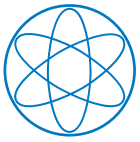
PHYSIK-DEPARTMENT

TECHNISCHE UNIVERSITÄT MÜNCHEN

Dissertation

Simulation of Biomolecular Binding

Fabian Tobias Zeller



PHYSIK-DEPARTMENT

TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Theoretische Biophysik T38 - Molekulardynamik

Simulation of Biomolecular Binding

Simulation biomolekularer Bindungsprozesse

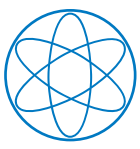
Vollständiger Abdruck der von der Fakultät der Physik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. Matthias Rief
Prüfer der Dissertation:	1. Univ.-Prof. Dr. Martin Zacharias
	2. Univ.-Prof. Dr. Carlo Camilloni

Die Dissertation wurde am 15.12.2016 bei der Technischen Universität München eingereicht und durch die Fakultät der Physik am 17.3.2017 angenommen.



I confirm that this dissertation is my own work and I have documented all sources and material used.

Fabian Tobias Zeller

Acknowledgments

Besonderer Dank gilt meinem Betreuer Martin Zacharias für die Einführung in die molekulare Biophysik und für die immerwährende Unterstützung während meiner Jahre am Lehrstuhl. Ich danke Manuel und Rainer für die sich teils am Rande des Wahnsinns bewegenden und mich möglicherweise genau deshalb voranbringenden Diskussionen, sowie die oft augenöffnende Kritik. Ich danke Florian, Christina, Giuseppe, Alexander, Sonja, Sjord und Mahmut für alles, womit sie mir während der Zeit am Lehrstuhl weitergeholfen haben. Außerdem möchte ich Florian und Christina als Wächter des Espresso hervorheben, und Florian und Giuseppe als Begründer des food-office.

Für die Zusammenarbeit bezüglich Adenylat-Kinase und Optical Tweezers bedanke ich mich bei Matthias Rief, Benjamin Pelz and Gabriel Zoldak.

Dank gilt außerdem der TUM für das Ermöglichen der Promotion, der Deutschen Foschungsgesellschaft für die Finanzierung und dem Leibnitz Rechenzentrum für bereitgestellte Rechenzeit.

Die Einleitung ist meiner Familie und meinen Freunden gewidmet.

Abstract

Molecular Dynamics simulations, complementary to experimental methods, allow atomistic and highly time resolved insight into details of a broad class of biomolecular processes which would otherwise remain inaccessible. In this thesis, several research projects are presented concerning the evaluation of different aspects of simulation models, the development of new simulation methodologies and the investigation of biological processes on the molecular level.

Molecular binding affinities play a crucial role for intracellular interactions and rational drug design. Two chapters of this thesis deal with the efficient calculation of binding affinities. Several current continuum solvent models, which significantly reduce the computational cost of solvent description, were tested regarding the quality of peptide-protein binding free energy prediction. It was found that, as long as no charged residues were located in the binding interfaces, these models yield reasonable values in comparison with an explicit solvent model and experimental results. In combination with the continuum solvent models, a perturbation based method was developed which allows the prediction of changes in binding affinities caused by small mutations of ligands, using only original wild-type simulations. This method can significantly reduce the computational effort for systematic ligand screening and optimization studies.

An intermediate role plays a study on adenine-derivate self-aggregation. Besides giving atomistic insight into base-stacking interactions, crucial for the structure of DNA, statistically converged relaxation times could be obtained. Remarkably, with fully flexible treatment of the water molecule hydrogen atoms, the related kinetic stacking rates were in close agreement with experimental measurements. This is of particular interest, as there was previously little evidence on the quality of kinetic predictions by MD simulations, while at the same time, kinetic properties are becoming increasingly important for drug design.

The mean residence time of the drug is today commonly accepted as a main factor determining drug efficacy. In this context, a multi-scale approach was designed for the efficient determination of kinetic binding rates of drug molecules to target protein sites. While the binding process into the binding site is simulated at atomistic detail, the diffusive approach of the drug molecule to the receptor is modeled at low resolution. This allowed the reconstruction of influenza neuraminidase inhibitor binding pathways

and the correct relative ranking of the binding rates.

In two separate projects current MD methodologies were used to study specific protein functions. The free energy landscapes of combined large-scale domain motions in Adenylate-Kinase (ADK) in dependence of bound adenosine-phosphate substrates were calculated by means of replica-exchange umbrella sampling simulations. The free energy landscapes allow insight into how the domains interplay in order to achieve efficient enzymatic throughput and, in particular, how unproductive bound substrate configurations might be avoided.

Plain MD simulations of several µs were used to investigate the recognition process of the O^6-methylguanine damage in DNA by Alkyltransferase-like (ATL) protein. The complete binding process and subsequent looping-out of the damaged base induced by specific ATL protein interactions could be reconstructed, revealing intermediate states and principles of the damage recognition mechanism.

Contents

1 Introduction

1.1 Principles of Living Organisms

The fundamentals of life happen on the microscopic scale under the reign of Brownian motion and thermal fluctuations. They can therefore only be understood within the framework of thermodynamics and the probabilistic predictions of statistical mechanics[1]. Living organisms elegantly combine the statistic properties of their individual components in order to achieve macroscopic effects: Eventually, the composition of a doctoral thesis.

It stands out that - although this is not necessarily true for all of their subunits - cells as a whole live far from thermodynamic equilibrium. The second law of thermodynamics states that the generation of non-equilibrium states cannot come without cost, and indeed, living cells need to consume free energy to stay alive[2]. This free energy is, at the origin of the food chain, harvested from sunlight.

The organization of the complex processes involved in free energy ingestion, propagation of information and ultimately self replication is encoded in DNA or related RNA molecules. It is known today that the four-letter code of the DNA is translated into sequences of different amino acids, basic building blocks, which are sequentially assembled into proteins, and that the sequence of amino acids determines how the proteins fold into their functional, three-dimensional shapes. The vast number of genetically encoded proteins constitutes the molecular machinery that provides catalytical enhancement of reactions and the structural basis of the cell[2].

It is the detailed molecular interactions between DNA, RNA, proteins and other molecules that ultimately ensure the cell's survival. Figure 1.1 shows an experimental, atom-level resolved static structure of a protein that interacts with a DNA double-strand[3]. Instead of depicting all the individual atoms of the complex, typically, a cartoon-like illustration is used in which three-dimensional, often occurring building patterns are symbolized: The DNA-backbone, the DNA bases which encode the genetic information, and the amino acid backbone chain of the protein with so-called α-helical subunits. The shown protein is involved in the detection and repair of base lesions in the DNA strand which can cause mutagenic damages in organisms. As can be seen from the experimental structure, binding of the protein involves a stable looped-out state of the damaged base. While experiments revealed the existence of such a state[3],

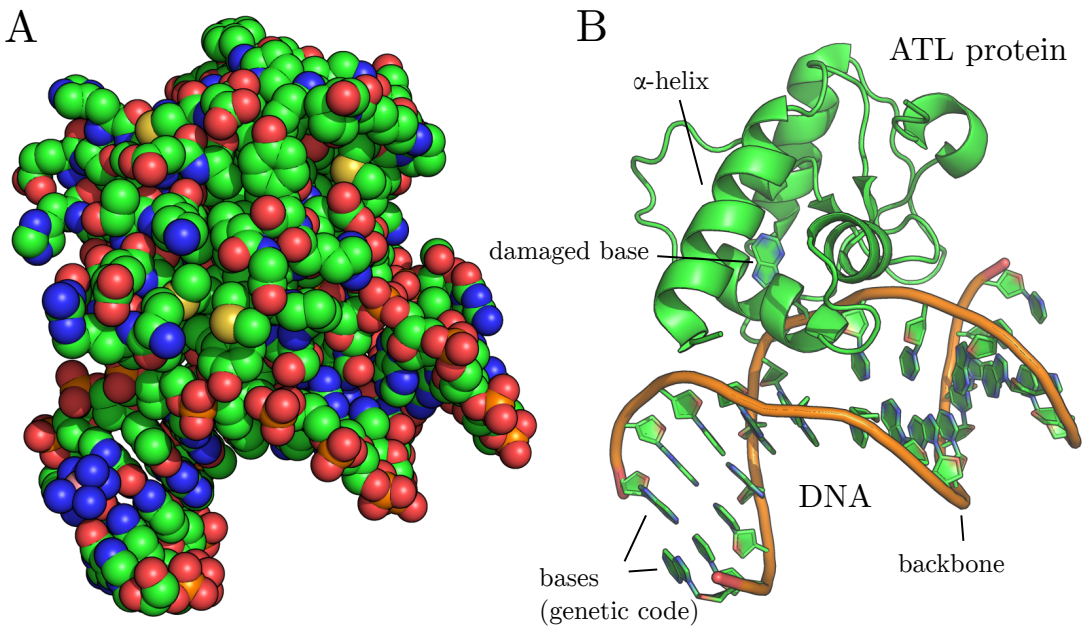


Figure 1.1: Crystal structure obtained by X-ray spectroscopy of Alkyltransferase-like (ATL) protein (PDB: 3gx4)[3] in complex with O^6-methylguanine-DNA. A) Depiction of all non-hydrogen atoms as spheres. B) Symbolic cartoon representation, illustrating basic structural features of DNA and protein.

many major questions remain unresolved: Does the protein actively push the base out of its original configuration or does it only bind to an already looped-out configuration? Does the protein find the damage by a completely random diffusive search or is there some kind of sliding mechanism along the DNA strand? In general, how is the protein able to efficiently locate the damaged bases among the vast amount of undamaged base pairs? This DNA-protein complex related to DNA-repair, which is subject of the research project presented in chapter 9, is only one fascinating example of vital molecular interactions within living organisms. Until today, only a tiny fraction of the immense number of such mechanisms has been fully understood.

1.2 The Computational Microscope

To gain access to how in detail DNA, RNA and proteins function and interplay within organisms, it is necessary to resolve their molecular structure and dynamics[5]. Figure 1.2 shows the accessible time and length scales of commonly used biophysical methods.

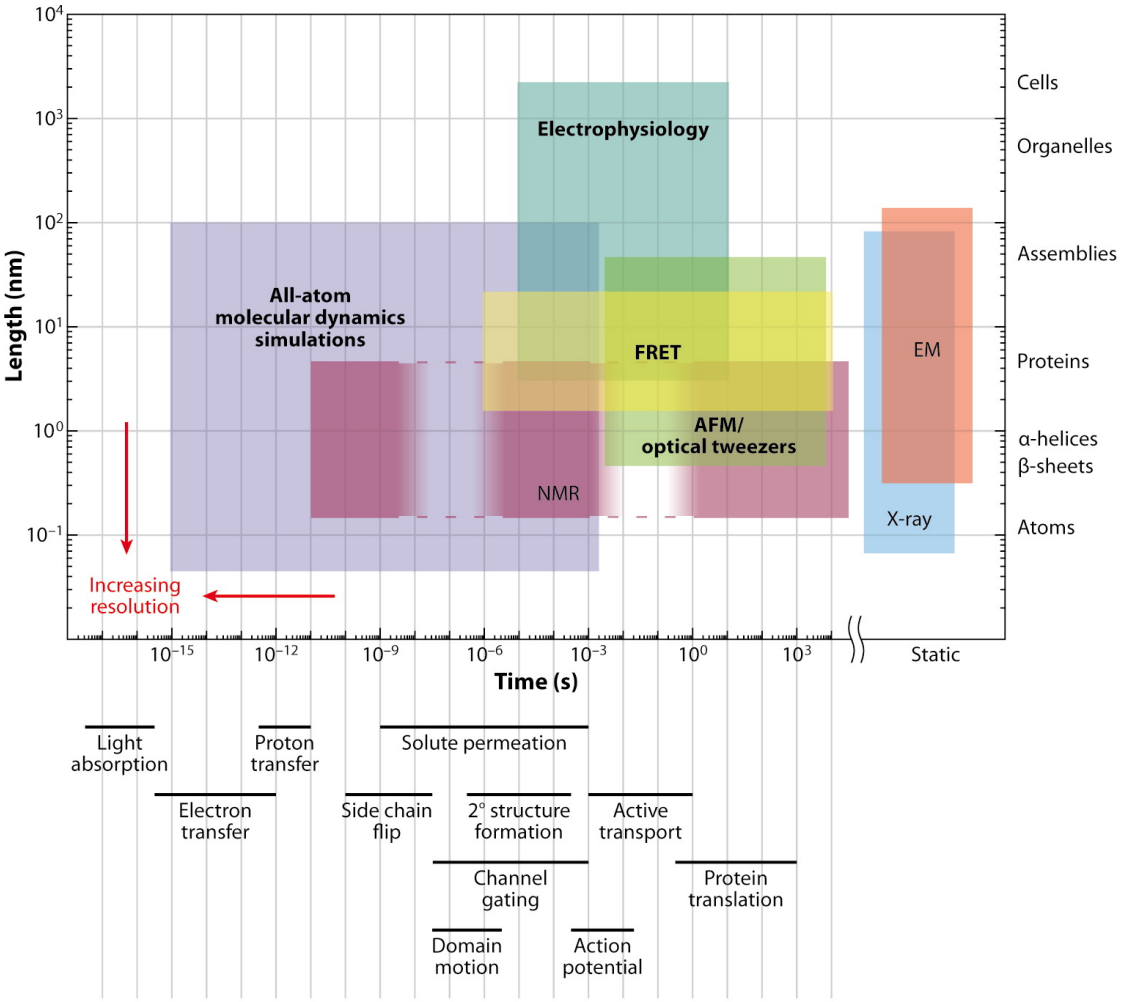


Figure 1.2: Commonly used biophysical methods and observable time and length scales (EM: Electron Microscopy, AFM: Atomic Force Microscopy, FRET: Förster Resonance Energy Transfer, NMR: Nuclear Magnetic Resonance). Associated molecular processes and biological objects are shown below and at the right of the graph. Methods which allow observation of single molecules are in boldface. Figure taken from Dror el al.[4].

X-ray spectroscopy in combination with protein crystallization can be used to obtain high-resolution atomic structures of biomolecules. Crystal spectroscopy data are however static in nature, and crystallization of proteins is often hard to achieve. Also, crystallization can be accompanied by structural artifacts not present at physiological conditions[6]. Electron microscopy (EM) or cryo-EM typically yields static information of lower resolution, but does not require prior protein crystallization. The exposure of the biological sample to radiation and high vacuum can be problematic[7]. Nuclear Magnetic Resonance (NMR) methods can be applied over a wide range of time scales to measure ensemble averages[8]. The Förster Resonance Energy Transfer (FRET) can be utilized to measure distances with high sensitivity. This technique often requires artificial attachment of fluorophores to the molecules of interest[9]. Electrophysiological methods allow the measurement of electrical properties down to the scale of single ion channel proteins[10]. Atomic Force Microscopy (AFM) and optical tweezers have evolved later and stand out as single molecule techniques capable of observing detailed molecular motion[11][12].

Figure 1.2 illustrates that Molecular Dynamics (MD) simulations, often referred to as a "Computational Microscope", provide access to spatial and temporal resolution not accessible by current experimental techniques. These simulations are based on classical approximations of bonded forces, van der Waals forces and electrostatic forces (see chapter 2). More precise quantum mechanical simulation methods have been developed[13][14][15][16], however, application to physiological time scales is currently mostly unfeasible.

Prominent studies indicate that classical treatment is suitable to accurately describe a whole range of biological processes. Several small proteins were folded in MD simulations to crystal-structure like conformations using the same set of classical parameters[17]. This shows that classical force fields can indeed be of sufficient generality to tackle the previously inaccessible folding problem. Also, binding free energies of a broad variety of ligands can often be predicted reliably[18]. Significant advances regarding the accessible simulation times have been made by computational parallelization, GPU usage[5] and the design of special purpose hardware[19]. All-atom MD simulations of small systems of up to the millisecond scale have been reported[19]. Also, the system size limits have been pushed ahead: Simulations with millions of atoms have been carried out on a virus[20] for several nanoseconds. Statistical mechanics based enhanced simulation methods for the calculation of equilibrium properties have been developed which drastically reduce required time scales[21].

It is worth to mention that MD simulations are an especially powerful tool when combined with experimental techniques. Ideally, the quality of the used simulation model is validated by experimental data in regimes with overlapping time and length scales. The validated simulations can then give valuable additional insight on scales

inaccessible to the experiment[22].

1.3 Project Overview

In this thesis, several research projects are presented, focusing the "Computational Microscope" on a broad range of biophysical processes. Chapters 4 and 5 are concerned with computationally efficient methods for the determination of binding affinities. Non-covalent, intermolecular binding is crucial for intercellular interactions and plays an important role in rational drug design. It has become increasingly appreciated that besides binding affinities, also binding kinetics are important for protein function and drug efficacy. Chapter 8 deals with the extraction of kinetic information from MD simulations in general, whereas in chapter 7, the binding rates of influenza neuraminidase inhibitors are investigated. In chapter 6, the role and coupling of large-scale domain motions of a kinase protein was investigated with respect to enzymatic efficiency. Finally, in chapter 9, the first crucial steps of DNA-damage recognition of a repair protein were reconstructed by MD simulations.

1.4 Bibliography

[1] Erich Sackmann and Rudolf Merkel. *Lehrbuch Der Biophysik*. Wiley-VCH, 2010.

[2] Bruce Alberts et al. *Molecular Biology of the Cell*. 5th ed. Garland Science, 2002.

[3] Julie L. Tubbs et al. "Flipping of Alkylated DNA Damage Bridges Base and Nucleotide Excision Repair." In: *Nature* 459.7248 (June 11, 2009), pp. 808–813.

[4] Ron O. Dror, Morten Ø Jensen, David W. Borhani, and David E. Shaw. "Exploring Atomic Resolution Physiology on a Femtosecond to Millisecond Timescale Using Molecular Dynamics Simulations." In: *J. Gen. Physiol.* 135.6 (June 2010), pp. 555–562.

[5] Ron O. Dror, Robert M. Dirks, J.P. Grossman, Huafeng Xu, and David E. Shaw. "Biomolecular Simulation: A Computational Microscope for Molecular Biology." In: *Annu. Rev. Biophys.* 41.1 (2012), pp. 429–452.

[6] Heping Zheng, Jing Hou, Matthew D. Zimmerman, Alexander Wlodawer, and Wladek Minor. "The Future of Crystallography in Drug Discovery." In: *Expert. Opin. Drug. Discov.* 2 (Feb. 28, 2014), pp. 125–137.

[7] Joachim Frank. "Single-Particle Imaging of Macromolecules by Cryo-Electron Microscopy." In: *Annu. Rev. Biophys. Biomol. Struct.* 31.1 (2002), pp. 303–319.

[8] Paul J. Barrett et al. “The Quiet Renaissance of Protein Nuclear Magnetic Resonance.” In: *Biochemistry (Mosc.)* 52.8 (2013), pp. 1303–1320.

[9] “Fluorescent Protein FRET: The Good, the Bad and the Ugly.” In: *Trends Biochem. Sci.* 32.9 (2007), pp. 407–414.

[10] Christof Grewer, Armanda Gameiro, Thomas Mager, and Klaus Fendler. “Electrophysiological Characterization of Membrane Transport Proteins.” In: *Annu. Rev. Biophys.* 42.1 (2013), pp. 95–120.

[11] Matthias Rief, Mathias Gautel, Filipp Oesterhelt, Julio M. Fernandez, and Hermann E. Gaub. “Reversible Unfolding of Individual Titin Immunoglobulin Domains by AFM.” In: *Science* 276.5315 (1997), pp. 1109–1112.

[12] Miklos S. Z. Kellermayer, Steven B. Smith, Henk L. Granzier, and Carlos Bustamante. “Folding-Unfolding Transitions in Single Titin Molecules Characterized With Laser Tweezers.” In: *Science* 276.5315 (1997), pp. 1112–1116.

[13] D. Xenides, B.R. Randolf, and B.M. Rode. “Hydrogen bonding in liquid water: An ab initio QM/MM MD Simulation Study.” In: *J. Mol. Liq.* 123.2–3 (2006), pp. 61–67.

[14] Haiyan Liu, Florian Müller-Plathe, and Wilfred F. van Gunsteren. “A Combined Quantum/Classical Molecular Dynamics Study of the Catalytic Mechanism of HIV Protease.” In: *J. Mol. Biol.* 261.3 (1996), pp. 454–469.

[15] Bora Karasulu, Mahendra Patil, and Walter Thiel. “Amine Oxidation Mediated by Lysine-Specific Demethylase 1: Quantum Mechanics/Molecular Mechanics Insights Into Mechanism and Role of Lysine 661.” In: *J. Am. Chem. Soc.* 135.36 (2013), pp. 13400–13413.

[16] A. Warshel and M. Levitt. “Theoretical Studies of Enzymic Reactions: Dielectric, Electrostatic and Steric Stabilization of the Carbonium Ion in the Reaction of Lysozyme.” In: *J. Mol. Biol.* 103.2 (1976), pp. 227–249.

[17] Kresten Lindorff-Larsen, Stefano Piana, Ron O. Dror, and David E. Shaw. “How Fast-Folding Proteins Fold.” In: *Science* 334.6055 (2011), pp. 517–520.

[18] Yuqing Deng and Benoit Roux. “Computations of Standard Binding Free Energies With Molecular Dynamics Simulations.” In: *J. Phys. Chem. B* 113.8 (2009), pp. 2234–2246.

[19] D.E. Shaw et al. “Millisecond-Scale Molecular Dynamics Simulations on Anton.” In: *High Performance Computing Networking, Storage and Analysis, Proceedings of the Conference On*. 2009, pp. 1–11.

[20] Peter L. Freddolino, Anton S. Arkhipov, Steven B. Larson, Alexander McPherson, and Klaus Schulten. "Molecular Dynamics Simulations of the Complete Satellite Tobacco Mosaic Virus." In: *Structure* 14.3 (2006), pp. 437–449.

[21] Fabian Zeller and Martin Zacharias. "Adaptive Biasing Combined With Hamiltonian Replica Exchange to Improve Umbrella Sampling Free Energy Simulations." In: *J. Chem. Theory Comput.* 10.2 (2014), pp. 703–710.

[22] Benjamin Pelz, Gabriel Zoldak, Fabian Zeller, Martin Zacharias, and Matthias Rief. "Subnanometre Enzyme Mechanics Probed by Single-Molecule Force Spectroscopy." In: *Nat. Commun.* 7 (Feb. 24, 2016), pages.

2 Simulation Methods

In this chapter, the principles of Molecular Dynamics (MD) and Brownian Dynamics (BD) simulations are introduced. Specific simulation details are described in the methods sections of the respective chapters.

2.1 Molecular Dynamics Simulations

2.1.1 Classical Atomistic Model

The MD simulations carried out in this thesis are based on a classical Hamiltonian in which atoms are treated as point particles with fixed partial charges. Electron wave functions are only represented as averages within classical approximations. The potential function used in this thesis has the following form[1]:

$$\begin{aligned} E_{\text{total}} = & \sum_{\text{bonds}} K_r \left(r - r_0\right)^2 \\ & + \sum_{\text{angles}} K_\Theta \left(\Theta - \Theta_0\right)^2 \\ & + \sum_{\text{dihedrals}} \frac{V_n}{2} \left[1 + cos\left(n\phi - \gamma\right)\right] \\ & + \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}}\right] \end{aligned} \tag{2.1}$$

Here, r, Θ and ϕ refer to distances, angles and dihedral angles, respectively, between covalently bound atoms. R refers to distances between non-bound atoms. Bond lengths and angles between covalently bound atoms are described by harmonic potentials with force constants K_r and K_Θ, as well as equilibrium values r_0 and Θ_0. Dihedral angles between covalently bound atoms are represented by cosine functions with scaling factors V_n, phases γ and multiplicity n. Lennard-Jones potentials with parameters A and B are used to describe the van der Waals interactions between non-bonded atom pairs. The long-range electrostatic interactions are calculated based on the partial charges of the atoms, according to Coulomb's law.

type	description
CT	any sp^3 carbon
C	any carbonyl sp^2 carbon
CA	any aromatic sp^2 carbon and Cϵ of Arg
CM	any sp^2 carbon, double bonded
CC	sp^2 aromatic in 5-membered ring with one substituent + next to nitrogen
CV	sp^2 aromatic in 5-membered ring next to carbon and lone pair nitrogen
CW	sp^2 aromatic in 5-membered ring next to carbon and NH
CR	sp^2 aromatic in 5-membered ring next to two nitrogens
CB	sp^2 aromatic at junction of 5- and 6-membered rings and both junction atoms in Ade and Gua
C^*	sp^2 aromatic in 5-membered ring next to two carbons
CN	sp^2 junction between 5- and 6-membered rings and bonded to CH and NH
CK	sp^2 carbon in 5-membered aromatic between N and N-R (C8 in purines)
CQ	sp^2 carbon in 6-membered ring between lone pair nitrogens

Table 2.1: Carbon atom types of the amber force field as an example for the atom classifications made in a classical, atomistic model.

The parameters used in the potential function are specific for atom types or atom type combinations, resulting in a relatively large set of parameters. Usually, several atom types are defined for atoms of the same element to account for specific chemical surroundings (see Table 2.1). A full list of the Amber atom types can be found in the corresponding publication[1]. The potential function in combination with the parameter set is referred to as "force field". Typically, the bonded parameters and the atomic partial charges are derived from approximative or semi-empirical quantum mechanical calculations, the non-bonded Lennard-Jones parameters are fitted to experimental solvation free energies. Up to date, a large number of force fields, optimized for different applications, has been developed[2][3].

2.1.2 Simulation of Thermodynamic Ensembles

Equations of Motion

The Amber MD code uses the Leap Frog[4][5] algorithm to integrate Newton's equations of motion over time:

$$\begin{aligned} \vec{x}_{\tau+1} &= \vec{x}_\tau + \vec{v}_{\tau+1/2}\Delta t \\ \vec{v}_{\tau+3/2} &= \vec{v}_{\tau+1/2} + \vec{a}(x_{\tau+1})\Delta t, \end{aligned} \tag{2.2}$$

Here, τ is the discrete index of the integration step, $\vec{x}_\tau$ and $\vec{v}_\tau$ are the coordinates and velocities of a particle at time $\tau\Delta t$, and $\vec{a}(x_{\tau+1})$ is the acceleration of the particle at time $(\tau+1)\Delta t$ as calculated from the Hamiltonian of the system (eq. 2.1).

Although there are more sophisticated integration algorithms with better short time energy conservation, the Leap Frog algorithm has several advantages. Firstly, it is fast to compute and only the first derivative of the particle coordinates is needed. More importantly, it is symplectic, implying the conservation of a given phase space volume over time. This is a necessary requirement for energy conservation. Overall, the Leap Frog algorithm conserves the total energy better than higher order algorithms over long periods of time[5].

Constant Temperature Simulations

Integration of Newton's equations of motion (2.2) yields constant energy trajectories (NVE) corresponding to the microcanonical ensemble. Under physiological conditions, biophysical systems are exposed to surrounding molecules which act as a heat bath. Therefore, it is more appropriate to carry out simulations in the canonical ensemble, in which the average temperature is constant (NVT). While simulation of a true heat bath is not feasible, several methods exist that allow simulation under approximate canonical conditions. A simple way to impose a constant average temperature is to rescale the velocities of all particles according to

$$\frac{dT}{dt} = \frac{T_0 - T}{\tau}. \tag{2.3}$$

This so called Berendsen thermostat[6] leads to fluctuations of the system temperature T around the heat-bath temperature T_0 with a relaxation time τ, approximating canonical distributions for large numbers of atoms. The Andersen thermostat[7] mimics the heat bath by introducing artificial collisions. Randomly selected particles are assigned a new velocity from a Maxwell-Boltzmann distribution at certain intervals. This yields stochastic trajectories and canonical distributions. A similar approach consists in adding a stochastic force term to equation 2.2, corresponding to a Langevin description[8] of Brownian motion. The stochastic force is related to the diffusional properties of the system by the Einstein relation (fluctuation-dissipation relation). Trajectories generating a true canonical distribution can also be simulated in a deterministic way by the Nosé-Hoover thermostat[9], in which the temperature is treated as an external variable.

Constant Pressure Simulations

Due to similar arguments as described in the previous subsection, the simulation at constant pressure instead of a constant volume is more adequate (NTP). This can be realized similarly to the Berendsen thermostat, rescaling the particle distances and thus the volume instead of the particle velocities. As most biophysical mechanisms

do not imply major volume changes, constant pressure conditions are of minor importance[10]. Therefore, for the simulations in this thesis, pressure regulation was carried out with the simple Berendsen barostat, or if indicated, neglected for the calculation of thermodynamic averages (yielding Helmholtz free energies instead of Gibbs free energies).

Periodic Boundary Conditions

The aim of an MD simulation is to simulate molecules at physiological conditions. This includes solvent molecules such as water molecules and ions surrounding proteins or DNA strands. The number of atoms that can be simulated is however computationally limited. A suitable way to mimic a large number of solvent molecules around a molecule of interest is the introduction of periodic boundary conditions. Space is divided into equal cuboids (or truncated octahedrons, both space filling), which are treated as equivalent images of an original central cuboid. The periodic boundary conditions imply that particles leaving the central simulation box re-enter it from the opposing side, and that all particles also interact with all particle images. With a sufficient box size respectively number of solvent atoms, this treatment introduces significantly less artifacts than artificial boundary potentials and is a standard method used in MD simulations[5].

Summation of pairwise short range interactions over an infinite number of particle images is avoided by the introduction of cut-off radii (2.1.3). For the treatment of long-range electrostatic interactions, see Particle Mesh Ewald method (2.1.3).

2.1.3 Approximations for the Increase of Simulation Throughput

Cut-off radii for Short Range Interactions

A major part of the computational cost arises from the calculation of pairwise non-bonded interactions. Summation over all particle pairs scales quadratically with system size. A significant speed-up of the simulations can therefore be achieved by neglecting non-bonded short-range interactions beyond a certain distance threshold. These are the Lennard-Jones interactions, decaying with r^6, and the short-ranged contribution of the coulombic interactions (see Particle Mesh Ewald method). It is widely accepted that a cut-off radius of $\approx 9\,\text{Å}$ implies only minor deviations from the original Hamiltonian function[11]. To avoid discontinuities in the energy function, the potentials are usually shifted to zero at the cut-off radius.

Particle Mesh Ewald Method

With periodic boundary conditions, long-range electrostatic interactions which cannot be neglected within a reasonable threshold are more complex to handle. A straightforward calculation would imply a slowly converging sum over all images of all particles. However, the periodic interactions can effectively by calculated in Fourier-space. This can be realized by the Ewald summation method. In this method, the point charges are screened by artificial Gaussian-distributed charges, which allows the separation of a short-ranged electrostatic contribution which can be truncated at a cut-off radius. The effect of the remaining periodically arranged opposite Gaussian-distributed screening charges can be efficiently calculated in Fourier space.

The Ewald sum, in principle, constitutes an exact calculation of the electrostatic contributions, but suffers from an inefficient $\mathcal{O}(N^2)$ scaling of the computational costs. Further acceleration of the calculations can be achieved by the interpolation of charge positions on a grid, referred to as the Particle Mesh Ewald method. The use of interpolated charge positions only yields approximative solutions, but allows the application of Fast Fourier Transforms, coming with a significantly lower computational cost and enhancing the scaling to $\mathcal{O}(N \log(N))$. Typically, the short-ranged contributions are calculated with high accuracy using the true charge positions, while the long-range contributions are calculated using a mesh based method[5].

Constraining Hydrogen Bonds

Of crucial importance for simulation throughput is the size of the integration time step. The integration time step is determined by the fastest motion in the system which has to be correctly represented in order to avoid numerical instabilities of the simulations. The fastest atomistic motion in biological systems is usually the fluctuation of hydrogen atom bonds, restricting the integration time step to 1 fs. At the same time, hydrogen bond length fluctuations are often not crucial for biophysical mechanisms. Therefore, the hydrogen bonds in the system are typically constrained to their equilibrium lengths by Lagrange multiplier based methods as SHAKE[12], allowing for a time step size of 2 fs.

Hydrogen Mass Repartitioning

Another approach to further enlarge the integration time step is based on the observation that the fast movements of the hydrogen atoms are due to their low mass. Recent studies indicate that shifting a fraction of the mass of hydrogen bond partner atoms to the hydrogen atoms yields simulation trajectories and observables close to original

mass simulations[13]. The hydrogen mass repartitioning scheme allows for a further increase of the integration time step to 4 fs or 4.5 fs.

2.1.4 Simulation Software and Hardware

For the MD simulations presented in this thesis, the respectively latest version of the Amber software package[14] was used. For projects 4, 5 and 6, CPU[1] based code was employed on an in-house cluster and the Leibnitz Rechen Zentrum (LRZ). When CUDA[2] based implementations became increasingly powerful, the GPU[3] version of the Amber code[15][16] was used (projects 8, 7 and 9) on an in-house cluster.

2.2 Brownian Dynamics Simulations

2.2.1 Simplified Model for Diffusive Regimes

The atomistic model described in 2.1 does not allow for sufficient simulation times and system sizes to describe the majority of diffusion controlled processes. This impedes for example the atomistic simulation of association pathway ensembles of ligands binding to receptor molecules. At the same time, atomistic details and internal flexibility of the molecules play only a minor part during the diffusive approach of two binding molecules. Up to certain distances, only long-range electrostatic interactions and the Brownian motion of the molecules dominate the process[17][18].

This can relatively accurately be described by a model that treats the molecules as rigid bodies undergoing a diffusive random walk, only interacting via electrostatic interactions. The solvent molecules are implicitly represented in the diffusion constants determining the random walk step sizes and in the screening of the electrostatic interactions. For electrostatic interactions in solution, the Poisson Boltzmann (PB) equation is an adequate description[19]. It is convenient to used the partial charges from the atomistic MD force fields for the electrostatic calculations.

Often, one is interested in the diffusive motion of a ligand in the complex electrostatic field of a receptor protein. It can then be sufficient to numerically solve the Poisson Boltzmann equation only for the electrostatic potential of the receptor and to calculate the force acting on the ligand from the receptor potential and the ligand charges. Typically, BD trajectories are stopped when two molecules come within a certain separation distance. The resulting molecule configurations may serve as starting points for more detailed MD simulations.

[1] central processing unit
[2] compute unified device architecture
[3] graphics processing unit

2.2.2 Simulation of Brownian Dynamics

General Formulation

A Brownian Dynamics trajectory can be obtained from computer simulations by integrating the following stochastic equations of motion for all molecules i[18]:

$$\Delta\vec{r}_i = \sum_j \frac{D_i^t \vec{F}_{ij}}{k_B T}\Delta t + \vec{R}\left(D_i^t, \Delta t\right) \tag{2.4}$$

$$\Delta\vec{\theta}_i = \sum_j \frac{D_i^o \vec{T}_{ij}}{k_B T}\Delta t + \vec{\Theta}\left(D_i^o, \Delta t\right). \tag{2.5}$$

$\Delta\vec{r}_i$ and $\Delta\vec{\theta}_i$ are the changes of the three-dimensional coordinates $\vec{r}_i$ and Euler angles $\vec{\theta}_i$ of molecule i after an integration time step of Δt. D_i^t and D_i^o are the translational respectively orientational diffusion constants of molecule i. $\vec{F}_{ij}$ and $\vec{T}_{ij}$ are the force respectively torque exerted by molecules j on molecule i. $\vec{R}$ and $\vec{\theta}$ are random displacements satisfying Gaussian distributions with $\langle\vec{R}(D_i^t, \Delta t)^2\rangle = 6D_i^t\Delta t$ and $\langle\vec{\Theta}(D_i^o, \Delta t)^2\rangle = 6D_i^o\Delta t$, respectively. T is the temperature and k_B is the Boltzmann constant. In this description, hydrodynamic interactions between the molecules are neglected.

For the Brownian dynamics description to hold, the integration time step Δt must satisfy the condition $\Delta t >> mD/k_B T$. The random walk is not meaningful at timescales shorter than the momentum relaxation times. In addition, Δt must be sufficiently small such that the changes in force and torque are approximately constant during Δt[18]. The assumptions of homogeneous diffusion constants, fast relaxation of momenta and smoothly changing forces break down when two molecules come into close contact. The physics of two closely interacting molecules are not meaningfully represented by BD simulations. In some cases it is convenient to take molecular structure into account by rejecting propagation steps which would lead to an overlap of two molecules[20].

In contrast to MD simulations, the BD integration time step Δt is not relevant for the numerical stability of the simulations. Typical BD integration time steps are in the order of ps, several orders of magnitude larger than the MD time steps.

Convenient Choice of the Reference System

A large part of the computational cost of BD simulation is spent on the generation of normally distributed random numbers. In this context, it is convenient to express equation 2.4 in the reference system of one of the simulated molecules. For simplicity, in the following, only two molecules 1 and 2 shall be considered. Taking into account $\langle\vec{R}(D^t, \Delta t)\rangle = 0$ and $\langle\vec{R}(D_2^t, \Delta t)^2\rangle + \langle\vec{R}(D_1^t, \Delta t)^2\rangle = \langle\vec{R}(D_1^t + D_2^t, \Delta t)^2\rangle$, the relative

translational diffusion can be expressed by the diffusion of molecule 2 in the reference system of molecule 1 with the relative diffusion constant $D_1^t + D_2^t$

$$\begin{aligned}\langle(\Delta\vec{r}_2 - \Delta\vec{r}_1)^2\rangle =& \langle\left(\frac{D_2^t\vec{F}_{12}}{k_BT}\Delta t\right)^2 + \left(\frac{D_1^t\vec{F}_{21}}{k_BT}\Delta t\right)^2 - 2\left(\frac{D_2^t\vec{F}_{12}}{k_BT}\Delta t\right)\left(\frac{D_1^t\vec{F}_{21}}{k_BT}\Delta t\right) \\ &+ \vec{R}^2\left(D_1^t, \Delta t\right) + \vec{R}^2\left(D_2^t, \Delta t\right)\rangle \\ =& \langle\left(\frac{\left(D_1^t + D_2^t\right)\vec{F}_{21}}{k_BT}\Delta t\right)^2 + \vec{R}(D_1^t + D_2^t, \Delta t)^2\rangle.\end{aligned} \tag{2.6}$$

Here, F_{12} was replaced by $-F_{21}$, according to Newton's third law. In this way, the generation of random numbers for the translational diffusion of molecule 1 is avoided without any loss of information. The orientational diffusion $\Delta\vec{\Theta}_1$ of molecule 1 relative to molecule 2 can directly be realized by rotating molecule 2 by $-\Delta\vec{\Theta}_1$ around the center of molecule 1.

