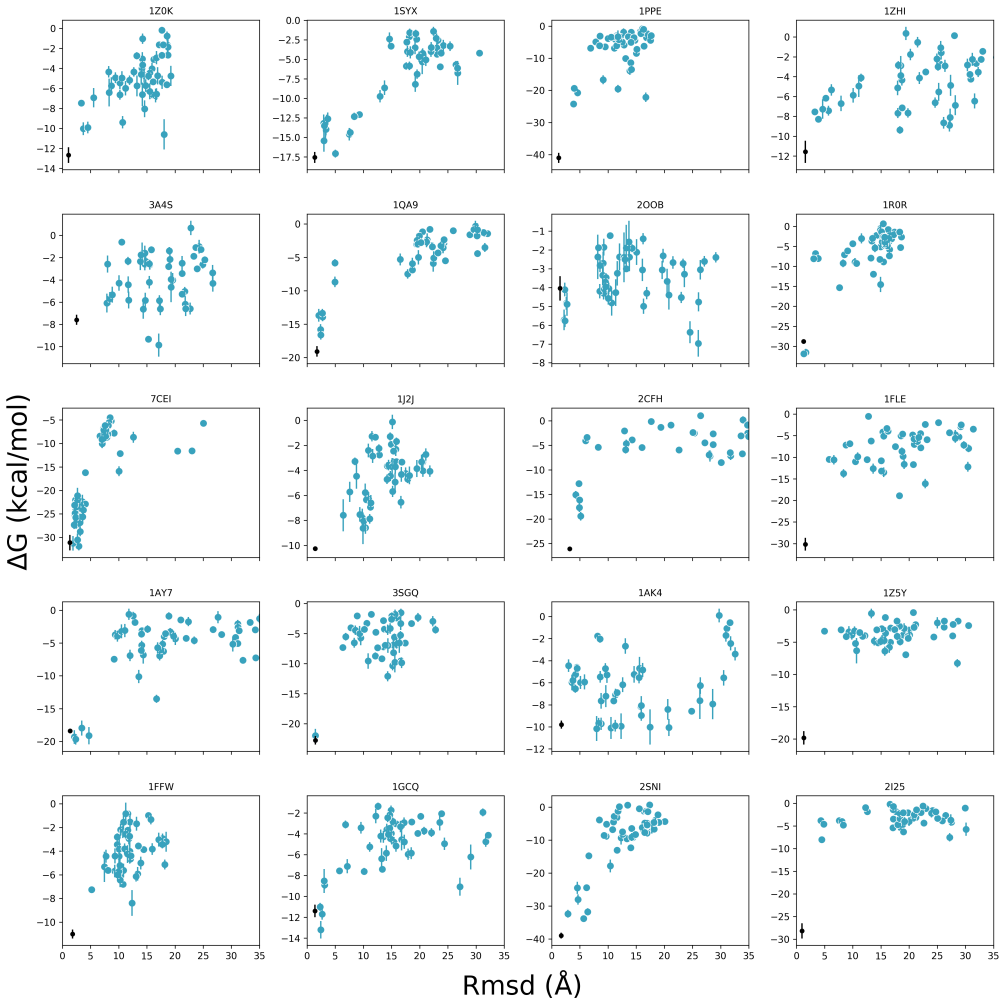


Figure C.3: RS score of all 50 poses for 20 protein-protein complexes vs. $\text{rmsd}_{\text{ligand}}$ (replica of minimal $\text{rmsd}_{\text{ligand}}$ in the first frame) from the native complex. The results of the explicit solvent RS-REMD simulations of the different poses (blue dots) and the native structures (black dots) are given with uncertainties.

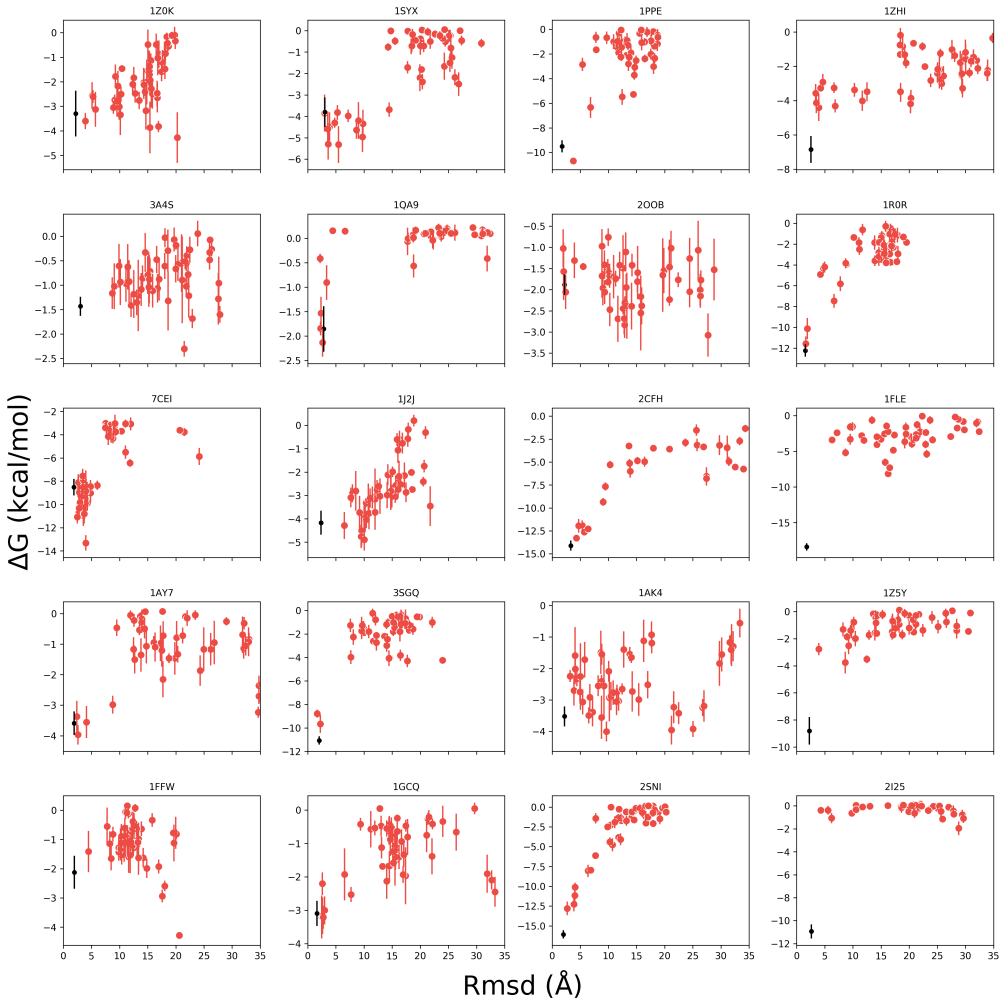


Figure C.4: RS score of all 50 poses for 20 protein-protein complexes vs. $\text{rmsd}_{\text{ligand}}$ (replica of minimal $\text{rmsd}_{\text{ligand}}$ in the first frame) from the native complex. The results of the implicit solvent RS-REMD simulations of the different poses (red dots) and the native structures (black dots) are given with uncertainties.

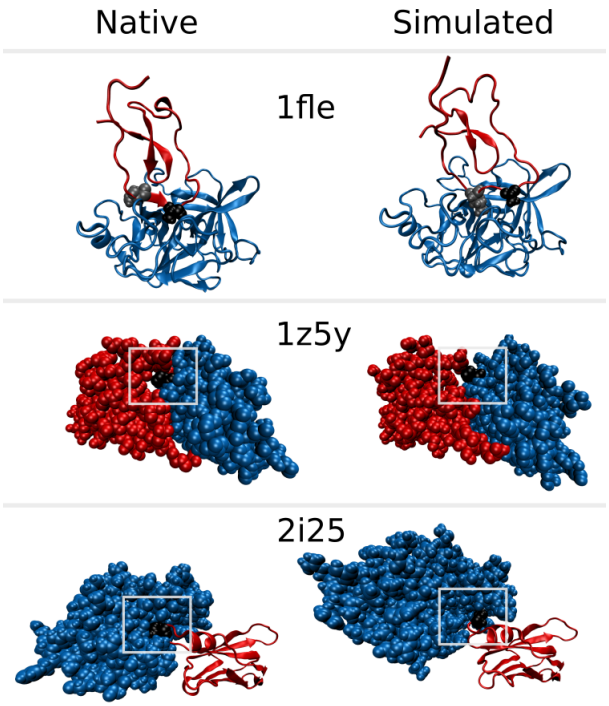


Figure C.5: Native complexes (left panel) and RS-REMD simulated complexes (right panel) in explicit solvent (reference replica) for three structures with only small deviations in the $rmsd_{ligand}$ but high differences in the corresponding scoring (see main text Figure 7.5). We identified the small structural details having a high impact on the binding affinity: (1) In case of 1fle in the native structure an alanine residue (black; vdw representation) of the ligand sits in the central pocket of the receptor. In the simulation the ligand is associated to the correct binding site but the binding is coordinated by different residues, an isoleucin residue sits in the mentioned receptor pocket (gray vdw representation). (2) For the second complex (1z5y) a proline residue of the receptor protein (black vdw spheres) fits deep into the native pocket of the ligand (red vdw representation). In the simulation this pocket is closed so that the proline residue is sterically hindered at the ligand surface and a compatible matching of ligand and receptor is not possible. (3) In case of 2i25 a serine amino acid of the native ligand protein fits deeply into the pocket of the receptor protein (black vdw spheres). In contrast, during the unbound simulations this residue sticks at the surface of the receptor (blue vdw representation) not being able to enter the pocket.

List of Publications

[1] **Till Siebenmorgen**, Michael Engelhard and Martin Zacharias. 'Prediction of protein–protein complexes using replica exchange with repulsive scaling'. In: Journal of Computational Chemistry 41.15 (2020), pp. 1436–1447.

[2] **Till Siebenmorgen** and Martin Zacharias. 'Computational prediction of protein–protein binding affinities'. In: Wiley Interdisciplinary Reviews: Computational Molecular Science (2019), e1448.

[3] **Till Siebenmorgen** and Martin Zacharias. 'Efficient Refinement and Free Energy Scoring of Predicted Protein–Protein Complexes Using Replica Exchange with Repulsive Scaling'. In: Journal of Chemical Information and Model- ing 60.11 (2020), pp. 5552–5562.

[4] **Till Siebenmorgen** and Martin Zacharias. 'Evaluation of predicted protein–protein complexes by binding free energy simulations'. In: Journal of Chemical Theory and Computation 15.3 (2019), pp. 2071–2086.

[5] **Till Siebenmorgen** and Martin Zacharias. 'Origin of ion specificity of telomeric DNA G-quadruplexes investigated by free-energy simulations'. In: Biophysical journal 112.11 (2017), pp. 2280–2290.

[6] Martyna Maszota-Zieleniak, Mateusz Marcisz, Małgorzata M. Kogut, **Till Siebenmorgen**, Martin Zacharias, Sergey A. Samsonov. 'Evaluation of Replica Exchange with Repulsive Scaling Approach for Docking Glycosaminoglycans.' In: Journal of Computational Chemistry (manuscript accepted).

Acknowledgments

Ich danke Martin Zacharias für die Einführung in die Welt der Protein-Simulationen und die Betreuung während der Promotion, sein stets offenes Ohr in allen Problemstellungen und seine Zuversicht bei schwierig erscheinenden Hindernissen. Außerdem danke ich dem gesamten bestehenden und ehemaligen T38-Team für viele Inspirationen und Hilfestellungen in den Projekten, aber vor allem für den tollen Zusammenhalt und die vielen Erlebnisse abseits der Arbeit. Insbesondere die Weihnachtsfeiern, Winterschulen und Ausflüge zur Isar werden mir noch lange im Gedächtnis bleiben. Mein besonderer Dank gilt Sonja für die stets gute Organisation, Danial für die vielen gemeinsamen Stunden im Büro, Julian für die Fahrrad-Expertise, Paul dafür, dass der Arbeitsalltag mit ihm nie langweilig wird und schließlich Max dafür, dass wir seit dem ersten Studientag den gleichen Weg beschreiten.

Ich danke der TUM für das Ermöglichen der Promotion und der deutschen Forschungsgesellschaft für die Finanzierung über den SFB 863. Für die Zusammenarbeit zum Interleukin-Projekt bedanke ich mich bei Isabel Aschenbrenner und Matthias Feige. Außerdem gilt mein Dank Sergey Samsonov und seinen Kollegen für die weiterführende Anwendung der RS-REMD Methode.

Für viele konstruktive Anmerkungen und Korrekturen verschiedener Abschnitte bedanke ich mich bei Massimo, Clio, Korbi, Max, Ralf, Brianda, Caro, Martha, Dagmar und Julian.

Mein abschließender Dank gilt meiner Familie und meinen Freunden. Ich danke Clio dafür, dass ich sie immer kontaktieren und ich mir ihrer Hilfe gewiss sein kann, auch abgesehen von chemischer Expertise. Ich danke meinen Eltern für ihre stete Unterstützung und dafür, dass sie mir diesen Weg ermöglicht haben. Zuletzt gilt mein ganz besonderer Dank Martha, für all die wunderbaren Momente, du bist mein Herz.

Bibliography

[1] Nicola GA Abrescia et al. 'Insights into assembly from structural analysis of bacteriophage PRD1'. In: *Nature* 432.7013 (2004), pp. 68–74.

[2] Stewart A Adcock and J Andrew McCammon. 'Molecular dynamics: survey of methods for simulating the activity of proteins'. In: *Chemical reviews* 106.5 (2006), pp. 1589–1615.

[3] Roman Affentranger, Ivano Tavernelli and Ernesto E Di Iorio. 'A novel Hamiltonian replica exchange MD protocol to enhance protein conformational space sampling'. In: *Journal of Chemical Theory and Computation* 2.2 (2006), pp. 217–228.

[4] Boris Aguilar, Richard Shadrach and Alexey V Onufriev. 'Reducing the secondary structure bias in the generalized Born model via R6 effective radii'. In: *Journal of Chemical Theory and Computation* 6.12 (2010), pp. 3613–3630.

[5] Mazen Ahmad et al. 'Adhesive water networks facilitate binding of protein interfaces'. In: *Nature communications* 2.1 (2011), pp. 1–7.

[6] Bruce Alberts et al. 'Molecular biology of the cell'. In: (2018).

[7] Matteo Aldeghi et al. 'Accurate calculation of the absolute free energy of binding for drug molecules'. In: *Chemical science* 7.1 (2016), pp. 207–218.

[8] Matteo Aldeghi et al. 'Statistical analysis on the performance of Molecular Mechanics Poisson–Boltzmann Surface Area versus absolute binding free energy calculations: Bromodomains as a case study'. In: *Journal of chemical information and modeling* 57.9 (2017), pp. 2203–2221.

[9] Rebecca F Alford et al. 'The Rosetta all-atom energy function for macromolecular modeling and design'. In: *Journal of chemical theory and computation* 13.6 (2017), pp. 3031–3048.

[10] Ramu Anandakrishnan et al. 'Speed of Conformational Change: Comparing Explicit and Implicit Solvent Molecular Dynamics Simulations'. In: *Biophysical Journal* 108.5 (Mar. 2015), pp. 1153–1164.

[11] Hans C Andersen. 'Molecular dynamics simulations at constant pressure and/or temperature'. In: *The Journal of chemical physics* 72.4 (1980), pp. 2384–2393.

[12] Nelly Andrusier et al. 'Principles of flexible protein–protein docking'. In: *Proteins: Structure, Function, and Bioinformatics* 73.2 (2008), pp. 271–289.

[13] Christian B Anfinsen. 'Principles that govern the folding of protein chains'. In: *Science* 181.4096 (1973), pp. 223–230.

[14] J Aqvist and John Marelius. 'The linear interaction energy method for predicting ligand binding free energies'. In: *Combinatorial chemistry & high throughput screening* 4.8 (2001), pp. 613–626.

[15] Ranjit Prasad Bahadur et al. 'A dissection of specific and non-specific protein–protein interfaces'. In: *Journal of molecular biology* 336.4 (2004), pp. 943–955.

[16] RP Bahadur and M Zacharias. 'The interface of protein-protein complexes: analysis of contacts and prediction of interactions'. In: *Cellular and Molecular Life Sciences* 65.7-8 (2008), pp. 1059–1072.

[17] Xiao-chen Bai, Greg McMullan and Sjors HW Scheres. 'How cryo-EM is revolutionizing structural biology'. In: *Trends in biochemical sciences* 40.1 (2015), pp. 49–57.

[18] Jacob B Bale et al. 'Accurate design of megadalton-scale two-component icosahedral protein complexes'. In: *Science* 353.6297 (2016), pp. 389–394.

[19] Nenad Ban et al. 'The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution'. In: *Science* 289.5481 (2000), pp. 905–920.

[20] Matthias GJ Baud et al. 'New synthetic routes to triazolo-benzodiazepine analogues: expanding the scope of the bump-and-hole approach for selective bromo and extra-terminal (BET) bromodomain inhibition'. In: *Journal of medicinal chemistry* 59.4 (2016), pp. 1492–1500.

[21] Ad Bax. 'Two-dimensional NMR and protein structure'. In: *Annual review of biochemistry* 58.1 (1989), pp. 223–256.

[22] Andrea Becchetti and Annarosa Arcangeli. *Integrins and ion channels: molecular complexes and signaling*. Vol. 674. Springer Science & Business Media, 2010.

[23] Ido Y Ben-Shalom et al. 'Efficient approximation of ligand rotational and translational entropy changes upon binding for use in MM-PBSA calculations'. In: *Journal of chemical information and modeling* 57.2 (2017), pp. 170–189.

[24] Charles H Bennett. 'Efficient estimation of free energy differences from Monte Carlo data'. In: *Journal of Computational Physics* 22.2 (1976), pp. 245–268.

[25] Herman JC Berendsen et al. 'Molecular dynamics with coupling to an external bath'. In: *The Journal of chemical physics* 81.8 (1984), pp. 3684–3690.

[26] HJC Berendsen, JR Grigera and TP Straatsma. 'The missing term in effective pair potentials'. In: *Journal of Physical Chemistry* 91.24 (1987), pp. 6269–6271.

[27] Helen M Berman et al. 'The protein data bank'. In: *Nucleic acids research* 28.1 (2000), pp. 235–242.

[28] Matthias Bischoff et al. 'Stereoselective construction of the 5-hydroxy diazabicyclo [4.3. 1] decane-2-one scaffold, a privileged motif for FK506-binding proteins'. In: *Organic letters* 16.20 (2014), pp. 5254–5257.

[29] Alexandre MJJ Bonvin. 'Flexible protein–protein docking'. In: *Current opinion in structural biology* 16.2 (2006), pp. 194–200.

[30] Stefan Boresch et al. 'Absolute binding free energies: a quantitative approach for their calculation'. In: *The Journal of Physical Chemistry B* 107.35 (2003), pp. 9535–9551.

[31] Gregory R Bowman, Vijay S Pande and Frank Noé. *An introduction to Markov state models and their application to long timescale molecular simulation*. Vol. 797. Springer Science & Business Media, 2013.

[32] Scott E Boyken et al. 'De novo design of protein homo-oligomers with modular hydrogen-bond network–mediated specificity'. In: *Science* 352.6286 (2016), pp. 680–687.

[33] Andreas Bracher et al. 'Structural characterization of the PPIase domain of FKBP51, a cochaperone of human Hsp90'. In: *Acta Crystallographica Section D: Biological Crystallography* 67.6 (2011), pp. 549–559.

[34] Ryan Brenke et al. 'Application of asymmetric statistical potentials to antibody–protein docking'. In: *Bioinformatics* 28.20 (2012), pp. 2608–2614.

[35] Anna Brückner et al. 'Yeast two-hybrid, a powerful tool for systems biology'. In: *International journal of molecular sciences* 10.6 (2009), pp. 2763–2788.