Time-independent Electrostatic Potential

If the influence of all but one molecule on the electrostatic potential can be neglected, as for example in the case of the diffusion of a charge neutral ligand in an electrostatic potential of a receptor, the Poisson Boltzmann equation can be calculated once and stored on a grid. As the simulation takes place in the reference system of the receptor, absolute coordinates of receptor and corresponding electrostatic potential remain unchanged during the whole simulation[20]. This significantly reduces the computational cost of the BD simulations, as the expensive Poisson Boltzmann calculation has to be carried out only once.

2.2.3 Simulation Software and Hardware

The BD code used in this thesis was written in python2.7. To solve the Poisson Boltzmann equation for receptor molecules, the APBS software[19] was used.

2.3 Validity of Thermodynamic Simulations

Besides the typical issues of comparing theory, models and experiment, observables obtained from thermodynamic simulations have to be handled with specific care due to their statistical nature. The observables depend on the quality and validity of the model Hamiltonian as well as on the sampling of phase space, respectively the statistical quality of obtained probability distributions. The so-called model and

sampling problems are intrinsically connected. As the quality of the model can only be validated or adapted via resulting statistical averages, sufficient sampling is necessary for model validation. At the same time, the sampled phase space regions depend on the Hamiltonian itself. Historically, with increasing computational possibilities, the models and sampling methods have iteratively been improved.

In principle, as all MD simulations are based upon some kind of undirected search in phase space, complete phase space coverage, in particular of all favorable free energy regions, cannot be assured from the simulations alone. Free energy barriers to undiscovered but relevant regions in phase space might not have been crossed by the trajectories. Statistical and mathematically rigorous error estimates therefore only make sense if sufficient phase space coverage is plausible. Estimation of the time scales relevant for a system or comparison of phase space visited by subensembles of simulations can be good indicators of sufficient phase space coverage. Ultimately, the simulations depend on experimental validation. Therefore, whenever possible, observables obtained from the simulations were at least in one representative case compared to experimental results in order to ensure that the simulated time scales in combination with the sampling methods were sufficient.

The thermodynamic systems usually studied with MD simulations are expected to show chaotic behavior due to the high number of degrees of freedom. This implies that minimal deviations in the starting conditions or errors introduced by the integration algorithm result in completely different trajectories after a certain time (see Lyapunov-instability). This renders an exact prediction of trajectories by MD simulations impossible. Within the context of thermodynamics, however, the interest lies only in a statistical description of the simulated systems, not in the simulation of an exactly true trajectory. There are indications, but no proofs, that for sufficiently small numerical errors, simulated trajectories can be a valid statistical representation of an ensemble of true trajectories[5].

2.4 Bibliography

[1] Wendy D. Cornell et al. “A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules.” In: *J. Am. Chem. Soc.* 117.19 (1995), pp. 5179–5197.

[2] Luca Monticelli and D. Peter Tieleman. “Force Fields for Classical Molecular Dynamics.” In: *Biomolecular Simulations: Methods and Protocols*. Ed. by Luca Monticelli and Emppu Salonen. Totowa, NJ: Humana Press, 2013, pp. 197–213.

[3] Viktor Hornak et al. "Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters." In: *Proteins: Struct., Funct., Bioinf.* 65.3 (2006), pp. 712–725.

[4] David A. Pearlman et al. "AMBER, a Package of Computer Programs for Applying Molecular Mechanics, Normal Mode Analysis, Molecular Dynamics and Free Energy Calculations to Simulate the Structural and Energetic Properties of Molecules." In: *Comput. Phys. Commun.* 91.1 (1995), pp. 1–41.

[5] Berend Smit Daan Frenkel. *Understanding Molecular Simulation*. Academic Press, 1996.

[6] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. "Molecular Dynamics With Coupling to an External Bath." In: *J. Chem. Phys.* 81.8 (1984), pp. 3684–3690.

[7] Hans C. Andersen. "Molecular Dynamics Simulations at Constant Pressure And/Or Temperature." In: *J. Chem. Phys.* 72.4 (1980), pp. 2384–2393.

[8] Richard W. Pastor, Bernard R. Brooks, and Attila Szabo. "An Analysis of the Accuracy of Langevin and Molecular Dynamics Algorithms." In: *Mol. Phys.* 65.6 (1988), pp. 1409–1419.

[9] Shuichi Nose. "A Unified Formulation of the Constant Temperature Molecular Dynamics Methods." In: *J. Chem. Phys.* 81.1 (1984), pp. 511–519.

[10] M. K. Gilson, J. A. Given, B. L. Bush, and J. A. McCammon. "The Statistical-Thermodynamic Basis for Computation of Binding Affinities: A Critical Review." In: *Biophys. J.* 72 (3 1997), pp. 1047–1069.

[11] Kresten Lindorff-Larsen, Stefano Piana, Ron O. Dror, and David E. Shaw. "How Fast-Folding Proteins Fold." In: *Science* 334.6055 (2011), pp. 517–520.

[12] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman J.C Berendsen. "Numerical Integration of the Cartesian Equations of Motion of a System With Constraints: Molecular Dynamics of N-Alkanes." In: *J. Comput. Phys.* 23.3 (1977), pp. 327–341.

[13] Chad W. Hopkins, Scott Le Grand, Ross C. Walker, and Adrian E. Roitberg. "Long-Time-Step Molecular Dynamics Through Hydrogen Mass Repartitioning." In: *J. Chem. Theory. Comput.* 11.4 (2015), pp. 1864–1874.

[14] D.A. Case et al. *Amber 14*. University of California, San Francisco, 2014.

[15] Andreas W. Götz et al. "Routine Microsecond Molecular Dynamics Simulations With AMBER on GPUs. 1. Generalized Born." In: *J. Chem. Theory. Comput.* 8.5 (May 8, 2012). 22582031[pmid], pp. 1542–1555.

[16] Romelia Salomon-Ferrer, Andreas W. Götz, Duncan Poole, Scott Le Grand, and Ross C. Walker. "Routine Microsecond Molecular Dynamics Simulations With AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald." In: *J. Chem. Theory Comput.* 9.9 (2013), pp. 3878–3888.

[17] Scott H. Northrup, Stuart A. Allison, and J. Andrew McCammon. "Brownian Dynamics Simulation of Diffusion-Influenced Bimolecular Reactions." In: *J. Chem. Phys.* 80.4 (1984), pp. 1517–1524.

[18] Donald L. Ermak and J. A. McCammon. "Brownian Dynamics With Hydrodynamic Interactions." In: *J. Chem. Phys.* 69.4 (1978), pp. 1352–1360.

[19] Nathan A. Baker, David Sept, Simpson Joseph, Michael J. Holst, and J. Andrew McCammon. "Electrostatics of Nanosystems: Application to Microtubules and the Ribosome." In: *Proc. Natl. Acad. Sci. USA.* 98.18 (2001), pp. 10037–10041.

[20] Jeffrey C. Sung, Adam W. Van Wynsberghe, Rommie E. Amaro, Wilfred W. Li, and J. Andrew McCammon. "Role of Secondary Sialic Acid Binding Sites in Influenza N1 Neuraminidase." In: *J. Am. Chem. Soc.* 132.9 (2010), pp. 2883–2885.

3 Thermodynamic Concepts

This chapter is intended for the recapitulation of general concepts the presented research projects are based upon. These concern mainly the thermodynamics of ligand-receptor binding and the connection of MD simulations to statistical mechanics. Also, the frequently employed Potential of Mean Force (PMF) and a selection of enhanced sampling methods are introduced.

3.1 Ligand-receptor binding: Thermodynamic characterization

3.1.1 Standard State and Standard Binding Free Energy

All research projects presented in this thesis are related to molecular binding processes. The thermodynamic basis of the binding of molecules at the limit of low concentrations is therefore shortly revised in this section. In the following, it is assumed that the activity coefficients of the involved molecules are essentially 1 and that higher order binding interactions can be neglected.

In general, the (Gibbs) free energy between two states A and B can directly be calculated by measuring the equilibrium probabilities p_A and p_B of the respective states in a given system[1]

$$\Delta G_{AB} = G_A - G_B = -RT \ln\left(\frac{p_A}{p_B}\right) = -RT \ln(K_{AB}), \tag{3.1}$$

where R is the gas constant and T is the absolute temperature of the system. The relative probability p_A/p_B can be interpreted as a corresponding equilibrium constant K_{AB}. As binding equilibria are strongly influenced by the concentrations of the molecules, which can differ by orders of magnitude depending on the specific systems or experimental techniques, a straight-forward equilibrium constant K_{AB} does not provide a meaningful, comparable measure for the binding affinity. It is therefore convenient to define a standard free energy of binding with respect to concentration normalized standard states.

For a system with receptor molecules R and ligand molecules L which can form the complex RL, the equilibrium condition

$$\mu_R + \mu_L = \mu_{RL} \tag{3.2}$$

is given by the equality of the sum of the chemical potentials μ of molecules R and L and the corresponding bound complex RL. For sufficiently low concentrations C_i, the chemical potentials of molecule i in solution can be expressed by[1]

$$\mu_i = \mu_i^0 + RT \ln \frac{C_i}{C_0}, \tag{3.3}$$

where μ_i^0 is defined as the chemical potential of species i when the concentration C_i equals a standard concentration C_0. Usually, $C_0 = 1\,\mathrm{M}$. Then, the free energy of binding with respect to the standard states ΔG_{RL}^0 is given by[2]

$$\Delta G_{RL}^0 = \mu_{RL}^0 - \mu_R^0 - \mu_L^0 = -RT \ln \left(\frac{C^0 C_{RL}}{C_R C_L} \right) = -RT \ln(K_{RL}^0 C_0). \ ^1 \tag{3.4}$$

This measure of binding affinities is now independent of the concentrations of the involved species.

For the analysis of MD simulations, it is often easier to extract the receptor and complex concentrations via the probability p_u to find the receptor in an unbound state and the probability p_b to find the receptor with a bound ligand ($p_u + p_b = 1$). The concentrations can then be expressed as $C_R = p_u C_R^{\mathrm{tot}}$ and $C_{RL} = p_b C_R^{\mathrm{tot}}$ and equation 3.4 then becomes[3]

$$\Delta G_{RL}^0 = -RT \ln \left(\frac{p_b}{p_u} \frac{C_0}{C_L} \right) = -RT \ln(K_{RL}^0 C_0). \tag{3.5}$$

The factor C_0/C_L can be interpreted as a normalization term corresponding to the volume entropy accessible to ligand in the unbound state.

3.1.2 Standard Binding Kinetics

The kinetic transition rate from a state A into a state B is defined by[4][5]

$$k_{A\to B} = \frac{N_{A\to B}}{T_A}, \tag{3.6}$$

where $N_{A\to B}$ is the number of observed transitions per total residence time T_A of the system in state A. Straight-forward application to a binding equilibrium with state probabilities p_A and p_B leads to

$$\begin{aligned} k_{A\to B} &= \frac{N_{A\to B}}{T_A} = \frac{N_{A\to B}}{p_A T_{\mathrm{tot}}} \\ k_{B\to A} &= \frac{N_{B\to A}}{T_B} = \frac{N_{B\to A}}{p_B T_{\mathrm{tot}}}, \end{aligned} \tag{3.7}$$

[1] in some publications K_{RL}^0 is dimensionless, including C_0

where T_{tot} is the total observation time. This is consistent with the equilibrium constant K_{AB}

$$\frac{k_{A\to B}}{k_{B\to A}} = \frac{p_A}{p_B} = K_{AB} \tag{3.8}$$

corresponding to equation 3.1. Again, this does not yield comparable kinetic rates. Taking into account that at low concentrations, the kinetic unbinding rate is not concentration dependent, the kinetic rates can be defined in accordance to the standard state definitions by

$$\begin{aligned} k^0_{u\to b} &= \frac{N_{u\to b}}{T^0_u} = \frac{N_{u\to b}}{p_u \frac{C_L}{C_0} T_{\text{tot}}} \\ k^0_{b\to u} &= \frac{N_{b\to u}}{T^0_b} = \frac{N_{b\to u}}{p_b T_{\text{tot}}} = k_{b\to u}, \end{aligned} \tag{3.9}$$

consistent with the definition of the standard state equilibrium constant defined in equation 3.5

$$\frac{k^0_{u\to b}}{k^0_{b\to u}} = \frac{p_u}{p_b}\frac{C_L}{C_0} = K^0_{RL} C_0. \tag{3.10}$$

3.1.3 Steady State Binding Kinetics

Besides the equilibrium kinetic rates, a steady state association rate can be defined. The steady state association rate is of particular interest for the analysis of the influence of the diffusion contribution to the binding rate, as the diffusion limit for the steady state binding rate can be obtained analytically. When for the unbound state a ligand concentration of $C_L = C_0$ is assumed at sufficiently far distances, the steady state association rate for a purely diffusion controlled binding process can be derived from the Smoluchowski diffusion equation[1]

$$k = 4\pi D r_0, \tag{3.11}$$

where D is the relative diffusion constant of receptor and ligand and r_0 is the distance between receptor and ligand at which complex formation is defined.

Northrup et al. have shown that in general, the steady state association rate is given by[6]

$$k = \frac{4\pi D b \beta}{1 - (1-\beta) b/q}, \tag{3.12}$$

where b is a radius at which interactions between receptor and ligand can be neglected (and at which configurations are isotropic) and β is the probability of ligands at a distance b to bind to the receptor before diffusing to distances greater than q. To calculate the rate k, the probability β can for example be obtained by MD or BD simulations.

3.2 Potential of Mean Force

3.2.1 Definition

The partition function Z for a classical system described by a Hamiltonian $H(q,p)$ is given by

$$Z = \int d\Gamma e^{-H(q,p)/k_B T}, \tag{3.13}$$

where $d\Gamma$ indicates an integral over phase space described by the generalized coordinates q and the generalized momenta p, k_B is the Boltzmann constant and T the temperature of the system.

The thermodynamic potential corresponding to the canonical ensemble is the Gibbs free energy G defined by [1]

$$G = -k_B T \ln(Z). \tag{3.14}$$

The average of a thermodynamic quantity $A(q,p)$ is given by [1]

$$\langle A \rangle = \int d\Gamma \rho_K(q,p) A(q,p), \tag{3.15}$$

where A is weighted by the density function

$$\rho_K(q,p) = \frac{e^{-H(q,p)/k_B T}}{Z}. \tag{3.16}$$

Often, rather than the thermodynamic average corresponding to the total system (Eq. 3.14), the thermodynamic average along a certain coordinate $\xi(q,p)$ in phase space is of interest. The projection of the phase space probabilities on the coordinate ξ can be expressed by means of the delta function δ, yielding the average probability distribution of the coordinate

$$\langle p(\xi) \rangle = \int d\Gamma \delta \left[\xi'(q,p) - \xi \right] \rho_K. \tag{3.17}$$

The potential of mean force (PMF) $W(\xi)$ along ξ, corresponding to the free energy along ξ, can be conveniently expressed based on the probability distribution $\langle p(\xi) \rangle$ by[7]

$$W(\xi) = -k_B T \ln \langle p(\xi) \rangle + W', \tag{3.18}$$

where W' is an arbitrary constant.

The coordinate ξ can for example be the distance between a receptor and ligand, the orientation of a nucleobase in a DNA double helix or a multidimensional measure of protein conformations. The PMF allows statements about equilibrium distributions and transition barriers and therefore plays an important role in computational biophysics[3]. Taking into account the ergodicity hypothesis, the (ensemble) average probability distribution $\langle p(\xi) \rangle$ can be directly obtained from the evolution of the system as generated by MD simulations[8].

3.2.2 Umbrella Sampling

In practice, the time scales achievable by MD simulations are often too short to satisfy the thermodynamic limit. This is due to free energy barriers along or orthogonal to the coordinate of interest, resulting in slow dynamics respectively long waiting times between transition events. Umbrella sampling is a computational method that can drastically enhance the efficiency of representative phase space sampling along a specific coordinate[9].

In order to achieve focused sampling along a coordinate ξ, an artificial biasing potential (umbrella potential) $w(\xi)$ can be introduced into the system. In particular, this allows efficient sampling across unfavorable free energy regions along ξ. Combining several appropriately biased simulations with potentials $w(\xi)_i$, the whole coordinate range of interest can be explored (see Figure 3.1).

The umbrella simulations yield biased probability distributions $\langle p(\xi)\rangle_i$. As the umbrella potentials $w(\xi)_i$ only modify the Hamiltonian along the coordinate ξ, the unbiased probability distribution $\langle p(\xi)\rangle$ from a single umbrella simulation can be obtained by [7]

$$\begin{aligned}\langle p(\xi)\rangle &= \int d\Gamma \frac{\delta\left[\xi'(q,p)-\xi\right] e^{-(H(q,p)+w(\xi)_i)/k_BT} e^{w(\xi)_i/k_BT}}{e^{-(H(q,p)+w(\xi)_i)/k_BT} e^{w(\xi)_i/k_BT}} \\ &= \langle p(\xi)\rangle_i \, e^{w(\xi)_i/k_BT} \int e^{-w(\xi)_i/k_BT}.\end{aligned} \tag{3.19}$$

The PMF corresponding to one umbrella simulation can then be directly calculated from the sampled biased probability distributions by[7]

$$W(\xi) = -k_BT \ln\langle p(\xi)\rangle_i - w(\xi)_i + C_i, \tag{3.20}$$

where C_i is a constant offset. To combine the sampling of the whole set of umbrella potentials $w(\xi)_i$, the weighted histogram analysis method (WHAM)[10][7] equations can be used.

$$\langle p(\xi)\rangle = \sum_{i=1}^{N_w} n_i \langle p(\xi)\rangle_i \times \left[\sum_{j=1}^{N_w} n_j e^{-\left[w_j(\xi)-C_j\right]/k_BT}\right]^{-1} \tag{3.21}$$

$$e^{-C_i/k_BT} = \int d\xi e^{-w_i(\xi)/k_BT} \langle p(\xi)\rangle. \tag{3.22}$$

Here, N_w refers to the total number of umbrella simulations and n to the number of sampling points in the individual umbrella simulations. This set of equations has to be solved iteratively and yields the optimal PMF estimate from all combined umbrella sampling simulations. In practice, to obtain reliable results, sufficient sampling overlap between the individual umbrella simulations is necessary.

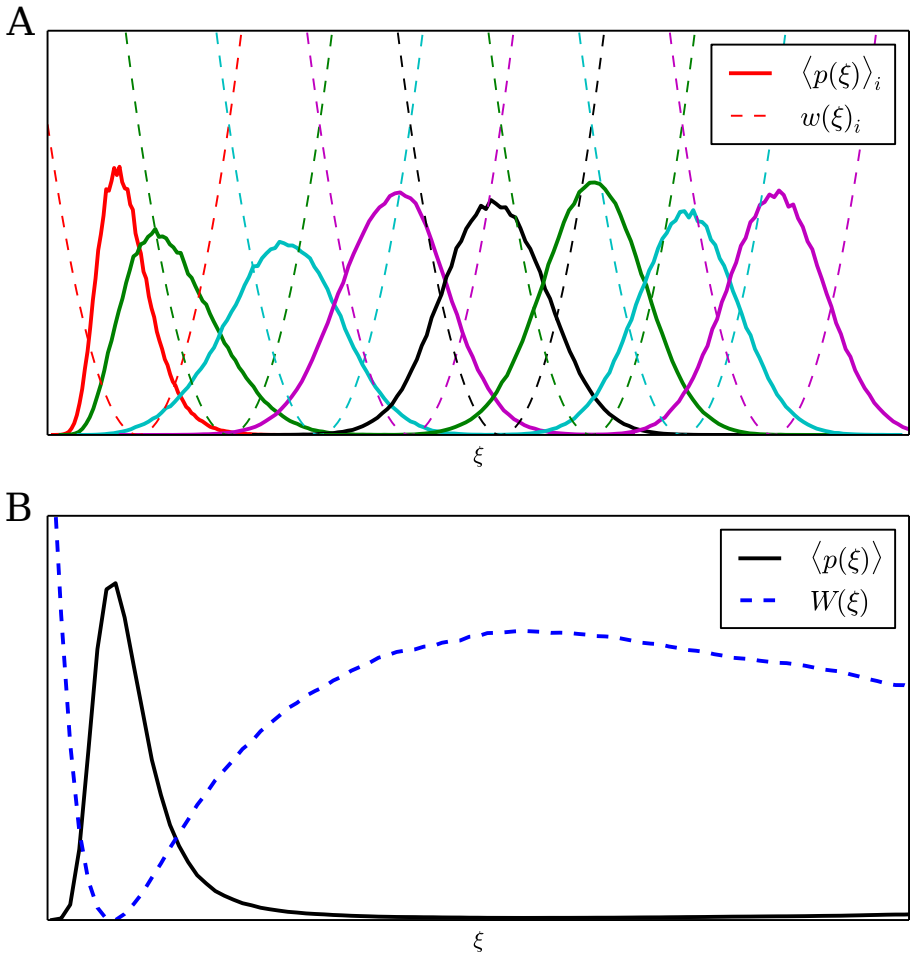


Figure 3.1: Schematic depiction of an US scheme for an arbitrary reaction coordinate ξ. A: Harmonic biasing potentials $w(\xi)_i$ and biased probability distributions $\langle p(\xi)\rangle_i$ as resulting from US simulations. B: Unbiased probability distribution $\langle p(\xi)\rangle$ and corresponding PMF $W(\xi)$ as determined by the WHAM equations.

3.2.3 Hamiltonian Replica Exchange

Umbrella potentials that are used to drive the system along a reaction coordinate range of interest can greatly enhance the sampling efficiency along the coordinate. Sampling of phase space orthogonal to the coordinate, however, is not enhanced and still requires standard diffusion. It is also possible that the umbrella potentials impede transitions between states orthogonal to the coordinate ξ due to the barriers imposed along ξ. In particular, configurations sampled in two neighboring umbrella windows might correspond to lower free energy states in other umbrella windows, but can be trapped in unfavorable states by the potential ξ.

In the Hamiltonian replica exchange (H-RE) method, at certain time intervals, the configurations sampled in the umbrella simulations are compared with respect to their internal energy. The replica exchange method was originally introduced in temperature space[11] before it was adapted also to simulations with differing Hamiltonians[12]. It can be shown that if the configurations are exchanged according to the Metropolis criterion[13],

$$p = \begin{cases} e^{-\Delta U / k_B T}, & \Delta U >= 0 \\ 1, & H_2 - H_1 < 0 \end{cases} \tag{3.23}$$

the individual umbrella simulations yield correct probability distributions according to the respective thermodynamic ensemble. Here, p is the acceptance probability of an exchange of configurations between two simulations and ΔU is the change in internal energy upon an exchange of configurations between two respective umbrella simulations. An exchange is always accepted when resulting in an overall lower energy. If the resulting overall energy change is unfavorable, the exchange is only accepted according to a Boltzmann weighted probability. Exchange attempts can in principle be applied to any pair of umbrella simulations, but in practice, attempts are often only applied to neighboring umbrella simulations, because the exchange probability of strongly differing Hamiltonians quickly decreases. H-RE umbrella sampling simulations are a convenient way to avoid trapping caused by the introduction of the umbrella potentials, to speed up equilibration of the umbrella sampling setup and to generally enhance sampling efficiency also orthogonal to ξ.

3.3 Bibliography

[1] Franz Schwabl. *Statistische Mechanik*. Springer, 2000.

[2] M. K. Gilson, J. A. Given, B. L. Bush, and J. A. McCammon. "The Statistical-Thermodynamic Basis for Computation of Binding Affinities: A Critical Review." In: *Biophys. J.* 72 (3 1997), pp. 1047–1069.

[3] Hyung-June Woo and Benoît Roux. "Calculation of Absolute Protein–ligand Binding Free Energy From Computer Simulations." In: *Proc. Natl. Acad. Sci. USA.* 102.19 (2005), pp. 6825–6830.

[4] Richard W. Pastor and Martin Karplus. "Inertial effects in butane stochastic dynamics." In: *J. Chem. Phys.* 91.1 (1989), pp. 211–218.

[5] Yuhong Zhang and Richard W Pastor. "A Comparision of Methods for Computing Transition Rates from Molecular Dynamics Simulation." In: *Mol. Simul.* 13.1 (1994), pp. 25–38.

[6] Scott H. Northrup, Stuart A. Allison, and J. Andrew McCammon. "Brownian dynamics simulation of diffusion-influenced bimolecular reactions." In: *J. Chem. Phys.* 80.4 (1984), pp. 1517–1524.

[7] Benoit Roux. "The calculation of the potential of mean force using computer simulations." In: *Comp. Phys. Com.* 91.1 (1995), pp. 275–282.

[8] Berend Smit Daan Frenkel. *Understanding Molecular Simulation*. Academic Press, 1996.

[9] G.M. Torrie and J.P. Valleau. "Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling." In: *J. Comput. Phys.* 23.2 (1977), pp. 187–199.

[10] Shankar Kumar, John M. Rosenberg, Djamal Bouzida, Robert H. Swendsen, and Peter A. Kollman. "The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method." In: *J. Comput. Chem.* 13.8 (1992), pp. 1011–1021.

[11] A. Mitsutake, Y. Sugita, and Y. Okamoto. "Generalized-Ensemble Algorithms for Molecular Simulations of Biopolymers." In: *Biopolymers* 60.2 (2001), pp. 96–123.

[12] Yilin Meng, Danial Sabri Dashti, and Adrian E. Roitberg. "Computing Alchemical Free Energy Differences With Hamiltonian Replica Exchange Molecular Dynamics (H-Remd) Simulations." In: *J. Chem. Theory Comput.* 7.9 (2011), pp. 2721–2727.

[13] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. "Equation of State Calculations by Fast Computing Machines." In: *J. Chem. Phys.* 21.6 (1953), pp. 1087–1092.

4 Evaluation of Generalized Born Model Accuracy for Absolute Binding Free Energy Calculations[1]

Generalized Born (GB) implicit solvent models are widely used in Molecular Dynamics simulations for evaluating the interactions of biomolecular complexes. The continuum treatment of the solvent results in significant computational savings in comparison to explicit solvent representation. It is, however, not clear how accurately the GB approach reproduces absolute free energies of biomolecular binding. Based on induced dissociation by means of umbrella sampling simulations absolute binding free energies of small proline-rich peptide ligands and a protein receptor were calculated. Comparative simulations according to the same protocol were performed employing an explicit solvent model and various GB-type implicit solvent models in combination with a nonpolar surface tension term. The peptide ligands differed in a key residue at the peptide-protein interface, including either a nonpolar, a neutral polar, a positively charged or a negatively charged group. For the peptides with a neutral polar or a nonpolar interface residue very good agreement between explicit solvent and GB implicit solvent results was found. Deviations in the main separation free energy contributions are smaller than 1 kcal/mol. In contrast, for the peptides with a charged interface residue significant deviations of 2-4 kcal/mol were observed. The results indicate that recent GB models can compete with explicit solvent representations in total binding free energy calculations as long as no charged residues are present at the binding interface.

4.1 Introduction

All atom Molecular Dynamics (MD) free energy simulations including explicit solvent are still too costly for routine applications in drug design or systematic evaluation of large sets of possible ligands binding to a receptor target molecule. Besides using

[1]This chapter has been previously published in similar form in The Journal of Physical Chemistry B, 118(24), 7467-7474, 2014. Some preliminary results have also been presented in the author's master thesis, "Binding Free Energy Calculations: Development of an Enhanced Sampling Technique and Evaluation of Implicit Solvent Model Accuracy", 2013.

coarse-grained force field models [1], a common strategy consists in treating the vast number of solvent molecules as a dielectric continuum. The combination of Generalized Born (GB) [2–5] models to calculate the polar and Solvent Accessible Surface Area (SASA) [6, 7] models to calculate the nonpolar solvation free energy contributions is frequently used in MD simulations (termed GB/SA). In comparison to explicit solvent simulations, the computational cost per MD time step is reduced and, as the solvent's degrees of freedom are eliminated within the continuum approach, typically less total sampling time is required to generate converged canonical ensembles. At the same time, especially for the calculation of binding free energies, it is not clear how well the GB/SA continuum approach compares to explicit solvent simulations.

GB model predictions on electrostatic solvation free energy contributions can be directly compared to Poisson-Boltzmann based (PB) models [2–5, 8]. It is possible to use GB as well as PB models to evaluate molecular binding based on post processing of trajectories that are generated in the presence of explicit solvent molecules. Such approaches have been utilized for comparative MM(Molecular Mechanics)/PBSA and MM/GBSA free energy approximations on ligand-protein complexes [9]. The accuracy of GB and GB/SA models in comparison to explicit solvent simulations has been tested on secondary or tertiary protein structure prediction [4, 5, 8], on thermal stability of folded proteins [5] and on salt-bridge strength [4, 5, 10]. Parametrization or comparison with respect to experimental solvation free energies has been carried out mostly for small molecules or amino acid side chains [2, 7]. In many cases, reasonable agreement with experiments could be achieved despite known limitations of the SASA approach[11] and the simplicity of the GB models. In particular, a tendency to over-stabilization of salt bridges has been observed for some GB models[10].

Implicit GB-type solvation models are widely utilized to investigate the binding of ligands to receptor molecules. However, the sensitivity of binding mechanisms to solvent models goes beyond the ambit the GB/SA models have been evaluated on. Careful validation of their applicability is therefore of critical importance. In this study, the performance of the GB/SA continuum models on binding processes has been evaluated in comparison to an explicit solvent model, based on rigorous free energy calculations.

Binding free energy calculations require a considerable amount of sampling, which prevents a statistical analysis of a comprehensive set of complexes. Instead, we focused on a small number of affordable cases by choosing four complexes that differ in the substitution of a key binding residue of the ligand. This allowed comparative calculations of converged free energy changes along a binding pathway for ligand-receptor dissociation. The modified residues correspond to either a charged, a polar or a non-polar chemical group. Several different GB models, namely the GB-HCT[2], GB-OBC2[3], GBn[4] and GBn2[5] models, were evaluated in combination with a nonpolar

SASA[6, 7] term. Since the chemical groups correspond to non-natural residues, they are not part of the original parametrization set of the GB models.

To calculate binding free energies, a radial potential of mean force (PMF) based approach was utilized [12]. In contrast to the alchemical double-decoupling scheme [13, 14], the calculation of large total solvation free energies, generally susceptible to errors, is avoided. The binding process is split into separation and conformation contributions, allowing further specification of the implicit solvent model performances. Additionally, some insight is given into the representation of the binding process, as it is imitated relatively closely[15].

The results suggest that for neutral ligands especially the GBn and GBn2 models predict changes in free energy that are very close to the free energy changes obtained from the explicit solvent simulations. The agreement of implicit and explicit solvent results is poorer in the case of charged key binding residues.

4.2 Methods

Simulation setups

Starting structures. Simulations were carried out on the SEM5 SH3 domain in complex with four proline rich PPPVXP ligands (Figure 4.1). The X residue indicates a substituted proline amino acid. The pyrrolidine ring of the wild type proline was mutated to N-cyclopropylmethyl (X_1, non-polar), N-(4-hydroxy) phenyl (X_2, polar), N-3-aminopropyl (X_3, positively charged) and N-2-carboxypropyl (X_4, negatively charged) (Figure 4.2). The crystallographic structure of the SEM5 SH3 domain in complex with the mutated PPPVX_1P(R)-peptide was taken from reference [16] (PDB:3SEM). An originally present arginine residue was removed from the ligand in order to obtain a neutral ligand in the case of the polar and non-polar variants. The PPPVX_1P structure was altered manually to obtain starting structures for the remaining ligands, since for these complexes no experimental structures were available.

Force fields. The ff12sb parameters [17] were used for the standard amino acid description. The modified residues were prepared using the GAFF[18] force field and antechamber, both part of AMBER12[19].

Explicit solvent simulations. The complexes were solvated in explicit TIP3P [20] water molecules in boxes of the size of $\sim 55 \times 55 \times 75\,\text{Å}^3$ and the isolated ligands in boxes of the size of $\sim 40 \times 45 \times 30\,\text{Å}^3$. Sodium or chloride ions, described by parameters from Joung and Cheatham[21], were added to the systems in order to neutralize the overall charge. A 2 fs integration time step was used with the SHAKE[22] algorithm applied to the hydrogen atoms. The short range interaction cutoff radius was 9 Å. The long range interactions were treated by the particle mesh Ewald method, using the default

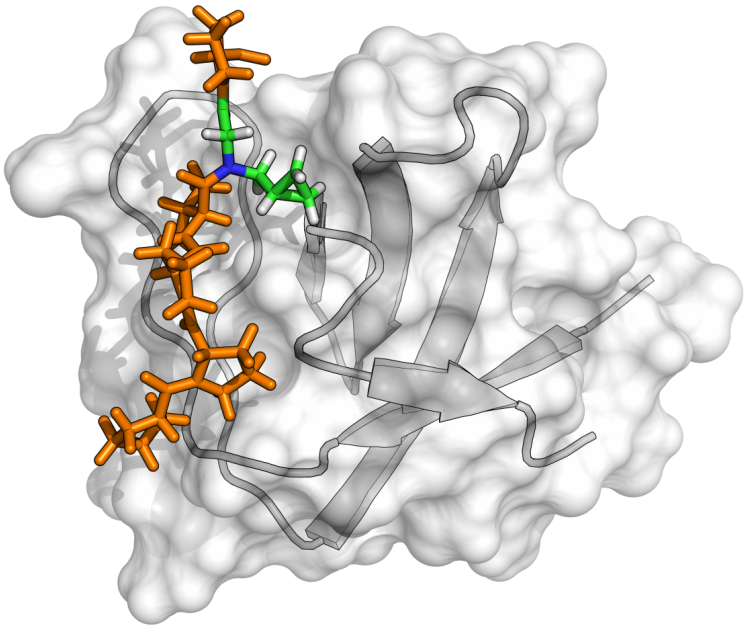
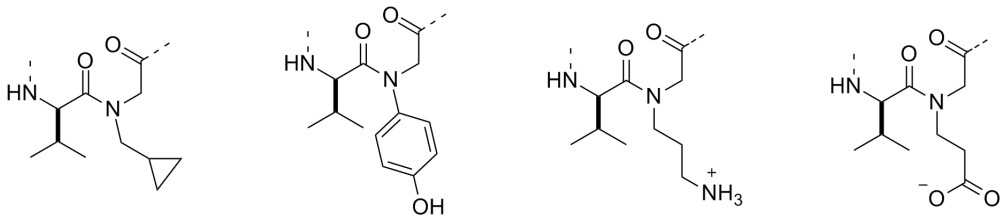


Figure 4.1: Equilibrated structure of the SEM SH3 domain (cartoon, surface) in complex with the nonpolar PPPVX$_1$P ligand (sticks, orange). The X_1 residue, which was systematically mutated to polar (X_2) and charged (X_3, X_4) residues, is illustrated with an element based color scheme.



X_1 X_2 X_3 X_4

Figure 4.2: Partial chemical structures of the investigated X_1, X_2, X_3 and X_4 ligands, differing in the mutation of an original proline residue.

AMBER12[19] parameters.

Implicit solvent simulations. Implicit solvent simulations were performed using four different GB models. a) the GB-HCT [2] model using mbondi radii and parameters from Tsui and Case [23] (this corresponds to the igb=1 option in AMBER12). b) the GB-OBC2 [3] model (igb=5), using mbondi2 [3] radii. c) the GBn [4] model (igb=7), using bondi [24] radii. d) the GBn2 [5] model (igb=8), using mbondi3 [5] radii. For a detailed description of the models we refer to the original publications. The implicit solvent simulations were carried out using a Langevin-thermostat [25] with a collision frequency of 10 ps^{-1}. As in the explicit solvent simulations, a 2 fs integration time step and SHAKE were used. The cutoff radius for the pairwise summation in the effective Born radius calculation was 25 Å, the interaction cutoff radius was 1000 Å. The external dielectric constant was 80, the internal dielectric constant was 1. The nonpolar contributions were accounted for by the solvent accessible surface area method [7], using the linear combination of pairwise overlaps (LCPO) [6] approach to estimate the surface area. The surface tension in the SASA model was 0.005 kcal/mol/$Å^2$.

In the explicit solvent simulations, neutralizing ions (the receptor is doubly negatively charged) and periodic mirror systems are present. This is not the case for the GB/SA model simulations, and the GB/SA systems are not necessarily neutral in charge. Salt concentration can be taken into account via the Debye-screening length, but it is unclear how well it compares to the presence of ions in the explicit solvent simulations. Here, for simplicity, the Debye-screening parameter was set to zero (no salt present) and the corresponding error was estimated (see Discussion).

Equilibration. For all individual systems, minimization and equilibration for a total of 1 ns was performed. During equilibration, the temperature was stepwise increased to 298 K, and positional restraints on the complexes with respect to the corresponding starting structures were gradually released.

Standard binding free energy determination via PMFs

The calculation of standard binding free energies is based on the potential of mean force (PMF) along a separation coordinate between ligand and protein, following a protocol by Woo and Roux [12] that has been successfully applied in several more studies [26–28].

In order to improve sampling efficiency, the thermodynamic binding pathway is split into several well defined intermediate states. In the bound state, the ligand's orientation (indexed by o) and movement along an axis (a) with respect to the protein as well as its conformation (c) are consecutively fixed via biasing potentials (u_o, u_a, u_c). These restraints reduce the effective phase space accessible to the ligand during a PMF calculation along a separation coordinate (r), reaching from the bound state to a

sufficiently distant bulk state (r^*). In the bulk state, all restraints are released in order to restore the free unbound system.

$$\begin{aligned} G_{\text{bind}} = & - G_a^{\text{bound},\, H_0}(u_a) - G_o^{\text{bound},\, H_0+u_a}(u_o) - G_c^{\text{bound},\, H_0+u_a+u_o}\,(W_c(\xi), u_c) \\ & - k_B T \ln \Big[I^{*\text{bound} \rightarrow \text{bulk},\, H_0+u_a+u_o+u_c}\,(r^*, W_r(r))\;\; S^{*\text{bulk},\, H_0+u_o+u_c}\,(r^*, u_a)\;\; C_0 \Big] \\ & + G_o^{\text{bulk},\, H_0+u_c}(u_o) + G_c^{\text{bulk},\, H_0}\,(W_c(\xi), u_c) \end{aligned} \tag{4.1}$$

The total standard binding free energy (Eq. 4.1) consists of the free energy changes along the constructed pathway, evaluated at the corresponding system states. The states of the system are defined by the ligand being located either in the binding site or well separated in the bulk, and the system's total Hamiltonian H, which is given as the sum of the forcefield/model potential H_0 and the applied restraining potentials u_a, u_o and u_c. The free energies G and the S^* contribution are related to introducing/releasing the restraints. I^* contains the separation PMF and corresponds to the energy cost of pulling the ligand from the bound state into the bulk. The volume accessible to the ligand in the bulk state is normalized with respect to a volume corresponding to the standard concentration C_0. The separation PMF $W_r(r)$ and two conformational PMFs $W_c(\xi)$, calculated in the bound and bulk states along a collective positional RMSD coordinate ξ, imply the main sampling effort. The G_o^{bound} and G_a^{bound} contributions were calculated via free energy perturbation (FEP).