[36] Axel Brünger, Charles L Brooks III and Martin Karplus. 'Stochastic boundary conditions for molecular dynamics simulations of ST2 water'. In: *Chemical physics letters* 105.5 (1984), pp. 495–500.

[37] Ignasi Buch, Toni Giorgino and Gianni De Fabritiis. 'Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations'. In: *Proceedings of the National Academy of Sciences* 108.25 (2011), pp. 10184–10189.

[38] James W Caldwell and Peter A Kollman. 'Structure and properties of neat liquids using nonadditive molecular dynamics: water, methanol, and N-methylacetamide'. In: *The Journal of Physical Chemistry* 99.16 (1995), pp. 6208–6219.

[39] Huaiqing Cao, Yongqi Huang and Zhirong Liu. 'Interplay between binding affinity and kinetics in protein–protein interactions'. In: *Proteins: Structure, Function, and Bioinformatics* 84.7 (2016), pp. 920–933.

[40] Phil Carter et al. 'Protein–protein docking using 3D-Dock in rounds 3, 4, and 5 of CAPRI'. In: *Proteins: Structure, Function, and Bioinformatics* 60.2 (2005), pp. 281–288.

[41] DA Case et al. 'AMBER 18. 2018'. In: *University of California, San Francisco* (2018).

[42] DA Case et al. 'AMBER 16'. In: *University of California, San Francisco* (2016).

[43] Kandala VR Chary and Girjesh Govil. *NMR in biological systems: from molecules to human*. Vol. 6. Springer Science & Business Media, 2008.

[44] Sidhartha Chaudhury et al. 'Benchmarking and analysis of protein docking performance in Rosetta v3. 2'. In: *PloS one* 6.8 (2011), e22477.

[45] Fu Chen et al. 'Assessing the performance of the MM/PBSA and MM/GBSA methods. 6. Capability to predict protein–protein binding free energies and re-rank binding poses generated by protein–protein docking'. In: *Physical Chemistry Chemical Physics* 18.32 (2016), pp. 22129–22139.

[46] Yifan Cheng. 'Single-particle cryo-EM—How did it get here and where will it go'. In: *Science* 361.6405 (2018), pp. 876–880.

[47] Jean-Baptiste Chéron et al. 'Update of the ATTRACT force field for the prediction of protein–protein binding affinity'. In: *Journal of Computational Chemistry* 38.21 (2017), pp. 1887–1890.

[48] Alexander S Cheung et al. 'Solvation effects in calculated electrostatic association free energies for the C3d-CR2 complex and comparison with experimental data'. In: *Biopolymers: Original Research on Biomolecules* 93.6 (2010), pp. 509–519.

[49] Christophe Chipot. 'Frontiers in free-energy calculations of biological systems'. In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 4.1 (2014), pp. 71–89.

[50] John D Chodera and Frank Noé. 'Markov state models of biomolecular conformational dynamics'. In: *Current opinion in structural biology* 25 (2014), pp. 135–144.

[51] Song-Ho Chong and Sihyun Ham. 'Dynamics of hydration water plays a key role in determining the binding thermodynamics of protein complexes'. In: *Scientific reports* 7.1 (2017), pp. 1–10.

[52] Song-Ho Chong and Sihyun Ham. 'New computational approach for external entropy in protein–protein binding'. In: *Journal of chemical theory and computation* 12.6 (2016), pp. 2509–2516.

[53] Cyrus Chothia and Joël Janin. 'Principles of protein–protein recognition'. In: *Nature* 256.5520 (1975), pp. 705–708.

[54] Gwo-Yu Chuang et al. 'DARS (Decoys As the Reference State) potentials for protein-protein docking'. In: *Biophysical journal* 95.9 (2008), pp. 4217–4227.

[55] Pietro Ciceri et al. 'Dual kinase-bromodomain inhibitors for rationally designed polypharmacology'. In: *Nature chemical biology* 10.4 (2014), pp. 305–312.

[56] Anthony J Clark et al. 'Relative binding affinity prediction of charge-changing sequence mutations with FEP in protein–protein interfaces'. In: *Journal of molecular biology* 431.7 (2019), pp. 1481–1493.

[57] G Marius Clore and Angela M Gronenborn. 'Determining the structures of large proteins and protein complexes by NMR'. In: *Trends in biotechnology* 16.1 (1998), pp. 22–34.

[58] Ryan G Coleman et al. 'Ligand pose and orientational sampling in molecular docking'. In: *PloS one* 8.10 (2013), e75992.

[59] Jeffrey Comer et al. 'The adaptive biasing force method: Everything you always wanted to know but were afraid to ask'. In: *The Journal of Physical Chemistry B* 119.3 (2015), pp. 1129–1151.

[60] Wendy D Cornell et al. 'A second generation force field for the simulation of proteins, nucleic acids, and organic molecules'. In: *Journal of the American Chemical Society* 117.19 (1995), pp. 5179–5197.

[61] Peter Csermely, Robin Palotai and Ruth Nussinov. 'Induced fit, conformational selection and independent dynamic segments: an extended view of binding events'. In: *Nature Precedings* (2010), pp. 1–1.

[62] Jeremy Curuksu, Jiri Sponer and Martin Zacharias. 'Elbow flexibility of the kt38 RNA kink-turn motif investigated by free-energy molecular dynamics simulations'. In: *Biophysical journal* 97.7 (2009), pp. 2004–2013.

[63] Tom Darden, Darrin York and Lee Pedersen. 'Particle mesh Ewald: An N log (N) method for Ewald sums in large systems'. In: *The Journal of chemical physics* 98.12 (1993), pp. 10089–10092.

[64] Eric Darve and Andrew Pohorille. 'Calculating free energies using average force'. In: *The Journal of Chemical Physics* 115.20 (2001), pp. 9169–9183.

[65] Fred P Davis and Andrej Sali. 'PIBASE: a comprehensive database of structurally defined protein interfaces'. In: *Bioinformatics* 21.9 (2005), pp. 1901–1907.

[66] LF Pineda De Castro and M Zacharias. 'DAPI binding to the DNA minor groove: a continuum solvent analysis'. In: *Journal of Molecular Recognition* 15.4 (2002), pp. 209–220.

[67] Sjoerd J De Vries et al. 'HADDOCK versus HADDOCK: new features and performance of HADDOCK2. 0 on the CAPRI targets'. In: *Proteins: structure, function, and bioinformatics* 69.4 (2007), pp. 726–733.

[68] Sjoerd J de Vries et al. 'A web interface for easy flexible protein–protein docking with ATTRACT'. In: *Biophysical Journal* 108.3 (2015), pp. 462–465.

[69] Sjoerd de Vries and Martin Zacharias. 'Flexible docking and refinement with a coarse-grained protein model using ATTRACT'. In: *Proteins: Structure, Function, and Bioinformatics* 81.12 (2013), pp. 2167–2174.

[70] Nanjie Deng et al. 'Comparing alchemical and physical pathway methods for computing the absolute binding free energy of charged ligands'. In: *Physical Chemistry Chemical Physics* 20.25 (2018), pp. 17081–17092.

[71] Yuqing Deng and Benoıt Roux. 'Computations of standard binding free energies with molecular dynamics simulations'. In: *The Journal of Physical Chemistry B* 113.8 (2009), pp. 2234–2246.

[72] Aalt D. J. van Dijk and Alexandre M. J. J. Bonvin. 'Solvated docking: introducing water into the modelling of biomolecular complexes'. en. In: *Bioinformatics* 22.19 (Oct. 2006), pp. 2340–2347.

[73] Cyril Dominguez, Rolf Boelens and Alexandre MJJ Bonvin. 'HADDOCK: a protein- protein docking approach based on biochemical or biophysical information'. In: *Journal of the American Chemical Society* 125.7 (2003), pp. 1731–1737.

[74] Qiwen Dong and Shuigeng Zhou. 'Novel nonlinear knowledge-based mean force potentials based on machine learning'. In: *IEEE/ACM transactions on computational biology and bioinformatics* 8.2 (2010), pp. 476–486.

[75] Ron O Dror et al. 'Biomolecular simulation: a computational microscope for molecular biology'. In: *Annual review of biophysics* 41 (2012), pp. 429–452.

[76] Lili Duan, Xiao Liu and John ZH Zhang. 'Interaction entropy: a new paradigm for highly efficient and reliable computation of protein–ligand binding free energy'. In: *Journal of the American Chemical Society* 138.17 (2016), pp. 5722–5728.

[77] J Dubochet and AW McDowall. 'Vitrification of pure water for electron microscopy'. In: *Journal of Microscopy* 124.3 (1981), pp. 3–4.

[78] JW Eastwood and RW Hockney. 'Computer Simulation using particles'. In: *New York: Mc GrawHill* (1981).

[79] Elaine A Elion. 'The ste5p scaffold'. In: *Journal of Cell Science* 114.22 (2001), pp. 3967–3978.

[80] Dominika Elmlund, Sarah N Le and Hans Elmlund. 'High-resolution cryo-EM: the nuts and bolts'. In: *Current opinion in structural biology* 46 (2017), pp. 1–6.

[81] Gerald D Fasman. *Circular dichroism and the conformational analysis of biomolecules*. Springer Science & Business Media, 2013.

[82] Xixi Feng et al. 'Structure–affinity relationship analysis of selective FKBP51 ligands'. In: *Journal of medicinal chemistry* 58.19 (2015), pp. 7796–7806.

[83] Juan Fernández-Recio, Maxim Totrov and Ruben Abagyan. 'ICM-DISCO docking by global energy optimization with fully flexible side-chains'. In: *Proteins: Structure, Function, and Bioinformatics* 52.1 (2003), pp. 113–117.

[84] Panagis Filippakopoulos. 'Crystal Structure of the first bromodomain of human BRD4 in complex with a 3,5-dimethylisoxazol ligand (Pdb entry: 3svg, to be published)'. In: ().

[85] Panagis Filippakopoulos et al. 'Benzodiazepines and benzotriazepines as protein interaction inhibitors targeting bromodomains of the BET family'. In: *Bioorganic & medicinal chemistry* 20.6 (2012), pp. 1878–1886.