A derivation of the expression for the total standard binding free energy from first principles can be found in the original study [12].

Free energy sampling

All simulations were carried out with the AMBER12[19] program package. The sampling trajectories were generated using the NVT ensemble at 298 K. Sampling data was recorded every 0.1 ps. Slight harmonic positional restraints (f_{pos} =0.002 kcal/mol/Å^2) to the starting structures were applied on the receptor atoms during the sampling runs in order to avoid drifting of the systems.

Bound state FEP. The free energy cost of consecutively introducing the five angular and dihedral angular harmonic potentials (summing up to G_o^{bound} and G_a^{bound}) on the complex in the bound state was calculated via FEP, based on 5 ns of total sampling. The complexes were sampled for 1 ns before introducing each respective restraint.

PMF calculations. The umbrella sampling (US) potentials of the individual windows i are given by $u_r^{US,i} = k_r^{US,i}(r - r_0^{US,i})^2$ and $u_c^{US,i} = k_c^{US,i} N(\xi - \xi_0^{US,i})^2$ (N indicates the number of atoms corresponding to the RMSD mask) for the radial (r) and the RMSD (ξ) US simulations, respectively. The starting configurations for the umbrella simulations

were generated from the equilibrated structures by simulating every umbrella window consecutively for 10 ps, starting from the umbrella window closest to the starting structure state. Hamiltonian replica exchanges of the configurations of neighboring umbrella windows were attempted every 2 ps according to a Metropolis-criterion, using the H-REMD routine implemented in AMBER12[19]. From all umbrella windows, data generated within the first 20% of sampling time was omitted in the PMF calculations in order to assure equilibration of the systems. The PMFs were calculated from the US probability distributions using the weighted histogram analysis method (WHAM). [29, 30] The separation PMFs were calculated using 48 umbrella potentials with r_0^{US} covering a range from 6.5 Å to 22.5 Å, with an intermediate spacing of 0.2 Å between 8.0 Å and 13.0 Å and a spacing of 0.5 Å everywhere else. The force constants k_r^{US} were 10 kcal/mol/Å^2 for the windows in the 0.5 Å spaced regions and 24 kcal/mol/Å^2 for the windows in the 0.2 Å spaced region. During the simulations, the restraining potentials u_a, u_o and u_c were applied. Simulation time was 6 ns per window. The PMFs for the bound state ligand RMSDs were calculated using 12 windows with ζ_0^{US} covering a range from 0.0 Å to 2.2 Å. Intermediate spacing was 0.2 Å and a force constant k_c^{US} of 0.3 kcal/mol/Å^2 was used. During the simulations, the restraining potentials u_a and u_o were applied. Sampling time was 3 ns per window. The isolated bulk state ligand RMSD PMFs were calculated using 28 windows with ζ_0^{US} covering a range from 0.0 Å to 5.4 Å. Intermediate spacing was 0.2 Å and a force constant k_c^{US} of 0.3 kcal/mol/Å^2 was used. Simulation time was 6 ns per window. In order to be able to compare the models along the same thermodynamic pathway, the reference angles optimal for the explicit solvent simulations were also used for the GB/SA systems. Similarly, an equilibrated explicit solvent structure was used as the ligand reference structure for the conformation restraints in all simulations. The reference structures were minimized for 10000 steps of steepest descent with the corresponding GB/SA models to avoid artificial internal tension. The effects of the minimization were local and did not lead to global changes in the reference structures.

Error estimates. All sampling runs were divided into four successive subruns. Statistical uncertainties were estimated as the root mean square deviation of the individual subrun free energies from the total run free energy. For the correlation times associated with the simulations performed in this study, this method yields more meaningful error estimates than bootstrap or block average procedures, which result in too small statistical uncertainties. The error bars indicated in the PMF plots correspond to the respective subrun PMFs showing the largest deviation.

4.3 Results and Discussion

Umbrella sampling simulations were used to calculate the binding free energies of four peptide-protein complexes with an explicit and different implicit solvent models. Each of the complexes consists of a systematically modified proline rich peptide ligand bound to a small SH3 protein domain (SEM-SH3). The fifth position of the peptide ligand (PPPVXP) corresponds to a non-natural amino acid that, in the bound state, is buried at the interface (Figure 4.1). Experimental studies on the system indicate that the substitution of this position by nonpolar (X_1), polar (X_2) or charged (X_3, X_4) chemical motives (Figure 4.2) can dramatically change the binding affinity [16]. Due to its sensitivity to specific ligand residue modifications, its small size and the availability of the crystal structure for the PPPVX_1P complex, the SH3 domain constitutes an ideal model system for a comparative analysis of binding free energy sampling based on different solvation models. Within the framework of the binding affinity calculation, several free energy contributions have to be determined (explained in the Methods section). These include contributions due to restraints on the ligand's relative orientation (G_o) and position (G_a, S^*) as well as on its internal conformation (G_c) in the bound and dissociated states. While the release of orientational and positional restraints in the bulk state can be handled analytically, all other contributions have to be determined by numerical sampling. The largest free energy contribution $W_r(r^*)$ arises from the dissociation of the peptide from the protein binding pocket and is discussed in the following paragraph.

Simulations employing the explicit TIP3P water model served as reference for all comparative considerations. In extensive comparative free energy simulations of the solvation of amino acid side chains, which represent a quite diverse set of polar and nonpolar organic molecules, little dependence of calculated free energies on the molecular water model was observed[31]. For example, average deviations in the calculated absolute solvation free energies of less than 0.25 kcal/mol between TIP3P or TIP4P simulations were found for all side chains. Since in binding free energy simulations amino acid side chains become only partially desolvated, it is to expect that the influence of the explicit water model is of the same order or less.

Radial separation PMFs. The separation PMFs for all systems are shown in Figure 4.3. For all systems, the TIP3P curves show a steeper rise in free energy around $r \approx 10$ Å. This presumably arises from the breaking of an explicitly present hydration shell around the binding site and is not accurately reproduced by the GB/SA continuum models. It is also observed in the case of the more recent GBn and GBn2 models that include a geometrical approach to better account for regions that are too small to accommodate explicit water molecules. The free energy differences were evaluated

between the binding site minimum and a bulk distance r^*. The expected asymptotic behavior with no interactions between protein and ligand is $-k_B T \ln(r^2) + const.$, due to the gain in entropy associated with the sphere shell element $\propto r^2$ accessible to the ligand. This asymptote is well reproduced by the reference explicit solvent curves at $r^* = 23\,\text{Å}$, which is therefore considered a sufficient bulk separation.

The explicit solvent free energy minimum regions at the binding sites are closely reproduced by all GB/SA models for the neutral ligands, whereas the predictions for the charged ligands are considerably less accurate. This is consistent with the binding site FEP results (see below).

For the neutral X_1 and X_2 ligands, at $r^* = 23\,\text{Å}$ all GB/SA models predict free energy differences that deviate less than 10% from the explicit solvent simulations, which is about the magnitude of the estimated sampling uncertainties. While the GB-HCT model slightly overestimates the change in free energy, the GBn2 model shows the closest agreement with the all-atom predictions (deviation < 5%). For the positively charged X_3 ligand, the GBn2 model underestimates the free energy change for the ligand separation by $\approx 2\,\text{kcal/mol}$ and the GB-HCT model gives the best agreement with the explicit solvent simulations. The other models underestimate the free energy change by $\approx 3\,\text{kcal/mol}$ compared to the TIP3P results. Similarly, in the case of the negatively charged side chain modification (X_4), the deviation of the GB/SA results from the explicit solvent calculations are larger than for the neutral X_1 and X_2 ligands. Here, all GB/SA variants underestimate the free energy change by $\approx$ 3-4 kcal/mol compared to the explicit solvent prediction, except for the GBn2 model which differs by $\approx 1.5\,\text{kcal/mol}$.

As mentioned above, for the separation PMFs the influence of the different screening characteristics on the explicit and implicit solvent simulations has to be estimated. For an upper estimate of deviations in the extreme case of complete screening of the electrostatic interactions in the explicit solvent simulations and no screening in the implicit solvent simulations, the coulombic interaction between receptor and ligand was approximated. To this end, corresponding point charge energies above the typical Debye-screening length were calculated. The estimate indicates that the difference in free energy is below $\approx 1.1\,\text{kcal/mol}$ for the charged ligands and presumably far lower for the neutral ligands. Thus, although different screening effects can have considerable influences on the separation PMFs, they do not affect the general conclusions about the GB/SA models in this study.

Bound state FEP on orientational and positional freedom of the ligand. The free energy cost of introducing the orientational and positional restraints in the bound state (Table 4.1) is an indirect measure for the ability of the GB/SA models to accurately represent the binding site properties. For the nonpolar and the polar ligands, the GB-OBC, GBn and GBn2 models achieved better agreement with explicit solvent

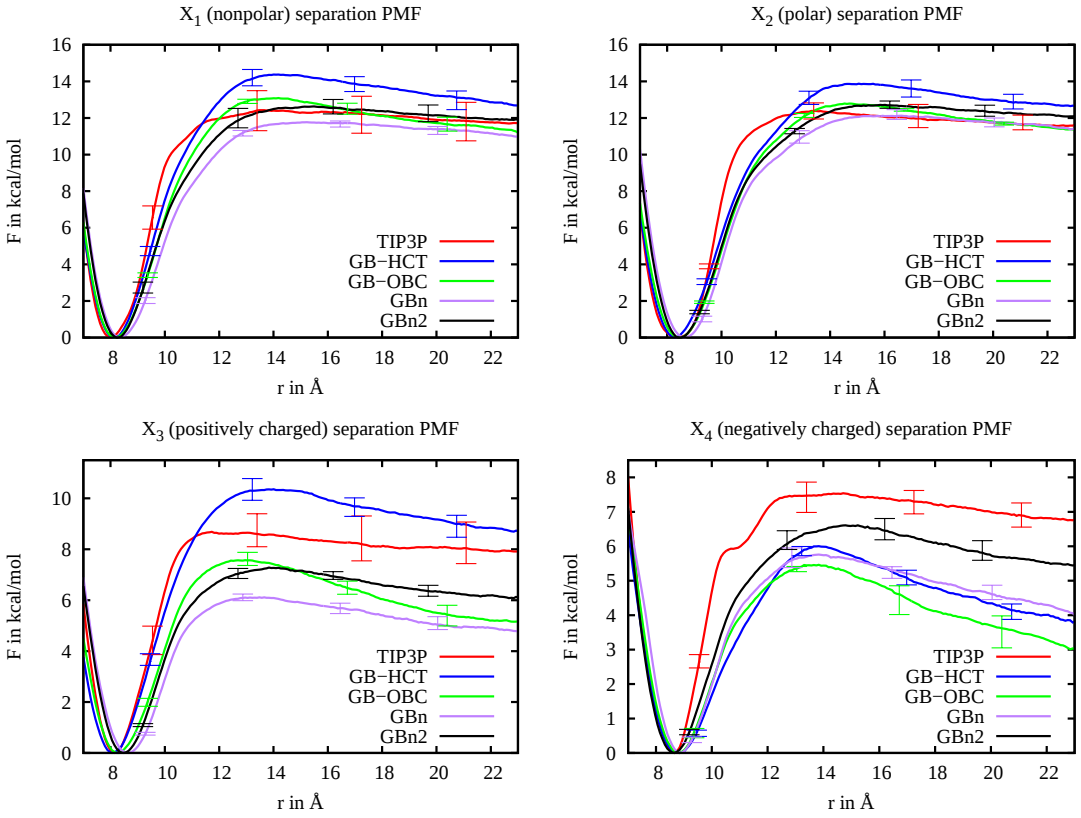


Figure 4.3: Separation PMFs for the four ligands, calculated with explicit TIP3P water representation and four different continuum solvent GB/SA models. In all PMFs, only few error bars are included to maintain legibility.

G_{o+a}^{bound}	TIP3P	GB-HCT	GB-OBC	GBn	GBn2
X_1	1.77±0.30	2.87±0.74	2.51±0.32	2.44±0.55	2.57±0.80
X_2	1.91±0.34	3.37±0.67	2.17±0.29	3.05±0.97	2.42±0.49
X_3	3.75±0.98	[12.00±43.54]	[6.52±4.63]	5.12±1.70	[14.17±121.25]
X_4	3.56±1.00	[4.90±7.17]	[5.17±10.37]	[27.58±5.02]	5.91±1.56
G_c^{bound}	TIP3P	GB-HCT	GB-OBC	GBn	GBn2
X_1	0.40±0.04	0.36±0.02	0.46±0.02	0.54±0.02	0.51±0.05
X_2	1.00±0.06	[0.91±0.04]	[1.27±0.04]	[0.86±0.04]	0.88±0.04
X_3	1.01±0.05	[1.43±0.12]	[1.73±0.21]	[1.64±0.06]	[1.94±0.26]
X_4	1.00±0.12	[1.49±0.11]	[1.11±0.08]	[1.03±0.10]	[1.02±0.14]
G_c^{bulk}	TIP3P	GB-HCT	GB-OBC	GBn	GBn2
X_1	1.76±0.28	2.26±0.34	2.38±0.35	1.95±0.24	1.82±0.13
X_2	2.08±0.15	2.98±0.17	2.99±0.19	2.78±0.26	2.58±0.35
X_3	2.29±0.08	[3.91±0.36]	[3.42±0.17]	2.61±0.21	2.91±0.28
X_4	2.31±0.14	[2.64±0.41]	[2.96±0.28]	2.35±0.12	[2.42±0.18]

Table 4.1: Binding free energy contributions $G_o^{\text{bound}} + G_a^{\text{bound}}$, G_c^{bound} and G_c^{bulk} in kcal/-mol. Results in square brackets for $G_o^{\text{bound}} + G_a^{\text{bound}}$ may not be completely converged, but are given for completeness. For the G_c^{bound} and G_c^{bulk} values in square brackets, the underlying GB/SA RMSD PMFs significantly differ from the reference TIP3P PMFs.

simulations than the GB-HCT model and the deviations are within the error estimates. For the charged ligands, however, the GB/SA models are in much worse agreement with explicit solvent simulations. The charged ligands are hardly restricted to their original orientation within the binding pocket. While for the TIP3P simulations the FEP calculations yield reasonable results, the apparent higher mobility of the ligands in the GB/SA models prevents converged binding site sampling, resulting in unreliable values. Contributions with estimated errors larger than the TIP3P reference values are therefore indicated by square brackets in Table 4.1. Note, that burying a charged chemical group at a protein binding interface represents an especially unfavorable situation because of the strong desolvation penalty and involves a delicate balance with respect to other favorable interactions to stabilize the peptide binding. Here, apparently even small differences in the GB models can significantly affect the mobility of the peptide in the binding pocket.

Bound state ligand RMSD PMFs. The free energy contribution of restraining the ligand conformation in the bound state was calculated from a potential of mean force

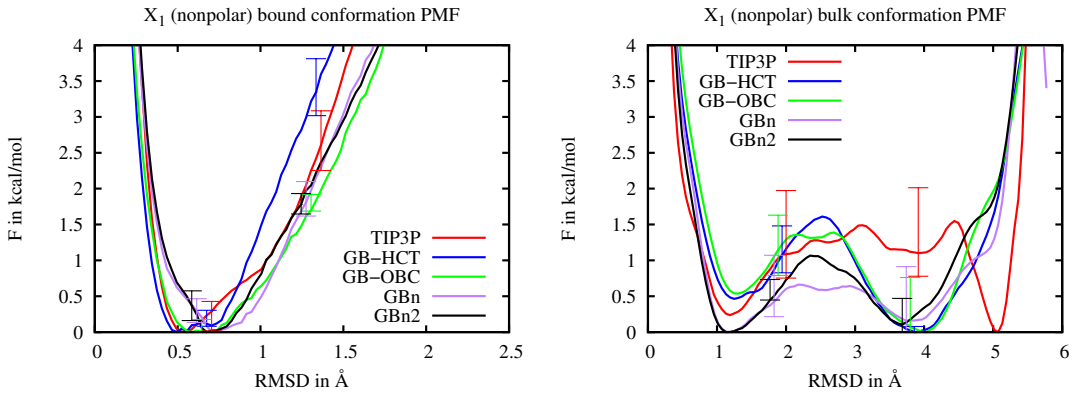


Figure 4.4: Bound and bulk ligand configuration RMSD PMF for the X_1 ligand, calculated with explicit TIP3P water representation and four different continuum solvent GB/SA models.

along a RMSD coordinate, using an equilibrated bound conformation as reference. As an example, the bound RMSD PMF is shown for the X_1 ligand (Figure 4.4). All GB/SA models reproduce the minimum region within a 0.2 Å deviation from the explicit solvent result. The location of the free energy barriers at small RMSD values (due to thermal fluctuations) and at large RMSD values is within 0.5 Å predicted by the GB/SA models. For the X_2 ligand, the majority of the GB/SA models perform similarly well, although partially predicting a second minimum at an RMSD value of about 2 Å not present in the calculated explicit solvent curve. For the charged X_3 and X_4 ligands, all GB/SA model predictions considerably differ from the TIP3P curves, especially at large RMSDs.

The free energy contributions calculated from the PMF curves are listed in Table 4.1. Due to the integration along the RMSD coordinate, deviations in the PMFs frequently cancel out and are not well reflected in the resulting single free energy values. Values for G_c^{bound} that should be carefully interpreted with regard to the underlying PMFs are indicated by square brackets (this also applies for the G_c^{bulk} contributions).

Bulk state ligand RMSD PMFs. The free energy contribution of restraining the peptide in the bulk was calculated from a PMF along the RMSD in the isolated state (illustrated for the X_1 ligand in Figure 4.4). As in the bound state case, the position of the first free energy minimum is well reproduced by all GB/SA models. Especially the GBn2 model, except for offsets, predicts the free energy curve quite correctly up to RMSDs of about 2.5 Å. At larger RMSDs, the GB/SA models show significant deviations from the TIP3P predictions. A second deep free energy minimum at 5 Å for X_1, predicted by the TIP3P simulation, is not reproduced by the GB/SA models. Inspection of the MD trajectories

indicated that it corresponds to a relatively stable hairpin-like peptide conformation with the two terminal proline residues in close proximity. The location of the upper RMSD free energy barriers is again well reproduced. GB/SA model performance is considerably worse for the charged ligands, partially showing little correlation with the TIP3P results. For the nonpolar and the polar ligands, the basic characteristics of the TIP3P PMFs are reproduced by the GB/SA models, and the corresponding free energy contributions (Table 4.1) coincide within about 20%.

Total binding free energies. The total binding free energies (see Table 4.2) are primarily given for completeness. As the total values can be affected by error compensation, the individual values for the different contributions are more significant. For the neutral ligands, the results of the GB/SA models are overall close to the explicit solvent simulations. All models except for GB-HCT predict an absolute binding free energy within the error margin of the explicit solvent simulations. Total GB/SA binding free energies were not calculated for the charged ligands, since complete convergence of the sampling for the G_a^{bound} and G_o^{bound} contributions could not be achieved.

In addition, comparable experimental binding free energies for the simulated ligands are listed. These values were determined for corresponding longer 12-mer peptides with an additional positive charge [16], contributing an attractive interaction with the negatively charged receptor protein. Hence, only the relative experimental binding free energies can be compared to the calculations. The explicit solvent model reproduces the experimentally observed ranking of the peptide variants very well.

Note, that any possible change of protonation states, especially of the charged ligands, are not included in the TIP3P or GB/SA solvent models. Furthermore, the discussed ranking of the models is based on combined GB/SA calculations with the respective standard parameterization, representing the most common application. Relative differences in the SASA-based nonpolar solvation term for the different GB/SA model simulations did not exceed 0.5 kcal/mol e.g. for the separation free energy contributions. Hence, variation of the surface tension parameter in the SASA based term can influence the calculated binding free energies by a few tenths of a kcal/mol but has only a minor influence on the relative ranking of similarly well performing GB models for ligand binding.

4.4 Conclusion

Absolute binding free energies of a series of ligands forming complexes with a SH3 domain have been determined using explicit solvent (TIP3P) and several GB/SA continuum solvent models. For each model, absolute binding free energies were

G_{bind}	TIP3P	GB-HCT	GB-OBC	GBn	GBn2	exp.[16]
X_1	-2.67 ± 1.40	-4.32 ± 1.44	-2.61 ± 1.02	-2.75 ± 1.04	-3.67 ± 1.40	-5.9
X_2	-3.11 ± 0.86	-4.60 ± 1.17	-2.57 ± 0.66	-3.32 ± 1.50	-3.46 ± 1.19	-7.4
X_3	-0.90 ± 1.83	-	-	-	-	-4.0
X_4	$+0.41 \pm 1.55$	-	-	-	-	-3.5

Table 4.2: Total binding free energies in kcal/mol calculated with explicit TIP3P water molecules and the GB/SA models. GB/SA model total binding free energies are not given for the charged ligands because convergence of some contributions was not achieved (see Results/Discussion for details). Experimental values for the ligand peptides with 12-mer background are given for relative comparison.

calculated following exactly the same protocol. This allows a direct comparison of each individual contribution along the binding pathway. For the neutral nonpolar and polar ligands, the GBn and GBn2 models showed surprisingly good agreement with explicit solvent. The predictions for the largest free energy contribution corresponding to the ligand-receptor separation deviate less than 10% from the TIP3P simulations, and the total calculated binding free energies agree within the sampling uncertainty. The present results suggest that recent GB/SA models are able to provide reasonable and computationally much cheaper binding affinity predictions, as long as no charged chemical groups are involved. For the charged ligands, however, the separation free energy predictions of all GB/SA models are up to 4 kcal/mol below the TIP3P results which is in the range of the total binding free energies of the investigated systems. Consequently, the GB/SA models predicted an even stronger decrease in binding affinity due to the introduction of charged side chains than the explicit solvent model. The binding site interactions are well reproduced by the newer GB/SA models for the neutral ligands, but less accurately for the peptide ligands with partially buried charges at the interface. Similarly, the configurational PMFs are reasonably reproduced by the GB/SA models for the neutral ligands, while for charged ligands significant deviations with respect to the explicit solvent reference simulations were observed.

It is important to note that the desolvation penalty for a charged chemical group at an interface is much larger than for a nonpolar or polar group. Hence, any inaccuracy or parameter dependence of the GB/SA models will be reflected to a greater extent in the free energy results for the charged ligand cases. In practical applications, accurate quantitative prediction of the effect of partial burying of charged groups, known to strongly destabilize ligand binding, will be of less importance than the prediction of changes due to polar or nonpolar groups at interfaces. In this regard,

the current results indicate that especially the GBn and GBn2 models compare very well with explicit solvent simulations. Finally, it should be emphasized that due to the large computational demands the conclusions drawn in this study are based on the investigation of only one receptor in complex with four different ligands. A future analysis of a broader range of ligand-receptor complexes is desirable to evaluate the generality of the presented results.

4.5 Bibliography

[1] Marissa G. Saunders and Voth A. Gregory. "Coarse-graining Methods for Computational Biology." In: *Annu. Rev. Biophys.* 42 (2013), pp. 73–93.

[2] Gregory D. Hawkins, Christopher J. Cramer, and Donald G. Truhlar. "Parametrized Models of Aqueous Free Energies of Solvation Based on Pairwise Descreening of Solute Atomic Charges from a Dielectric Medium." In: *J. Phys. Chem.* 100.51 (1996), pp. 19824–19839.

[3] 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: Struct., Func., and Bioinf.* 55.2 (2004), pp. 383–394.

[4] John Mongan, Carlos Simmerling, J. Andrew McCammon, David A. Case, and Alexey Onufriev. "Generalized Born Model with Simple, Robust Molecular Volume Correction." In: *J. Chem. Theory Comput.* 3 (2007), pp. 156–169.

[5] Hai Nguyen, Daniel R. Roe, and Carlos Simmerling. "Improved Generalized Born Solvent Model Parameters for Protein Simulations." In: *J. Chem. Theory Comput.* 9.4 (2013), pp. 2020–2034.

[6] Jörg Weiser, Peter S. Shenkin, and W. Clark Still. "Approximate Atomic Surfaces from Linear Combinations of Pairwise Overlaps (LCPO)." In: *J. Comput. Chem.* 20.2 (1999), pp. 217–230.

[7] Doree Sitkoff, Kim A. Sharp, and Barry Honig. "Accurate Calculation of Hydration Free Energies Using Macroscopic Solvent Models." In: *J. Phys. Chem.* 98.7 (1994), pp. 1978–1988.

[8] Michael Feig et al. "Performance Comparison of Generalized Born and Poisson Methods in the Calculation of Electrostatic Solvation Energies for Protein Structures." In: *J. Comput. Chem.* 25.2 (2004), pp. 265–284.

[9] Tingjun Hou, Junmei Wang, Youyong Li, and Wei Wang. "Assessing the Performance of the MM/PBSA and MM/GBSA Methods. 1. The Accuracy of Binding Free Energy Calculations Based on Molecular Dynamics Simulations." In: *J. Chem. Inf. Model.* 51.1 (2011), pp. 69–82.

[10] Raphaël Geney, Melinda Layten, Roberto Gomperts, Viktor Hornak, and Carlos Simmerling. "Investigation of Salt Bridge Stability in a Generalized Born Solvent Model." In: *J. Chem. Theory Comput.* 2.1 (2006), pp. 115–127.

[11] Jianhan Chen and Charles L. Brooks. "Implicit Modeling of Nonpolar Solvation for Simulating Protein Folding and Conformational Transitions." In: *Phys. Chem. Chem. Phys.* 10 (4 2008), pp. 471–481.

[12] Hyung-June Woo and Benoit Roux. "Calculation of Absolute Protein–Ligand Binding Free Energy from Computer Simulations." In: *Proc. Natl. Acad. Sci. U.S.A.* 102.19 (2005), pp. 6825–6830.

[13] M. K. Gilson, J. A. Given, B. L. Bush, and J. A. McCammon. "The Statistical-Thermodynamic Basis for Computation of Binding Affinities: A Critical Review." In: *Biophys. J.* 72 (1997), pp. 1047–1069.

[14] Hideaki Fujitani et al. "Direct Calculation of the Binding Free Energies of FKBP Ligands." In: *J. Chem. Phys.* 123.8 (2005), p. 084108.

[15] Yuqing Deng and Benoit Roux. "Computations of Standard Binding Free Energies with Molecular Dynamics Simulations." In: *J. Phys. Chem. B* 113.8 (2009), pp. 2234–2246.

[16] Jack T. Nguyen, Christoph W. Turck, Fred E. Cohen, Ronald N. Zuckermann, and Wendell A. Lim. "Exploiting the Basis of Proline Recognition by SH3 and WW Domains: Design of N-Substituted Inhibitors." In: *Science* 282.5396 (1998), pp. 2088–2092.

[17] Wendy D. Cornell et al. "A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules." In: *J. Am. Chem. Soc.* 117.19 (1995), pp. 5179–5197.

[18] Junmei Wang, Romain M. Wolf, James W. Caldwell, Peter A. Kollman, and David A. Case. "Development and Testing of a General Amber Force Field." In: *J. Comput. Chem.* 25.9 (2004), pp. 1157–1174.

[19] D. A. Case et al. *AMBER12*. University of California, San Francisco, CA, 2012.

[20] William L. Jorgensen, Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. "Comparison of Simple Potential Functions for Simulating Liquid Water." In: *J. Chem. Phys.* 79.2 (1983), pp. 926–935.

[21] In Suk Joung and Thomas E. Cheatham. "Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations." In: *J. Phys. Chem. B* 112.30 (2008), pp. 9020–9041.

[22] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman J.C Berendsen. "Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes." In: *J. Comput. Phys.* 23.3 (1977), pp. 327–341.

[23] Vickie Tsui and David A. Case. "Theory and Applications of the Generalized Born Solvation Model in Macromolecular Simulations." In: *Biopolymers* 56.4 (2000), pp. 275–291.

[24] A. Bondi. "Van der Waals Volumes and Radii." In: *J. Phys. Chem.* 68.3 (1964), pp. 441–451.

[25] Richard W. Pastor, Bernard R. Brooks, and Attila Szabo. "An Analysis of the Accuracy of Langevin and Molecular Dynamics Algorithms." In: *Mol. Phys.* 65.6 (1988), pp. 1409–1419.

[26] James C. Gumbart, Benoit Roux, and Christophe Chipot. "Standard Binding Free Energies from Computer Simulations: What is the Best Strategy?" In: *J. Chem. Theory Comput.* 9.1 (2013), pp. 794–802.

[27] James C. Gumbart, Benoit Roux, and Christophe Chipot. "Efficient Determination of Protein–Protein Standard Binding Free Energies from First Principles." In: *J. Chem. Theory Comput.* 9.8 (2013), pp. 3789–3798.

[28] Fabian Zeller and Martin Zacharias. "Adaptive Biasing Combined with Hamiltonian Replica Exchange to Improve Umbrella Sampling Free Energy Simulations." In: *J. Chem. Theory Comput.* 10.2 (2014), pp. 703–710.

[29] Shankar Kumar, John M. Rosenberg, Djamal Bouzida, Robert H. Swendsen, and Peter A. Kollman. "The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method." In: *J. Comput. Chem.* 13.8 (1992), pp. 1011–1021.

[30] Alan Grossfield. *WHAM: The weighted histogram analysis method 2.0.7*. Grossfield Lab: Rochester, NY, 2013.

[31] Michael R. Shirts and Vijay S. Pande. "Solvation Free Energies of Amino Acid Side Chain Analogs for Common Molecular Mechanics Water Models." In: *J. Chem. Phys.* 122.13 (2005), p. 134508.

5 Efficient Calculation of Relative Binding Free Energies by Umbrella Sampling Perturbation[1]

An important task of biomolecular simulation is the calculation of relative binding free energies upon chemical modification of partner molecules in a biomolecular complex. The potential of mean force (PMF) along a reaction coordinate for association or dissociation of the complex can be employed to estimate binding affinities. A free energy perturbation approach, termed umbrella sampling perturbation, has been designed that allows an efficient calculation of the change of the PMF upon modification of a binding partner based on the trajectories obtained for the wild type reference complex. The approach was tested on the interaction of modified water molecules in aqueous solution and applied to in silico alanine scanning of a peptide-protein complex. For the water interaction test case, excellent agreement with an explicit PMF calculation for each modification was obtained as long as no long range electrostatic perturbations were considered. For the alanine scanning, the experimentally determined ranking and binding affinity changes upon alanine substitutions could be reproduced within 0.6 – 2.0 kcal/mol. In addition, good agreement with explicitly calculated PMFs was obtained mostly within the sampling uncertainty. The combined umbrella sampling and perturbation approach yields, under the condition of sufficiently small system modifications, rigorously derived changes in free energy and is applicable to any PMF calculation.

5.1 Introduction

Molecular dynamics (MD) simulations in combination with umbrella sampling (US)[1] are widely used in computational biophysics for the calculation of potentials of mean force (PMFs) along specific coordinates[2–5]. Typically, a series of simulations is set up with different harmonic potentials added to the force field in order to achieve an efficient free energy sampling in selected regions of the reaction coordinate. In

[1]This chapter has been previously published in similar form in the Journal of Computational Chemistry, 35(31), 2256-2262, 2014.

many cases a detailed investigation of the effects due to specific modifications of systems, such as the substitution of single atoms, chemical groups or force field parameters, by means of PMF calculations is of interest for a mechanistic understanding or molecular design applications. For example, in the case of ligand-receptor binding studies, a central contribution to the binding affinity can be calculated as a PMF along a reaction coordinate separating the ligand and receptor molecules. The effect of a ligand modification on binding can then in principle be obtained by sampling a separate PMF for the modified ligand-receptor complex. The systematic sampling of PMFs is, however, computationally costly since for every modification a whole set of additional simulations for all umbrella windows along the reaction coordinate is required. Alternatively, the effect of a chemical ligand modification on its binding affinity to a receptor can be evaluated by computational alchemical transformations in the bound and unbound states of the ligand[6]. Again, this is a computationally expensive approach because typically for every modification a series of simulations of intermediate states is needed. The Bennett's Acceptance Ratio[7] and one-step perturbation approaches [8, 9] have been used to avoid the simulation of intermediate states during alchemical transformations. Such approaches are limited by the degree of sampling overlap between the end-states. Methods aiming at more rapid and less costly free energy estimates have been developed but are usually limited to special applications or do not yield rigorous free energies. For instance, in the molecular mechanics Poisson-Boltzmann surface area (MM/PBSA)[10] or the linear interaction energy (LIE)[11] methods pseudo-free energies of binding are calculated by comparing average energies of the bound complex and the isolated partners in single MD simulations.

In this study, we present a perturbation method that allows one to estimate the change of a PMF upon a modification of a system's Hamiltonian, based on previously generated US trajectories for the original system. To this end, the change in potential energy corresponding to the modification is calculated for each stored configuration and the US probability distributions are reweighted accordingly. As in practice sampling intervals much larger than the MD integration time step are sufficient, the computational cost for the post-processing is only a small fraction of the effort necessary for a complete recalculation of the PMF based on MD simulations for each umbrella window. The approach can be considered an extension of free energy perturbation (FEP)[12] to US simulations and PMF calculations. Similarly, the necessary condition for reasonable free energy estimates is sufficient phase space overlap between the original and the modified system. This usually implies that the system does not undergo significant configurational changes upon the perturbation and that the dynamics of the perturbed system are close to the dynamics of the originally simulated system.

We first tested the applicability of the method on the PMF associated with the separation of two water molecules in aqueous solution and the effect of force field

modifications. For these systems with a low number of degrees of freedom the agreement of the PMFs obtained by the reweighting scheme with standard US results was excellent in the cases of an implicit solvent model and an explicit solvent model in combination with mutations not involving a change of charges. Using an explicit solvent representation, limitations were observed for modifications that involved long ranged changes of the electric field. We then employed the method for an estimation of the influence of substituting peptide amino acids by alanine on the binding of the PMI-MDM2 complex. The PMI-MDM2 complex is formed by a peptide fragment of the protein p53 (TSFAEYWNLLSP, termed PMI) and the MDM2 receptor protein[13]. For this purpose, the main binding free energy contribution was calculated for the original PMI ligand by means of a PMF along a restrained separation pathway between peptide and receptor, using an implicit GB/SA (Generalized Born/Surface Area) solvent model. Based on the obtained US trajectories, the PMFs corresponding to the PMI alanine mutations were calculated using the perturbation scheme. Here, the application of the method is challenging in the sense that the mutations of original tryptophan or phenylalanine residues at the binding site to alanine constitute considerable modifications of the original system. For four specific mutations known to have significant effects on binding, the PMFs obtained by reweighting the original PMI trajectories were compared with corresponding directly sampled PMFs. An agreement within $1.0-1.6\,\mathrm{kcal/mol}$ was observed. The relative changes in separation free energy for all mutations agreed within $\pm 1.6\,\mathrm{kcal/mol}$ with experimental binding free energy results. These deviations are in the order of the sampling uncertainty of the original PMF calculation.

The US perturbation method constitutes an efficient tool for free energy based studies. For sufficiently small perturbations it is generally applicable to any type of US simulations and yields rigorously derived free energies.

5.2 Methods

Reweighted probability distributions in the WHAM equations

The WHAM equations[14, 15] can be used to calculate a free energy change (PMF) along a specific coordinate ξ based on several umbrella sampling (US)[16] simulations. For a set of N_w US windows i with biasing potentials $w_i(\xi)$, they can be written as[17]

$$\langle \rho(\xi) \rangle = \frac{\sum_{i=1}^{N_w} n_i \langle \rho(\xi) \rangle_i}{\sum_{j=1}^{N_w} n_j e^{-\left[w_j(\xi) - F_j\right]/k_B T}} \tag{5.1}$$

$$e^{-F_i/k_B T} = \int d\xi e^{-w_i(\xi)/k_B T} \langle \rho(\xi) \rangle. \tag{5.2}$$

The unbiased probability distribution $\langle \rho(\xi) \rangle$, and hence the related changes in free energy, is expressed as a weighted sum over the biased probability distributions $\langle \rho(\xi) \rangle_i$ obtained by each US simulation. Here, F_i is the free energy associated with each umbrella window and n_i denotes the total number of sampling points obtained for each window. The equations can be solved self-consistently using an iteration procedure. As in the following reweighting approach non-discrete reweighted probability distributions are used, the integral over the probability distribution is employed for normalization instead of the total number of sampling points: $n_i \hat{=} \int d\xi \langle \rho(\xi) \rangle_i$.

We now consider a small perturbation ΔH. The probability distributions in Eq. 5.1 and n_i for a system with Hamiltonian $H + \Delta H$ and umbrella potential $w_i(\xi)$ can be written as

$$\langle \rho(\xi) \rangle_{H+\Delta H,i} = \frac{\int dR \delta[\xi'(R) - \xi] e^{-[H(R)+\Delta H(R)+w_i(\xi'(R))]/k_B T}}{\int dR e^{-[H(R)+\Delta H(R)+w_i(\xi'(R))]/k_B T}}, \tag{5.3}$$

with δ denoting the delta-distribution and $\int dR$ indicating a phase space integral over all degrees of freedom of the system. This Boltzmann weighted average is usually obtained by numerical sampling methods, as for instance MD simulations. Similarly to standard free energy perturbation[12] theory, the probability distributions can be written as reweighted Boltzmann averages corresponding to Hamiltonian H:

$$\begin{aligned} \langle \rho(\xi) \rangle_{H+\Delta H,i} &= \frac{\int dR \delta[\xi'(R) - \xi] e^{-[H(R)+w_i(\xi'(R))]/k_B T} e^{-\Delta H(R)/k_B T}}{\int dR e^{-[H(R)+w_i(\xi'(R))]/k_B T}} \\ &\times \frac{\int dR e^{-[H(R)+w_i(\xi'(R))]/k_B T}}{\int dR e^{-[H(R)+\Delta H(R)+w_i(\xi'(R))]/k_B T}} \\ &= \frac{\langle \rho(\xi) e^{-\Delta H/k_B T} \rangle_{H,i}}{\langle e^{-\Delta H(R)/k_B T} \rangle_{H,i}} \end{aligned} \tag{5.4}$$

With sampling data existent for H, the ΔH can be computed for every sampling point from the stored trajectories. Thus, the probability distributions $\langle \rho(\xi) \rangle_{H+\Delta H,i}$ are obtained without additional simulations. In practice, for a limited extent of sampling, the reweighted distributions will be a good estimate in the case of sufficient phase space overlap, implying that the perturbation ΔH of the system is small.