[86] Panagis Filippakopoulos et al. 'Selective inhibition of BET bromodomains'. In: *Nature* 468.7327 (2010), pp. 1067–1073.

[87] Sébastien Fiorucci and Martin Zacharias. 'Binding site prediction and improved scoring during flexible protein-protein docking with ATTRACT'. In: *Proteins: Structure, Function, and Bioinformatics* 78.15 (2010), pp. 3131–3139.

[88] Paul V Fish et al. 'Identification of a chemical probe for bromo and extra C-terminal bromodomain inhibition through optimization of a fragment-derived hit'. In: *Journal of medicinal chemistry* 55.22 (2012), pp. 9831–9837.

[89] Federico Fogolari, Alessandro Brigo and Henriette Molinari. 'The Poisson–Boltzmann equation for biomolecular electrostatics: a tool for structural biology'. In: *Journal of Molecular Recognition* 15.6 (2002), pp. 377–392.

[90] Oriol Fornes et al. 'On the use of knowledge-based potentials for the evaluation of models of protein–protein, protein–DNA, and protein–RNA interactions'. In: *Advances in protein chemistry and structural biology*. Vol. 94. Elsevier, 2014, pp. 77–120.

[91] Joachim Frank. 'Time-resolved cryo-electron microscopy: Recent progress'. In: *Journal of structural biology* 200.3 (2017), pp. 303–306.

[92] Daan Frenkel and Berend Smit. *Understanding molecular simulation: from algorithms to applications*. Vol. 1. Elsevier, 2001.

[93] Richard A Friesner et al. 'Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy'. In: *Journal of medicinal chemistry* 47.7 (2004), pp. 1739–1749.

[94] Haian Fu. *Protein-protein interactions: methods and applications*. Vol. 261. Springer Science & Business Media, 2004.

[95] Hiroaki Fukunishi, Osamu Watanabe and Shoji Takada. 'On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction'. In: *The Journal of chemical physics* 116.20 (2002), pp. 9058–9067.

[96] Steffen Gaali et al. 'Rapid, structure-based exploration of pipecolic acid amides as novel selective antagonists of the FK506-binding protein 51'. In: *Journal of Medicinal Chemistry* 59.6 (2016), pp. 2410–2422.

[97] Steffen Gaali et al. 'Selective inhibitors of the FK506-binding protein 51 by induced fit'. In: *Nature chemical biology* 11.1 (2015), pp. 33–37.

[98] Vytautas Gapsys et al. 'Accurate and rigorous prediction of the changes in protein free energies in a large-scale mutation scan'. In: *Angewandte Chemie International Edition* 55.26 (2016), pp. 7364–7368.

[99] Alfonso T Garcıa-Sosa and Ricardo L Mancera. 'Free energy calculations of mutations involving a tightly bound water molecule and ligand substitutions in a ligand-protein complex'. In: *Molecular Informatics* 29.8-9 (2010), pp. 589–600.

[100] Cunliang Geng et al. 'Finding the ΔΔG spot: Are predictors of binding affinity changes upon mutations in protein–protein interactions ready for it?' In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 9.5 (2019), e1410.

[101] Francesco Luigi Gervasio, Alessandro Laio and Michele Parrinello. 'Flexible docking in solution using metadynamics'. In: *Journal of the American Chemical Society* 127.8 (2005), pp. 2600–2607.

[102] Michael K Gilson et al. 'The statistical-thermodynamic basis for computation of binding affinities: a critical review'. In: *Biophysical journal* 72.3 (1997), pp. 1047–1069.

[103] Holger Gohlke and David A Case. 'Converging free energy estimates: MM-PB (GB) SA studies on the protein–protein complex Ras–Raf'. In: *Journal of computational chemistry* 25.2 (2004), pp. 238–250.

[104] Holger Gohlke, Christina Kiel and David A Case. 'Insights into protein–protein binding by binding free energy calculation and free energy decomposition for the Ras–Raf and Ras–RalGDS complexes'. In: *Journal of molecular biology* 330.4 (2003), pp. 891–913.

[105] Javier Gomez et al. 'The heat capacity of proteins'. In: *Proteins: Structure, Function, and Bioinformatics* 22.4 (1995), pp. 404–412.

[106] Ranganath Gopalakrishnan et al. 'Evaluation of synthetic FK506 analogues as ligands for the FK506-binding proteins 51 and 52'. In: *Journal of medicinal chemistry* 55.9 (2012), pp. 4114–4122.

[107] Ronald D Gorham, Chris A Kieslich and Dimitrios Morikis. 'Electrostatic clustering and free energy calculations provide a foundation for protein design and optimization'. In: *Annals of biomedical engineering* 39.4 (2011), pp. 1252–1263.

[108] Andreas W Gotz et al. 'Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. Generalized born'. In: *Journal of chemical theory and computation* 8.5 (2012), pp. 1542–1555.

[109] Jeffrey J Gray. 'High-resolution protein–protein docking'. In: *Current Opinion in Structural Biology* 16.2 (2006), pp. 183–193.

[110] Jeffrey J Gray et al. 'Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations'. In: *Journal of Molecular Biology* 331.1 (2003), pp. 281–299.

[111] Norma J Greenfield. 'Circular dichroism (CD) analyses of protein-protein interactions'. In: *Protein-Protein Interactions*. Springer, 2015, pp. 239–265.

[112] Norma J Greenfield. 'Methods to estimate the conformation of proteins and polypeptides from circular dichroism data'. In: *Analytical biochemistry* 235.1 (1996), pp. 1–10.

[113] Sam Z Grinter and Xiaoqin Zou. 'A Bayesian statistical approach of improving knowledge-based scoring functions for protein–ligand interactions'. In: *Journal of computational chemistry* 35.12 (2014), pp. 932–943.

[114] Paweł Grochowski and Joanna Trylska. 'Continuum molecular electrostatics, salt effects, and counterion binding—a review of the Poisson–Boltzmann theory and its modifications'. In: *Biopolymers: Original Research on Biomolecules* 89.2 (2008), pp. 93–113.

[115] M Michael Gromiha, K Yugandhar and Sherlyn Jemimah. 'Protein–protein interactions: scoring schemes and binding affinity'. In: *Current opinion in structural biology* 44 (2017), pp. 31–38.

[116] Alan Grossfield. 'WHAM: an Implementation of the Weighted Histogram Analysis Method. Version 2.0. 9'. In: *Rochester University: Rochester, NY* (2013).

[117] Tomasz Grycuk. 'Deficiency of the Coulomb-field approximation in the generalized Born model: An improved formula for Born radii evaluation'. In: *The Journal of Chemical Physics* 119.9 (2003), pp. 4817–4826.

[118] Jenny Gu and Philip E Bourne. *Structural bioinformatics*. Vol. 44. John Wiley & Sons, 2009.

[119] James C Gumbart, Benoıt Roux and Christophe Chipot. 'Efficient determination of protein–protein standard binding free energies from first principles'. In: *Journal of chemical theory and computation* 9.8 (2013), pp. 3789–3798.

[120] James C Gumbart, Benoıt Roux and Christophe Chipot. 'Standard binding free energies from computer simulations: What is the best strategy?' In: *Journal of chemical theory and computation* 9.1 (2013), pp. 794–802.

[121] Peter Güntert. 'Automated structure determination from NMR spectra'. In: *European Biophysics Journal* 38.2 (2009), p. 129.

[122] Hugo Guterres and Wonpil Im. 'Improving protein-ligand docking results with high-throughput molecular dynamics simulations'. In: *Journal of Chemical Information and Modeling* 60.4 (2020), pp. 2189–2198.

[123] Tomas Hansson, John Marelius and Johan Åqvist. 'Ligand binding affinity prediction by linear interaction energy methods'. In: *Journal of computer-aided molecular design* 12.1 (1998), pp. 27–35.

[124] Edward Harder et al. 'OPLS3: a force field providing broad coverage of drug-like small molecules and proteins'. In: *Journal of chemical theory and computation* 12.1 (2016), pp. 281–296.

[125] David S Hewings et al. 'Optimization of 3, 5-dimethylisoxazole derivatives as potent bromodomain ligands'. In: *Journal of medicinal chemistry* 56.8 (2013), pp. 3217–3227.

[126] Chad W. Hopkins et al. 'Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning'. In: *Journal of Chemical Theory and Computation* 11.4 (2015), pp. 1864–1874.

[127] Jozef Hritz and Chris Oostenbrink. 'Hamiltonian replica exchange molecular dynamics using soft-core interactions'. In: *Journal of Chemical Physics* 128.14 (2008), p. 144121.

[128] Sheng-You Huang. 'Exploring the potential of global protein–protein docking: an overview and critical assessment of current programs for automatic ab initio docking'. In: *Drug discovery today* 20.8 (2015), pp. 969–977.

[129] Sheng-You Huang. 'Search strategies and evaluation in protein-protein docking: principles, advances and challenges'. In: *Drug Discovery Today* 0 (2014), In press.

[130] Po-Ssu Huang, Scott E Boyken and David Baker. 'The coming of age of de novo protein design'. In: *Nature* 537.7620 (2016), pp. 320–327.

[131] Gary A Huber and Sangtae Kim. 'Weighted-ensemble Brownian dynamics simulations for protein association reactions'. In: *Biophysical journal* 70.1 (1996), pp. 97–110.

[132] Howook Hwang et al. 'Protein–protein docking benchmark version 3.0'. In: *Proteins: Structure, Function, and Bioinformatics* 73.3 (2008), pp. 705–709.

[133] Wonpil Im et al. 'Challenges in structural approaches to cell modeling'. In: *Journal of molecular biology* 428.15 (2016), pp. 2943–2964.

[134] Y Inbar, R Nussinov, HJ Wolfson et al. 'PatchDock and SymmDock: servers for rigid and symmetric docking'. In: *Nucleic Acids Res* 33 (2005), W363–W367.

[135] Saeed Izadi, Ramu Anandakrishnan and Alexey V Onufriev. 'Building water models: a different approach'. In: *The journal of physical chemistry letters* 5.21 (2014), pp. 3863–3871.