MD-Simulations

Water test systems. Two water molecules (W1 and W2) were set up manually and solvated either by standard TIP3P water molecules or using the implicit GBn2[18] solvent model via Leap (Amber12[19]). Three different parameterization sets were used

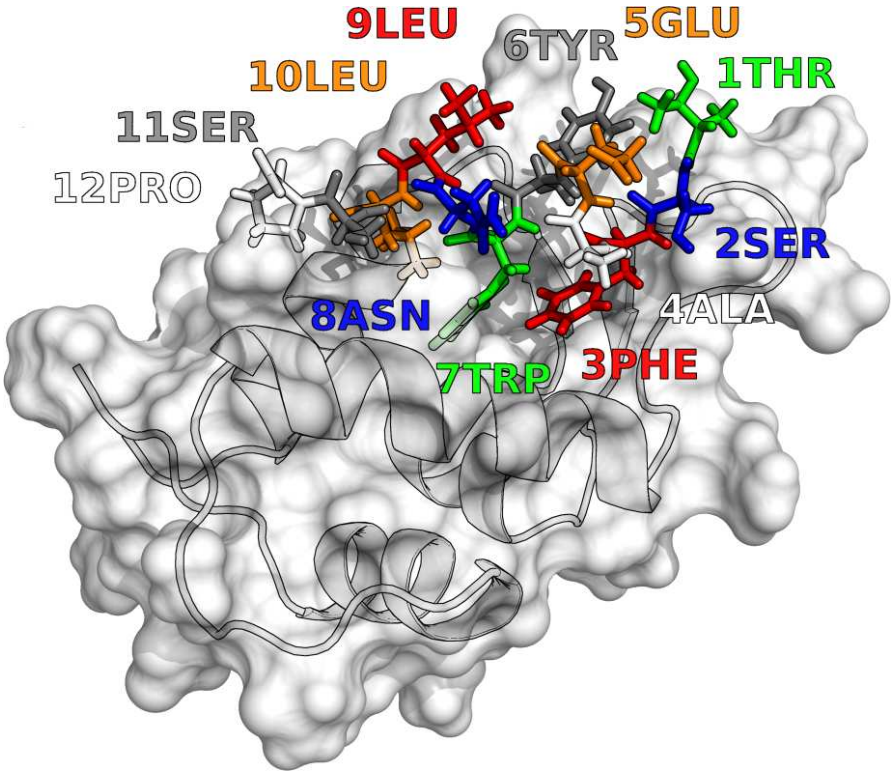


Figure 5.1: Equilibrated configuration of the PMI ligand (sticks) in complex with the MDM2 domain (surface, cartoon). For the direct evaluation of the PMF reweighting estimates, reference PMFs were calculated with standard US simulations for alanine substitutions of residues 3PHE, 7TRP and 10LEU which are deeply buried in the MDM2 binding pocket and for residue 8ASN located at the rim of the interface.

for W1 and W2, resulting in a total of six test systems: the standard TIP3P[20] water model (Top1), the TIP3P model with the charge of the oxygen atom of W1 changed by +0.3 e and the charge of the oxygen atom of W2 changed by -0.3 e (Top2), and the TIP3P model with the Lennard-Jones interactions between W1 and W2 multiplied by a factor of two (Top3). The periodic box size for the explicit TIP3P simulations was $\approx (23\,\text{Å})^3$, containing 1213 water molecules in total.

MDM2-PMI complex. The crystallographic structure (3EQS.pdb) of the investigated PMI-MDM2 complex was taken from Pazgier et al.[13]. The 12PRO residue of the PMI ligand missing in the structure was added manually. Residues were mutated to alanine by removing the corresponding side chains from the backbone and constructing the

alanine side chain by Leap (AMBER12[19]). The ff12SB force field [21] was used for the parametrization of ligand and receptor. For the solvent description, the implicit GBn2 [18] model was employed, in combination with the solvent accessible surface area (SASA) approximation [22, 23] accounting for nonpolar contributions.
Simulation parameters. All simulations were performed using AMBER12[19]. Explicit solvent simulations were carried out in the NVT ensemble using a weak temperature coupling algorithm[24] at 298 K. The Particle Mesh Ewald (PME) method was used to calculate the electrostatic interactions. Implicit solvent simulations were carried out with a Langevin thermostat [25] at 298K, using a collision frequency of 10 ps^{-1} and the AMBER12 default GB/SA model parameters. The integration time step was 2 fs. Bonds involving hydrogen atoms were restrained with the SHAKE[26] algorithm. H-REMD exchanges were attempted between adjacent US windows every 2 ps.
PMF sampling: Water test systems. Hamiltonian Replica Exchange Umbrella Sampling simulations (H-REMD-US)[5, 16, 27, 28] were performed along the center of mass separation of molecules W1 and W2 for all systems. To cover the range between 2.0 Å and 7.0 Å, 6 windows with an intermediate spacing of 1.0 Å were used. The harmonic potential force constant was 1 kcal/mol/$Å^2$. The configuration resulting from the equilibration was used as starting configuration for all US windows. The total sampling time was 10 ns per window, sampling data was recorded every 0.1 ps after 1 ns of further equilibration of each US simulation.
PMF sampling: MDM2-PMI complex. H-REMD-US simulations along a ligand-receptor separation coordinate were performed for the original PMI ligand and for four PMI ligands with the 3PHE, 7TRP, 8ASN or 10LEU residue mutated to alanine. During the simulations, restraints on the relative orientation and position of receptor and ligand were applied, as well as on the backbone conformation of the ligand with respect to its bound configuration. By the restraints, the effective accessible phase space during the separation PMF calculation is significantly reduced in order to enhance convergence of the computationally demanding sampling. We thereby follow an approach originally introduced by Woo and Roux [3].

A total of 24 umbrella windows were set up between 12.0 Å and 27.0 Å. The range between 12.0 Å and 19.5 Å was covered by 16 windows with an intermediate spacing of 0.5 Å and a harmonic force constant of 16 kcal/mol/$Å^2$, whereas the range between 20.0 Å and 27.0 Å was covered by 8 windows with an intermediate spacing of 1.0 Å and a harmonic force constant of 8 kcal/mol/$Å^2$. Starting configurations for the umbrella simulations were generated based on the equilibrated bound configuration by consecutively applying the corresponding umbrella potentials for 0.1 ns. Total sampling time was 32 ns per window. In order to assure equilibration of the simulations, the first 8 ns of sampling were omitted for the PMF calculation. Sampling data was stored every 0.1 ps.

PMF calculation. The PMFs were obtained by solving the WHAM equations for the original or reweighted US probability distributions, using the code from Grossfield [29].
PMF error estimates. The uncertainties of the PMFs were estimated by taking into account the largest deviations of subset PMFs obtained by dividing the sampling data into four successive subgroups.

In silico alanine scanning

Potential energies corresponding to the modified Hamiltonians were calculated by reprocessing the original US trajectories with altered topologies. For the water test systems, US simulations for three topologies were performed (see MD simulations section). As in this case, the modifications only consisted of the change of topology parameters, the reweighting scheme could be directly applied to each of the original trajectories using the two different topologies for a mutual comparison of the results.

For the PMI mutations (compare Alanine Scan [30]), the respective residue was replaced by an alanine residue in the topologies. Additional hydrogen atoms introduced by the alanine residue were assigned to the positions of closest former carbon or oxygen atoms belonging to the replaced residue. The new C-H bond lengths were set to the corresponding C-H equilibrium distances by the SHAKE algorithm used during the energy recalculation. Redundant side chain atoms were removed from the trajectories. Possible angular and dihedral reweighting terms involving the C-H bond were neglected.

As single sampling points are reweighted by a corresponding Boltzmann factor, some frames corresponding to very low energy differences can be extremely overweighted. This can be due to the relatively gross mutation protocol or occasional unfavorable configurations of the non-mutated side chains. In order to avoid noise stemming from such sampling points, a threshold was introduced by excluding points with potential energies lower than 4 kcal/mol below $-k_B T \ln\langle e^{-\Delta H(R)/k_B T}\rangle_{H,i}$ for each US window. This lead to an exclusion of less than 0.01 % of the sampling points.

5.3 Results and Discussion

The change of a PMF upon perturbation of a system can be calculated by a probability distribution reweighting scheme that only requires post-processing of previously stored US trajectories instead of additional simulations. This is achieved by calculating corresponding potential energy differences for the sampled configurations in order to obtain new probability distributions, according to Eq. 5.4 (Methods section). The modified probability distributions are then inserted into the WHAM equations (Eq. 5.1 and 5.2) in order to estimate the changes of the PMFs. As it is sufficient to evaluate the

modified potentials for every sampling frame instead of calculating the total potential energies and forces for every MD step, the perturbation method comes with significant computational savings in comparison to additional MD simulations.

The approach was first tested on PMFs that correspond to the separation of two selected water molecules in an aqueous environment. H-REMD-US simulations were carried out with either standard TIP3P parameters (Top1) or two different parameter modifications (Top2: change in charge, Top3: change of LJ parameters, see Methods) for the two selected water molecules. Here, the modifications of the Hamiltonians only consisted of simple parameter changes. The perturbed PMFs corresponding to the two other topologies were estimated using the reweighting method. The calculations were performed using two different solvation models: Solvation with explicit standard TIP3P molecules and with an implicit GBn2 model, which implies a reduced number of degrees of freedom. The directly sampled PMFs and the PMFs obtained by reweighting the original probability distributions for all three topologies are shown in Fig. 5.2. The PMFs for the corresponding topologies are in excellent agreement as long as no long ranged changes of the electric field (Top2) in combination with explicit solvent water molecules are involved. Note, that in the Top2 case the two water molecules W1 and W2 are no longer neutral but have total charges of the same magnitude with opposite sign. This corresponds to the generation of a large dipole that increases with the distance between W1 and W2. Presumably, the orientational polarization of the surrounding explicit water molecules stored in the original trajectories does not align very well with the electrostatic field created by the perturbation. In contrast, phase space overlap is significantly enhanced if an implicit solvent model is used, in which the degrees of freedom of the solvent are treated within a continuum approximation (with an instantaneous solvent response in each evaluated trajectory frame). For the test systems, we conclude that the method works very well even for significant free energy changes in the close region ($2.5\,\text{Å} < r < 3.5\,\text{Å}$), but application to perturbations including explicit water molecules and at the same time long ranged electrostatic changes can be problematic.

The approach was also employed for a computational alanine-scanning investigation of the PMI-MDM2 complex [13], illustrated in Figure 1. A biased separation PMF between peptide and receptor was calculated by means of H-REMD-US simulations (see Methods) for the original PMI ligand, using an implicit GBn2/SA solvation model. The resulting separation free energy contribution corresponds to the largest binding free energy contribution in a framework originally introduced by Roux and Woo[3]. The separation PMFs for PMI peptides with single residues mutated to alanine were then calculated using the reweighting scheme based on the original PMI peptide trajectories. To enable a direct evaluation of the perturbation estimates, the PMFs for four specific alanine substitutions were also directly calculated via H-REMD-US simulations. These

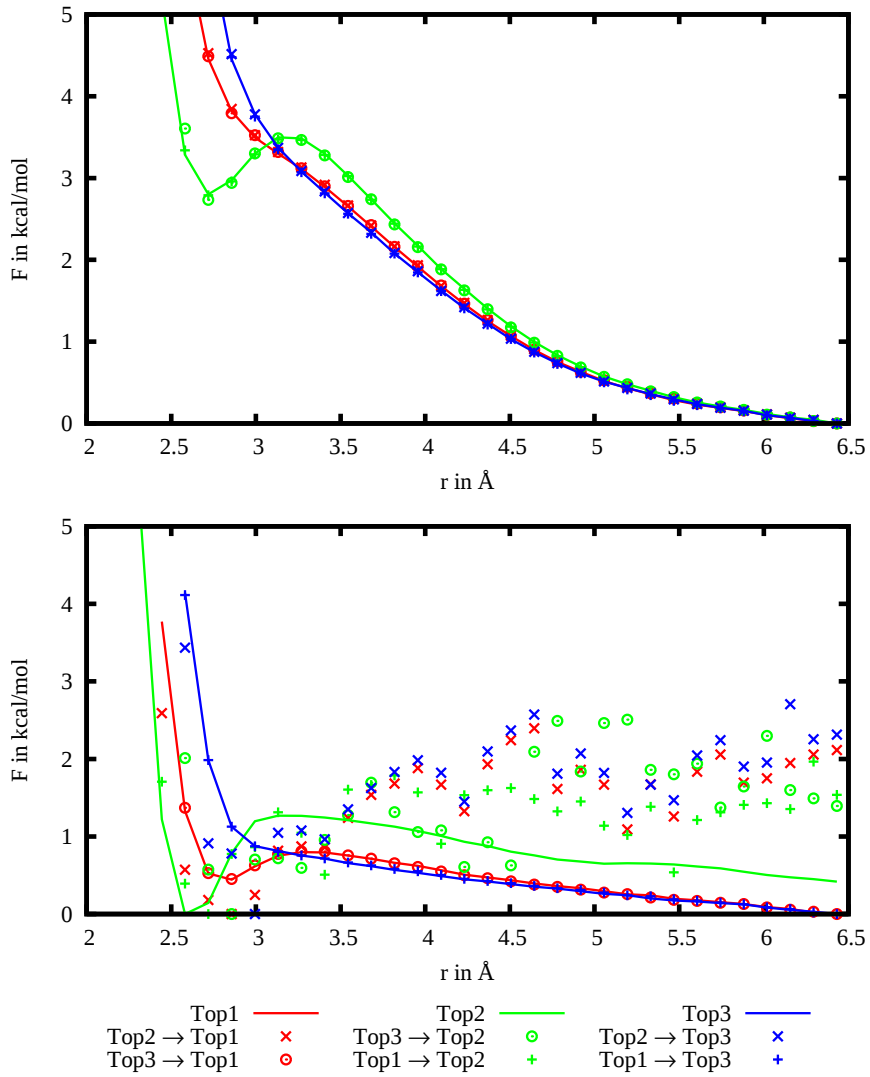


Figure 5.2: Water separation PMFs for three different topologies (indicated by different colors), simulated using the implicit GBn2 solvation model (top) and in the presence of explicit solvent TIP3P molecules (bottom). PMFs obtained by standard US are plotted with lines. PMFs obtained using the reweighting approach based on the trajectories generated with different topologies are plotted with corresponding symbols (original sampling topology → post-processing topology).

four specific mutations were selected because they are known to come with significant changes in free energy (3PHE, 7TRP, 10LEU) or because it was suggested that the change in free energy is largely due to conformational contributions (8ASN), which might be challenging for a perturbation approach [31].

The directly sampled PMFs and the reweighing estimates for the four specific mutations are plotted in Fig 5.3. The reweighted PMFs deviate between 0.6 kcal/mol and 2.0 kcal/mol from the reference PMFs and agree within the sampling uncertainties. Figure 5.4 shows the backbone RMSD of trajectories generated in the US windows with respect to the wild type PMI equilibrated bound structure. For the US windows located between about 14 and 19 Å the backbone RMSD of the 3PHE ligand differs significantly from the wild type RMSD, which is also where the corresponding reweighted PMF shows the largest fluctuations. Different dynamics in the direct simulations that are not precisely covered by the trajectories obtained with the original PMI ligand are likely to be a main source of errors. Furthermore, a general tendency towards underestimation of the binding free energies of the modified ligands can be observed. Presumably, this is due to a possible favorable adaptation of the modified ligands to the binding pocket that stabilizes the bound structure and is not captured within the reweighting approach. Also, the increase in separation free energy upon the 8ASN mutation (due to stabilization of the bound configuration of the ligand) is not reflected by the reweighted PMF. In this case however, the statistical uncertainties do not allow definite conclusions. Except for the above discussed limitations of the reweighting approach, the impact on the calculated binding free energy caused by the removal of the complete tryptophan or phenylalanine side chains is quite well reproduced. Although not explicitly tested, it was assumed that in order to achieve sufficient phase space overlap, an implicit solvation model has to be used. The removal of a large side chain in the re-evaluation of the original trajectories would result in empty (vacuum) regions if explicit water molecules were included. In the case of an implicit solvent model, the resulting empty space is assigned a high dielectric solvent-like constant in the perturbation calculation.

The differences in separation free energy between the minimum and $r = 26$ Å were compared to experimentally observed changes in binding free energy (Table 5.1). The values agree within less than 2.0 kcal/mol. The ranking of the alanine scan mutations is correctly reproduced for the four residues contributing the most. Interpretation of values for mutations that lead to relative changes of about 1 kcal/mol or less is unreliable because this is in the order of the statistical uncertainty of the original PMF calculation. Note that comparing the results with the experiment, the force field, solvation model or eventual error cancellations may also play a role. Also, for simplicity, additional contributions to the complete standard binding free energy were neglected, assuming that these contributions to relative free energies of binding are small. For example, this concerns possible free energy differences in forming the

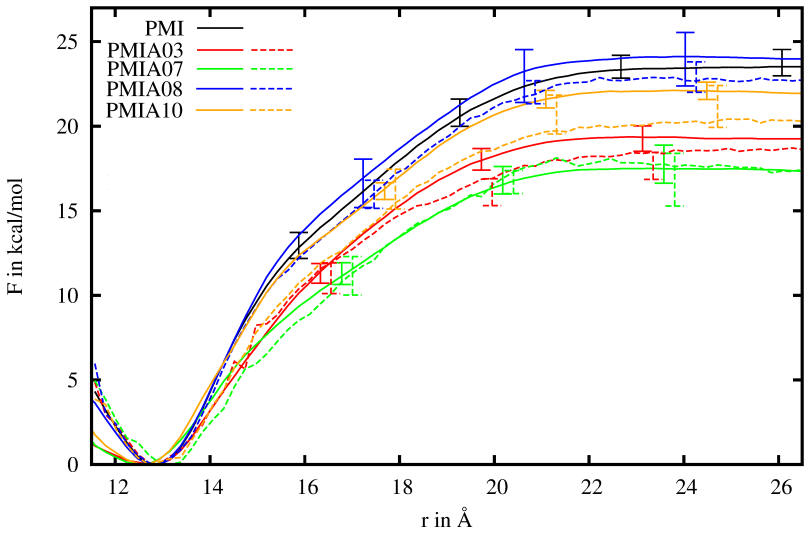


Figure 5.3: Directly sampled separation PMFs for the PMI ligand and four specific mutations (solid lines), in comparison with the PMFs obtained by the reweighting scheme based on the original PMI ligand sampling (dashed lines).

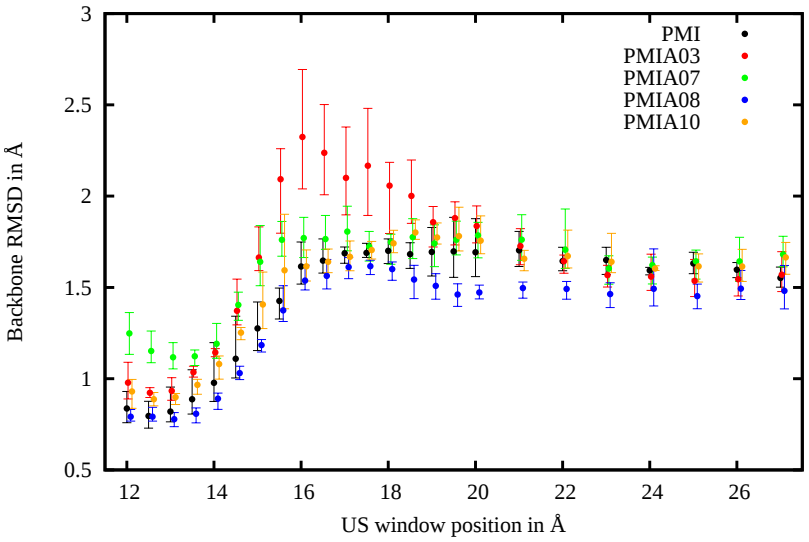


Figure 5.4: Ligand backbone RMSD with respect to a bound configuration of the original PMI ligand, sampled in the respective US windows.

Residue		ΔΔF US perturb.	ΔΔF exp.[31]	ΔΔF Amber Scan[30, 32]	ΔΔF Mult. traj. MM/GBSA[32]
1	THR	0.22	0.39	-0.17	2.00
2	SER	0.12	1.24	0.19	0.25
3	PHE	4.88	5.46	7.21	4.42
4	ALA	-	-	-	-
5	GLU	0.52	1.10	0.23	0.60
6	TYR	1.12	3.06	4.17	4.56
7	TRP	6.19	6.31	11.19	7.07
8	ASN	0.82	-1.10	-0.27	-0.13
9	LEU	-0.51	-0.17	0.41	1.31
10	LEU	3.18	3.28	4.78	3.14
11	SER	-0.12	0.12	-0.31	0.93
12	PRO	-	-0.25	-	0.50

Table 5.1: Relative changes of the binding affinity upon alanine mutations (in kcal/mol). The ΔΔF computed in this study are only the changes in separation free energy. MM/GBSA based alanine-scan results are given for comparison.

alpha-helical peptide bound structure for each mutation. In principle, all additional relative contributions to binding could be obtained by PMF calculations and employing the reweighting approach. Overall, the results indicate that the reweighting approach is able to provide valuable results for alanine scanning, as long as the absolute PMF accuracy is sufficient to cover the changes in free energy due to the mutations. For comparison, results of MM/GBSA based approaches (Massova and Kollman [30], Liu et al.[32]) on the same system are given in Table 5.1. The results of the reweighting scheme are overall in slightly better agreement with the experimental values compared to the MM/GBSA method.

Most importantly, the method yields rigorously derived free energies. It should be emphasized that, as in the separation PMF approach only the relevant relative solvation free energies of binding are calculated, the corresponding errors are relatively small. In the MM/PBSA or MM/GBSA methods, binding energies are obtained by the subtraction of high average absolute energies with large associated fluctuations.

For all PMF calculations, a sampling interval of 0.1 ps was used, which is 50 times larger than the MD time step. For the post-processing, only potential energies and no forces have to be calculated. This in principle implies a significantly reduction of the computational effort by a factor of about 1/100 in comparison to a full PMF calculation. In practice, the computational cost was about 1/30 using the imin=5

routine of AMBER12, presumably because the handling of large trajectories implies some additional cost. Furthermore, considering that for the reweighting only the changed interactions involving the few atoms affected by the perturbation have to be recalculated rather than the full new potential energy (as done with the imin=5 routine in AMBER12), there is still much room for improvement of the computational efficiency by modifying the corresponding computational routines.

5.4 Conclusion

A reweighting method is presented that can be used to estimate the changes of PMFs upon small perturbations of a system by post-processing existing trajectories, avoiding the need to perform new computationally costly MD simulations. In a first evaluation on a simple system with a low number of degrees of freedom, excellent free energy predictions were obtained, as long as no long ranged electrostatic changes in combination with explicit solvent molecules were included. The estimation of the MDM2-PMI complex separation free energy upon the mutation of individual ligand residues to alanine constitutes a relevant application. Good estimates for the separation PMFs were achieved even upon significant modifications, agreeing with direct PMF sampling within between 0.6 and 2.0 kcal/mol. The experimentally determined ranking of the binding affinity by changes in the separation free energy contributions could mostly be reproduced. In future studies it would be interesting to compare the US perturbation approach with one-step alchemical perturbation methods that have also been used for in silico alanine scanning [9]. In contrast to other free energy estimation methods as MM/GBSA or LIE, the present perturbation method yields rigorously derived changes in free energy, under the condition of sufficient phase space overlap. We expect that with an optimization of the routines further computational savings can be obtained. As the method is generally applicable to any PMF undergoing small perturbations, it might prove an interesting instrument especially for the investigation of atomistic mechanisms and molecular design studies. For example, for the study of protein-protein or protein-DNA interactions it is possible to perform a complete in silico alanine scanning of all interface residues based on a single PMF calculation for the original (wild type) complex. This offers the possibility to systematically investigate free energy contributions to biomolecular complex formation and could also help to design specific interactions.

5.5 Bibliography

[1] G. M. Torrie and J. P. Valleau. "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling." In: *J. Comput. Phys.* 23.2 (1977), pp. 187–199.

[2] William L. Jorgensen. "Interactions between amides in solution and the thermodynamics of weak binding." In: *J. Am. Chem. Soc.* 111.10 (1989), pp. 3770–3771.

[3] Hyung-June Woo and Benoît Roux. "Calculation of Absolute Protein–Ligand Binding Free Energy from Computer Simulations." In: *Proc. Natl. Acad. Sci. USA* 102.19 (2005), pp. 6825–6830.

[4] James C. Gumbart, Benoît Roux, and Christophe Chipot. "Efficient Determination of Protein–Protein Standard Binding Free Energies from First Principles." In: *J. Chem. Theory Comput.* 9.8 (2013), pp. 3789–3798.

[5] Fabian Zeller and Martin Zacharias. "Adaptive Biasing Combined with Hamiltonian Replica Exchange to Improve Umbrella Sampling Free Energy Simulations." In: *J. Chem. Theory Comput.* 10.2 (2014), pp. 703–710.

[6] M.K. Gilson, J.A. Given, B.L. Bush, and J.A. McCammon. "The Statistical-Thermodynamic Basis for Computation of Binding Affinities: A Critical Review." In: *Biophys. J.* 72.3 (1997), pp. 1047–1069.

[7] Gerhard König, Stefan Bruckner, and Stefan Boresch. "Unorthodox uses of Bennett's acceptance ratio method." In: *J. Comp. Chem.* 30.11 (2009), pp. 1712–1718.

[8] Haiyan Liu, Alan E. Mark, and Wilfred F. van Gunsteren. "Estimating the Relative Free Energy of Different Molecular States with Respect to a Single Reference State." In: *J. Phys. Chem.* 100.22 (1996), pp. 9485–9494.

[9] Zhixiong Lin, Jörgen Kornfeld, Markus Mächler, and Wilfred F. van Gunsteren. "Prediction of Folding Equilibria of Differently Substituted Peptides Using One-Step Perturbation." In: *J. Am. Chem. Soc.* 132.21 (2010), pp. 7276–7278.

[10] Jayashree Srinivasan, Thomas E. Cheatham, Piotr Cieplak, Peter A. Kollman, and David A. Case. "Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate-DNA Helices." In: *J. Am. Chem. Soc.* 120.37 (1998), pp. 9401–9409.

[11] Johan Åqvist, Carmen Medina, and Jan-Erik Samuelsson. "A New Method for Predicting Binding Affinity in Computer-Aided Drug Design." In: *Protein Eng. Des. Sel.* 7.3 (1994), pp. 385–391.

[12] Robert W. Zwanzig. "High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases." In: *J. Chem. Phys.* 22.8 (1954), pp. 1420–1426.

[13] Marzena Pazgier et al. "Structural basis for high-affinity peptide inhibition of p53 interactions with MDM2 and MDMX." In: *Proc. Natl. Acad. Sci. USA* 106.12 (2009), pp. 4665–4670.

[14] Alan M. Ferrenberg and Robert H. Swendsen. "Optimized Monte Carlo Data Analysis." In: *Phys. Rev. Lett.* 63 (12 Sept. 1989), pp. 1195–1198.

[15] Shankar Kumar, John M. Rosenberg, Djamal Bouzida, Robert H. Swendsen, and Peter A. Kollman. "The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method." In: *J. Comput. Chem.* 13.8 (1992), pp. 1011–1021.

[16] William L. Jorgensen. "Interactions Between Amides in Solution and the Thermodynamics of Weak Binding." In: *J. Am. Chem. Soc.* 111.10 (1989), pp. 3770–3771.

[17] Benoît Roux. "The Calculation of the Potential of Mean Force Using Computer Simulations." In: *Comput. Phys. Commun.* 91.1–3 (1995), pp. 275–282.

[18] Hai Nguyen, Daniel R. Roe, and Carlos Simmerling. "Improved Generalized Born Solvent Model Parameters for Protein Simulations." In: *J. Chem. Theory Comput.* 9.4 (2013), pp. 2020–2034.

[19] D. A. Case et al. *AMBER12*. University of California, San Francisco, CA, 2012.

[20] William L. Jorgensen, Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. "Comparison of Simple Potential Functions for Simulating Liquid Water." In: *J. Chem. Phys.* 79.2 (1983), pp. 926–935.

[21] Wendy D. Cornell et al. "A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules." In: *J. Am. Chem. Soc.* 117.19 (1995), pp. 5179–5197.

[22] Doree Sitkoff, Kim A. Sharp, and Barry Honig. "Accurate Calculation of Hydration Free Energies Using Macroscopic Solvent Models." In: *J. Phys. Chem.* 98.7 (1994), pp. 1978–1988.

[23] Jörg Weiser, Peter S. Shenkin, and W. Clark Still. "Approximate Atomic Surfaces from Linear Combinations of Pairwise Overlaps (LCPO)." In: *J. Comput. Chem.* 20.2 (1999), pp. 217–230.

[24] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. "Molecular Dynamics with Coupling to an External Bath." In: *J. Chem. Phys.* 81.8 (1984), pp. 3684–3690.

[25] Richard W. Pastor, Bernard R. Brooks, and Attila Szabo. "An Analysis of the Accuracy of Langevin and Molecular Dynamics Algorithms." In: *Mol. Phys.* 65.6 (1988), pp. 1409–1419.

[26] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman J.C Berendsen. "Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes." In: *J. Comput. Phys.* 23.3 (1977), pp. 327–341.

[27] Yuji Sugita, Akio Kitao, and Yuko Okamoto. "Multidimensional Replica-Exchange Method for Free-Energy Calculations." In: *J. Chem. Phys.* 113.15 (2000), pp. 6042–6051.

[28] James C. Gumbart, Benoît Roux, and Christophe Chipot. "Standard Binding Free Energies from Computer Simulations: What is the Best Strategy?" In: *J. Chem. Theory Comput.* 9.1 (2013), pp. 794–802.

[29] Alan Grossfield. *WHAM: The weighted histogram analysis method 2.0.7*. Grossfield Lab: Rochester, NY, 2013.

[30] Irina Massova and Peter A. Kollman. "Computational Alanine Scanning To Probe Protein-Protein Interactions: A Novel Approach To Evaluate Binding Free Energies." In: *J. Am. Chem. Soc.* 121.36 (1999), pp. 8133–8143.

[31] Chong Li et al. "Systematic Mutational Analysis of Peptide Inhibition of the p53–MDM2/MDMX Interactions." In: *J. Mol. Biol.* 398.2 (2010), pp. 200–213.

[32] Yun Liu, DavidP. Lane, and ChandraS. Verma. "Systematic Mutational Analysis of an Ubiquitin Ligase (MDM2)-Binding Peptide: Computational Studies." In: *Theor. Chem. Acc.* 130.4-6 (2011), pp. 1145–1154.

6 Substrate Binding Specifically Modulates Domain Arrangements in Adenylate Kinase[1]

The enzyme Adenylate Kinase (ADK) features two substrate binding domains which undergo large-scale motions during catalysis. In the apo state, the enzyme preferentially adopts a globally open state with accessible binding sites. Binding of two substrate molecules (AMP + ATP or ADP + ADP) results in a closed domain conformation, allowing efficient phosphoryl-transfer catalysis. We employed Molecular Dynamics (MD) simulations to systematically investigate how the individual domain motions are modulated by the binding of substrates. Two-dimensional free energy landscapes were calculated along the opening of the two flexible lid domains for apo and holo ADK as well as for all single natural substrates bound to one of the two binding sites of ADK. The simulations reveal a strong dependence of the conformational ensembles on type and binding position of the bound substrates and a non-symmetric behavior of the lid domains. Altogether, the ensembles suggest that upon initial substrate binding to the corresponding lid site the opposing lid is maintained open and accessible for subsequent substrate binding. In contrast, ATP binding to the AMP-lid induces global domain closing, preventing further substrate binding to the ATP-lid site. This might constitute a mechanism by which the enzyme avoids the formation of a stable but enzymatically unproductive state.

6.1 Introduction

Enzymes have developed elaborate strategies in order to efficiently catalyze reactions fundamental for cell function. As such, large scale domain motions coupled to substrate binding are essential for the activity of a large number of enzymes[1]. A prominent example is Adenylate Kinase (ADK), catalyzing the reversible interconversion of ATP-Mg^{2+} + AMP and ADP-Mg^{2+} + ADP. The enzyme is involved in the regulation of energy homeostasis in bacteria and eukaryotic cells [2][3].

[1]This chapter has been previously published in similar form in Biophysical Journal, 109(9), 1978-1985, 2015.

ADK possesses two binding cavities which are specifically occupied by AMP and ATP, or alternatively by two ADPs, in the functional states. The cavities are each formed by a large flexible domain, termed AMP-lid and ATP-lid, in combination with the protein core. Several crystal structures of ADK have been determined, showing that these flexible domains can adopt an open conformation in the absence of substrates[4][5] and a closed conformation when substrates or inhibitors are bound [6][7][8] (Fig. 6.1). In the closed conformation, active site residues and substrates are shielded from the aqueous environment and arranged in a configuration appropriate for the chemical reaction[9]. The importance of the global domain rearrangements in the functional cycle has been emphasized by a study identifying the domain opening as the rate limiting step for catalysis [10]. However, despite being crucial for the biological activity of the enzyme the mechanistic coupling between binding of different substrates and large-scale domain motion has not been fully understood.

Nuclear magnetic resonance (NMR) spectroscopy and fluorescence energy transfer (FRET) experiments indicate a high degree of conformational flexibility of ADK[11][5][12]. Closing or partial closing of lids upon binding of the respective substrates has been observed[12][13][14]. Evidence for different possible conformational substates is also given by crystallography. For example, a partially closed structure of an inactive variant of ADK in complex with an ATP analog has been determined that shows a closed ATP-lid and a fully opened AMP-lid[15]. The question arises how in detail the domain motions and possible substates are modulated or induced by the binding of substrates in order to achieve efficient catalysis. As the substrates can potentially bind to different binding sites and to a broad ensemble of conformational states, accessibility of the pathways is substantially complicated.

In addition to the experiments, the dynamics of ADK and the possible influence of different substrates have been addressed in computational studies. Pathways between different conformations of apo ADK have been investigated making use of network models[16] or other non-atomistic models[17][18]. Several atomistic continuous Molecular Dynamics (MD) simulations have been performed[19][20][21][22]. However, such simulations are usually limited to timescales on which the domain rearrangements are rare events. Alternatively, coarse-grained models allow more extensive sampling [18], but the accuracy of such models and the implicit treatment of the solvent may not be sufficient to analyze the influence of different substrates on the dynamics of ADK. More recently, progress has been made by employing advanced sampling methods in atomistic MD simulations on ADK [23][24][25][26][27][28]. Significant contributions have emerged from these studies, but as they focused on ADK in the apo state or bound to an inhibitor, the dynamics in the presence of natural substrates have not been covered.

In this study, we employ atomistic MD simulations controlling substrate type and

binding location in order to systematically expand the picture that has been outlined by experimental and previous simulation studies. Sampling of the domain configurations is achieved via Umbrella Sampling simulations in combination with Hamiltonian replica exchanges. Using these sampling techniques, 2D free energy landscapes along the separate domain opening motions were calculated for apo ADK and nine different cases of bound natural substrates. These include also non-reactive states with substrates bound to the non-designated lid domain. Each 2D free energy landscape is based on extensive sampling of more than 0.75 µs. The free energy estimates from these MD simulations contribute insight into how evolution might have optimized the protein for its specific task, as they allow us to predict dominant features of binding and domain movement as well as to propose a mechanism that possibly avoids unproductive substrate bound states.

6.2 Methods

Force Field

The Amber ff14SB force field[29] was used for the protein description. The TIP3P model[30] was used for the water molecules. The GAFF force field[31] in combination with Antechamber[32] was used to describe the AMP molecule. Specific force field parameters were employed for ADP and ATP[33], the magnesium ion [34] and the potassium ions[35].

Starting Structures

A total of ten starting structures were prepared for E.Coli ADK: apo ADK, ADK in complex with two ADPs and Mg, and ADK with only one of the two binding sites occupied by AMP, ADP, ADP+Mg or ATP+Mg. The closed ADK configuration was taken from a crystal structure of ADK in complex with the inhibitor AP5A (PDB:1AKE[6]). For the simulations on apo ADK, the inhibitor was removed from the structure. In order to construct starting structures for the closed E. Coli ADK in complex with natural substrates, the configuration of two ADP molecules and a magnesium ion bound to mycobacterium tuberculosis ADK was taken from PDB:2CDN[36]. The AP5A inhibitor was replaced by this substrate configuration by least RMSD fitting[37] the positions of the adenosine and adjacent phosphate groups of the ADP molecules onto the corresponding atom groups of AP5A (Fig. 6.1). Redundant atoms were removed from the 2ADP+Mg complex in the cases of AMP, ADP and ADP+Mg, and an additional phosphate was added to ADP in the cases of ATP+Mg. The resulting ten complexes were solvated with $\approx$ 8500 water molecules in a periodic octahedral box with edge

length of $\approx$ 28 Å. The overall charge was neutralized by adding potassium ions. The starting structures were energy minimized, heated up and equilibrated for a total of 2 ns in the NTP ensemble, using Amber14[32].

2D H-REMD-US simulations

All sampling simulations were performed in the NVT ensemble, using Amber14. The opening coordinates were defined by the center of mass distances between the C_α atoms of residues 112-121,160-175 and 33-58 (AMP-lid), and of residues 1-28 and 125-152 (ATP-lid). Umbrella windows were set up with an intermediate spacing of 1 Å in both dimensions, between 20 Å and 31 Å for the AMP-lid opening and between 17 Å and 28 Å for ATP-lid opening distance. Based on the equilibrated structures, starting structures for the different umbrella windows were generated by successively applying the umbrella potentials with a force of 4 kcal/mol/Å^2 for 100 ps, first along the AMP-lid coordinate, then along the ATP-lid coordinate. A second set of starting structures was generated by repeating the process in the reverse order, starting from the previously obtained completely open structure. Positional restraints with a force of 0.025 kcal/mol/Å^2 were applied to the substrates during the generation of the initial umbrella window configurations in order to prevent dissociation due to the enforced fast domain motions. Sampling of the configurational space of two bound ligands in the case of 2ADP+Mg bound is considerably more demanding than sampling of only single bound substrates. In order to also incorporate states into the sampling in which the 2ADP+Mg complex is broken apart (a process which does not happen during affordable simulation times), before the generation of the second set of starting structures the ADP molecule bound to the AMP-lid site was driven to the open AMP-lid in the open ADK conformation. This was accomplished by applying an RMSD restraint of 1 kcal/mol/Å^2 on AMP-lid and the corresponding ADP molecule with respect to an open configuration from simulations of only ADP bound to the open AMP-lid for additional 20 ps. In the simulations of only ADP bound to the AMP-lid, this configuration occurred naturally. Finally, two strands of 2D H-REMD-US[38][39] simulations were performed in parallel, starting from the structures generated based on the closed and on the open state. During the sampling, no restraints besides the umbrella potentials were applied. Neighboring umbrella windows were allowed to exchange configurations every 1 ps according to the metropolis criterion alternately within the strands and between the two strands. In this way, all umbrella simulations at the different positions and initiated from different structures were connected. The regions in the reaction coordinate plane corresponding to one closed and one open lid were omitted in the H-REMD-US runs in order to obtain a computationally advantageous number of 256 umbrella simulations, focusing sampling on the region of interest. During the sampling

runs, the umbrella potential force was 2 kcal/mol/Å^2. Each window was sampled for 3 ns, corresponding to 6 ns of sampling for each umbrella window position, summing up to more than 0.75 μs of simulation time for each of the ten ligation states. During the US simulations, the average internal RMSDs of the atom groups that defined the center-of-mass distances with respect to the starting structures were lower than 1.6 Å. Only at very open configurations, the AMP-Lid domain atom group reached a maximum internal RMSD value of 4.0 Å. This indicates that the center-of-mass distances are dominated by global domain arrangements and not by internal deformations. The rates of successful exchanges between the individual windows varied between 0.1 and 0.5. For the final PMF calculation via WHAM[40][41], the sampling data of the first of the 3 ns was skipped as further equilibration. Convergence of the sampling was evaluated by calculating the PMFs for several subsets of the total sampling data. For the substrate RMSD calculations[32], the structures from all US trajectories were superimposed onto the starting structures with respect to the positions of all protein atoms (least RMSD fit). These structures were used to calculate the RMSD of all substrate atoms with respect to the starting structures. Figures 1-3 show the mean substrate RMSD values for 50×50 2D bins within the reaction coordinate plane.

6.3 Results

2D Potentials of Mean Force (PMFs) have been calculated along the center-of-mass distance between the mobile AMP-lid and ATP-lid domains and subsets of atoms of the protein core of E.coli ADK (Fig. 1). This allows an independent treatment of the movement of both domains and yields a 2D free energy projection in which the experimentally known open and closed ADK conformations are well separated regions. The PMF calculations are based on Hamiltonian Replica Exchange Umbrella Sampling (H-REMD-US) simulations, in which the sampling is distributed across the reaction coordinate plane by 2D harmonic biasing potentials. In this way, the systems are driven into regions of phase space that might otherwise be rarely sampled. Unfavorable trapping of simulations in local free energy minima is avoided by allowing configurations of neighboring windows to exchange according to a Monte Carlo scheme. Every umbrella window was seeded with starting configurations originating from a completely closed and a completely open configuration. This further enhances convergence of the sampling, given that due to the exchange scheme the configurations can rapidly diffuse along the whole reaction coordinate space and reach an equilibrated distribution (see Methods).

Apo ADK and ADK fully occupied by 2ADP+Mg

In the absence of substrates the calculated free energy landscape along the ADK domain opening coordinates (Fig. 6.1) indicates a broad minimum in the vicinity of the known experimental crystal structure conformation (upper right corner). The broad free energy basin allows considerable global motions of both lid domains towards the closed holo ADK conformation without significant changes in free energy. States with a single closed lid are disfavored only by a few kcal/mol. However, a state within few Å from the completely closed crystal structure configuration is disfavored by a free energy penalty of ≈ 8 kcal/mol and hence hardly accessible. Analysis of the trajectories did not indicate significant rearrangements of the positively charged key residues in the apo form in comparison to the substrate bound form. During the simulations, the positively charged active sites remained partially hydrated also in the closed conformations. Apparently, in absence of compensating negatively charged substrates, a complete closing of the lids creates a significant free energy penalty due to the electrostatic repulsion of the buried basic amino acids that surround the nucleotide binding sites.

The calculated free energy landscape in the presence of two ADP molecules and a magnesium ion located at the AMP and ATP binding sites predicts a global free energy minimum coinciding with an experimental holo crystal structure (with a bound AP5A inhibitor). Indeed, NMR solution studies on E.coli ADK [12] but also crystal structures of ADK with two bound ADP molecules [8] indicate close similarity of inhibitor bound and double ADP bound closed ADK structures. Interestingly, a free energy plateau can be observed for intermediate configurations. Only in the close vicinity of the global minimum complete closing is induced by a steep free energy gradient, allowing initiation of the enzymatic reaction. Importantly, the simulations predict greater possible fluctuations and flexibility in the ATP-lid direction than in the AMP-lid direction. This is supported by recent crystal structures of ADK in complex with two ADP molecules[8], which in comparison to the AP5A bound crystal structure show very similar arrangements of core and AMP-lid domain, but a slightly more open ATP-lid domain.

It should be noted that during the sampled timescales, the 2ADP+Mg complex stays close to the ADK closed form configuration (indicated by the small variation of the substrate root mean square deviation (RMSD) with respect to its configuration in the closed form (Fig. 6.1). In a realistic scenario, at a certain degree of domain opening, the substrate complex will eventually partly dissociate. The PMF for holo ADK therefore provides a good estimate of the free energy landscape only in the vicinity of closed configurations.

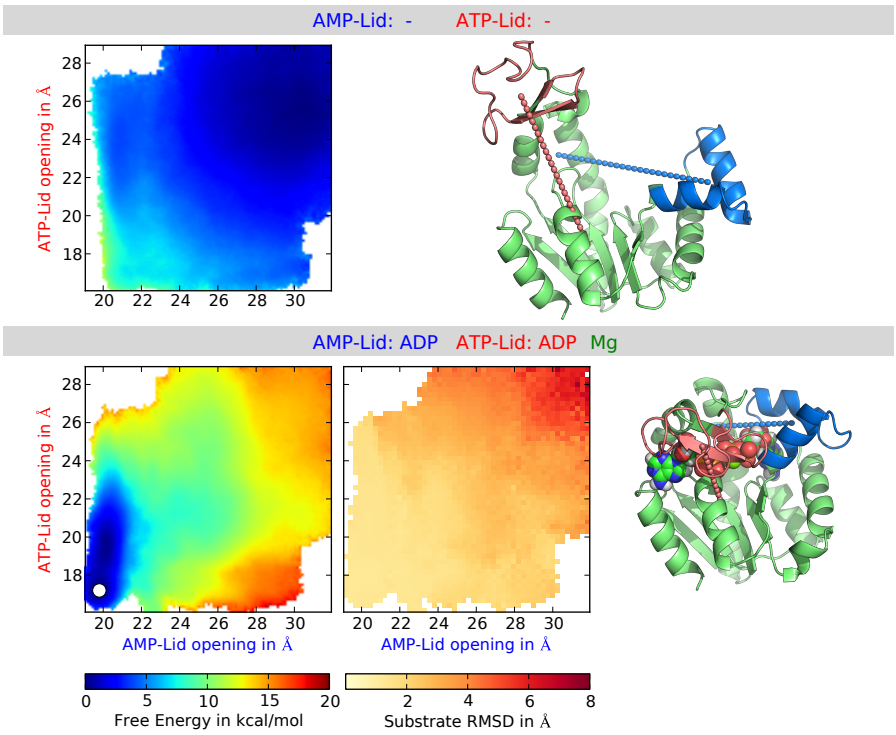


Figure 6.1: Left panels: 2D free energy landscapes along the opening of the AMP-lid and ATP-lid of ADK in the apo form (top) and in complex with 2ADP+Mg (bottom). The top/left and bottom/right regions of the free energy landscapes were excluded from sampling (see Methods). Middle panel, bottom, for ADK in complex with 2ADP+Mg: Root mean square deviation (RMSD) of the substrates with respect to their configuration/position in the initial closed state of ADK for the different domain opening configurations during the simulations. Right panels: Cartoon representations[37] corresponding to the crystal structures of apo ADK (PDB:4AKE [4])/ fully occupied ADK (PDB:1AKE [6]). The AMP-lid and ATP-lid residues and the corresponding opening coordinates (dashed lines) are shown in blue and red, respectively. The originally present AP5A inhibitor of the closed E.coli Adenylate Kinase in PDB:1AKE was replaced by two ADP molecules and a magnesium ion (spheres), taken from PDB:2CDN[36] (see Methods). The lid configuration corresponding to the crystal structure in the closed state is indicated in the PMF by a circle. The crystal structure configuration of open apo ADK is located at (30.9, 29.8).

Single site occupation of the ATP-lid by different substrates

Using the same H-REMD-US methodology, 2D free energy landscapes for the ADK domain motions were calculated for four different substrates (AMP, ADP, ADP+Mg, ATP+Mg) bound to the ATP-lid binding site (Fig. 6.2). Binding of AMP, ADP or ADP+Mg does not result in drastic changes of the free energy landscape in comparison to the apo form, only a shift of the ensembles towards more closed ATP-lid configurations and decreased lid-lid distances can be observed. Facilitation of AMP-lid closing is not indicated. The presence of Mg in the ADP case leads to a slightly more open ATP-lid ensemble. This is in line with the reported acceleration of lid opening upon addition of Mg[8].

In contrast, binding of ATP+Mg allows open and closed ATP-lid configurations and additionally influences the AMP-lid motion. A secondary free energy minimum at a completely closed configuration can be identified. The free energy basins around the minima are broad, indicating that the mobilities of the lids are maintained. These findings are strongly supported by a recent NMR study: Upon binding of ATP to the ATP-lid, interconversion between open and closed ATP-lid states and an equilibrium between open and closed AMP-lid states with a slightly higher population of open states was observed[12].

Notably, ADP and ATP vary only slightly in position and orientation during ATP-lid opening and remain at the core part of the ATP binding site (Fig. 2, low substrate RMSD, snapshots). This finding coincides with a recent NMR study indicating initial binding of ATP to the core part of the ATP-binding site, not to the residues in the mobile ATP-lid domain[42]. AMP, presumably due to the very low binding affinity to the ATP binding site (Kd=1.7 mM [12]), exhibits significant mobility within the binding site upon domain opening (Fig. 6.2).

Single site occupation of the AMP-lid by different substrates

In another series of H-REMD-US simulations, the ADK domain motion upon binding of AMP, ADP, ADP+Mg and ATP+Mg to the AMP-lid site was investigated (Fig.3). For substrates binding to the AMP-lid, coupling to the ATP-lid dynamics can be observed. Binding of AMP to the AMP-lid results in a relatively flat domain opening free energy landscape. The global minimum of the lid configurations is shifted towards the closed state and the free energy penalty to sample the completely closed state is significantly lowered in comparison with apo ADK. Also in NMR experiments fluctuations between closed and open conformations upon binding of AMP to the AMP-lid were reported[12]. Partial closing of the ATP-lid was also observed in energy transfer experiments upon addition of AMP[13]. Binding of ADP or ADP+Mg drastically changes the free energy

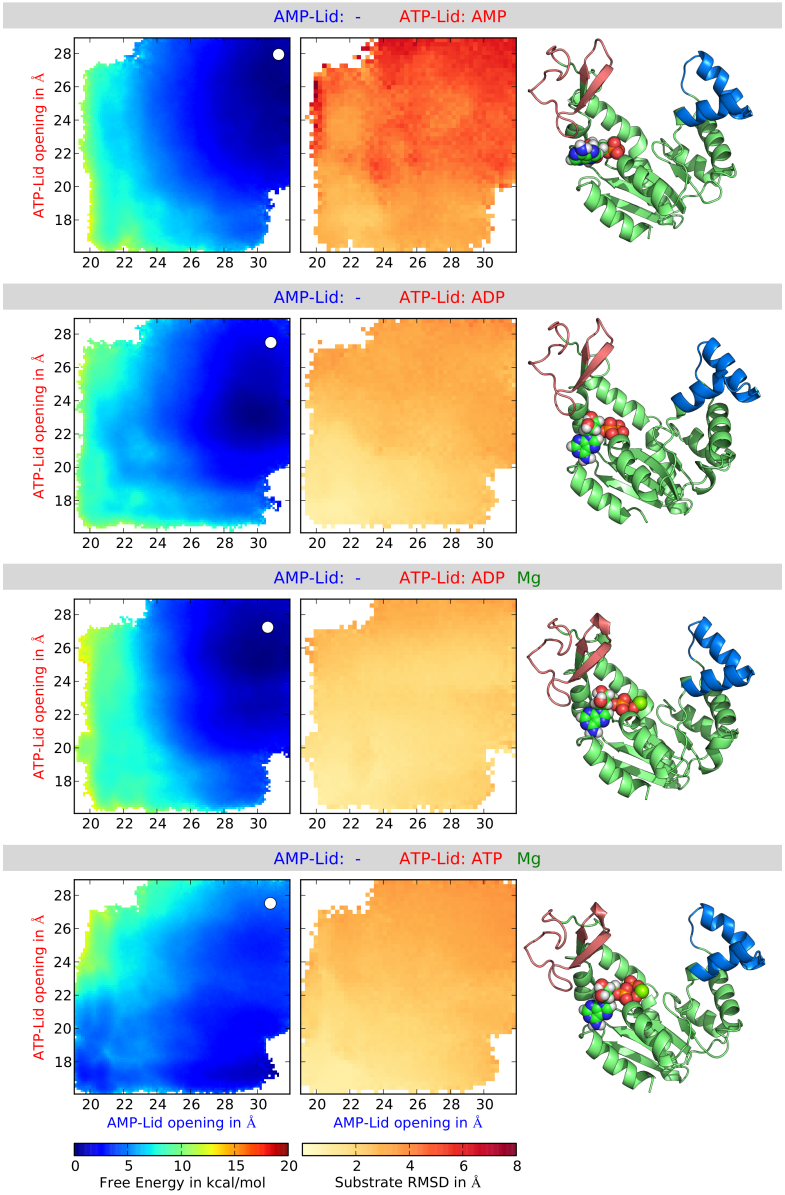


Figure 6.2: Left panels: 2D free energy landscapes along the opening of the AMP-lid and ATP-lid of ADK with AMP, ADP, ADP+Mg or ATP+Mg bound to the ATP-lid binding site (see panel title). Middle panels: Mean RMSD of the substrate with respect to its configuration in the initial closed state of ADK at the different domain opening configurations during the simulations. Right panels: Exemplary snapshots from the sampled trajectories. The corresponding domain configurations are indicated as circles in the free energy plots.

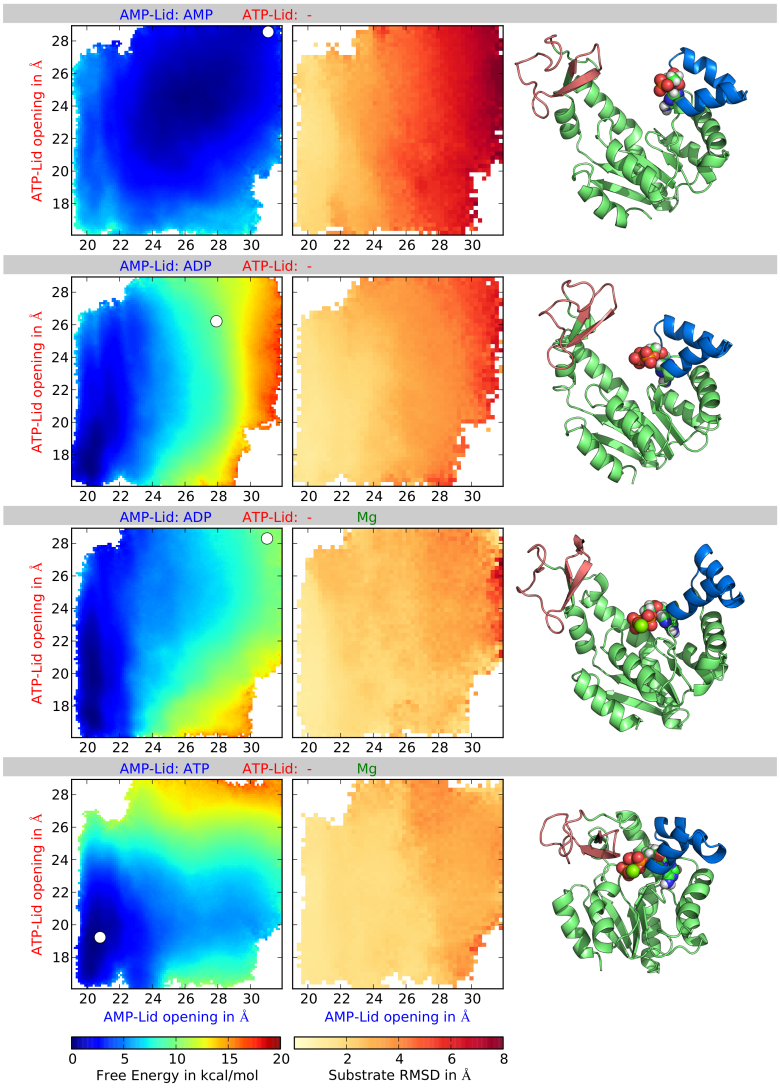


Figure 6.3: Left panels: 2D free energy landscapes along the opening of the AMP-lid and ATP-lid of ADK with AMP, ADP, ADP+Mg or ATP+Mg bound to the AMP-lid binding site. Middle panels: Mean RMSD of the substrate with respect to its configuration in the initial closed state of ADK at the different domain opening configurations during the simulations. Right panels: Exemplary snapshots from the sampling trajectories. The corresponding domain configurations are indicated as circles in the free energy plots.

landscape, leading to a free energy minimum at the closed state. The free energy gradient along the AMP-lid coordinate is steeper, but also ATP-lid closing is favored. As in the ATP-lid site case, ADP+Mg leads to slightly more open configurations than ADP alone, again in line with faster lid opening[8]. For ATP bound to the AMP-lid site, additionally, a strong tendency towards ATP-lid closing is indicated.

In contrast to the ATP-lid binding pocket, substrate arrangement in the AMP-lid pocket changes significantly upon domain opening. AMP and ADP remain attached to the AMP-lid and partially follow its opening movement (Fig. 6.3, snapshots). This is also shown by the increased substrate RMSD with respect to the equilibrated closed form configuration.

6.4 Discussion

Despite many efforts in experimental and simulation studies, a full understanding of the global domain motion of ADK and its coupling to substrate binding is still lacking. In this study, free energy landscapes along the changes in conformation of ADK for basically all possible ligation states with natural substrates and the apo form have been calculated. In order to adequately capture the conformational rearrangements the two lid domains were controlled individually by means of two-dimensional US simulations. This comprehensive approach allows a systematic analysis of the interplay between substrate binding and lid motion. Significant differences in the free energy landscapes were obtained not only for the apo and holo forms of ADK but also between complexes that included only single substrates at the different binding sites.

The calculated 2D PMF for apo ADK agrees qualitatively with experimental findings [10][11][5], which indicate a high degree of global flexibility and the possibility of adopting closed-like states even in the absence of substrates. However, our simulations indicate a significant free energy barrier for adopting a completely closed state with both lid opening coordinates within few Å from the crystal structure conformation of the holo enzyme. In particular, in FRET studies[11] closing of ADK in the absence of substrate or inhibitors was observed. It is, however, not clear if the experimental spatial resolution is indeed sufficient to distinguish completely closed states from closed-like states that do not significantly differ in hydration of the polar and charged residues near the substrate binding sites. The calculated free energy landscape for the apo form argues against a pure conformational selection mechanism for substrate binding. This is also supported by NMR studies on single substrate binding to ADK [42]. Previous MD simulation studies based either on an implicit solvation model [24] or performed in explicit solvent also predicted a significant penalty for complete closing in the apo form [20][28]. The simulations on holo ADK suggest that opening of the ATP-lid is the

most likely first step in the release of the substrates (or products) as the free energy landscape indicates a higher mobility of the ATP-lid compared to the AMP-lid in the holo form. Since domain opening has been determined to be the rate limiting step for catalysis[10], it appears reasonable that the opening motion of one of the lids in the holo state is not disfavored by a steep free energy gradient.

Strikingly, the free energy landscapes for single substrate bound states of ADK strongly depend on the type and position of the substrate. For all single substrate bound states for which to our knowledge experimental data is available, the free energy landscapes are compatible with the experimental findings (see Results). In general, we observe that upon binding of a substrate to its corresponding lid (which can initiate an enzymatically productive occupation state of ADK) flexibility of the opposing lid is maintained with significant population of open configurations. In this way efficient binding of the second substrate can be achieved which in a closed state would be sterically hindered.

Of special interest is the possible binding of substrates to the non-designated lid. Presumably, binding of AMP to the ATP-lid binding site is of little relevance since AMP binds to the ATP binding site only weakly compared to ATP ($K_D(\mathrm{AMP}) = 1700\,\mu\mathrm{M}$ vs. $K_D(\mathrm{ATP}) = 50\,\mu\mathrm{M}$ [12]). In line with this weak binding, in our simulations the AMP molecule showed large mobility in the ATP-lid site. However, the situation is different for binding of ATP to the AMP binding site. A significant fraction of ADK molecules may bind ATP in the incorrect binding site under equilibrium conditions ($K_D(\mathrm{ATP}) = 750\,\mu\mathrm{M}$ vs. $K_D(\mathrm{AMP}) = 210\,\mu\mathrm{M}$ [42]). Note, that this case is difficult to access (or isolate) experimentally because once the ATP-binding site is occupied, binding of a second ATP to the AMP site is not possible due to sterical reasons. In contrast to the productive initiation states, ATP bound incorrectly to the AMP-lid site results in closing and restricted mobility of both the AMP-lid and the ATP-lid. The closing of both lids may prevent binding of a second incoming nucleotide to the ATP-lid site which would result in a stable but unproductive blocking state. Possible processes following initial binding of ATP+Mg (which has a much higher physiological concentration than ADP or AMP[43]) to each of the lids are exemplary illustrated in Figure 6.4.

More generally, we observe that the lid behavior is intrinsically asymmetric. While for the ATP-lid, the substrates appear to bind with a high fraction of their affinity to the open state and only slightly promote lid closing, the same substrates induce a considerable stronger tendency towards closing when bound to the AMP-lid. Given that the ATP+AMP bound state is asymmetric, it appears necessary that an enzyme which needs to distinguish such states evolves asymmetric binding site lids. Additionally, providing a stable initiation state in form of an open ATP-lid might be beneficial for the overall catalysis rate.

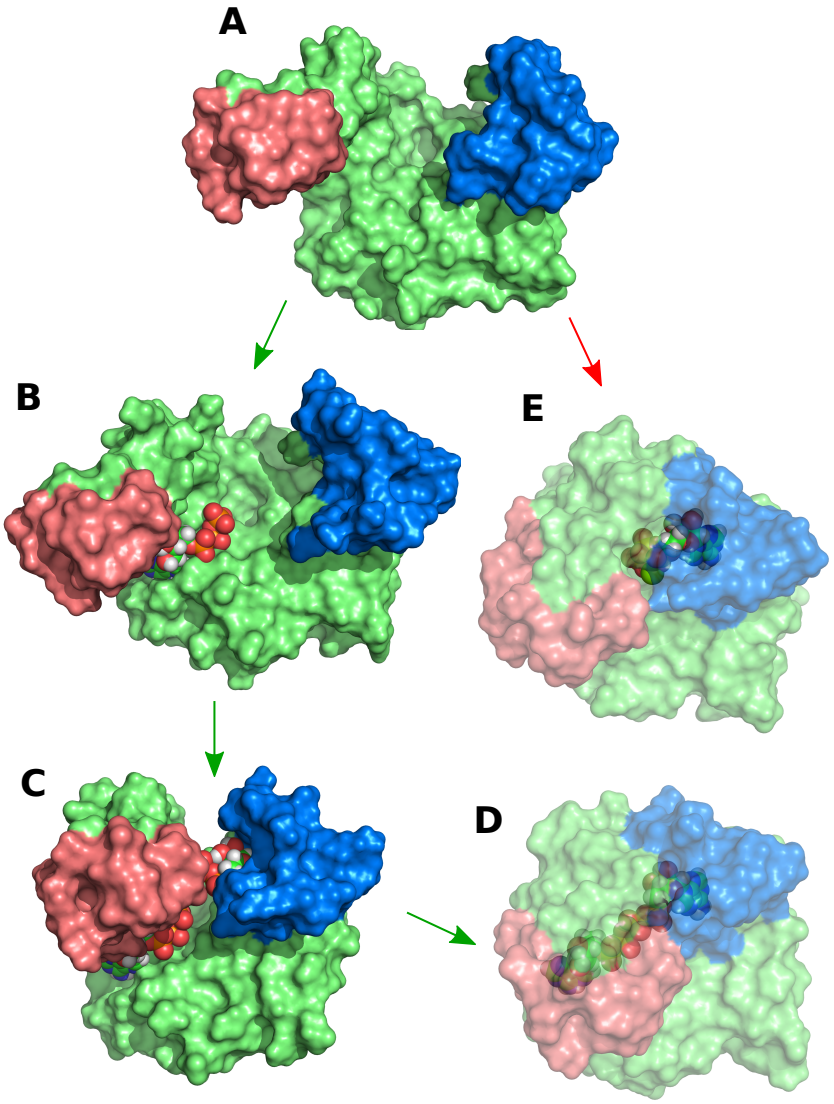


Figure 6.4: Schematic illustration of possible pathways initiated by ATP+Mg binding on the basis of ADK configurations obtained from the simulations. **A:** Open, flexible ADK without substrates. **B:** Upon binding of ATP+Mg to the ATP-lid site flexibility of both lids is maintained. The AMP-lid is most probably found in open states. **C:** AMP (added to the figure) can bind to the open AMP-lid. **D:** Once two adequate substrates are bound complete binding is induced, facilitating the chemical reaction. **E:** Binding of ATP+Mg to the AMP-lid site induces a global shift to closed conformations. This may hinder further occupation of the ATP-lid site which would result in a stable but unproductive state.

In conclusion, the simulations indicate that ADK has evolved in a way that initial substrate binding corresponding to productive occupation states still maintains the opposing lid open and accessible for the subsequent substrate. Such a behavior could hardly be realized with only one movable lid. Furthermore, incorrect single substrate bound states might be hindered to move towards stable but unproductive fully occupied states by closing of both lids, sterically preventing further occupation. For enzymes like ADK this strategy might be necessary because considerable binding affinity for unproductive initial occupation states cannot always be avoided by the composition of the binding sites due to the structural similarity of the substrates.

6.5 Bibliography

[1] Gordon G. Hammes. "Multiple Conformational Changes in Enzyme Catalysis." In: *Biochemistry.* 41.26 (2002), pp. 8221–8228.

[2] Petras Dzeja and Andre Terzic. "Adenylate kinase and AMP Signaling Networks: Metabolic Monitoring, Signal Communication and Body Energy Sensing." In: *Int. J. Mol. Sci.* 10 (2009), pp. 1729–1772.

[3] J. R. Knowles. "Enzyme-catalyzed phosphoryl transfer reactions." In: *Annual. Rev.* 49 (1980), pp. 877–919.

[4] C W Müller, G J Schlauderer, J Reinstein, and G E Schulz. "Adenylate kinase motions during catalysis: An energetic counterweight balancing substrate binding." In: *Structure* 4.2 (1996), pp. 147–156.

[5] Katherine A Henzler-Wildman et al. "Intrinsic motions along an enzymatic reaction trajectory." In: *Nature* 450.7171 (2007), pp. 838–844.

[6] Christoph W Müller and Georg E Schulz. "Structure of the complex between adenylate kinase from Escherichia coli and the inhibitor Ap5A refined at 1.9 Åresolution: A model for a catalytic transition state." In: *J. Mol. Biol.* 224.1 (1992), pp. 159–177.

[7] U. Abele and G. E. Schulz. "High-resolution structures of adenylate kinase from yeast ligated with inhibitor Ap5A, showing the pathway of phosphoryl transfer." In: *Prot. Sci.* 4 (1995), pp. 1262–1271.

[8] S Jordan Kerns et al. "The energy landscape of adenylate kinase during catalysis." In: *Nat. Struct. Mol. Biol.* 22.2 (2015), pp. 124–131.

[9] G. E. Schulz. "Induced-fit movements in adenylate kinases." In: *Farady Discuss.* 93 (1996), pp. 85–93.

[10] M Wolf-Watz et al. “Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair.” In: *Nat. Struct. Mol. Biol.* 11.10 (2004), pp. 945–949.

[11] Jeffrey A Hanson et al. “Illuminating the mechanistic roles of enzyme conformational dynamics.” In: *Pro. Nat. Acad. Sci. USA* 104.46 (2007), pp. 18055–18060.

[12] Jörgen Ådén and Magnus Wolf-Watz. “NMR identification of transient complexes critical to adenylate kinase catalysis.” In: *J. Am. Chem. Soc.* 129.45 (2007), pp. 14003–14012.

[13] T Bilderback, T Fulmer, WW Mantulin, and M Glaser. “Substrate binding causes movement in the ATP binding domain of Escherichia coli adenylate kinase.” In: *Biochemistry* 35.19 (1996), pp. 6100–6106.

[14] Michael A. Sinev, Elena V. Sineva, Varda Ittah, and Elisha Haas. “Domain closure in adenylate kinase.” In: *Biochemistry* 35.20 (1996), pp. 6425–6437.

[15] G.J. Schlauderer, K. Proba, and G.E. Schulz. “Structure of a Mutant adenylate kinase Ligated with an ATP-analogue Showing Domain Closure Over ATP.” In: *J. Mol. Biol.* 256.2 (1996), pp. 223–227.

[16] Paul Maragakis and Martin Karplus. “Large Amplitude Conformational Change in Proteins Explored with a Plastic Network Model: Adenylate Kinase.” In: *J. Mol. Biol.* 352.4 (2005), pp. 807–822.

[17] P C Whitford, S Gosavi, and J N Onuchic. “Conformational transitions in Adenylate Kinase. Allosteric communication reduces misligation.” In: *J. Biol. Chem.* 283.4 (2008), pp. 2042–2048.

[18] Divesh Bhatt and Daniel M. Zuckerman. “Heterogeneous Path Ensembles for Conformational Transitions in Semiatomistic Models of Adenylate Kinase.” In: *J. Chem. Theory. Comput.* 6.11 (2010), pp. 3527–3539.

[19] J Ping, P Hao, YX Li, and JF Wang. “Molecular dynamics studies on the conformational transitions of adenylate kinase: A computational evidence for the conformational selection mechanism.” In: *BioMed. Res. Int.* 2013 (2013), e628536.

[20] Hyun Deok Song and Fangqiang Zhu. “Conformational dynamics of a ligand-free adenylate kinase.” In: *PLoS ONE* 8.7 (2013), e68023.

[21] Jason B. Brokaw and Jhih Wei Chu. “On the roles of substrate binding and hinge unfolding in conformational changes of adenylate kinase.” In: *Biophys. J.* 99.10 (2010), pp. 3420–3429.

[22] Francesco Pontiggia, Andrea Zen, and Cristian Micheletti. “Small- and large-scale conformational changes of adenylate kinase: A molecular dynamics study of the subdomain motion and mechanics.” In: *Biophys. J.* 95.12 (2008), pp. 5901–5912.

[23] Hongfeng Lou and Robert I. Cukier. "Molecular dynamics of apo-adenylate kinase: A principal component analysis." In: *J. Phys. Chem. B* 110.25 (2006), pp. 12796–12808.

[24] Karunesh Arora and Charles L Brooks. "Large-scale allosteric conformational transitions of adenylate kinase appear to involve a population-shift mechanism." In: *Proc. Nat. Acad. Sci. USA* 104.47 (2007), pp. 18496–18501.

[25] Davit a. Potoyan, Pavel I. Zhuravlev, and Garegin a. Papoian. "Computing free energy of a large-scale allosteric transition in adenylate kinase using all atom explicit solvent simulations." In: *J. Phys. Chem. B* 116.5 (2012), pp. 1709–1715.

[26] Jinan Wang et al. "Exploring transition pathway and free-energy profile of large-scale protein conformational change by combining normal mode analysis and umbrella sampling molecular dynamics." In: *J. Phys. Chem. B* 118.1 (2014), pp. 134–143.

[27] Yasuhiro Matsunaga, Hiroshi Fujisaki, Tohru Terada, and Furuta. "Minimum free energy path of ligand-induced transition in adenylate kinase." In: *PLoS Comput. Biol.* 8.6 (2012), e1002555.

[28] Elena Formoso, Vittorio Limongelli, and Michele Parrinello. "Energetics and structural characterization of the large-scale functional motion of adenylate kinase." In: *Sci. Rep.* 5.8425 (2015).

[29] Viktor Hornak et al. "Comparison of multiple Amber force fields and development of improved protein backbone parameters." In: *Proteins* 65.3 (2006), pp. 712–725.

[30] William L. Jorgensen, Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. "Comparison of Simple Potential Functions for Simulating Liquid Water." In: *J. Chem. Phys.* 79.2 (1983), pp. 926–935.

[31] Junmei Wang, Romain M. Wolf, James W. Caldwell, Peter A. Kollman, and David A. Case. "Development and Testing of a General Amber Force Field." In: *J. Comput. Chem.* 25.9 (2004), pp. 1157–1174.

[32] D.A. Case et al. *AMBER 14*. University of California, San Francisco, 2014.

[33] Kristin L. Meagher, Luke T. Redman, and Heather A. Carlson. "Development of polyphosphate parameters for use with the AMBER force field." In: *J. Comput. Chem.* 24.9 (2003), pp. 1016–1025.

[34] Olof Allnér, Lennart Nilsson, and Alessandra Villa. "Magnesium Ion–Water Coordination and Exchange in Biomolecular Simulations." In: *J. Chem. Theory. Comput.* 8.4 (2012), pp. 1493–1502.

[35] In Suk Joung and Thomas E. Cheatham. "Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations." In: *J. Phys. Chem. B* 112.30 (2008), pp. 9020–9041.

[36] M. Bellinzoni et al. "The Crystal Structure of Mycobacterium Tuberculosis adenylate kinase in Complex with Two Molecules of Adp and Mg2+ Supports an Associative Mechanism for Phosphoryl Transfer." In: *Prot. Sci.* 15 (2006), p. 1489.

[37] Schrödinger, LLC. "The PyMOL Molecular Graphics System, Version 1.7.4." Aug. 2010.

[38] G M Torrie and J P Valleau. "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling." In: *J. Comput. Phys.* 23.2 (1977), pp. 187–199.

[39] 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: *J. Chem. Phys.* 116.20 (2002), p. 9058.

[40] Shankar Kumar, John M. Rosenberg, Djamal Bouzida, Robert H. Swendsen, and Peter A. Kollman. "The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method." In: *J. Comput. Chem.* 13.8 (1992), pp. 1011–1021.

[41] Alan Grossfield. *WHAM: The weighted histogram analysis method 2.0.7*. 2013.

[42] Jörgen Ådén, Christoph F. Weise, Kristoffer Brännström, Anders Olofsson, and Magnus Wolf-Watz. "Structural topology and activation of an initial adenylate kinase-substrate complex." In: *Biochemistry* 52.6 (2013), pp. 1055–1061.

[43] I Beis and E A Newsholme. "The contents of adenine nucleotides, phosphagens and some glycolytic intermediates in resting muscles from vertebrates and invertebrates." In: *Biochem. J.* 152.1 (1975), pp. 23–32.

7 Multi-Scale Calculation of Binding Rates for Neuraminidase Inhibitors

A detailed understanding of the drug-receptor association process is of fundamental importance for drug design. Due to the long timescales of typical binding kinetics, the atomistic simulation of the ligand traveling from bulk solution into the binding site is still computationally challenging. In this work, we apply a multi-scale approach of combined molecular dynamics (MD) and Brownian dynamics (BD) simulations to investigate association pathway ensembles for the two prominent H1N1 neuraminidase inhibitors oseltamivir and zanamivir. Including knowledge of the approximate binding site location allows for the confinement of detailed but expensive MD simulations to its immediate vicinity and the use of less demanding BD simulations for the diffusion controlled part of the association pathway. This approach permits the use of commonly available hardware and is intrinsically parallelizable. We apply a binding criterion based on the residence time of the inhibitor in the binding pocket, thereby avoiding definitions of geometric criteria that typically require prior knowledge about the binding mechanism. The method ranks the association rates of both inhibitors in agreement with experiment and yields reasonable absolute values. The simulated association pathway ensembles reveal that initially the ligands are oriented in the electrostatic field of the receptor. Subsequently, a salt bridge is formed between the inhibitors' carboxyl group and neuraminidase residue Arg368, followed by the adoption of the native binding poses. Unexpectedly, despite of oseltamivir's higher overall association rate, the rate into the intermediate salt-bridge state was found to be higher for zanamivir.

7.1 Introduction

Until recently, the focus in drug design was on determining and optimizing equilibrium binding affinities to increase the inhibition efficacy of potential drug molecules[1]. There is, however, evidence that drug efficacy is also determined by binding kinetics[2]. Thus, the detailed association pathways play an increasingly important role in drug design. Knowledge of the complete pathway ensemble may additionally reveal alternative binding modes, which remain disregarded when optimizing only the affinity of a single binding pose[3, 4]. Computer simulations are a promising tool to investigate drug

molecule binding at different levels of detail. However, the atomistic simulation of complete association pathways is still computationally demanding. In contrast to the calculation of equilibrium quantities which can be obtained from enhanced sampling methods, reliable determination of the kinetics requires simulation at timescales of the physical association process.