[136] Saeed Izadi et al. 'Accuracy comparison of generalized Born models in the calculation of electrostatic binding free energies'. In: *Journal of chemical theory and computation* 14.3 (2018), pp. 1656–1670.

[137] Joël Janin. 'Assessing predictions of protein–protein interaction: the CAPRI experiment'. In: *Protein Scienceence* 14.2 (2005), pp. 278–283.

[138] Lin Jiang et al. 'Potential of mean force for protein–protein interaction studies'. In: *Proteins: Structure, Function, and Bioinformatics* 46.2 (2002), pp. 190–196.

[139] Zhifeng Jing et al. 'Polarizable force fields for biomolecular simulations: Recent advances and applications'. In: *Annual Review of biophysics* 48 (2019), pp. 371–394.

[140] Sunhwan Jo et al. 'CHARMM-GUI: a web-based graphical user interface for CHARMM'. In: *Journal of computational chemistry* 29.11 (2008), pp. 1859–1865.

[141] Graham T Johnson et al. 'cellPACK: a virtual mesoscope to model and visualize structural systems biology'. In: *Nature methods* 12.1 (2015), pp. 85–91.

[142] William L Jorgensen. 'The many roles of computation in drug discovery'. In: *Science* 303.5665 (2004), pp. 1813–1818.

[143] William L Jorgensen and Jeffry D Madura. 'Temperature and size dependence for Monte Carlo simulations of TIP4P water'. In: *Molecular Physics* 56.6 (1985), pp. 1381–1392.

[144] William L Jorgensen et al. 'Comparison of simple potential functions for simulating liquid water'. In: *The Journal of chemical physics* 79.2 (1983), pp. 926–935.

[145] Agnieszka A Kaczor et al. 'Protein–Protein Docking in Drug Design and Discovery'. In: *Computational Drug Discovery and Design*. Springer, 2018, pp. 285–305.

[146] Hiqmet Kamberaj. *Molecular Dynamics Simulations in Statistical Physics: Theory and Applications*. Springer, 2020.

[147] Srinivasaraghavan Kannan and Martin Zacharias. 'Enhanced sampling of peptide and protein conformations using replica exchange simulations with a peptide backbone biasing-potential'. In: *Proteins: Structure, Function, and Bioinformatics* 66.3 (2007), pp. 697–706.

[148] Martin Karplus and J Andrew McCammon. 'Molecular dynamics simulations of biomolecules'. In: *Nature structural biology* 9.9 (2002), pp. 646–652.

[149] Johannes Kästner and Walter Thiel. 'Bridging the gap between thermodynamic integration and umbrella sampling provides a novel analysis method:"Umbrella integration"'. In: *The Journal of chemical physics* 123.14 (2005), p. 144104.

[150] Panagiotis L Kastritis and Alexandre MJJ Bonvin. 'Are scoring functions in protein- protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark'. In: *Journal of proteome research* 9.5 (2010), pp. 2216–2225.

[151] Panagiotis L Kastritis and Alexandre MJJ Bonvin. 'Erratum: Are scoring functions in protein-Protein Docking Ready to predict interactomes? Clues from a novel binding affinity benchmark (Journal of Proteome Research (2010) 9 (2216-2225)'. In: *Journal of Proteome Research* 10.2 (2011), pp. 921–922.

[152] Panagiotis L Kastritis and Alexandre MJJ Bonvin. 'On the binding affinity of macromolecular interactions: daring to ask why proteins interact'. In: *Journal of The Royal Society Interface* 10.79 (2013), p. 20120835.

[153] Panagiotis L Kastritis et al. 'A structure-based benchmark for protein–protein binding affinity'. In: *Protein Scienceence* 20.3 (2011), pp. 482–491.

[154] Panagiotis L Kastritis et al. 'Proteins feel more than they see: fine-tuning of binding affinity by properties of the non-interacting surface'. In: *Journal of molecular biology* 426.14 (2014), pp. 2632–2652.

[155] Ephraim Katchalski-Katzir et al. 'Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques'. In: *Proceedings of the National Academy of Sciences* 89.6 (1992), pp. 2195–2199.

[156] Ozlem Keskin, Buyong Ma and Ruth Nussinov. 'Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues'. In: *Journal of molecular biology* 345.5 (2005), pp. 1281–1294.

[157] Maximilian Kienlein and Martin Zacharias. 'Ligand binding and global adaptation of the GlnPQ substrate binding domain 2 revealed by molecular dynamics simulations'. In: *Protein Science* 29.12 (2020), pp. 2482–2494.

[158] Young C Kim et al. 'Replica exchange simulations of transient encounter complexes in protein–protein association'. In: *Proceedings of the National Academy of Sciences* 105.35 (2008), pp. 12855–12860.

[159] Neil P King et al. 'Accurate design of co-assembling multi-component protein nanomaterials'. In: *Nature* 510.7503 (2014), pp. 103–108.

[160] John G Kirkwood. 'Statistical mechanics of fluid mixtures'. In: *The Journal of chemical physics* 3.5 (1935), pp. 300–313.

[161] Pavel V Klimovich, Michael R Shirts and David L Mobley. 'Guidelines for the analysis of free energy calculations'. In: *Journal of computer-aided molecular design* 29.5 (2015), pp. 397–411.

[162] Małgorzata M Kogut et al. 'Computational insights into the role of calcium ions in protein–glycosaminoglycan systems'. In: *Physical Chemistry Chemical Physics* (2021).

[163] Gergely Kohut et al. 'Protein–Ligand interaction energy-based entropy calculations: fundamental challenges for flexible systems'. In: *The Journal of Physical Chemistry B* 122.32 (2018), pp. 7821–7827.

[164] Hironori Kokubo, Toshimasa Tanaka and Yuko Okamoto. 'Ab Initio prediction of protein–ligand binding structures by replica-exchange umbrella sampling simulations'. In: *Journal of computational chemistry* 32.13 (2011), pp. 2810–2821.

[165] Peter A Kollman et al. 'Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models'. In: *Accounts of chemical research* 33.12 (2000), pp. 889–897.

[166] Tanja Kortemme and David Baker. 'Computational design of protein–protein interactions'. In: *Current opinion in chemical biology* 8.1 (2004), pp. 91–97.

[167] Daniel E Koshland Jr. 'Das Schüssel-Schloß-Prinzip und die Induced-fit-Theorie'. In: *Angewandte Chemie* 106.23-24 (1994), pp. 2468–2472.

[168] Marcin Król, Alexander L. Tournier and Paul A. Bates. 'Flexible relaxation of rigid-body docking solutions'. In: *Proteins: Structure, Function, and Bioinformatics* 68.1 (2007), pp. 159–169.

[169] Shankar Kumar et al. 'The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method'. In: *Journal of computational chemistry* 13.8 (1992), pp. 1011–1021.

[170] Sriram Kumaraswamy and Renee Tobias. 'Label-free kinetic analysis of an antibody–antigen interaction using biolayer interferometry'. In: *Protein-Protein Interactions*. Springer, 2015, pp. 165–182.

[171] Petras J Kundrotas et al. 'Dockground: a comprehensive data resource for modeling of protein complexes'. In: *Protein Scienceence* 27.1 (2018), pp. 172–181.

[172] Petras J Kundrotas et al. 'Templates are available to model nearly all complexes of structurally characterized proteins'. In: *Proceedings of the National Academy of Sciences* 109.24 (2012), pp. 9438–9441.

[173] John Kuriyan, Boyana Konforti and David Wemmer. *The molecules of life: Physical and chemical principles*. Garland Science, 2012.

[174] Daisuke Kuroda and Jeffrey J Gray. 'Pushing the backbone in protein-protein docking'. In: *Structure* 24.10 (2016), pp. 1821–1829.

[175] John E Ladbury and Babur Z Chowdhry. 'Sensing the heat: the application of isothermal titration calorimetry to thermodynamic studies of biomolecular interactions'. In: *Chemistry & biology* 3.10 (1996), pp. 791–801.

[176] Cheng-Tsung Lai et al. 'Rational modulation of the induced-fit conformational change for slow-onset inhibition in Mycobacterium tuberculosis InhA'. In: *Biochemistry* 54.30 (2015), pp. 4683–4691.

[177] Alessandro Laio and Michele Parrinello. 'Escaping free-energy minima'. In: *Proceedings of the National Academy of Sciences* 99.20 (2002), pp. 12562–12566.

[178] Andrew R Leach. *Molecular modelling: principles and applications*. Pearson education, 2001.

[179] Hui Sun Lee et al. 'Application of binding free energy calculations to prediction of binding modes and affinities of MDM2 and MDMX inhibitors'. In: *Journal of chemical information and modeling* 52.7 (2012), pp. 1821–1832.

[180] Marc F Lensink, Raúl Méndez and Shoshana J Wodak. 'Docking and scoring protein complexes: CAPRI 3rd Edition'. In: *Proteins: Structure, Function, and Bioinformatics* 69.4 (2007), pp. 704–718.

[181] Marc F Lensink, Sameer Velankar and Shoshana J Wodak. 'Modeling protein–protein and protein–peptide complexes: CAPRI 6th edition'. In: *Proteins: Structure, Function, and Bioinformatics* 85.3 (2017), pp. 359–377.

[182] Marc F Lensink and Shoshana J Wodak. 'Docking and scoring protein interactions: CAPRI 2009'. In: *Proteins: Structure, Function, and Bioinformatics* 78.15 (2010), pp. 3073–3084.

[183] Huei-Jiun Li et al. 'A structural and energetic model for the slow-onset inhibition of the Mycobacterium tuberculosis enoyl-ACP reductase InhA'. In: *ACS chemical biology* 9.4 (2014), pp. 986–993.

[184] Lin Li et al. 'ASPDock: protein-protein docking algorithm using atomic solvation parameters model'. In: *BMC bioinformatics* 12.1 (2011), p. 36.

[185] Hai Lin and Donald G. Truhlar. 'QM/MM: what have we learned, where are we, and where do we go from here?' In: *Theoretical Chemistry Accounts* 117.2 (2006), p. 185.