A prominent effort of simulating ligand-receptor complex formation at atomistic detail was based on ultra-long molecular dynamics (MD) simulations without imposing prior knowledge of possible binding sites and binding modes. This approach used specialized hardware allowing simulation times of up to milliseconds. For the G-protein-coupled receptor, final stable bound poses were observed coinciding with the geometry found in experimental crystal structures for several small molecules[5]. The allosteric modulation of inhibition sites by bound drug molecules has been investigated using a similar approach[6]. As an alternative, the application of Markov state models allows for the combination of independent short MD simulations. This approach has been used to investigate the association of the charged inhibitor benzamidine to trypsin[7, 8]. However, using ultra-long MD simulations or Markov state models, a large sampling effort is spent on the diffusive search of the binding site. This sampling effort is expected to be particularly large for charge-neutral inhibitors that are not strongly attracted by the binding site through long-range electrostatic interactions.

For these diffusion controlled regimes, Brownian dynamics (BD)[9] simulations are an efficient representation and have been previously used to simulate entire association pathways[10–16]. However, BD simulations ignore the flexibility of receptor and ligand, which can play a role in association processes. Additionally, effects of the explicit solvent molecules are neglected and short range interactions between receptor and ligand are only roughly modeled by a collision criterion.

In the search for possible inhibitors or inhibitor optimization, frequently, the target binding site is already known and an atomistic simulation is essential only for the final steps of the association process. In this study, we explore a multi-scale combination of MD and BD simulations to efficiently investigate the association process of two charge-neutral inhibitors (oseltamivir and zanamivir) to an important drug target, influenza H1N1 neuraminidase. We limit the computationally expensive atomistic MD simulations to the vicinity of the known binding site (Figure 7.1, blue area) and connect the MD regime to less demanding BD simulations in the diffusion controlled regime (Figure 7.1, dashed sphere into red encounter surface), similar to previous studies on comparatively small systems with charged ligands[17, 18].

To calculate kinetic k_{on} rates from the pathway ensemble, previous studies relied on a variety of reaction criteria to define the bound state of the ligand in the binding site. Often a reaction surface defined by the distance between centers of masses of ligand and receptor[17, 19] or the number of native contacts[12, 20] was used. These criteria

can in principle be arbitrarily adjusted and require additional knowledge about the complex as for instance an X-ray structure. In contrast, we use a minimal residence time of 2 µs in the MD covered area to define a bound event. Thereby, we avoid a priori geometrical or knowledge based binding criteria and no model parameters have to be fitted to obtain absolute association rates.

From the simulations, detailed association pathways for oseltamivir and zanamivir were obtained, revealing a single, common intermediate state. They predict a higher binding rate for oseltamivir, coinciding with experimental findings. Surprisingly, zanamivir reaches the intermediate state with a higher rate.

7.2 Methods Summary

Association pathways and on-rate calculation

The diffusion dominated part of the association pathway was connected to the computationally expensive MD simulations at an encounter surface (see Figure 7.1, red circle) in the vicinity of the binding site. The surface was defined by a 12 Å distance of the center of mass of the C_α atoms of Asn248 and Pro431. This point is located approximately at the center of the entrance to the neuraminidase binding site.

The core of this study are atomistic MD simulations with an explicit solvent representation starting from the encounter surface. For each drug molecule, more than 600 starting configurations were generated and propagated until they left the MD area (Figure 7.1, blue area) or after a maximum simulation time of 2 µs. Trajectories staying within the MD area for 2 µs were considered to represent bound inhibitors. For each drug molecule, the MD start ensemble was generated by starting $\sim$ 5000 short MD simulations at a distance between 30 Å and 32 Å from the center of the binding site, approximately at the outer shell of the blue MD region in Figure 7.1.

The MD simulations starting at the vicinity of the binding site were complemented by Brownian dynamics simulations starting outside of this region. For each drug molecule, 10^6 BD trajectories were started from a spherical shell of radius $b = 60$ Å to determine the probability of arriving at the ES before diffusing to distances larger than $q = 100$ Å. Also, 10^6 BD trajectories were started from the MD configurations terminated outside of the MD area to determine the probability to re-enter the vicinity of the binding site before diffusing to distances larger than b. From the isotropic rate into a sphere of radius b, the results of the BD simulations and the fraction of binding MD trajectories, the on-rate was calculated using the method of Northrup et al.[13] (see Methods Details).

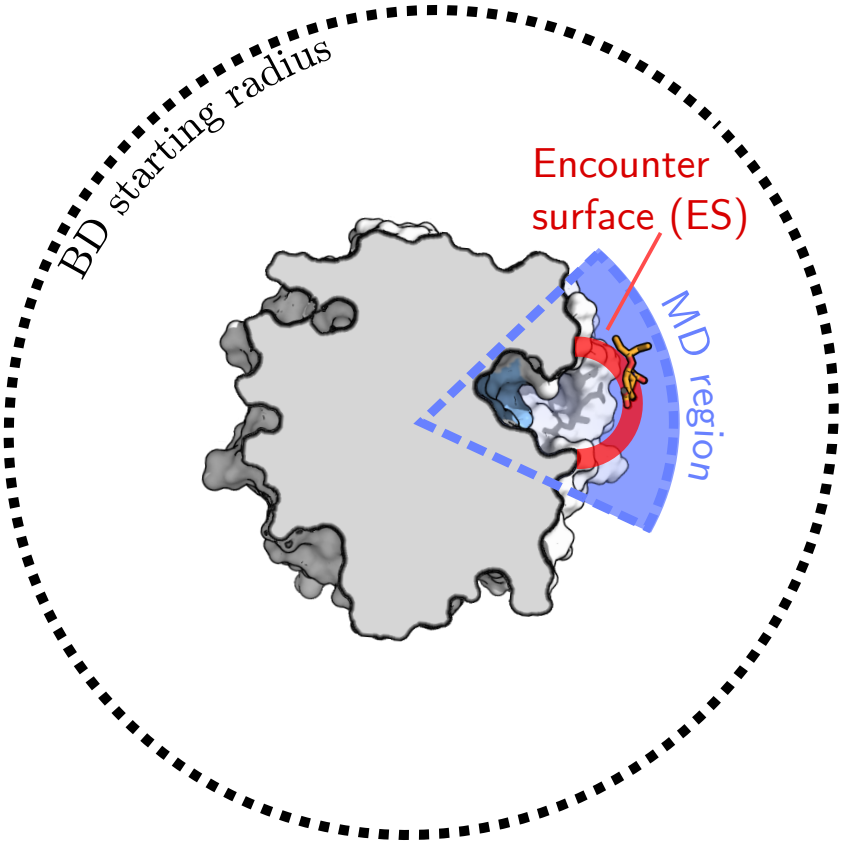


Figure 7.1: Schematic depiction of the simulation setup. The isotropic rate from infinity into a sphere of radius 60 Å (dotted circle) can be obtained by continuum theory. The diffusion and long-range electrostatics dominated part of the association pathways from the sphere into the encounter surface (ES, red) was treated with Brownian dynamics simulations. Close range interactions were covered by MD simulations with explicit water representation, starting from a separately generated atomistic ensemble at the ES. The propagation was terminated either when the ligand left the MD region (blue segment) or after a minimal residence time of 2 µs. MD trajectories in which the ligand left the MD region were continued as an ensemble of BD trajectories to determine their probability to re-enter the ES.

Structures

To avoid any bias towards a bound conformation of the binding site, the neuraminidase apo structure was used for the association simulations (monomeric apo H1N1 neuraminidase from PDB:4B7M[21]). For comparison with the bound conformation, reference crystal structures of oseltamivir and zanamivir bound to H1N1 neuraminidase were taken from PDBs 4B7Q and 4B7R[21].

Molecular dynamics simulations

The MD simulations were carried out with the pmemd.cuda[22] module of the Amber14 package[23] at atomistic detail[24, 25], including explicit solvent. The whole receptor[26–28], the ions[29] and the drug molecules[30, 31] were solvated in 8150(oseltamivir) or 8214(zanamivir) TIP3P[32] water molecules in a octahedral box. Equilibration was carried out in the NPT ensemble using the Berendsen thermostat and barostat[33]. The association trajectories were propagated in the NVT ensemble using the Andersen thermostat[34].

Brownian dynamics simulations

In the Brownian dynamics simulations, the molecules were treated as rigid structures undergoing translational and rotational diffusion in an implicit solvent[9]. Electrostatic forces on the drug molecule in the field of the receptor were taken into account via the Poisson-Boltzmann equation[35]. The topology of the molecules was imposed by rejecting propagation steps leading to an overlap of the structures.

Inhibitor bound structures

The most stable bound configurations were determined by clustering the ligand poses from all combined 2 µs binding MD trajectories with the gromos[36] clustering method as implemented in GROMACS 4.6[37].

7.3 Results

The approximate location of the binding site was incorporated into the study of the inhibitor association pathways in the form of a geometrically defined encounter surface (ES) (see Figure 7.1). The ES has to be designed such that it extends over the entire binding site entrance and that it has to be crossed by approaching inhibitors in order to bind. Far away from the binding site, up to the ES, the association pathways are represented by coarse BD simulations. Starting from the ES, the association pathways

are represented by atomistic MD simulations. The ES is ideally located in the region where specific short–ranged interactions between inhibitors and binding site start to play a role. In this study, the ES was defined by a distance of 12 Å between the inhibitors and the approximate center of the binding site entrance (see Methods Details). Exemplary trajectory snapshots of inhibitors at the ES are illustrated in Figure 7.2A).

Rate into encounter surface (BD)

The steady-state rate of oseltamivir and zanamivir from infinite distances into the ES $k^{\mathrm{BD}}_{\infty\to\mathrm{ES}}$ (Table 7.1) was calculated from the isotropic continuum rate into a sphere surface of radius 60 Å and the fraction of BD trajectories starting at this sphere surface and, subsequently, reaching the ES (see Methods). As expected for the two inhibitors with similar diffusion constants, at distances outside the ES, where the binding site does not directly interact with the inhibitors, the rates are roughly equal.

Pathways from encounter surface into binding site (MD)

The transitions from the encounter surface into the actual binding site were studied with molecular dynamics (MD) simulations. MD trajectories were started with an inhibitor at the encounter surface and terminated when it left the MD area (Figure 7.1, blue area). Trajectories were considered bound when the inhibitor resided in the MD region for at least 2 µs (see Methods).

Exemplary inhibitor configurations serving as starting points for the MD simulations at the encounter surface are shown in Figure 7.2A. Analyzing all binding trajectories of both oseltamivir and zanamivir, the formation of the salt bridge between the carboxyl group of the drug molecule and Arg368 of neuraminidase (distance smaller than 3 Å, Figure 7.2B) was identified as a common first transient interaction with the binding site. This key interaction was observed in a variety of orientations of the ligand relative to the binding site and was only restricted sterically by the shape of the binding site entrance. The probability of the MD trajectories to form the carboxyl salt bridge was found to be higher for zanamivir (101 of 606) than for oseltamivir (64 of 676), predicting a higher rate into the salt-bridge ensemble for zanamivir. This can be explained by the orientational bias caused by the higher electrostatic dipole moment of zanamivir (28.3 D) interacting with the electrostatic field of the positively charged binding site in comparison to oseltamivir (23.0 D) (see Figure 7.3).

Of all trajectories in which the salt-bridge was formed, only a small fraction (19 for oseltamivir, 4 for zanamivir) transitioned into stable modes remaining in the MD region for 2 µs (see 7.2C). Although zanamivir adopted the intermediate salt-bridge state more frequently, it reaches final bound modes with a lower probability than

Molecule	$k^{BD}_{\infty\to ES}$ [1/μMs]	$\alpha^{MD}_{ES\to bound}$	$k^{calc.}_{on}$ [1/μMs]	$k^{exp.}_{on}$ [1/μMs]
Oseltamivir	794	19/676	25.5	3.1
Zanamivir	737	4/606	5.5	1.3

Table 7.1: Rate from infinite distance into encounter surface (ES) obtained from continuum theory and BD simulations ($k^{BD}_{\infty\to ES}$), number of binding MD trajectories over total number of MD trajectories started at the ES ($\alpha^{MD}_{ES\to bound}$), total calculated on-rate ($k^{calc.}_{on}$) and experimental on-rate[21] ($k^{exp.}_{on}$) for both ligands.

oseltamivir, agreeing with the experimentally determined ratio of on-rates. Thus, the subsequent specific short range interactions and detailed binding site rearrangements must be the reason for a higher overall on-rate for oseltamivir compared to zanamivir. For the transition from the salt-bridge ensemble into the stable modes, no distinct predominant intermediate states could be identified. The 150-loop (residues 146-152), known to be very flexible[38], remained in a closed conformation in all oseltamivir binding trajectories. For zanamivir, in two of the 2 µs trajectories, the flexible 150-loop opened up (shown in Figure 7.4).

Total association rates

MD trajectories terminated at the border of the MD region were continued as an ensemble of BD trajectories to determine the probability of re-arriving at the encounter surface. The total on-rate k_{on} (Table 7.1) was calculated from the BD simulations and the fraction of bound MD trajectories $\alpha^{MD}_{ES\to bound}$ by the method of Northrup[13] (see Methods Details). The calculated total association rates coincide with the experimental ranking of the inhibitors, overestimating the absolute values by a factor of 4-8.

Binding modes

To analyze the most stable binding modes indicated by the simulations, the bound trajectories were clustered with respect to ligand RMSD. In Figure 7.2C the most populated clusters are shown with their respective population. For oseltamivir, in nearly 80% of the frames the inhibitor adopted a binding geometry very similar to the crystal structure, with an RMSD of only 1.0 Å (see Co1). This pose was observed to exchange with two secondary binding modes. The diethane group was frequently observed to flip out of the corresponding binding cavity (Co2) and vice versa. For the reverse transition (Co2 to Co1), the opening of the Asn295 side-chain appears to be crucial. Also, the carboxyl group detaches and the salt bridge occasionally opens (Co3),

with all other binding features remaining closed. The 150-loop remained closed for all binding modes similar to the X-ray structure.

In the zanamivir simulations binding geometries very similar to the crystal structure were sampled, but the most frequently observed modes were deviating from the crystal structure pose by at least 3.7 Å. In modes Cz1 and Cz5 (Figure 7.2), the carboxyl group of zanamivir formed a stable salt bridge with Arg118 and the guanidinium group formed a salt bridge with Asp151. In modes Cz2 and Cz3, Arg118 and Asp151 were displaced as a result of the partially opened 150-loop, not allowing interactions with zanamivir. The carboxyl Arg368 salt-bridge was formed. Mode Cz4 shows a closed carboxyl-Arg368 salt bridge with the glycerol group of zanamivir pointing in the direction of the 150-loop.

7.4 Discussion

During association processes, the initial approach of a ligand towards the receptor is governed by a diffusive random walk and possibly guided by long range electrostatic interactions. This regime can be accurately represented by BD simulations[9], treating molecules as rigid bodies with their translational and rotational movement modeled as a diffusion process under the influence of electrostatic forces. The solvent is represented by a stochastic force and effects of the molecular solvent structure are neglected. These simplifying assumptions allow far longer simulation times compared to MD, but break down when the ligand comes into contact with the binding site. In order to extract association rates from such approaches, in previous studies, geometric reaction criteria were applied[13, 17, 19, 40]. However, as resulting kinetics are highly sensitive to the choice of these criteria, the predictive value is limited. For the binding site vicinity, a more detailed model with atomistic resolution and an explicit representation of the solvent, shown to significantly influence ligand binding mechanisms[41, 42], is advantageous. Atomistic MD simulations provide such a high level of detail at the cost of increased computational demands. Multi-scale approaches have been proposed that combine the BD simulations in the diffusive regime with expensive but more accurate MD simulations for close range interactions[17].

In the present study, we employed a combination of Brownian dynamics (BD) and explicit solvent molecular dynamics (MD) simulations to investigate the association of the two clinically relevant inhibitors, oseltamivir and zanamivir, to H1N1 neuraminidase. Only knowledge of the rough location of the binding site was used and all simulations were started from the apo form of the receptor to avoid a bias towards the bound form. BD simulations were used to estimate the rate into the ES. From the ES, MD simulations were started and run only in the vicinity of the known binding site. This

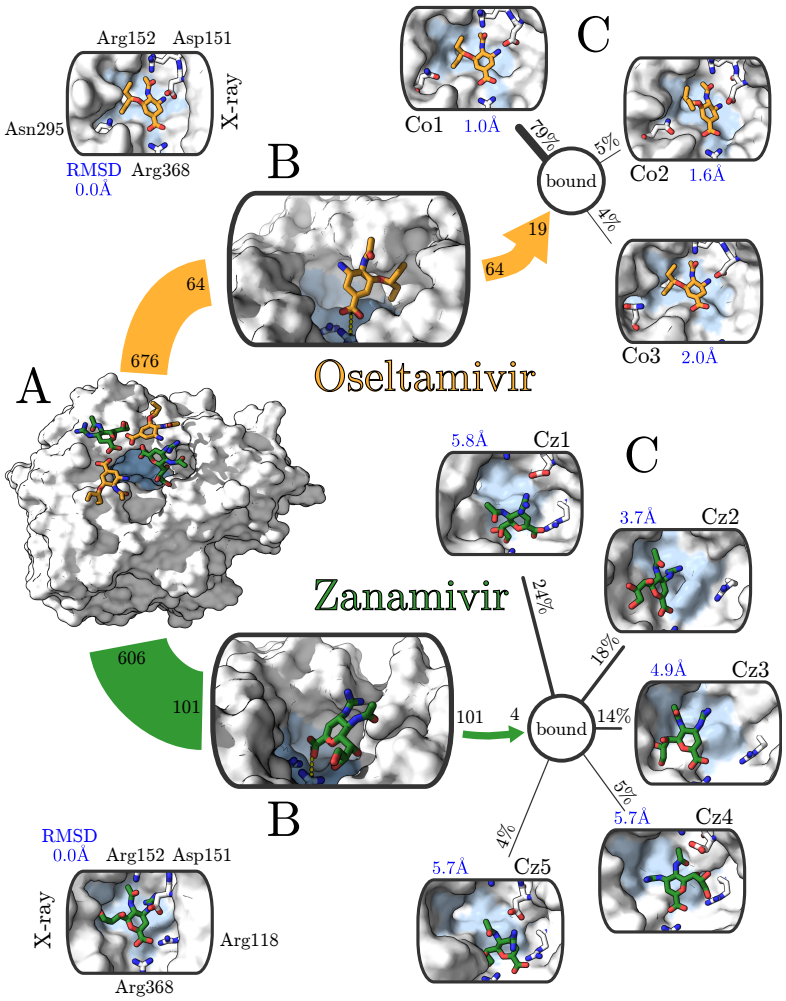


Figure 7.2: Predominant association pathways for oseltamivir (orange sticks) and zanamivir (green sticks) to the binding site (light blue area) of neuraminidase (white). MD Simulations were started from ligand positions at the ES (A, exemplary ligand positions). In all trajectories that resulted in binding, the ligand's carboxyl group first formed a key contact with Arg368 (B). The largest clusters from bound trajectories (C) are depicted with population percentage and ligand RMSD from X-ray structure (blue numbers). Colored arrows indicate transition probabilities from ES (A) via initial salt bridge state (B) to bound modes (C). Numbers associated with these arrows represent the number of trajectories that started from or reached the respective state. Bound structures from X-ray crystallography[21] are shown on the left. Important binding site residues of neuraminidase are shown in white stick representation.

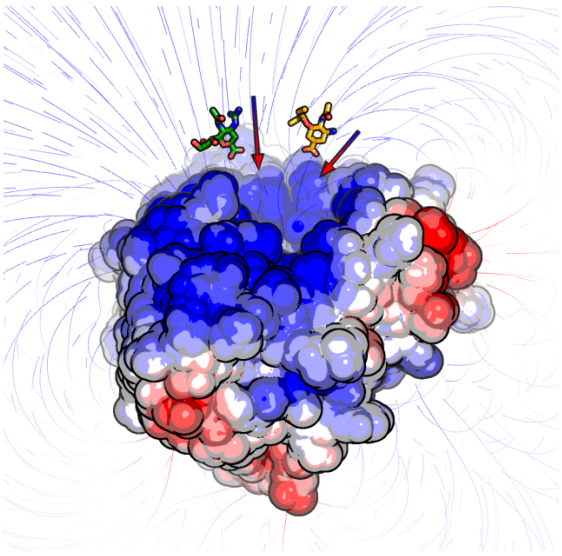


Figure 7.3: Illustration of electrostatic properties of neuraminidase and inhibitors. The surface potential (color coded, blue positive, red negative) and electrostatic field lines of neuraminidase were obtained from Poisson Boltzmann calculations[39]. The field lines guide the two ligands oseltamivir (orange) and zanamivir (green) into the positively charged binding site according to their respective dipole moment, represented as arrows.

confinement of the MD simulations allows for long simulation of the relevant parts of the association pathway and the application of a more generic binding criterion in the form of a minimal residence time. This avoids geometric definitions requiring previous assumptions about the binding and, thus, in contrast to previous approaches, no fitting of a binding criterion to experimental data is required[13, 19, 40, 43].

As expected for two inhibitors with similar diffusion constants, BD simulations yielded equal association rates into the encounter surface (Table 7.1). Due to the change to an atomistic water representation, BD simulations were not directly continued at MD resolution. Instead, MD simulations were started from an independently generated ensemble at the ES (red sphere shell in Figure 7.1). Our MD simulations reveal one single, essential intermediate state in which a salt bridge between the carboxyl group of the inhibitors and Arg368 of neuraminidase is formed. From the salt-bridge state, further association proceeds as a broad ensemble of pathways and is governed by complex short–ranged interactions and rearrangements of the binding site. Interestingly, for the association processes of other receptor-inhibitor complexes, simulation studies indicated multiple distinct intermediate states[8]. Although zanamivir has a higher

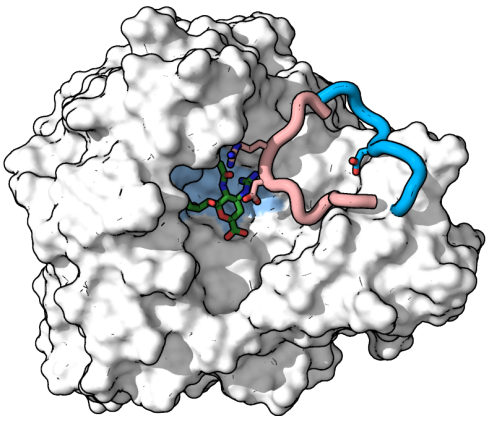


Figure 7.4: Opening of the 150-loop of neuraminidase with bound zanamivir. The closed loop conformation of the X-ray crystal structure[21] (light red cartoon) is shown in comparison with a sampled open loop (blue cartoon) conformation. The bound X-ray crystal structure of zanamivir (green) and contacting side chains of the loop are depicted as sticks, the binding site is colored in light blue.

association rate towards the salt-bridge state due to the guiding effect of its higher dipole moment, oseltamivir has a higher overall on-rate. The probability of oseltamivir reaching a bound state from the intermediate state is about 10 times higher than that of zanamivir. This is in good agreement with experiment and highlights the importance of full atomistic coverage of the final steps of association.

In the case of zanamivir, large scale rearrangements of the 150-loop, also found to be flexible in experiments[21] and other simulations[38], affected key features of the binding site. This is a prime example that the time scales of all relevant conformational rearrangements need to be accessed by the simulation when extracting binding kinetics. We speculate, that a closed loop facilitates binding for oseltamivir as interactions of the ligand with loop residues contribute to binding and do not sterically impede the binding process. For zanamivir, we observed a broader spectrum of bound modes and an opening of the 150-loop in two of four cases. The opening could play an important role in zanamivir association, as the guanidinium group did not bind under the 150-loop residue Asp151 while the loop was closed. The simulation time of 2 µs might be insufficient to capture the full process involving opening and closing. The lowest energy bound mode, possibly a crystal-structure-like orientation of the ligand under a closed loop, might not have been found. While for oseltamivir, the association rate was found to be very robust with respect to the choice of the minimal residence time, for zanamivir the reduction to 1 µs increased the number of bound trajectories by

a factor of two.

The closed loop conformation is overrepresented in the starting ensemble, as all simulations start from the same crystal structure receptor conformation. This could be a reason for the overestimation of the association rate, at least for oseltamivir. In future studies, one could consider using an equilibrated ensemble of receptor starting configurations. Apart from known limitations of fixed charge force fields and water models, it is also possible that the protonation states of residues change during drug molecule association, which could influence the association and has not been covered in the present study.

From the obtained kinetic on-rates, kinetic off-rates or the average thermodynamic residence time can also be obtained when the equilibrium binding affinity is known. Efficient free energy methods exist for the calculation of such absolute binding free energies[44, 45]. The direct calculation of dissociation constants would require simulation times comparable to the mean residence times and is thus computationally unfeasible with current hardware.

The present methodology makes optimal use of the strengths of both BD and MD simulations. In particular, the computationally demanding MD simulations are limited to an essential minimum. Atomistic resolution of the association pathway in the vicinity of the binding site is mandatory to account for side chain rearrangements and solvent molecule effects, while in the bulk state a coarse grained representation is sufficient. Apart from the approximate location of the binding site of neuraminidase, no further knowledge is required. The method complements the available techniques for calculating equilibrium quantities in the drug discovery process by an efficient and highly parallelizable method for the investigation of the dynamics of binding processes on readily available hardware.

7.5 Methods Details

Definitions

Radius r

The coordinate r, used in the BD simulations and to define the MD region, was defined as the center of mass (COM) distance between all receptor atoms and all drug molecule atoms.

Encounter Surface (ES)

The ES was defined as the distance of 12 Å between the COM of the drug molecule and the COM of the C_α atoms of residues Asn248 and Pro431 of the receptor. The COM of

Asn248 and Pro431 lies approximately at the center of the entrance to the binding site.

MD region

The MD region in the form of a cone was defined by the radius r (as above), the angle θ and the dihedral angle φ in combination with the conditions $r < 32\,\text{Å}$, $56^\circ < \theta < 116^\circ$, $65^\circ < \varphi < 125^\circ$. θ was defined as the angle between the COMs of C_α atoms of residues: 97 to 232, 97 to 448, and the COM of the drug molecule. φ was defined as the dihedral angle between the COMs of C_α atoms of residues: 369 to 448, 97 to 232, 97 to 448, and the COM of the drug molecule.

Calculation of the on-rate

The on-rate of the drug molecules was calculated as proposed by Northrup et. al.[13]:

$$k_{\text{on}} = \frac{k_D \beta^\infty \alpha^{\text{MD}}_{\text{ES}\to\text{bound}}}{1 - (1 - \alpha^{\text{MD}}_{\text{ES}\to\text{bound}})\,[\Delta^{\text{BD}} + \beta^\infty(1 - \Delta^{\text{BD}})]} \tag{7.1}$$

with

$$k_D = 4\pi D b \tag{7.2}$$

$$\beta^\infty = \frac{\beta^{\text{BD}}}{1 - (1 - \beta^{\text{BD}})\,\Omega} \tag{7.3}$$

$$\Omega = \frac{b}{q} \tag{7.4}$$

The nomenclature is adopted from the original publication. Here, α refers to the probability of trajectories starting from the ES to not diffuse out of the MD region within $\tau = 2\,\mu\text{s}$. This probability was determined at full atomistic detail by MD simulations (Section 7.5). The probabilities β and Δ were determined by BD simulations (Section 7.5). β is the probability of trajectories starting at $b = 60\,\text{Å}$ to arrive at the ES before diffusing to $q = 100\,\text{Å}$. Δ is the probability of trajectories that left the MD region to re-arrive at the ES before diffusing to b. At b and q, any anisotropy in the system had vanished. D is the relative translational diffusion coefficient of the receptor and respective ligand (Table 7.3).

$k_D\beta^\infty = k^{\text{BD}}_{\infty\to\text{ES}}$ can be interpreted as the rate into the ES. Table 7.2 shows the obtained values for oseltamivir and zanamivir.

Brownian dynamics simulations

Brownian dynamics simulations were carried out on the ligand-receptor system to calculate the diffusion dominated contribution to the association rate from the bulk

Molecule	$\alpha^{\mathrm{MD}}_{\mathrm{ES}\to\mathrm{bound}}$	β^{BD}	Δ^{BD}	$k^{\mathrm{BD}}_{\infty\to\mathrm{ES}}$ [$1/\mu$Ms]	k_{on} [$1/\mu$Ms]
Oseltamivir	0.0281	0.0085 [0.0083,0.0088]	0.108 [0.105,0.112]	794 [776,822]	25.5
Zanamivir	0.0066	0.0082 [0.0080,0.0084]	0.105 [0.102,0.109]	737 [519,754]	5.5

Table 7.2: Transition probabilities from MD and BD simulations and resulting total rate constants. The errors of β^{BD}, Δ^{BD} and $k^{\mathrm{BD}}_{\infty\to\mathrm{ES}}$ are 95% confidence intervals obtained by the bootstrap method. A statistical error estimate of $\alpha^{\mathrm{MD}}_{\mathrm{ES}\to\mathrm{bound}}$ and thus the total on-rate k_{on} is not meaningful due to the small number of sampled events.

into the ES. In the simulations, the drug molecule and the receptor are treated as rigid structures that undergo translational and rotational diffusion[13]. The diffusion coefficients were calculated according to the Stokes-Einstein equation, using the radius of gyration of the molecules as resulting from the MD simulations (Table 7.3).

Molecule	R_{gyr} [Å]	D_{rot} [$\mathrm{rad^2\,ns^{-1}}$]	D_{trans} [$\mathrm{nm^2\,ns^{-1}}$]
Oseltamivir	3.6	4.0	0.70
Zanamivir	3.8	3.5	0.67
Neuraminidase	19.4	0.026	0.13

Table 7.3: Diffusion constants calculated according to the Stokes-Einstein relation, based on the radius of gyration of the molecules.

The propagation time step was 100 ps. Propagation steps leading to collisions between the molecules were rejected. Electrostatic interactions were taken into account by calculating the force acting on the drug molecule in the electrostatic field of the receptor. The electrostatic potential of the receptor was calculated according to the Poisson-Boltzmann equation using APBS[35] at a temperature of 310 K. For the simulations, the electrostatic potential was pre-calculated and stored in a grid with a spacing of 0.33 Å. The influence of the drug molecule on the electrostatic field was neglected. Hydrodynamic interactions between the protein and the drug molecule were neglected.

To calculate the probability β, for both drug molecules 10^6 trajectories were generated starting from a random configuration at a radius of $r = b = 60$ Å from the receptor and terminated when the drug molecules diffused to radii larger then $r > q = 100$ Å or came within the ES of the binding site (Figure 7.5).

To calculate the probability Δ, for both drug molecules a total of 10^6 trajectories were started uniformly from the ≈ 600 configurations at which the respective MD trajectories had been stopped and were terminated when the drug molecules re-arrived at the ES or diffused to radii $r > b$ (Figure 7.5).

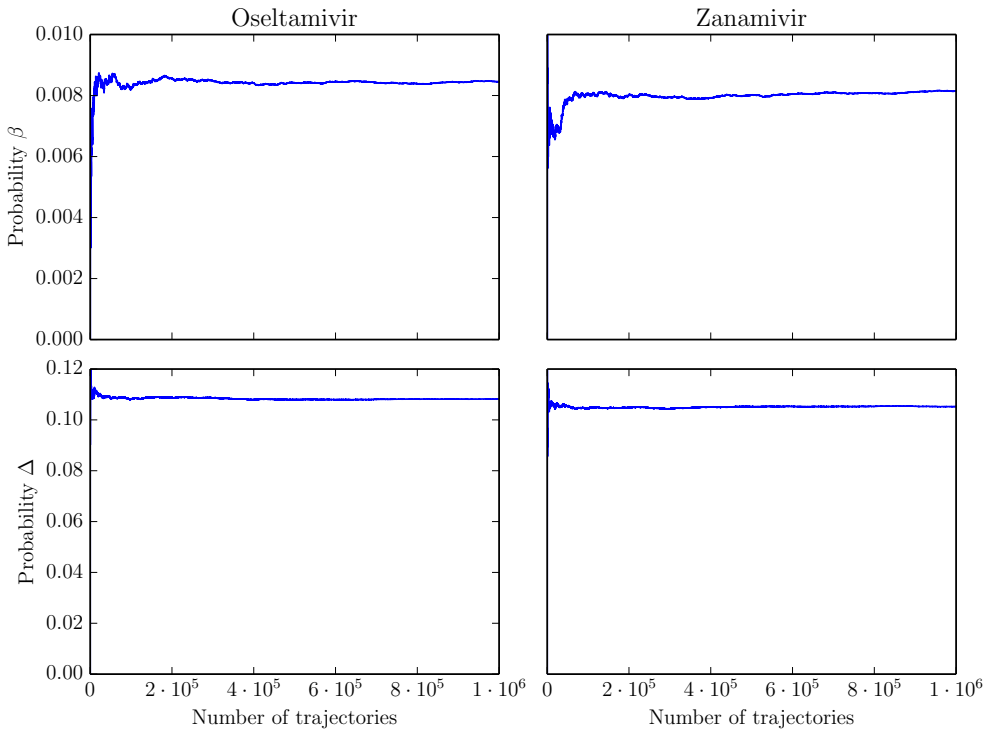


Figure 7.5: Cumulative mean values of the probabilities β and Δ obtained from BD simulations for oseltamivir and zanamivir.

Molecular dynamics simulations

System preparation

Neuraminidase of pathogen H1N1 virus A/California/07/09 (Cal07) [21] was used for the atomistic MD simulations.The apo-configuration of neuraminidase is only available for a mutant (PDBCode 4B7M) [21]. The wild type amino acid sequence of H1N1 Cal07 was restored by in silico reverse mutating residues Ile106Val, Arg223Ile, and Asp248Asn in the starting structure. Residual phosphate ions and mono saccharides were removed from the structure but the crystallographic water and one CA^{2+} ion bound close to the receptor site was conserved. To reduce the computational overhead only one monomer

of the tetrameric neuraminidase complex was used, the others were removed. The protonation state of neuraminidase was calculated with karlsberg+ software package [27, 28] for both complexes with and without ligand bound. It was found that the protonation state of residue Glu119 which is located directly in the ligand binding site changes from unprotonated to protonated upon ligand association. As an on-the-fly adaption of protonation states during the MD simulations was not within the scope of this study, the protonation of the bound form was chosen for the MD topologies. Topologies for the drug molecules oseltamivir/zanamivir were parametrized in the GAFF force field[30] with the antechamber package[31]. Calcium ion parameters were taken from Bradbrook et al. [29]. The protein was solvated with the TIP3P[32] explicit water model in a periodic box and sodium (1) and chloride (5) ions were added to render the net charge neutral. Energy minimization was performed for 2000 steps with the steepest descent algorithm on the box and subsequently the system was equilibrated in the NPT ensemble for 2 ns with a time step of 0.2 fs. The reference temperature of 310 K was controlled with the Berendsen thermostat and a pressure of 1.01 bar was adapted with the Berendsen barostat[33]. A cut-off radius of 0.9 nm was used and all solute heavy atoms were position restrained during the equilibration phase. MD simulations and energy minimization were performed with the Amber14 package [23].

Generation of MD ensembles at the ES

Starting structures for the MD simulations were generated by placing the drug molecules in the MD region (see Section 7.5) at distances of 31 Å. Trajectories were propagated from these starting structures using a non-deterministic Langevin dynamics integrator with a collision frequency of $5\,\mathrm{ps}^{-1}$. As long as the respective drug molecule resided within the MD region, the configurations from these trajectories were saved every 10 ps. This procedure was repeated until ≈ 5000 configurations had been generated for both inhibitors.

The resulting configurations were propagated until they left the MD region or entered the ES. This resulted in an ensemble of ≈ 600 configurations for each drug molecule with the drug molecule located at the ES. Further association simulations were started from these ensembles.

Drug molecule association simulations

The hydrogen mass repartitioning scheme[25] and the SHAKE method[24] were applied to the non-solvent hydrogen atoms. This allowed for a MD integration time step of 4 fs. The simulations were performed in the NVT ensemble using the Andersen thermostat[34] with a collision frequency of $2\,\mathrm{ps}^{-1}$.

Bound mode determination via clustering

Stable bound modes were extracted by means of a clustering approach. For each ligand, all trajectories that resulted in a bound drug molecule were concatenated. The concatenated trajectory with a time resolution of 1 ns was position aligned with respect to the β-sheet forming protein backbone of neuraminidase. The drug molecules poses were then clustered based on their respective all atom RMSD with a 0.15 nm cutoff and the gromos method[36] as implemented in GROMACS v4.6.5[37].

7.6 Bibliography

[1] Robert A. Copeland, David L. Pompliano, and Thomas D. Meek. “Drug-target residence time and its implications for lead optimization.” In: *Nat. Rev. Drug Discov.* 5.9 (2006), pp. 730–739.

[2] Albert C. Pan, David W. Borhani, Ron O. Dror, and David E. Shaw. “Molecular determinants of drug–receptor binding kinetics.” In: *Drug Discov. Today* 18.13–14 (2013), pp. 667–673.

[3] Manuel P Luitz and Martin Zacharias. “Protein–Ligand Docking Using Hamiltonian Replica Exchange Simulations with Soft Core Potentials.” In: *J. Chem. Inf. Model.* 54.6 (2014), pp. 1669–1675.

[4] Hideaki Fujitani et al. “Direct calculation of the binding free energies of FKBP ligands.” In: *J. Chem. Phys.* 123.8, 084108 (2005), p. 084108.

[5] Ron O. Dror et al. “Pathway and mechanism of drug binding to G-protein-coupled receptors.” In: *Proc. Natl. Acad. Sci. USA* 108.32 (2011), pp. 13118–13123.

[6] Ron O. "Dror et al. “"Structural basis for modulation of a G-protein-coupled receptor by allosteric drugs".” In: *”Nature”* "503" ("2013"), "295–299".

[7] Ignasi Buch, Toni Giorgino, and Gianni De Fabritiis. “Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations.” In: *Proc. Natl. Acad. Sci. USA* 108.25 (2011), pp. 10184–10189.

[8] Nuria Plattner and Frank Noe. “Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models.” In: *Nat. Commun.* 6 (2015).

[9] Donald L. Ermak and J. A. McCammon. “"Brownian dynamics with hydrodynamic interactions".” In: *”J. Chem. Phys.”* "69"."4" ("1978"), "1352–1360".

[10] Huan Xiang Zhou. “Kinetics of diffusion-influenced reactions studied by Brownian dynamics.” In: *J. Phys. Chem.* 94.25 (1990), pp. 8794–8800.

[11] Huan-Xiang Zhou. "Brownian dynamics study of the influences of electrostatic interaction and diffusion on protein-protein association kinetics." In: *Biophys. J.* 64.6 (1993), p. 1711.

[12] Sanbo Qin, Xiaodong Pang, and Huan-Xiang Zhou. "Automated prediction of protein association rate constants." In: *Structure* 19.12 (2011), pp. 1744–1751.

[13] Scott H. Northrup, Stuart A. Allison, and J. Andrew McCammon. "Brownian dynamics simulation of diffusion-influenced bimolecular reactions." In: *J. Chem. Phys.* 80.4 (1984), pp. 1517–1524.

[14] Xiaofeng Yu et al. "webSDA: A web server to simulate macromolecular diffusional association." In: *Nucleic Acids Res.* 43.W1 (2015), W220–W224.

[15] Rebecca C Wade et al. "Simulation of enzyme-substrate encounter with gated active sites." In: *Nat. Struct. Mol. Biol.* 1.1 (1994), pp. 65–69.

[16] Alexander Spaar, Christian Dammer, Razif R Gabdoulline, Rebecca C Wade, and Volkhard Helms. "Diffusional encounter of barnase and barstar." In: *Biophys. J.* 90.6 (2006), pp. 1913–1924.

[17] Brock A Luty, Samir El Amrani, and J Andrew McCammon. "Simulation of the bimolecular reaction between superoxide and superoxide dismutase: synthesis of the encounter and reaction steps." In: *J. Am. Chem. Soc.* 115.25 (1993), pp. 11874–11877.

[18] Lane W. Votapka and Rommie E. Amaro. "38 Multiscale estimation of binding kinetics using molecular dynamics, brownian dynamics, and milestoning." In: *J. Biomol. Struct. Dyn.* 33.sup1 (2015), pp. 26–27.

[19] Ali S. Saglam and Lillian T. Chong. "Highly Efficient Computation of the Basal kon using Direct Simulation of Protein-Protein Association with Flexible Molecular Models." In: *J. Phys. Chem. B* 120.1 (2016), pp. 117–122.

[20] Ramzi Alsallaq and Huan-Xiang Zhou. "Electrostatic rate enhancement and transient complex of protein–protein association." In: *Proteins* 71.1 (2008), pp. 320–335.

[21] Erhard Van Der Vries et al. "H1N1 2009 pandemic influenza virus: resistance of the I223R neuraminidase mutant explained by kinetic and structural analysis." In: *PLoS Pathog.* 8.9 (2012), e1002914.

[22] Romelia Salomon-Ferrer, Andreas W Götz, Duncan Poole, Scott Le Grand, and Ross C Walker. "Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald." In: *J. Chem. Theory Comput.* 9.9 (2013), pp. 3878–3888.

[23] D.A. Case et al. *AMBER 14*. University of California, San Francisco, 2014.

[24] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman J.C Berendsen. "Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes." In: *J. Comput. Phys.* 23.3 (1977), pp. 327–341.

[25] Chad W. Hopkins, Scott Le Grand, Ross C. Walker, and Adrian E. Roitberg. "Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning." In: *J. Chem. Theory Comput.* 11.4 (2015), pp. 1864–1874.

[26] Kresten Lindorff-Larsen et al. "Improved side-chain torsion potentials for the Amber ff99SB protein force field." In: *Proteins* 78.8 (2010), pp. 1950–1958.

[27] Björn Rabenstein and Ernst-Walter Knapp. "Calculated pH-dependent population and protonation of carbon-monoxy-myoglobin conformers." In: *Biophys. J.* 80.3 (2001), pp. 1141–1150.

[28] Gernot Kieseritzky and Ernst-Walter Knapp. "Optimizing pKa computation in proteins with pH adapted conformations." In: *Proteins* 71.3 (2008), pp. 1335–1348.

[29] Gail M Bradbrook et al. "X-Ray and molecular dynamics studies of concanavalin-A glucoside and mannoside complexes Relating structure to thermodynamics of binding." In: *J. Chem. Soc. Faraday. T.* 94.11 (1998), pp. 1603–1611.

[30] Junmei Wang, Romain M Wolf, James W Caldwell, Peter A Kollman, and David A Case. "Development and testing of a general amber force field." In: *J. Comput. Chem.* 25.9 (2004), pp. 1157–1174.

[31] Junmei Wang, Wei Wang, Peter A Kollman, and David A Case. "Automatic atom type and bond type perception in molecular mechanical calculations." In: *J. Mol. Graph. Model.* 25.2 (2006), pp. 247–260.

[32] William L. Jorgensen, Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. "Comparison of simple potential functions for simulating liquid water." In: *J. Chem. Phys.* 79.2 (1983), pp. 926–935.

[33] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. "Molecular dynamics with coupling to an external bath." In: *J. Chem. Phys.* 81.8 (1984), pp. 3684–3690.

[34] Hans C. Andersen. "Molecular dynamics simulations at constant pressure and/or temperature." In: *J. Chem. Phys.* 72.4 (1980), pp. 2384–2393.

[35] Nathan A. Baker, David Sept, Simpson Joseph, Michael J. Holst, and J. Andrew McCammon. "Electrostatics of nanosystems: Application to microtubules and the ribosome." In: *Proc. Natl. Acad. Sci. USA* 98.18 (2001), pp. 10037–10041.

[36] Xavier Daura et al. “Peptide folding: when simulation meets experiment.” In: *Angew. Chem. Int. Edit.* 38.1-2 (1999), pp. 236–240.

[37] Berk Hess, Carsten Kutzner, David Van Der Spoel, and Erik Lindahl. “GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation.” In: *J. Chem. Theory Comput.* 4.3 (2008), pp. 435–447.

[38] Rommie E. Amaro et al. “Remarkable Loop Flexibility in Avian Influenza N1 and Its Implications for Antiviral Drug Design.” In: *J. Am. Chem. Soc.* 129.25 (2007), pp. 7764–7765.

[39] Nathan A Baker, David Sept, Simpson Joseph, Michael J Holst, and J Andrew McCammon. “Electrostatics of Nanosystems: Application to Microtubules and the Ribosome.” In: *Proc. Natl. Acad. Sci. USA* 98.18 (2001), pp. 10037–10041.

[40] Jeffrey C. Sung, Adam W. Van Wynsberghe, Rommie E. Amaro, Wilfred W. Li, and J. Andrew McCammon. “Role of Secondary Sialic Acid Binding Sites in Influenza N1 Neuraminidase.” In: *J. Am. Chem. Soc.* 132.9 (2010), pp. 2883–2885.

[41] Gerhard Klebe. “Applying thermodynamic profiling in lead finding and optimization.” In: *Nat. Rev. Drug Discov.* 14.2 (2015), pp. 95–110.

[42] John E. Ladbury. “Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design.” In: *Chem. Biol.* 3.12 (1996), pp. 973–980.

[43] Ly Le, Eric H. Lee, David J. Hardy, Thanh N. Truong, and Klaus Schulten. “Molecular Dynamics Simulations Suggest that Electrostatic Funnel Directs Binding of Tamiflu to Influenza N1 Neuraminidases.” In: *PLoS Comput. Biol.* 6.9 (Sept. 2010), pp. 1–13.

[44] Hyung-June Woo and Benoit Roux. “Calculation of absolute protein-ligand binding free energy from computer simulations.” In: *Proc. Natl. Acad. Sci. USA* 102.19 (2005), pp. 6825–6830.

[45] Michael S Lee and Mark A Olson. “Calculation of absolute protein-ligand binding affinity using path and endpoint approaches.” In: *Biophys. J.* 90.3 (2006), pp. 864–877.

8 Nucleobase Stacking Thermodynamics and Kinetics from MD Simulations

We present simulations of *N*6*N*9-dimethyladenine aggregation yielding statistically converged stacking thermodynamics and kinetics. The simulations confirm the validity of the random isodesmic stacking model without cooperative effects. The rigid-body TIP3P water model leads to significantly accelerated dynamics and underestimation of the entropy contribution. Inclusion of full flexibility of the water atoms by the SPCFW model yields values in close agreement with experimental equilibrium constant, entropy contribution as well as, importantly, kinetic observables. In addition, we report stacking kinetics based upon dimerization simulations using different commonly used water models and simulation techniques.

8.1 Introduction

Evaluation of kinetic properties for example of drug molecule binding[1] or self-aggregation processes by MD simulations becomes increasingly important[2]. Mean residence times of molecules in binding sites or residence times of intermediate states of proteins which cannot be observed experimentally have been and are being investigated using MD simulations. At the same time, little statistically relevant data on the reliability of kinetic predictions by MD simulations is available. MD force-fields are usually parametrized and validated via equilibrium values like solvation free energies[3], and except for the self diffusion of water models[4], dynamic properties have hardly been compared to experiments. Reliable kinetic data is in general difficult to obtain, as in contrast to equilibrium values, the real time scales of transitions have to be sampled a statistically relevant number of times.

To investigate MD predictions of binding kinetics, we chose a self aggregating system with available experimental data and associated timescales that can be covered by current MD simulation capabilities. We present 2 µs simulations of ensembles of 25 residues of the adenine derivate *N*6*N*9-dimethyladenine, which allow statistically converged analysis of the stacking kinetics. The stacking of this dimethyladenine derivate was previously investigated experimentally based on the random isodesmic model[5][6], yielding an equilibrium stacking constant by evaluation of the stack

concentration, and kinetic dissociation and recombination rates by evaluation on the relaxation time of the stack concentration. These observables can also be obtained from MD simulations, allowing direct comparison of equilibrium distributions and kinetics with the experiments. Simulation at different temperatures also allowed determination of the entropy of stack formation. The simulations indicate that flexibility of the water model is a crucial factor for correct predictions of stacking entropy contribution and stacking kinetics.

8.2 Random Isodesmic Stacking Model

In experiments using NMR[5], vapor pressure osometry or sound absorption[6][7], base-stacking thermodynamics were analyzed using the random isodesmic stacking model. In this model, all stacking interactions are considered to be equivalent with one single equilibrium stacking constant K[6]

$$A_i A_j K = A_{i+j} \tag{8.1}$$

and multimer concentrations

$$A_i = A_1^i K^{i-1},\ A_1 K < 1, \tag{8.2}$$

where A_i is the concentration of $N6N9$-dimethyladenine multimers consisting of i residues, involving $i-1$ stacks. Furthermore we note

$$C_0 = \sum_{i=1}^{\infty} A_i i,\ C_m = \sum_{i=1}^{\infty} A_i,\ C_s = \sum_{i=1}^{\infty} A_i (i-1), \tag{8.3}$$

where C_0 is the concentration of $N6N9$-dimethyladenine residues, C_m is the total concentration of multimers (including monomers) and C_s is the concentration of formed stacking interactions. Experiments found the stacking behavior of $N6N9$-dimethyladenine to be in reasonable agreement with the isodesmic stacking model.

From equilibrium observations alone, a sequential isodesmic stacking behavior, in which stacks can only be formed or broken at the ends of the multimers, cannot be distinguished from random isodesmic stacking behavior, in which stacks can be formed between any multimers or be broken at any position[6]. The random isodesmic stacking model however predicts one single relaxation time τ according to[6]

$$1/\tau = C_m K_R + 2K_D, \tag{8.4}$$

where K_R and K_D are, respectively, the recombination and dissociation rate constants, while the sequential isodesmic stacking model predicts a broad spectrum of relaxation times. Sound absorption measurements of the relaxation behavior found only one dominant relaxation time, indicating random isodesmic stacking behavior.

8.3 Methods

Force fields

The *N*6*N*9-dimethyladenine structure was adopted from the ff14SB[8] adenonsine residue library entry. The ff14SB force field was used for the *N*6*N*9-dimethyladenine molecule. The point charges were obtained by a semi-empirical qm calculation on the BCC level using the antechamber module of Amber14[9]. The TIP3P[10], TIP4P[11], TIP5P[12], SPC/E[13] and SPCFW[4] water models were used as implemented in Amber14.

Simulation parameters

All simulations were carried out with Amber14[9]. The short-range interactions were cut off at 9 Å, the long-range interactions were treated with the PME method using the standard Amber14 parameters. The *N*6*N*9-dimethyladenine hydrogen atoms were constrained with the SHAKE algorithm [14] and modified by hydrogen mass repartitioning (HMR)[15] for the simulations using a 4 fs time step. The water molecules were not modified by HMR. Only the SHAKE algorithm was used for the 2 fs time step simulations. For the simulations in which all hydrogen atoms were fully flexible, a 1 fs time step was used. For minimization, the steepest decent method was employed. For equilibration, the langevin thermostat[16] was used with a collision frequency of $5\,ps^{-1}$. In the subsequent simulations, the temperature was controlled with the Andersen thermostat[17], using a collision frequency of $5\,ps^{-1}$. For a comparative simulation, the Berendsen thermostat[18] was used with a coupling time of 1 ps. The pressure of 1 bar was regulated with the Berendsen barostat in all simulations, using a coupling time of 1 ps. The nucleobase COM separations were saved at intervals of 10 ps.

Bulk simulations

For the bulk simulations, 25 *N*6*N*9-dimethyladenine molecules were randomly placed in a box with edge length 50 Å, imposing a minimum inter-molecule distance of 3 Å. The systems were solvated with $\approx$ 8000 TIP3P or SPCFW water molecules, energy minimized for 1000 steps and equilibrated at 298 K or 313 K for 40 ps. The four simulations were then run for 2 µs. A stacking interaction was defined to be established for center of mass distances smaller than 5 Å. If more than two inter-molecular distances were within the stacking threshold for one molecule, only the two closest interactions were defined as stacking interactions. The statistical errors of the autocorrelation functions and equilibrium concentrations were calculated by the block bootstrap method[19] in the form of 95% confidence intervals, using a block

size corresponding to the relaxation times of the respective simulations. The statistical errors of the relaxation times and rates correspond to to exponential fits to the total lower/upper CIs of the autocorrelation functions. The statistical errors of the isodesmic model are standard deviations, treating the CIs as individual uncertainties in a non-linear least square fit. Snapshots were prepared using pymol[20].

Dimerization simulations

For each investigated water model and simulations setup, 5000 simulations were run. The starting structure of each simulation was generated by placing two *N*6*N*9-dimethyladenine structures with random relative orientation at a distance of $r_{\text{start}} = 21\,\text{Å}$. The systems were solvated with ≈ 3000 water molecules and subsequently minimized for 1000 steps and equilibrated for 40 ps with a harmonic distance restraint of $k = 5\,\text{kcal/mol/Å}^2$, keeping the two molecules at the distance of r_{start} during the equilibration. The systems were then simulated until the center of mass separation of the two molecules was larger than $r_{\text{max}} = 26\,\text{Å}$. This lead to the simulation of at least 750 binding events for all simulation setups. The simulated trajectories within separations of 21 Å correspond to equilibrium conditions, as every trajectory that entered the corresponding phase space was simulated until leaving it.

For the calculation of the equilibrium kinetic rates, the bound state was defined by nucleobase center of mass separations smaller than 5 Å. Molecules were defined to be in the unbound state between distances of 12 Å and 21 Å. To obtain kinetic rates and equilibrium constants corresponding to standard states, the residence times in the unbound state were scaled by a factor $(1/(c_0 V_{\text{unbound}}))$ that accounts for the volume accessible at the standard state concentration c_0 of 1 M.

95% confidence intervals of the kinetic rates and equilibrium constants of the dimerization simulations were calculated by standard bootstrapping, as the observables are based on independent simulations which are not correlated.

8.4 Results and Discussion

Multimerization simulations

We simulated a system of 25 *N*6*N*9-dimethyladenine molecules at a concentration of $C_0 = 0.16\,\text{M}$ at 298 K, which is within in the concentration and temperature range of the experiments, using the widely used rigid body TIP3P[10] water model (with HMR and SHAKE) and the fully flexible SPCFW[4] water model. In the simulations, which were started from a random arrangement of the 25 monomers, multimers form and dissociate spontaneously (Fig. 8.1).

		TIP3P	SPCFW	Osom./Sound[6]	NMR[5]
ΔH	[kcal/mol]	-5.30 ± 0.14	-7.3 ± 0.9	-8.7 ± 1.5	-7.2 ± 0.6
ΔS	[cal/mol/K]	-12.8 ± 0.5	-17.0 ± 3.0	-21.6 ± 3	-17.9 ± 1.8
				298K	
K	[1/M]	12.30 ± 0.05	41.6 ± 1.1	45.6	23.6
τ	[ns]	0.9 [0.8, 1.2]	4.9 [4.2, 6.2]		
k_R	[10^8/Ms]	46 [33, 54]	16 [12, 20]	9.3	
k_D	[10^7/s]	38 [28, 44]	3.8 [2.8, 4.6]	5.0	
				313K	
K	[1/M]	8.01 ± 0.05	23.1 ± 0.5	18.0	15.0
τ	[ns]	0.8 [0.7, 1.1]	2.7 [2.3, 3.8]		
k_R	[10^8/Ms]	41 [29, 49]	21 [14, 26]	20	
k_D	[10^7/s]	51 [37, 61]	9.3 [6.3, 11]	6.6	

Table 8.1: Thermodynamic quantities of stack formation obtained from 25 residue ensemble MD simulations with the TIP3P and the SPCFW water models. Comparison to equilibrium quantities obtained from vapor pressure osometry[6] and NMR spectroscopy[5], and kinetic data obtained by sound absorption measurements[6]. All measured quantities are based on the random isodesmic stacking model. Simulations were performed at a residue concentration of $C_0 = 0.16\,\mathrm{M}$, experiments were performed within a residue concentration range of $C_0 = 0.05\,\mathrm{M} - 0.5\,\mathrm{M}$. Values in square brackets are statistical 95% confidence intervals, the errors for the equilibrium values are standard deviations as resulting from fits of the isodesmic stacking model.

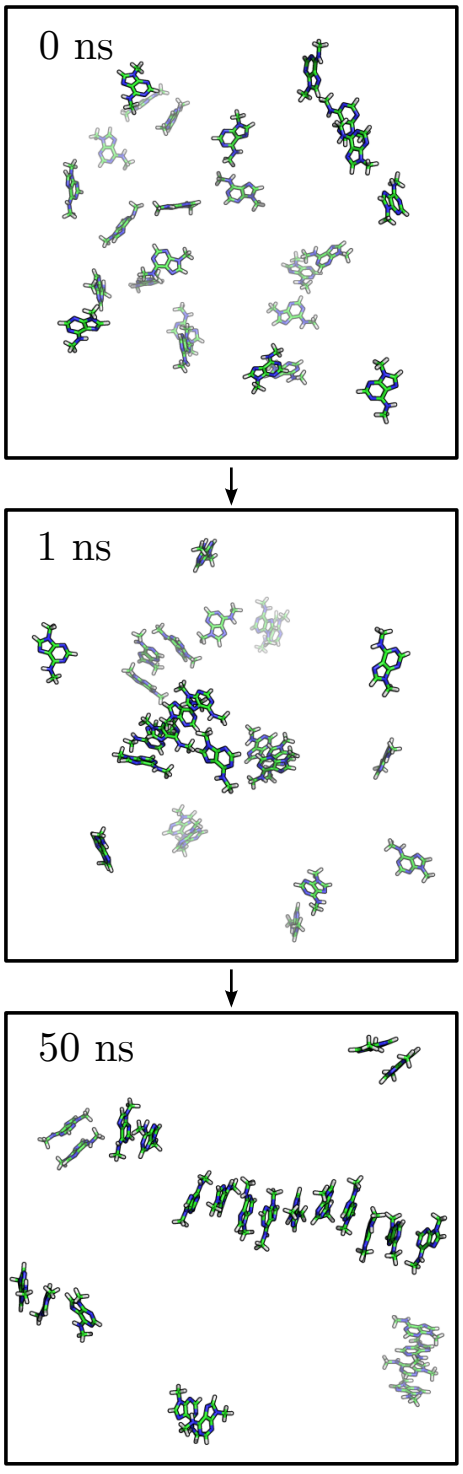


Figure 8.1: Trajectory snapshots of the 2 µs 0.16 M *N*6*N*9-dimethyladenine simulation with the fully flexible SPCFW water model at 298 K, showing spontaneous multimerization of the residues.

For a direct comparison of the simulations with experiments, we also used the random isodesmic stacking model for analysis. A stacking interaction was defined to be established at a nucleobase center-of-mass distance of less than 5 Å, which is about the range of sensitivity of NMR distance measurements. While in the experiments, individual concentrations of the multimers of different order can usually not be observed and several experiments at different residue concentrations C_0 have to be performed, in MD-simulations, they can directly be extracted and simulation of the system at only one residue concentration C_0 is sufficient. The equilibrium constant K was obtained by fitting the isodesmic model to the average multimer concentrations (Fig. 8.2). Observation of the cumulative average multimer concentration C_m (Figure 8.3) shows excellent convergence of the equilibrium properties within the simulation time of 2 µs. The statistical errors are larger for the SPCFW water model due to its associated slower relaxation times (see below).

As in the experiments, despite its simplicity, the isodesmic model appears to reasonably well describe the equilibrium distribution of the different multimer orders. We observe a slight underestimation of the monomer concentration using the SPCFW water model, but no significant (anti-)cooperativity could be observed. Experimental analysis taking cooperativity into account did also not find (anti-)-cooperativity within experimental error. While the TIP3P water model underestimates the equilibrium stacking constant K, the SPCFW water model predicts a value lying between the experimental result of a NMR study and a study using vapor pressure osometry (Table 8.1).

To obtain the enthalpy and entropy contributions to the equilibrium constant K, the systems were also simulated at 313 K. Entropy and enthalpy were separated using the van't Hoff equation, which was also done in the experiments. While the TIP3P model underestimates the enthalpy as well as the negative entropy change (which largely cancels out), the entropy contribution is well predicted by the flexible SPCFW model. This is not an unexpected result, as rigid water models neglects a large amount of degrees of freedom in the water bulk.

To extract kinetic rates, in the experiment, the relaxation behavior was measured by ultra-sound absorption, and found to be in reasonable agreement with only one relaxation time. The relaxation time was extracted from the MD simulations using the auto-correlation function of the stack concentration, following the Onsager fluctuation theorem. As observed in sound-absorption experiments, the relaxation behavior can be reasonably well described by a single relaxation time (Fig. 8.4). We observe a quasi instantaneous decay of the stack concentration auto-correlation within the time scale of the used trajectory time resolution of 10 ps to about 0.6-0.8. We attribute this decay to fast fluctuation of the stacking concentrations due to the numerical stacking threshold of 5 Å and therefore used a scaling factor for the exponential fit. The relaxation time and the rates predicted by the TIP3P model (with HMR) are up to one order of

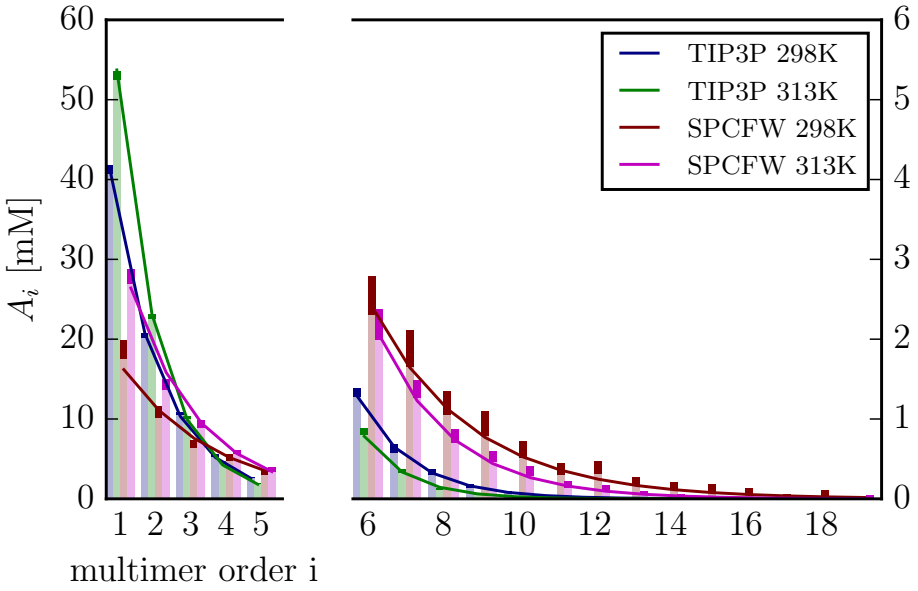


Figure 8.2: Average concentrations A_i of multimers of order i, at a residue concentration of $C_0 = 0.16\,\mathrm{M}$. Dark bars indicate 95% confidence intervals. The lines are the fits of the isodesmic stacking model.

magnitude faster than the experimental results. The relaxation time resulting from the SPCFW model is only about a factor of 1.5 shorter than the experimental value, which remarkably leads to kinetic rates which agree with the experimental values (factor 2) within the errors.

The simulations allow for a detailed analysis of the stacking configurations. Figure 8.5 shows that relative fluctuations of the residues mainly occur along translations or rotations parallel to the base plane. Principal stacking and dissociation pathways are likely to follow these low free energy regions. Principal relative orientations of two stacked residues are dominated by the positions of the methyl groups. No significant population of configurations occurring in B-DNA (about 34° relative rotation) could be observed.

Dimerization simulations

The two water models and non-water hydrogen treatments used for the bulk simulations showed significant differences with respect to equilibrium and kinetic properties. For a detailed comparison of a larger set of water models and MD parameters, we chose to

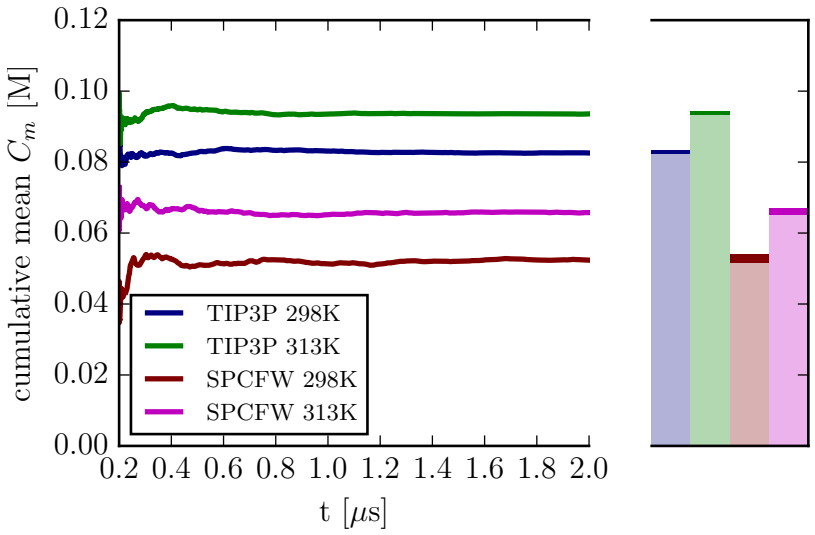


Figure 8.3: Cumulative average multimer (including monomers) concentration C_m, indicating excellent convergence of the equilibrium characteristics. Dark bars indicate 95% confidence intervals.

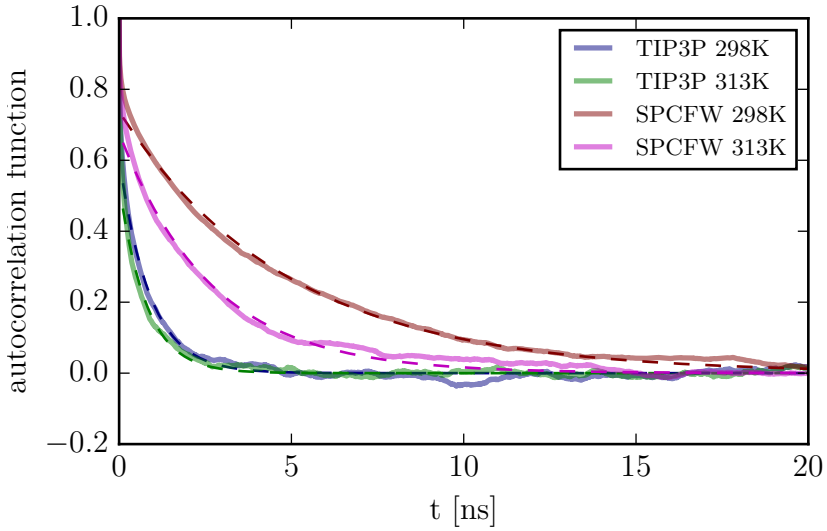


Figure 8.4: Autocorrelation function of the stack-concentration C_s at a residue concentration of $C_0 = 0.16\,\mathrm{M}$. The lines are fits to an exponential decay $ce^{t/\tau}$ with scaling factor c.

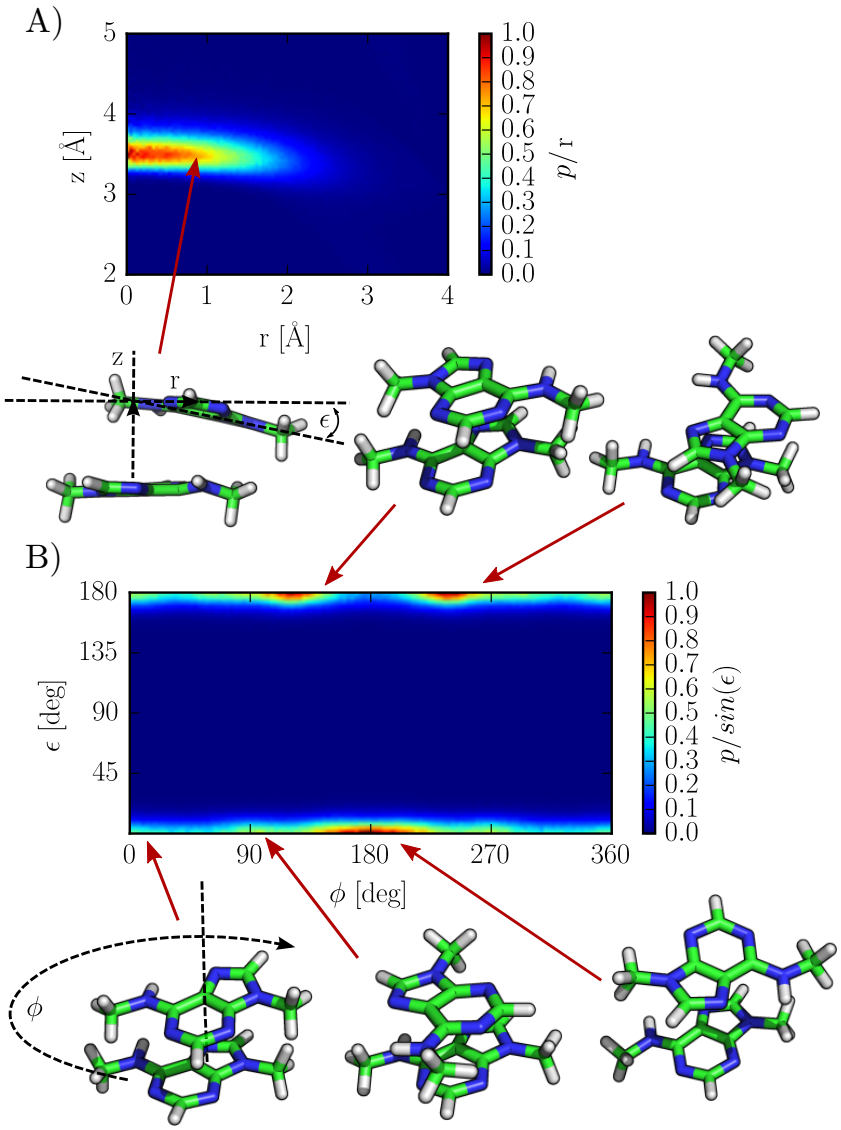


Figure 8.5: Configuration probability histograms from the SPCFW simulation at 298 K projected on relative A) vertical (z) and horizontal (r) translation and B) torsion (ϕ) and tilt (ϵ) rotation of two stacking residues. Trajectory snapshots illustrate prevalently populated configurations.

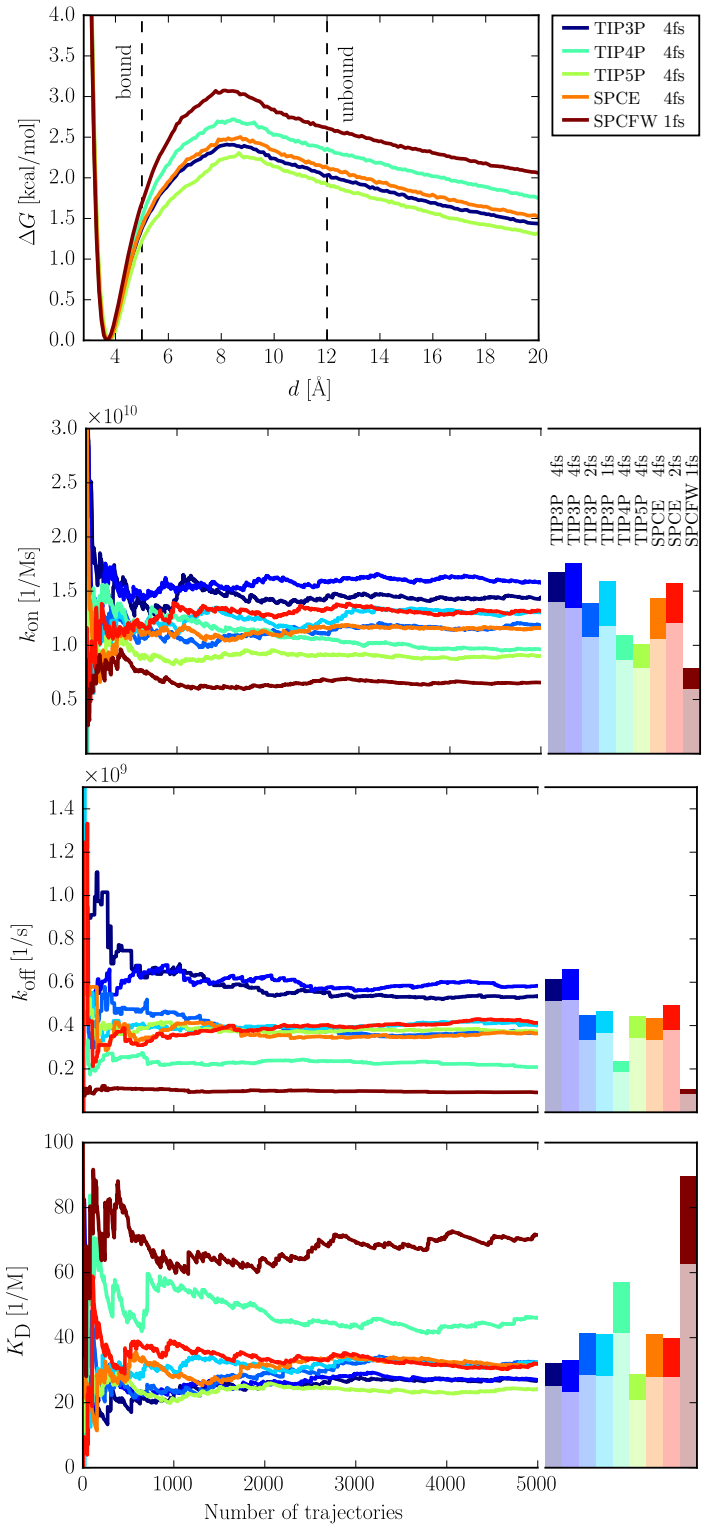


Figure 8.6: Dimerization simulations with different water models. The bound state was defined in these simulations by residue distances smaller than 5 Å, the unbound state by residue distances larger than 12 Å. For each water model, 5000 trajectories were started at a residue-residue distance of 21 Å and terminated when the distance became larger than 26 Å. This lead to the simulation of at least 750 stacking events, yielding statistically converged binding and dissociation rates. The used MD time step corresponds to different simulation conditions: HMR and SHAKE applied to non-water hydrogen atoms (4 fs), SHAKE applied to non-water hydrogen atoms (2 fs), fully flexible non-water hydrogen atoms (1 fs). All water models except SPCFW are rigid body water models.

limit the simulations to computationally easier affordable dimerization reactions. (This is usually not possible in experiments, as very low concentrations would be required) Instead of a single bulk ensemble simulation, independent trajectories were simulated in parallel for the dimerization.

We observe that with the same state definitions (stacked for center of mass distance smaller than 5 Å), the equilibrium constants for *N*6*N*9-dimethyladenine dimerization are significantly higher for TIP3P (24 /M) and SPCFW (60 /M) than the equilibrium constants resulting from the isodesmic model bulk simulations. This is not in contradiction to the observed equivalence of different stacking interactions according to the isodesmic stacking model, but can be attributed to the fact that in the bulk, beyond the threshold of 5 Å, favorable interactions with other residues exists that lower the free energy of the unstacked state (see PMF, attraction up to ca. 10 Å). It shows, however, that the equilibrium constant resulting from the conditions in the bulk experiments/simulations are not completely concentration independent and do not correspond to true standard binding free energies. Thus, as the dimerization simulations and the bulk simulations/experiments are not directly comparable, for a relative comparison of the water models we chose to analyze the dimerization with respect to the more general standard binding free energy[21] and corresponding kinetic rates of two residues (stacked state closer than 5 Å, unstacked state larger than 12 Å, where attractive interactions are negligible, see PMF). The kinetic dimerization rates were now defined in the standard way according to

$$k_{\text{on}} = \frac{N}{T_{\text{bulk}}}, \quad k_{\text{off}} = \frac{N}{T_{\text{bound}}}, \tag{8.5}$$

where N is the number of observed transitions, T_{bulk} is the total time spent in the standard state and T_{bound} is the total time spend in the bound/stacked state.