[186] Song Liu et al. 'A physical reference state unifies the structure-derived potential of mean force for protein folding and binding'. In: *Proteins: Structure, Function, and Bioinformatics* 56.1 (2004), pp. 93–101.

[187] Xiao Liu, Long Peng and John ZH Zhang. 'Accurate and Efficient Calculation of Protein–Protein Binding Free Energy-Interaction Entropy with Residue Type-Specific Dielectric Constants'. In: *Journal of chemical information and modeling* 59.1 (2018), pp. 272–281.

[188] HA Lorentz. 'Ueber die Anwendung des Satzes vom Virial in der kinetischen Theorie der Gase'. In: *Annalen der Physik* 248.1 (1881), pp. 127–136.

[189] Hongfeng Lou and Robert I Cukier. 'Molecular dynamics of apo-adenylate kinase: a distance replica exchange method for the free energy of conformational fluctuations'. In: *The journal of physical chemistry B* 110.47 (2006), pp. 24121–24137.

[190] Hui Lu, Long Lu and Jeffrey Skolnick. 'Development of unified statistical potentials describing protein-protein interactions'. In: *Biophysical journal* 84.3 (2003), pp. 1895–1901.

[191] Mingyang Lu, Athanasios D Dousis and Jianpeng Ma. 'OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing'. In: *Journal of molecular biology* 376.1 (2008), pp. 288–301.

[192] Manuel P Luitz and Martin Zacharias. 'Protein–ligand docking using hamiltonian replica exchange simulations with soft core potentials'. In: *Journal of Chemical Information and Modeling* 54.6 (2014), pp. 1669–1675.

[193] Manuel P Luitz and Martin Zacharias. 'Role of tyrosine hot-spot residues at the interface of colicin E9 and immunity protein 9: A comparative free energy simulation study'. In: *Proteins: Structure, Function, and Bioinformatics* 81.3 (2013), pp. 461–468.

[194] Manuel Luitz et al. 'Exploring biomolecular dynamics and interactions using advanced sampling methods'. In: *Journal of Physics: Condensed Matter* 27.32 (2015), p. 323101.

[195] Jiankun Lyu et al. 'Ultra-large library docking for discovering new chemotypes'. In: *Nature* 566.7743 (2019), pp. 224–229.

[196] Alex D MacKerell Jr et al. 'All-atom empirical potential for molecular modeling and dynamics studies of proteins'. In: *The journal of physical chemistry B* 102.18 (1998), pp. 3586–3616.

[197] Irene Maffucci and Alessandro Contini. 'Improved computation of protein–protein relative binding energies with the Nwat-MMGBSA method'. In: *Journal of chemical information and modeling* 56.9 (2016), pp. 1692–1704.

[198] James A. Maier et al. 'ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB'. In: *Journal of Chemical Theory and Computation* 11.8 (2015), pp. 3696–3713.

[199] Siewert J Marrink et al. 'The MARTINI force field: coarse grained model for biomolecular simulations'. In: *The journal of physical chemistry B* 111.27 (2007), pp. 7812–7824.

[200] Joseph A Marsh and Sarah A Teichmann. 'Structure, dynamics, assembly, and evolution of protein complexes'. In: *Annual review of biochemistry* 84 (2015), pp. 551–575.

[201] Marc A Martı-Renom et al. 'Comparative protein structure modeling of genes and genomes'. In: *Annual Review of Biophysics and Biomolecular Structure* 29.1 (2000), pp. 291–325.

[202] Andreas M März et al. 'Large FK506-binding proteins shape the pharmacology of rapamycin'. In: *Molecular and cellular biology* 33.7 (2013), pp. 1357–1367.

[203] Nicholas A Marze et al. 'Efficient flexible backbone protein–protein docking for challenging targets'. In: *Bioinformatics* 34.20 (2018), pp. 3461–3469.

[204] Efrat Mashiach, Ruth Nussinov and Haim J Wolfson. 'FiberDock: Flexible induced-fit backbone refinement in molecular docking'. In: *Proteins: Structure, Function, and Bioinformatics* 78.6 (2010), pp. 1503–1519.

[205] Efrat Mashiach, Ruth Nussinov and Haim J. Wolfson. 'FiberDock: a web server for flexible induced-fit backbone refinement in molecular docking'. In: *Nucleic Acids Research* 38.suppl 2 (2010), W457–W461.

[206] Efrat Mashiach et al. 'An integrated suite of fast docking algorithms'. In: *Proteins: Structure, Function, and Bioinformatics* 78.15 (2010), pp. 3197–3204.

[207] Martyna Maszota-Zieleniak et al. ‘Evaluation of Replica Exchange with Repulsive Scaling Approach for Docking Glycosaminoglycans’. In: *Journal of Computational Chemistry* (2021).

[208] Ali May et al. ‘Coarse-grained versus atomistic simulations: realistic interaction free energies for real proteins’. In: *Bioinformatics* 30.3 (2014), pp. 326–334.

[209] Andreas May and Martin Zacharias. ‘Accounting for global protein deformability during protein–protein and protein–ligand docking’. In: *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* 1754.1-2 (2005), pp. 225–231.

[210] Andreas May and Martin Zacharias. ‘Energy minimization in low-frequency normal modes to efficiently allow for global flexibility during systematic protein–protein docking’. In: *Proteins: Structure, Function, and Bioinformatics* 70.3 (2008), pp. 794–809.

[211] Andreas May and Martin Zacharias. ‘Protein-ligand docking accounting for receptor side chain and global flexibility in normal modes: evaluation on kinase inhibitor cross docking’. In: *J Med Chem* 51.12 (2008). PMID: 18517186, pp. 3499–3506.

[212] Jarek Meller. ‘Molecular dynamics’. In: *e LS* (2001).

[213] William M Menzer et al. ‘Simple entropy terms for end-point binding free energy calculations’. In: *Journal of chemical theory and computation* 14.11 (2018), pp. 6035–6049.

[214] Nicholas Metropolis et al. ‘Equation of state calculations by fast computing machines’. In: *The journal of chemical physics* 21.6 (1953), pp. 1087–1092.

[215] C.L. Meyerkord and H. Fu. *Protein-Protein Interactions: Methods and Applications*. Methods in Molecular Biology. Springer New York, 2016.

[216] Julien Michel, Nicolas Foloppe and Jonathan W Essex. ‘Rigorous free energy calculations in structure-based drug design’. In: *Molecular informatics* 29.8-9 (2010), pp. 570–578.

[217] Julian Mintseris et al. ‘Integrating statistical pair potentials into protein complex prediction’. In: *Proteins: Structure, Function, and Bioinformatics* 69.3 (2007), pp. 511–520.

[218] Sanzo Miyazawa and Robert L Jernigan. ‘An empirical energy potential with a reference state for protein fold and sequence recognition’. In: *Proteins: Structure, Function, and Bioinformatics* 36.3 (1999), pp. 357–369.

[219] Iain H Moal and Paul A Bates. ‘SwarmDock and the use of normal modes in protein-protein docking’. In: *International journal of molecular sciences* 11.10 (2010), pp. 3623–3648.

[220] David L Mobley and Michael K Gilson. 'Predicting binding free energies: frontiers and benchmarks'. In: *Annual review of biophysics* 46 (2017), pp. 531–558.

[221] John Mongan et al. 'Generalized Born model with a simple, robust molecular volume correction'. In: *Journal of chemical theory and computation* 3.1 (2007), pp. 156–169.

[222] Luca Monticelli and D Peter Tieleman. 'Force fields for classical molecular dynamics'. In: *Biomolecular simulations: Methods and protocols* (2013), pp. 197–213.

[223] Irina S Moreira, Pedro A Fernandes and Maria J Ramos. 'Protein–protein docking dealing with the unknown'. In: *Journal of computational chemistry* 31.2 (2010), pp. 317–342.

[224] Roberto Mosca, Arnaud Céol and Patrick Aloy. 'Interactome3D: adding structural details to protein networks'. In: *Nature methods* 10.1 (2013), p. 47.

[225] Roberto Mosca et al. '3did: a catalog of domain-based interactions of known three-dimensional structure'. In: *Nucleic acids research* 42.D1 (2014), pp. D374–D379.

[226] Roberto Mosca et al. 'Towards a detailed atlas of protein–protein interactions'. In: *Current opinion in structural biology* 23.6 (2013), pp. 929–940.

[227] Demetri T Moustakas et al. 'Development and validation of a modular, extensible docking program: DOCK 5'. In: *Journal of computer-aided molecular design* 20.10-11 (2006), pp. 601–619.

[228] Asher Mullard. *Protein–protein interaction inhibitors get into the groove*. 2012.

[229] Lenka Munoz. 'Non-kinase targets of protein kinase inhibitors'. In: *Nature Reviews Drug Discovery* 16.6 (2017), p. 424.

[230] Hai Nguyen, Daniel R Roe and Carlos Simmerling. 'Improved generalized born solvent model parameters for protein simulations'. In: *Journal of chemical theory and computation* 9.4 (2013), pp. 2020–2034.

[231] Zaneta Nikolovska-Coleska. 'Studying protein-protein interactions using surface plasmon resonance'. In: *Protein-Protein Interactions*. Springer, 2015, pp. 109–138.

[232] Irene MA Nooren and Janet M Thornton. 'Diversity of protein–protein interactions'. In: *The EMBO journal* 22.14 (2003), pp. 3486–3492.

[233] Shuichi Nosé. 'A unified formulation of the constant temperature molecular dynamics methods'. In: *The Journal of chemical physics* 81.1 (1984), pp. 511–519.

[234] Alexey Onufriev. 'Continuum electrostatics solvent modeling with the generalized Born model'. In: *Modeling Solvent Environments: Applications to Simulations of Biomolecules* (2010), pp. 127–165.

[235] Alexey Onufriev. 'Implicit solvent models in molecular dynamics simulations: A brief overview'. In: *Annual Reports in Computational Chemistry* 4 (2008), pp. 125–137.

[236] Alexey Onufriev, Donald Bashford and David A Case. 'Exploring protein native states and large-scale conformational changes with a modified generalized born model'. In: *Proteins: Structure, Function, and Bioinformatics* 55.2 (2004), pp. 383–394.

[237] Alexey Onufriev, Donald Bashford and David A. Case. 'Modification of the Generalized Born Model Suitable for Macromolecules'. In: *Journal of Physical Chemistry B* 104.15 (Apr. 2000), pp. 3712–3720.

[238] Katja Ostermeir and Martin Zacharias. 'Accelerated flexible protein-ligand docking using Hamiltonian replica exchange with a repulsive biasing potential'. In: *PloS one* 12.2 (2017), e0172072.

[239] Katja Ostermeir and Martin Zacharias. 'Rapid alchemical free energy calculation employing a generalized born implicit solvent model'. In: *The Journal of Physical Chemistry B* 119.3 (2015), pp. 968–975.

[240] Albert C Pan et al. 'Atomic-level characterization of protein–protein association'. In: *Proceedings of the National Academy of Sciences* 116.10 (2019), pp. 4244–4249.

[241] Richard W Pastor, Bernard R Brooks and Attila Szabo. 'An analysis of the accuracy of Langevin and molecular dynamics algorithms'. In: *Molecular Physics* 65.6 (1988), pp. 1409–1419.

[242] RW Pastor. 'Techniques and applications of Langevin dynamics simulations'. In: *The Molecular Dynamics of Liquid Crystals*. Springer, 1994, pp. 85–138.

[243] Jan Walther Perthold and Chris Oostenbrink. 'GroScore: Accurate Scoring of Protein–Protein Binding Poses Using Explicit-Solvent Free-Energy Calculations'. In: *Journal of Chemical Information and Modeling* 59.12 (2019), pp. 5074–5085.

[244] Jan Walther Perthold and Chris Oostenbrink. 'Simulation of reversible protein–protein binding and calculation of binding free energies using perturbed distance restraints'. In: *Journal of chemical theory and computation* 13.11 (2017), pp. 5697–5708.

[245] Gregory A Petsko and Dagmar Ringe. *Protein structure and function*. New Science Press, 2004.

[246] Marharyta Petukh, Minghui Li and Emil Alexov. 'Predicting binding free energy change caused by point mutations with knowledge-modified MM/PBSA method'. In: *PLoS computational biology* 11.7 (2015), e1004276.

[247] Sarah Picaud et al. 'RVX-208, an inhibitor of BET transcriptional regulators with selectivity for the second bromodomain'. In: *Proceedings of the National Academy of Sciences* 110.49 (2013), pp. 19754–19759.

[248] Nuria Plattner et al. 'Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling'. In: *Nature chemistry* 9.10 (2017), p. 1005.

[249] Vladimir Potapov, Mati Cohen and Gideon Schreiber. 'Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details'. In: *Protein engineering, design & selection* 22.9 (2009), pp. 553–560.

[250] Sanbo Qin, Xiaodong Pang and Huan-Xiang Zhou. 'Automated prediction of protein association rate constants'. In: *Structure* 19.12 (2011), pp. 1744–1751.

[251] Ronald T Raines. 'Fluorescence polarization assay to quantify protein-protein interactions: an update'. In: *Protein-Protein Interactions*. Springer, 2015, pp. 323–327.

[252] Seesandra V Rajagopala et al. 'The binary protein-protein interaction landscape of Escherichia coli'. In: *Nature biotechnology* 32.3 (2014), pp. 285–290.

[253] Giulio Rastelli et al. 'Binding estimation after refinement, a new automated procedure for the refinement and rescoring of docked ligands in virtual screening'. In: *Chemical Biology Drug Design* 73.3 (2009), pp. 283–286.

[254] Sereina Riniker et al. 'Calculation of relative free energies for ligand-protein binding, solvation, and conformational transitions using the GROMOS software'. In: *The Journal of Physical Chemistry B* 115.46 (2011), pp. 13570–13577.

[255] David W Ritchie and Vishwesh Venkatraman. 'Ultra-fast FFT protein docking on graphics processors'. In: *Bioinformatics* 26.19 (2010), pp. 2398–2405.

[256] F Anthony Romero et al. 'Disrupting acetyl-lysine recognition: progress in the development of bromodomain inhibitors'. In: *Journal of medicinal chemistry* 59.4 (2016), pp. 1271–1298.

[257] Bernhard Rupp. *Biomolecular crystallography: principles, practice, and application to structural biology*. Garland Science, 2009.

[258] Jean-Paul Ryckaert, Giovanni Ciccotti and Herman JC Berendsen. 'Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes'. In: *Journal of Computational Physics* 23.3 (1977), pp. 327–341.

[259] Romelia Salomon-Ferrer et al. 'Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald'. In: *Journal of chemical theory and computation* 9.9 (2013), pp. 3878–3888.

[260] Alexander Sasse et al. 'Rapid design of knowledge-based scoring potentials for enrichment of near-native geometries in protein-protein docking'. In: *PloS one* 12.1 (2017), e0170625.

[261] Sjors HW Scheres. 'RELION: implementation of a Bayesian approach to cryo-EM structure determination'. In: *Journal of structural biology* 180.3 (2012), pp. 519–530.

[262] Christina EM Schindler et al. 'Large-scale assessment of binding free energy calculations in active drug discovery projects'. In: *Journal of Chemical Information and Modeling* (2020).

[263] Christina EM Schindler, Sjoerd J de Vries and Martin Zacharias. 'iATTRACT: Simultaneous global and local interface optimization for protein–protein docking refinement'. In: *Proteins: Structure, Function, and Bioinformatics* 83.2 (2015), pp. 248–258.

[264] Tamar Schlick. *Molecular modeling and simulation: an interdisciplinary guide: an interdisciplinary guide*. Vol. 21. Springer Science & Business Media, 2010.

[265] Daniel V Schroeder. *An introduction to thermal physics*. 1999.

[266] Yibing Shan et al. 'How does a drug molecule find its target binding site?' In: *Journal of the American Chemical Society* 133.24 (2011), pp. 9181–9183.

[267] David E Shaw et al. 'Anton 2: raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer'. In: *SC'14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE. 2014, pp. 41–53.

[268] Bradley Sherborne et al. 'Collaborating to improve the use of free-energy and other quantitative methods in drug discovery'. In: *Journal of computer-aided molecular design* 30.12 (2016), pp. 1139–1141.

[269] C David Sherrill. 'An introduction to Hartree-Fock molecular orbital theory'. In: *School of Chemistry and Biochemistry Georgia Institute of Technology* (2000).

[270] Ai Shinobu et al. 'Refining evERdock: Improved selection of good protein-protein complex models achieved by MD optimization and use of multiple conformations'. In: *Journal of Chemical Physics* 149.19 (Nov. 2018), p. 195101.

[271] Michael R Shirts and John D Chodera. 'Statistically optimal analysis of samples from multiple equilibrium states'. In: *The Journal of chemical physics* 129.12 (2008), p. 124105.

[272] Michael R Shirts and Vijay S Pande. 'Comparison of efficiency and bias of free energies computed by exponential averaging, the Bennett acceptance ratio, and thermodynamic integration'. In: *The Journal of chemical physics* 122.14 (2005), p. 144107.

[273] Benjamin A Shoemaker and Anna R Panchenko. 'Deciphering protein–protein interactions. Part II. Computational methods to predict protein and domain interaction partners'. In: *PLoS Comput Biol* 3.4 (2007), e43.

[274] Matthew D Shoulders and Ronald T Raines. 'Collagen structure and stability'. In: *Annual review of biochemistry* 78 (2009), pp. 929–958.

[275] Andrew Shrake and John A Rupley. 'Environment and exposure to solvent of protein atoms. Lysozyme and insulin'. In: *Journal of molecular biology* 79.2 (1973), pp. 351–371.

[276] Till Siebenmorgen, Michael Engelhard and Martin Zacharias. 'Prediction of protein–protein complexes using replica exchange with repulsive scaling'. In: *Journal of Computational Chemistry* 41.15 (2020), pp. 1436–1447.

[277] Till Siebenmorgen and Martin Zacharias. 'Computational prediction of protein–protein binding affinities'. In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* (2019), e1448.

[278] Till Siebenmorgen and Martin Zacharias. 'Efficient Refinement and Free Energy Scoring of Predicted Protein–Protein Complexes Using Replica Exchange with Repulsive Scaling'. In: *Journal of Chemical Information and Modeling* 60.11 (2020), pp. 5552–5562.

[279] Till Siebenmorgen and Martin Zacharias. 'Evaluation of predicted protein–protein complexes by binding free energy simulations'. In: *Journal of Chemical Theory and Computation* 15.3 (2019), pp. 2071–2086.

[280] Till Siebenmorgen and Martin Zacharias. 'Origin of ion specificity of telomeric DNA G-quadruplexes investigated by free-energy simulations'. In: *Biophysical journal* 112.11 (2017), pp. 2280–2290.

[281] Daniel-Adriano Silva et al. 'De novo design of potent and selective mimics of IL-2 and IL-15'. In: *Nature* 565.7738 (2019), pp. 186–191.

[282] Inês CM Simões et al. 'New parameters for higher accuracy in the computation of binding free energy differences upon Alanine Scanning Mutagenesis on protein–protein interfaces'. In: *Journal of Chemical Information and Modeling* 57.1 (2017), pp. 60–72.

[283] Manfred J Sippl and Sabine Weitckus. 'Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations'. In: *Proteins: Structure, Function, and Bioinformatics* 13.3 (1992), pp. 258–271.

[284] Pär Söderhjelm, Gareth A Tribello and Michele Parrinello. 'Locating binding poses in protein-ligand systems using reconnaissance metadynamics'. In: *Proceedings of the National Academy of Sciences* 109.14 (2012), pp. 5170–5175.

[285] Neelesh Soni and MS Madhusudhan. 'Computational modeling of protein assemblies'. In: *Current opinion in structural biology* 44 (2017), pp. 179–189.

[286] Marc Souaille and Benoıt Roux. 'Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations'. In: *Computer physics communications* 135.1 (2001), pp. 40–57.

[287] Einat Sprinzak, Shmuel Sattath and Hanah Margalit. 'How reliable are experimental protein–protein interaction data?' In: *Journal of molecular biology* 327.5 (2003), pp. 919–923.

[288] Amelie Stein et al. 'A systematic study of the energetics involved in structural changes upon association and connectivity in protein interaction networks'. In: *Structure* 19.6 (2011), pp. 881–889.

[289] W Clark Still et al. 'Semianalytical treatment of solvation for molecular mechanics and dynamics'. In: *J. Am. Chem. Soc* 112.16 (1990), pp. 6127–6129.

[290] Yuji Sugita and Yuko Okamoto. 'Replica-exchange molecular dynamics method for protein folding'. In: *Chemical Physics Letters* 314.1-2 (1999), pp. 141–151.

[291] Huiyong Sun et al. 'Assessing the performance of MM/PBSA and MM/GBSA methods. 7. Entropy effects on the performance of end-point binding free energy calculation approaches'. In: *Physical Chemistry Chemical Physics* 20.21 (2018), pp. 14450–14460.

[292] Zhaoxi Sun et al. 'Interaction entropy for protein-protein binding'. In: *The Journal of Chemical Physics* 146.12 (2017), p. 124124.

[293] Kazuhiro Takemura, Nobuyuki Matubayasi and Akio Kitao. 'Binding free energy analysis of protein-protein docking model structures by evERdock'. In: *The Journal of chemical physics* 148.10 (2018), p. 105101.

[294] Zhiqiang Tan. 'On a likelihood approach for Monte Carlo integration'. In: *Journal of the American Statistical Association* 99.468 (2004), pp. 1027–1036.

[295] Seiji Tanaka and Harold A Scheraga. 'Medium-and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins'. In: *Macromolecules* 9.6 (1976), pp. 945–950.

[296] Maxim Totrov and Ruben Abagyan. 'Flexible ligand docking to multiple receptor conformations: a practical alternative'. In: *Current opinion in structural biology* 18.2 (2008), pp. 178–184.

[297] Abdulnour Y Toukmaji and John A Board Jr. 'Ewald summation techniques in perspective: a survey'. In: *Computer physics communications* 95.2-3 (1996), pp. 73–92.

[298] Johannes Träg and Dirk Zahn. 'Improved GAFF2 parameters for fluorinated alkanes and mixed hydro-and fluorocarbons'. In: *Journal of Molecular Modeling* 25.2 (2019), p. 39.

[299] Oleg Trott and Arthur J Olson. 'AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading'. In: *Journal of computational chemistry* 31.2 (2010), pp. 455–461.

[300] Mark E. Tuckerman. *Statistical Mechanics: Theory and Molecular Simulation*. Oxford Graduate Texts, 2011.

[301] Ozlem Ulucan, Tanushree Jaitly and Volkhard Helms. 'Energetics of hydrophilic protein–protein association and the role of water'. In: *Journal of chemical theory and computation* 10.8 (2014), pp. 3512–3524.

[302] Sandor Vajda, David R Hall and Dima Kozakov. 'Sampling and scoring: A marriage made in heaven'. In: *Proteins: Structure, Function, and Bioinformatics* 81.11 (2013), pp. 1874–1884.

[303] Sandor Vajda and Dima Kozakov. 'Convergence and combination of methods in protein–protein docking'. In: *Current opinion in structural biology* 19.2 (2009), pp. 164–170.

[304] Anna Vangone and Alexandre MJJ Bonvin. 'Contacts-based prediction of binding affinity in protein–protein complexes'. In: *elife* 4 (2015), e07454.

[305] Adrian Velazquez-Campoy, Stephanie A Leavitt and Ernesto Freire. 'Characterization of protein-protein interactions by isothermal titration calorimetry'. In: *Protein-Protein Interactions*. Springer, 2004, pp. 35–54.

[306] Vishwesh Venkatraman and David W. Ritchie. 'Flexible protein docking refinement using pose-dependent normal mode analysis'. In: *Proteins: Structure, Function, and Bioinformatics* 80.9 (2012), pp. 2262–2274.

[307] Marcel L Verdonk et al. 'Improved protein–ligand docking using GOLD'. In: *Proteins: Structure, Function, and Bioinformatics* 52.4 (2003), pp. 609–623.

[308] Loup Verlet. 'Computer" experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules'. In: *Physical review* 159.1 (1967), p. 98.

[309] Thom Vreven et al. 'Updates to the integrated protein–protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2'. In: *Journal of molecular biology* 427.19 (2015), pp. 3031–3041.

[310] Karine Vuignier et al. 'Drug–protein binding: a critical review of analytical tools'. In: *Analytical and bioanalytical chemistry* 398.1 (2010), pp. 53–66.

[311] Changhao Wang et al. 'Recent developments and applications of the MMPBSA method'. In: *Frontiers in molecular biosciences* 4 (2018), p. 87.

[312] Chu Wang, Philip Bradley and David Baker. 'Protein–protein docking with backbone flexibility'. In: *Journal of Molecular Biology* 373.2 (2007), pp. 503–519.

[313] Junmei Wang et al. 'Development and testing of a general amber force field'. In: *Journal of computational chemistry* 25.9 (2004), pp. 1157–1174.

[314] Kai Wang et al. 'Identifying ligand binding sites and poses using GPU-accelerated Hamiltonian replica exchange molecular dynamics'. en. In: *Journal of Computer-Aided Molecular Design* 27.12 (Dec. 2013), pp. 989–1007.

[315] Lee-Ping Wang et al. 'Systematic improvement of a classical molecular model of water'. In: *The Journal of Physical Chemistry B* 117.34 (2013), pp. 9956–9972.

[316] Lingle Wang et al. 'Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field'. In: *Journal of the American Chemical Society* 137.7 (2015), pp. 2695–2703.

[317] Yansong Wang et al. 'Increasing the efficiency of ligands for FK506-binding protein 51 by conformational control'. In: *Journal of medicinal chemistry* 56.10 (2013), pp. 3922–3935.

[318] Mark Nicholas Wass et al. 'Towards the prediction of protein interaction partners using physical docking'. In: *Molecular systems biology* 7.1 (2011), p. 469.

[319] Thomas R Weikl and Fabian Paul. 'Conformational selection in protein binding and function'. In: *Protein Scienceence* 23.11 (2014), pp. 1508–1518.

[320] M Willander and S Al-Hilli. *Micro and Nano Technologies in Bioanalysis: Methods and Protocols*. 2009.

[321] Magnus Willander and Safaa Al-Hilli. 'Analysis of biomolecules using surface plasmons'. In: *Micro and Nano Technologies in Bioanalysis*. Springer, 2009, pp. 201–229.

[322] Christof Winter et al. 'SCOPPI: a structural classification of protein–protein interfaces'. In: *Nucleic acids research* 34.suppl_1 (2006), pp. D310–D314.

[323] Shoshana J Wodak et al. 'Protein–protein interaction networks: the puzzling riches'. In: *Current opinion in structural biology* 23.6 (2013), pp. 941–953.

[324] Hyung-June Woo and Benoıt Roux. 'Calculation of absolute protein–ligand binding free energy from computer simulations'. In: *Proceedings of the National Academy of Sciences* 102.19 (2005), pp. 6825–6830.

[325] Yinghao Wu et al. 'OPUS-Ca: A knowledge-based potential function requiring only Cα positions'. In: *Protein Science* 16.7 (2007), pp. 1449–1463.

[326] Martin Zacharias. 'Accounting for conformational changes during protein–protein docking'. In: *Current opinion in structural biology* 20.2 (2010), pp. 180–186.

[327] Martin Zacharias. 'ATTRACT: protein–protein docking in CAPRI using a reduced protein model'. In: *Proteins: Structure, Function, and Bioinformatics* 60.2 (2005), pp. 252–256.

[328] Martin Zacharias. 'Continuum solvent modeling of nonpolar solvation: Improvement by separating surface area dependent cavity and dispersion contributions'. In: *The Journal of Physical Chemistry A* 107.16 (2003), pp. 3000–3004.

[329] Martin Zacharias. 'Protein–protein docking with a reduced protein model accounting for side-chain flexibility'. In: *Protein Scienceence* 12.6 (2003), pp. 1271–1282.

[330] Martin Zacharias. *Protein-protein complexes: Analysis, modeling and drug design*. World Scientific, 2010.

[331] Fabian Zeller and Martin Zacharias. 'Adaptive biasing combined with Hamiltonian replica exchange to improve umbrella sampling free energy simulations'. In: *Journal of chemical theory and computation* 10.2 (2014), pp. 703–710.

[332] Chi Zhang et al. 'A knowledge-based energy function for protein- ligand, protein- protein, and protein- DNA complexes'. In: *Journal of medicinal chemistry* 48.7 (2005), pp. 2325–2335.

[333] Zheng Zheng and Kenneth M Merz Jr. 'Development of the knowledge-based and empirical combined scoring algorithm (kecsa) to score protein–ligand interactions'. In: *Journal of chemical information and modeling* 53.5 (2013), pp. 1073–1083.

[334] Yaoqi Zhou et al. 'What is a desirable statistical energy functions for proteins and how can it be obtained?' In: *Cell biochemistry and biophysics* 46.2 (2006), pp. 165–174.

[335] Daniel M Zuckerman. 'Equilibrium sampling in biomolecular simulations'. In: *Annual review of biophysics* 40 (2011), pp. 41–62.

[336] Daniel M Zuckerman and Lillian T Chong. 'Weighted ensemble simulation: review of methodology, applications, and software'. In: *Annual review of biophysics* 46 (2017), pp. 43–57.

[337] Erik RP Zuiderweg. 'Mapping protein- protein interactions in solution by NMR spectroscopy'. In: *Biochemistry* 41.1 (2002), pp. 1–7.

[338] Robert W Zwanzig. 'High-temperature equation of state by a perturbation method. I. Nonpolar gases'. In: *The Journal of Chemical Physics* 22.8 (1954), pp. 1420–1426.

[339] Matthew C Zwier et al. 'WESTPA: An interoperable, highly scalable software package for weighted ensemble simulation and analysis'. In: *Journal of chemical theory and computation* 11.2 (2015), pp. 800–809.