Figure 8.6 shows the Potential of Mean force (PMF), equilibrium binding constants and on- and off-rates of *N*6*N*9-dimethyladenine dimerization for the set of tested water models and simulation parameters. The cumulative average values indicate statistical convergence of the thermodynamic properties within 5000 simulations (observing ca 750 binding events). The used MD time step corresponds to different simulation conditions: Hydrogen Mass Repartitioning (HMR)[15] and hydrogen bond constraints (SHAKE)[14] applied to non-water hydrogen atoms (4 fs), SHAKE applied to non-water hydrogen atoms (2 fs), fully flexible non-water hydrogen atoms (1 fs). TIP3P[10] 4 fs and SPCFW[4] 1 fs dimerization shows the same relative behavior as for the bulk simulations. TIP4P[11], TIP5P[12] and SPCE[13] 4 fs show rates between TIP3P 4 fs and SPCFW 1 fs. To test whether the slower and more realistic behavior of SPCFW 1 fs is due to the flexible water model or due to the different treatment of the non-water hydrogen atoms, we also performed simulations with TIP3P and SPC 2 fs, without hydrogen mass repartitioning. Indeed, HMR significantly accelerates kinetics, but the

rigid body water models yield considerably higher rates than SPCFW also without HMR. Additional switching off of the SHAKE algorithm, leading to full flexibility of the non-water hydrogen atoms as in SPCFW 1 fs does not yield different rates within statistical errors. The different treatment of the non-water hydrogen atoms, within statistical errors, results in equal binding constants.

8.5 Conclusion

As before in experiments, good agreement of simulations of an ensemble of *N*6*N*9-dimethyladenine residues with the random isodesmic stacking model was observed. Simulation under experimental conditions with the fully flexible SPCFW water model yields thermodynamic equilibrium constants, enthalpy and entropy contributions and kinetic rates in close agreement with experiment. The rigid body TIP3P water model in combination with the hydrogen mass repartitioning scheme yields significantly accelerated kinetics. Detailed comparison by means of dimerization simulations with TIP4P, TIP5P and SPC indicates that the same trend also for other rigid body water models. The simulations show that HMR can considerably accelerate kinetics, but also without HMR the rigid body water models yield too overestimated rates.

In conclusion, we observe that full flexibility and therefore increased friction as in the SPCFW water model is crucial to accurately describe the thermodynamics of the system, which is not to the same degree achieved with rigid body water models. With the fully flexible SPCFW water model, this MD study on *N*6*N*9-dimethyladenine stacking performed under experimental-like conditions, yields stacking free energies, entropy contributions and relaxation times in close agreement with the experiment. The use of a flexible water model is likely to be important also for the simulation of other binding and aggregation processes.

8.6 Bibliography

[1] Albert C. Pan, David W. Borhani, Ron O. Dror, and David E. Shaw. “Molecular determinants of drug–receptor binding kinetics.” In: *Drug Discov. Today* 18.13–14 (2013), pp. 667–673.

[2] Ron O. Dror et al. “Pathway and mechanism of drug binding to G-protein-coupled receptors.” In: *Proc. Natl. Acad. Sci. USA* 108.32 (2011), pp. 13118–13123.

[3] Michael R. Shirts, Jed W. Pitera, William C. Swope, and Vijay S. Pande. “Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins.” In: *J. Chem. Phys.* 119.11 (2003), pp. 5740–5761.

[4] Yujie Wu, Harald L. Tepper, and Gregory A. Voth. “Flexible simple point-charge water model with improved liquid-state properties.” In: *J. Chem. Phys.* 124.2, 024503 (2006).

[5] W. Schimmack, H. Sapper, and W. Lohmann. “Stacking Interactions of nucleobases: NMR-Investigations I. Selfassociation of N6,N9-Dimethyladenine and N6-Dimethyl-N9-Ethyladenine.” In: *Biophys. Struct. Mechanism* 1.2 (1975), pp. 113–120.

[6] Dietmar Pörschke and Frieder Eggers. “Thermodynamics and Kinetics of Base-Stacking Interactions.” In: *Eu. J. Biochem.* 26.4 (1972), pp. 490–498.

[7] M. P. Heyn, C. U. Nicola, and G. Schwarz. “Kinetics of the base-stackingSc reaction of N6-dimethyladenosine. An ultrasonic absorption and dispersion study.” In: *J. Phys. Chem.* 81.17 (1977), pp. 1611–1617.

[8] Viktor Hornak et al. “Comparison of multiple Amber force fields and development of improved protein backbone parameters.” In: *Proteins* 65.3 (2006), pp. 712–725.

[9] D.A. Case et al. *AMBER 14*. University of California, San Francisco, 2014.

[10] William L. Jorgensen, Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. “Comparison of Simple Potential Functions for Simulating Liquid Water.” In: *J. Chem. Phys.* 79.2 (1983), pp. 926–935.

[11] William L. "Jorgensen, Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L." Klein. “"Comparison of simple potential functions for simulating liquid water".” In: *"J. Chem. Phys."* "79"."2" ("1983"), "926–935".

[12] Michael W. Mahoney and William L. Jorgensen. “A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions.” In: *J. Chem. Phys.* 112.20 (2000), pp. 8910–8922.

[13] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma. “The missing term in effective pair potentials.” In: *J. Phys. Chem.* 91.24 (1987), pp. 6269–6271.

[14] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman J.C Berendsen. “Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes.” In: *J. Comput. Phys.* 23.3 (1977), pp. 327–341.

[15] Chad W. Hopkins, Scott Le Grand, Ross C. Walker, and Adrian E. Roitberg. "Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning." In: *J. Chem. Theory. Comput.* 11.4 (2015), pp. 1864–1874.

[16] Richard W. Pastor, Bernard R. Brooks, and Attila Szabo. "An Analysis of the Accuracy of Langevin and Molecular Dynamics Algorithms." In: *Mol. Phys.* 65.6 (1988), pp. 1409–1419.

[17] Hans C. Andersen. "Molecular dynamics simulations at constant pressure and/or temperature." In: *J. Chem. Phys.* 72.4 (1980), pp. 2384–2393.

[18] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. "Molecular Dynamics with Coupling to an External Bath." In: *J. Chem. Phys.* 81.8 (1984), pp. 3684–3690.

[19] Hans R. Kunsch. "The Jackknife and the Bootstrap for General Stationary Observations." In: *Ann. Statist.* 17.3 (Sept. 1989), pp. 1217–1241.

[20] Schrödinger, LLC. "The PyMOL Molecular Graphics System, Version 1.7.4." Aug. 2010.

[21] M. K. Gilson, J. A. Given, B. L. Bush, and J. A. McCammon. "The Statistical-Thermodynamic Basis for Computation of Binding Affinities: A Critical Review." In: *Biophys. J.* 72 (1997), pp. 1047–1069.

9 MD Simulation of Guanine Methylation Damage Recognition by ALT

We used unbiased BD and atomistic MD simulations to study the binding and recognition process of O^6-methylguanine base damage in DNA by Alkyltransferase-like (ALT) protein. The simulations indicate an unproductive trapped state with the ALT helix-turn-helix (HTH) motif binding to the major grove. In minor grove binding trajectories, we observe looping out of the damaged base and intercalation of a crucial ARG residue to the base position within several µs of simulation time, in close agreement with the respective crystal structure configuration. The observed trajectories are compatible with a recognition mechanism that is based on lower base pairing strength of the O^6-methylguanine. The simulations allow predictions on the time scales of intermediate recognition states and indicate that, once the protein is located at the base pair damage, complete rotation around the backbone is the rate limiting binding step.

9.1 Introduction

O^6-alkylguanine-DNA alkyltransferase (AGT) proteins repair guanine alkylation, preventing mutagenic damages caused by environmental or endogenous alkylating agents[1]. This is achieved by irreversibly transferring the alkyl to a reactive cysteine[2]. Alkyltransferase-like proteins (ALTs) are similar in structure and binding motif, however, lack the ability of reactive alkyltransfer[3]. However, they also provide alkylation protection by binding to guaninealkylation damaged DNA with high affinity[4] and processing the damage to nucleotide excision repair[5]. The alkylation mutation resistance, at the same time, interferes with alkylation based chemotherapies for cancer treatment[6]. This makes AGTs and ALTs essential for DNA damage repair but also a drug target in combination with chemotherapies.

Fundamental insight into how these proteins recognize alkylation damages has been given by crystal structures which have been determined for both AGT and ALT in complex with O^6-alkylguanine-DNA[5, 7]. Notably, as ALTs have no reactive binding site, the crystal structure of ALT bound to damaged DNA[5] could be obtained without mutation of the protein or alteration of the alkylation damage. It reveals the key features of the recognition process, which are binding of the HTH motif to the minor

grove, widening of the minor grove, intercalation of an arginine residue at the O^6-alkylguanine position and complete looping out of the damaged base into the ATL binding site (Figure 9.1). Also, a thyrosine residue has been found to contribute to the repair rate and is suggested to facilitate phosphate rotation by repulsive interactions[5]. Due to the static nature of the crystal structures, however, detailed recognition pathways and kinetically aspects still remain unclear. The actual recognition process is in general difficult to observe experimentally.

In light of the experimental ALT-DNA crystal structure (Figure 9.1), we used unbiased Brownian Dynamics (BD) and Molecular Dynamics (MD) simulations of up to 6 µs to investigate the recognition process, starting from well separated apo protein and damaged B-DNA configurations. In the simulations, the complete protein binding process and ARG128 finger induced looping out of the damaged base could be observed. This state was stable on the µs scale. Complete turning and binding of the damaged base into the protein binding site presumably happens on slower time-scales, as this state was not reached in the free simulations. However, when rotation of the phosphate backbone was externally induced, binding of O^6-alkylguanine into the ATL binding site could be observed within 100 ns. The binding trajectories are in agreement with the conclusions drawn from the crystal structure, and allow further predictions about intermediate recognition states and associated time scales.

9.2 Results

The apo configuration of ALT[5] and a 13 base pair B-DNA double strand with an O^6-methylguanine - cytosine base pair in the central position were used as the starting point of the simulations. Brownian Dynamics simulations, treating DNA and protein as rigid structures undergoing translational and rotational diffusion in an electrostatic field, were started from protein - DNA distances of 100 Å and used to generate 1000 configurations with a separation distance of 40 Å (example: Figure 9.4A). The influence of the electrostatic field on the relative orientation of the protein was found to be negligible at these distances. From each of the obtained configurations, a MD simulation was started to investigate the protein-DNA association process at atomistic detail. The 1000 simulations were run for 100 ns at 340 °K, or aborted when the protein diffused away from the DNA or attached towards the terminal ends of the DNA strand. This resulted in a total of 64 trajectories of 100 ns with end configurations in which the protein is bound to the central region of the DNA double strand.

In all of these simulations, the protein encloses the DNA phosphate backbone, introducing the helix-turn-helix (HTH) binding motif into the DNA groves. In 20 of the 64 stable configurations, the protein binds with the HTH motif to the DNA minor

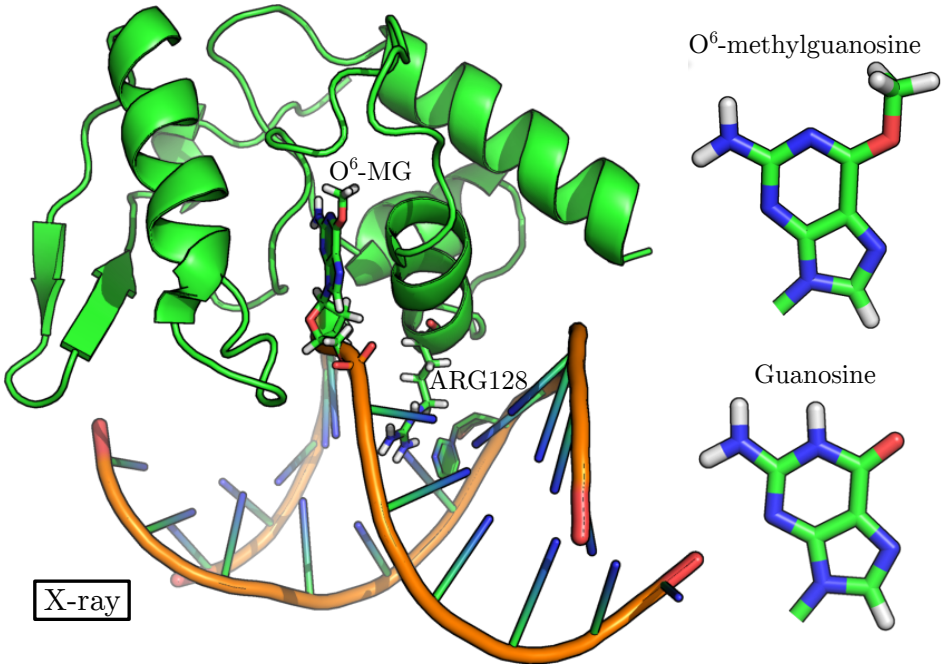


Figure 9.1: Crystal structure (PDB: 3gx4)[5] of Alkyltransferase-like (ATL) protein in complex with O^6-alkylguanine-DNA. The methylated guanine looped out of the DNA and bound to the protein binding site and the intercalated ARG128 residue are shown in stick representation. O^6-methylguanine is illustrated in comparison to guanine (without backbone). The methylation weakens hydrogen bonding of the damaged base pair.

grove as in the crystal structure (Figure 9.2A). The other 44 configurations show the HTH motif binding to DNA major grove (Figure 9.2B), corresponding to a rotation of roughly 180° of the protein with respect to the DNA backbone in comparison to the crystal structure. In the 20 minor grove binding protein configurations, 9 times the protein was bound to the undamaged strand and 11 times to the damaged strand, showing no statistically significant preference towards the damaged strand. Also, the trajectories showed no significant bias of the initial attachment position towards the damaged base. The 11 configurations in which the protein was bound to the minor grove of the damaged strand (at different positions along the strand) were continued for at least 4 µs of simulation time in a slightly smaller simulation box to investigate the damaged base recognition process after the initial binding step in detail.

Figure 9.4 shows an exemplary 6 µs MD trajectory, representative for all prolonged trajectories with respect to the observations described in the following. The initial association process is fast: Attachment (without looping out and corresponding residue rearrangements) to the DNA (protein RMSD to crystal structure < 3 Å) happens within several ns. This attachment is accompanied by an equally fast distortion of the DNA, inducing minor grove widening, as seen in the crystal structure (decrease of DNA RMSD). The simulations indicate that this distortion indeed locally and unspecifically destabilizes base pairing. The fraction of looped out bases (see Methods) in the protein-DNA complex (without ARG128 intercalation) is higher by a factor of ≈ 6 than the fraction of spontaneously flipped out bases observed in a comparative 5 µs simulation of the DNA without protein (example illustrated in Figure 9.5). The trajectories show some diffusivity of the protein along the DNA. The maximum sliding distance along the 13-base pair DNA was about 7 Å (example see Figure 9.3), the average sliding distance was 2 Å within 4 µs. We observed a slight tendency to slide towards the damaged central base pair (1.2 Å in average).

In two of the 11 prolonged MD simulations, which were continued to > 6 µs of simulation time (example Figure 9.4), we observe repeated, stable looping out of the methylated guanine supported by ARG128 intercalation. While the initial binding of the protein happens on the ns time scale, the ARG128 induced looping out of the damaged base requires several µs of waiting time, during which the protein shows significant flexibility with respect to its DNA binding configuration (sampling RMSDs of up to 10 Å). Occasional looping out of the damaged base without the ARG128 support was observed, but considerably less stable (some ns). We also observed a single ARG128 stabilized looping out event of an undamaged GC base pair next to the damaged pair.

Before further turning of the damaged guanine towards the binding site of the protein, we observe flipping back of the guanine base in all ARG128 supported looped out cases. This indicates that there is a significant free energy barrier towards the final bound state, which was not overcome within the 6 µs of simulation time. When an external force of

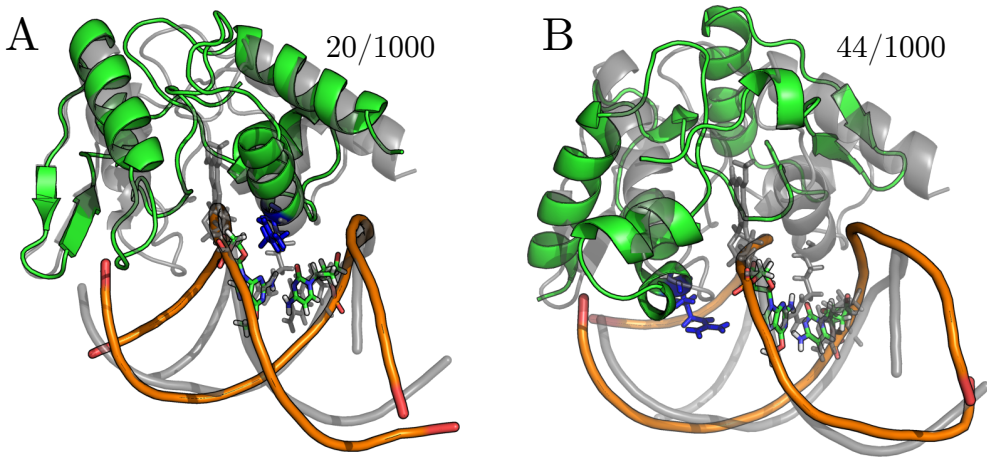


Figure 9.2: Exemplary snapshots of binding poses after 1000 100 ns MD simulations, starting from well separated protein-DNA configurations. In 20 trajectories (A), the protein binds with the HTH motif to the minor grove (at different positions), similar to the crystal structure (gray). In 44 trajectories (B), the protein binds with the HTH motif to the major grove. Presumably, this binding pose can not directly lead to damage recognition and constitutes a kinetically trapped state. We did not observe significantly differing other binding poses. The base pair containing the O^6-methlyguanine is shown as sticks, the ARG128 finger residue is shown as blue sticks.

few kcal/mol/Å^2 was used to unspecifically push the damaged base towards further outward rotated states, the O^6-methylguanine bound to the ALT binding site in a pose very close to the crystal structure within 50 ns (see Figure 9.4E). This indicates that the major free energy barrier is indeed associated with complete phosphate group rotation around the backbone, and that no major structural rearrangements of the protein are necessary for the final binding step.

9.3 Discussion

We used BD and MD simulations to study the binding and damage recognition of the ATL protein. ATL possesses a HTH motif that is known for preferential major grove binding. In the ATG crystal structure, for the first time the HTH motif was found to also bind to the minor grove, which was thought of being beneficial for base looping. Interestingly, our simulations indicate that within 100 ns simulations ATL binds to the major grove more often (64 to 20), presumably owed to the better accessibility of the

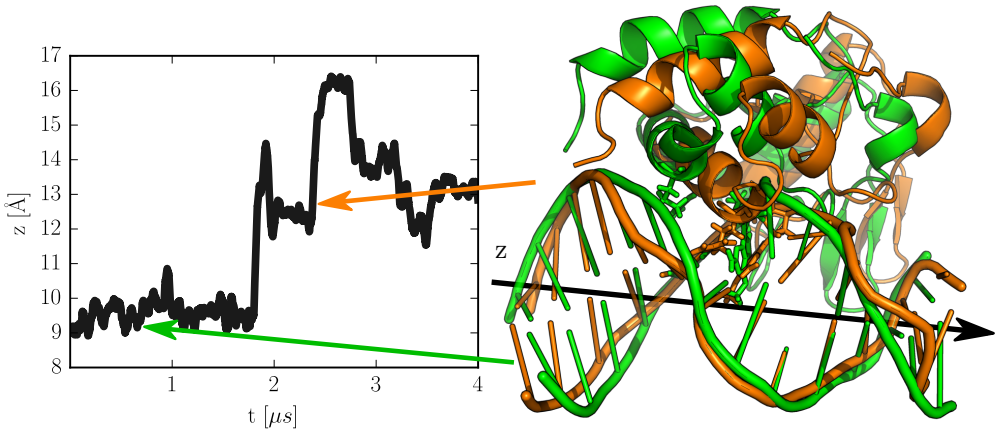


Figure 9.3: Example of sliding motion of ATL along a coordinate (z) parallel to the DNA axis during a simulation time of 4 µs. Respective snapshots are drawn in green and orange, with the ARG128 residue and the damaged base represented as sticks. The later orange structure shows deformation and a looped out (non-damaged) base in combination with the protein movement.

wider major grove. There are no indications that ATL can recognize the alkylated guanine in the major grove bound state, and looping of bases through the minor grove is generally supposed to be an unpreferential pathway. We therefore presume that the major grove bound state constitutes an unproductive trapped state, arising from the structural similarities of minor and major grove.

It was proposed that ATL recognizes damaged DNA regions by facilitated binding to a less base-pair stabilized minor grove. Within the number of affordable binding simulations and thus limited statistical significance, we did not observe any significant preference of initial binding to the damaged strand (11 to 9), or at the alkylated guanine position within the damaged strand. The simulations thus suggest that facilitation by O^6-alkylguanine of initial protein binding has a relatively weak effect. However, we observe that DNA binding in the binding vicinity significantly weakens base pairing interactions in comparison to DNA alone.

Regarding one-dimensional diffusion along the DNA, again with statistical limitations in mind, the simulations allow an estimate of an average displacement of 2 Å without dissociation of the protein, with some trajectories showing displacement of one or two base pair positions within 4 µs. A slight directional bias towards the central damaged base pair could be observed (1.2 Å μs^{-1} in contrast to 0 for no directional bias), however, we cannot exclude that in the relatively short DNA strand the central positions are

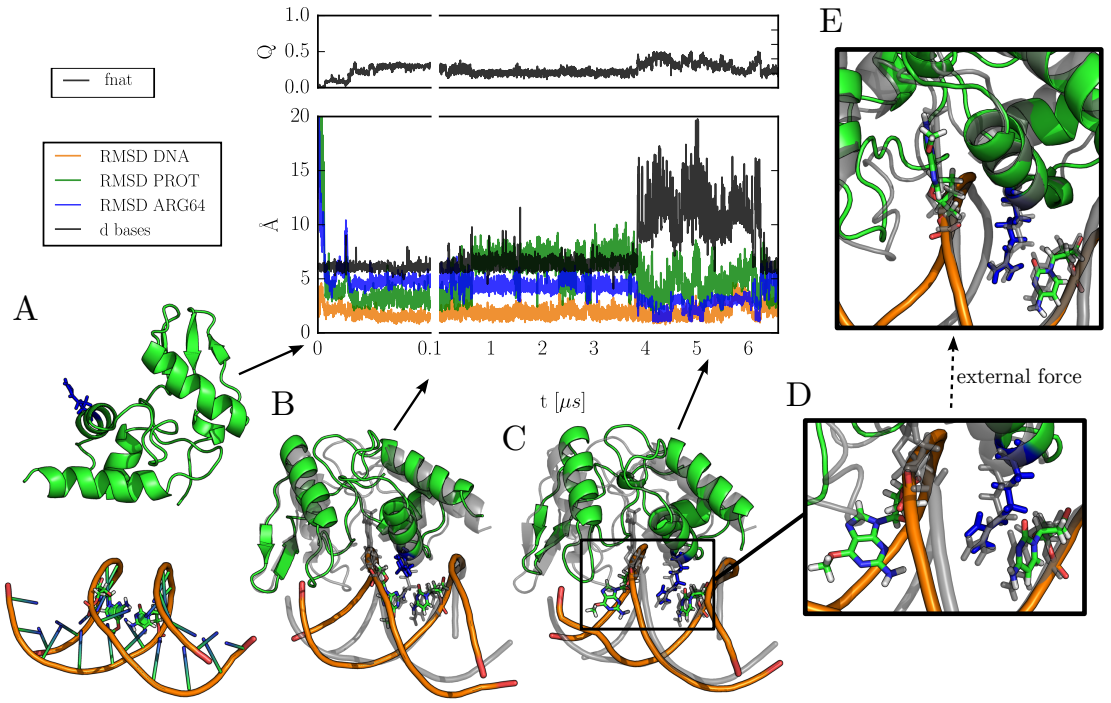


Figure 9.4: Exemplary minor grove binding trajectory of 6 µs, starting from well separated apo protein and B-DNA configurations (A). Initial binding and associated DNA deformation (see protein RMSD and DNA backbone RMSD) are relatively fast. Within several ns, protein conformations close to the crystal structure (gray) (protein RMSD < 3 angstrom) are sampled (B). On the µs time scale, the ARG128 residue (blue sticks) induces looping out of the O^6-methylguanine base (C and D). Complete rotation of the damaged base and binding to the ATL binding site is presumably associated with a significant free energy barrier, as it could not be observed in free simulations within several µs. Induction of complete base rotation by an external force however lead to fast binding of the base into the binding site close to the crystal structure (E) without significant rearrangements of the protein.

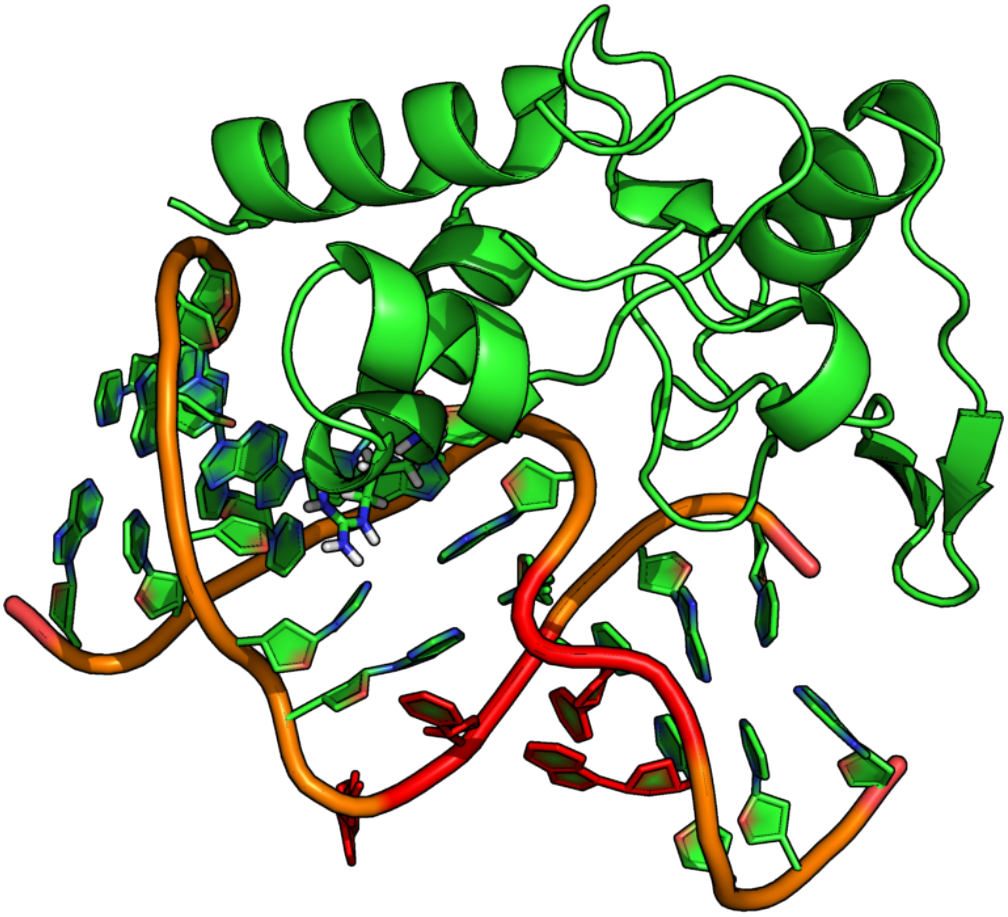


Figure 9.5: Exemplary snapshot of two disrupted (undamaged) base pair interactions (red), induced as a result of the strain put on the DNA strands by protein binding. A comparative 5 µs simulations of the DNA without protein showed significantly less spontaneous looping out events.

favored by other factors than the damaged base pair. The simulations at 340 °K suggest an upper sliding speed limit of 1 base pair position per µs. This is very slow, indicating that correlated DNA damage search is likely to involve some degree of dissociation of the protein from the DNA.

The simulations show that the protein is indeed able to actively induce looping out of the damaged base, and that binding does not require a previously spontaneously looped out base. Although the simulations were carried out at 340 °K and MD simulations are known to yield slightly overestimated kinetics, they give a reasonable estimate of the involved time scales. While the initial binding process after initial diffusion into the vicinity of the DNA happens within several ns, looping out of the damaged base requires several µs. The looped out state was stable between 0.5 and 2.5 µs when stabilized by the ARG128 residue. Final binding of the damaged base into the ATL binding site is presumably associated with a considerable free energy barrier, as it could not be observed within 6 µs of simulation time. Externally induced complete rotation of the alkylated guanine however lead to binding events within less than 100 ns, indicating that phosphate rotation constitutes the limiting step.

In the simulations, we were able to retrace the first crucial steps of O^6-methylguanine recognition from completely unbiased starting conditions, identifying kinetic traps associated with the complex role of the HTH motif, intermediate states and relevant time scales. Due to the short DNA strand (13-base pairs) and the computationally limited statistics, especially 1-dimensional sliding along the DNA and final binding of the damaged base to the protein binding site remain open questions which might be tackled when more powerful computational tools are available.

9.4 Methods

Structures

DNA with the damaged 6MG base was built from B-DNA with the following sequence: GCCATG-O^6-mg-CTAGCG. Antechamber[8] was used to parametrize the O^6-methylguanine residue. Structures of apo-ALT (PDB:3gva) and the ALT-DNA (PDB:3gx4) complex were taken from[5]. Protonation states of apo-ALT were calculated using PROPKA[9], which predicted standard protonation states for all amino acids.

Brownian Dynamics Simulations

Brownian Dynamics (BD) simulations were carried out as described in [10][11], treating the molecules as rigid bodies. The translational and orientational diffusion constants were calculated from the respective Stokes law, using the radius of gyration of the

molecules as the spherical radius. To approximate the electrostatic long distance interactions, the electrostatic field of the DNA was calculated by the Poisson Boltzmann equation (at 340 K) using APBS[12], acting on the point charges of the protein. The charges of the molecules were taken from the MD parametrizations.

The BD time step was 100 ps. The temperature was 340 K. 10^6 BD simulations were started from configurations in which the protein was placed randomly at a center of mass distance of 100 Å from the DNA with random relative orientation, and terminated when the protein diffused to center of mass distances larger than 100 Å or to closest protein - DNA distances smaller than 40 Å. This yielded slightly more than 1000 configurations with the protein positioned in the vicinity of the DNA strand, of which the first 1000 structures served as starting structures for the subsequent MD simulations.

Molecular Dynamics Simulations

The protein was parametrized with the ff14sb force field[13]. The DNA was described with the corresponding up-to-date corrections[14][15]. Na and Cl ions were described by the parameters from [16]. The starting structures resulting from the BD simulations were solvated with $\approx$ 14000 TIP3P water molecules. The SHAKE[17] algorithm and the hydrogen mass repartitioning scheme[18] was applied to the DNA and protein hydrogen atoms. The integration time step was 4 fs. A short-range cutoff radius of 9 Å was used. Electrostatics were treated by the PME method (default Amber14 parameters[8]). The starting structures were energy minimized with the steepest-decent method for 1000 steps and equilibrated for 100 ps in the NPT ensemble at a temperature of 340 K using a Langevin-thermostat[19] with a collision frequency of 5 ps and at a pressure of 1 bar using the Berendsen barostat[20]. For the subsequent NVT-ensemble simulations, the weak coupling Berendsen thermostat[20] was used with a coupling time of 1 ps with a reference temperature of 340 K. In all simulations, the distances between the terminal DNA base pairs were restrained with an harmonic potential centered at 2 Å with a force constant of 1 kcal/Å^2 to avoid melting of the DNA at the termini. The 1000 BD starting structures were simulated for a maximum time of 100 ns, or aborted and rejected when the DNA-protein center of mass distance was larger than 45 Å. The simulations were also aborted and rejected when the protein approached the DNA strand towards the strand ends (angle between one of strand ends, center of mass of DNA and center of mass of protein $<$ 45 degree), as spending simulation time on protein-DNA end interactions was undesirable. 11 trajectories with the protein attached to the damaged DNA strand in a crystal-structure like orientation (see Results) were then prepared for long-time simulations. As the protein-DNA bound systems were now smaller, a smaller simulation box could be used to enhance simulation efficiency. To this end, the closest 3000 water molecules, the closest 18 Na+ ions and the closest Cl-

ion were kept. The system was resolvated in a smaller box with in total $\approx$ 8000 water molecules and reequilibrated. The resulting structures were then simulated for at least 4 µs.

Analysis

Base pair distances were defined between the N1 atoms of the bases, a base pair interaction was considered to be broken at distances larger than 7 Å. All RMSDs with respect to the crystal structure were measured after superpositioning the structures on the DNA phosphorus atoms of the 7 central DNA base pairs. The protein RMSD refers to all protein heavy atoms, the ARG128 RMSD to all ARG atoms, the DNA RMSD to all DNA backbone atoms. Native contacts were defined as inter protein-DNA heavy atom distances smaller than 5 Å. The z-coordinate was defined by the projection of the protein center-of-mass (COM) on the vector between COMs of the N1 atoms of bases 3,4,23,24 and 10,11,16,17. All snapshots were prepared using pymol[21].

9.5 Bibliography

[1] Tomas Lindahl and Richard D. Wood. "Quality Control by DNA Repair." In: *Science* 286.5446 (1999), pp. 1897–1905.

[2] Anthony E Pegg. "Repair of O6-alkylguanine by alkyltransferases." In: *Res./Rev. Mut. Res.* 462.2–3 (2000). Special Issue in Honor of Ruggiero Montesano, pp. 83–100.

[3] Geoffrey P. Margison et al. "Alkyltransferase-like proteins." In: *DNA Repair* 6.8 (2007), pp. 1222–1228.

[4] Oliver J. Wilkinson et al. "Alkyltransferase-like protein (Atl1) distinguishes alkylated guanines for DNA repair using cation-pi interactions." In: *Proc. Natl. Acad. Sci. USA* 109.46 (Nov. 13, 2012). 201209451[PII], pp. 18755–18760.

[5] Julie L. Tubbs et al. "Flipping of alkylated DNA damage bridges base and nucleotide excision repair." In: *Nature* 459.7248 (June 11, 2009), pp. 808–813.

[6] Julie L Tubbs, Anthony E Pegg, and John A Tainer. "DNA binding, nucleotide flipping, and the helix-turn-helix motif in base repair by O6-alkylguanine-DNA alkyltransferase and its implications for cancer chemotherapy." In: *DNA repair* 6.8 (Aug. 2007), pp. 1100–1115.

[7] Douglas S. Daniels et al. "DNA binding and nucleotide flipping by the human DNA repair protein AGT." In: *Nat. Struct. Mol. Biol.* 11.8 (Aug. 2004), pp. 714–720.

[8] D.A. Case et al. *AMBER 14*. University of California, San Francisco, 2014.

[9] Hui Li, Andrew D. Robertson, and Jan H. Jensen. "Very Fast Empirical Prediction and Rationalization of Protein pKa Values." In: *Prot. Struct. Func. Bioinf.* 61.4 (2005), pp. 704–721.

[10] Scott H. Northrup, Stuart A. Allison, and J. Andrew McCammon. "Brownian dynamics simulation of diffusion-influenced bimolecular reactions." In: *J. Chem. Phys.* 80.4 (1984), pp. 1517–1524.

[11] Donald L. Ermak and J. A. McCammon. "Brownian dynamics with hydrodynamic interactions." In: *J. Chem. Phys.* 69.4 (1978), pp. 1352–1360.

[12] Nathan A. Baker, David Sept, Simpson Joseph, Michael J. Holst, and J. Andrew McCammon. "Electrostatics of nanosystems: Application to microtubules and the ribosome." In: *Proc. Natl. Acad. Sci. USA* 98.18 (2001), pp. 10037–10041.

[13] Viktor Hornak et al. "Comparison of multiple Amber force fields and development of improved protein backbone parameters." In: *Proteins* 65.3 (2006), pp. 712–725.

[14] Marie Zgarbova et al. "Toward Improved Description of DNA Backbone: Revisiting Epsilon and Zeta Torsion Force Field Parameters." In: *J. Chem. Theory Comput.* 9.5 (May 14, 2013). 24058302[pmid], pp. 2339–2354.

[15] Miroslav Krepl et al. "Reference simulations of noncanonical nucleic acids with different chi variants of the AMBER force field: quadruplex DNA, quadruplex RNA and Z-DNA." In: *J. Chem. Theory Comput.* 8.7 (July 10, 2012). 23197943[pmid], pp. 2506–2520.

[16] In Suk Joung and Thomas E. Cheatham. "Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations." In: *J. Phys. Chem. B* 112.30 (2008), pp. 9020–9041.

[17] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman J.C Berendsen. "Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes." In: *J. Comput. Phys.* 23.3 (1977), pp. 327–341.

[18] Chad W. Hopkins, Scott Le Grand, Ross C. Walker, and Adrian E. Roitberg. "Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning." In: *J. Chem. Theory Comput.* 11.4 (2015). PMID: 26574392, pp. 1864–1874.

[19] Richard W. Pastor, Bernard R. Brooks, and Attila Szabo. "An Analysis of the Accuracy of Langevin and Molecular Dynamics Algorithms." In: *Mol. Phys.* 65.6 (1988), pp. 1409–1419.

[20] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. "Molecular Dynamics with Coupling to an External Bath." In: *J. Chem. Phys.* 81.8 (1984), pp. 3684–3690.

[21] Schrödinger, LLC. “The PyMOL Molecular Graphics System, Version 1.7.4.” Aug. 2010.

10 Outlook

With the enhanced methods for standard binding free energy determination and for binding rate calculations described in this thesis, also the mean residence time and thus a complete computational characterization of possible inhibitors is accessible. With the expected increase in computational capabilities in the next years, future works could combine these approaches in order to rank larger sets of previously untested inhibitors and suggest them for experimental screening. In chapter 4 it was shown that cost-efficient continuum solvent models are in many cases capable of viable binding affinity predictions. In combination with continuum solvent models, a perturbation based method for the prediction of binding affinities upon small mutations was presented in chapter 5. This enables optimization and analysis of potential inhibitors on the basis of only original inhibitor simulations. Regarding large-scale inhibitor screening, it might be convenient to make use of these results, performing an initial implicit solvent perturbation based screening of an extensive set of possible inhibitors and then proceeding with higher precision model calculations for promising candidates. Expensive comprehensive experimental tests of inhibitor binding affinities could in this way be limited to a significantly smaller, preselected set of ligand molecules.

Usually, for the computational prediction of binding affinities by means of MD simulations, an experimentally determined bound inhibitor pose is required. In chapter 7, a multiscale approach was presented which allows the reconstruction of ligand binding pathways, requiring only knowledge of the rough location of the receptor binding site. With this approach, inhibitor bound poses can be computationally predicted, which is particularly attractive for the design of entirely new inhibitors.

Increasing attention is put on the mean residence time of clinical inhibitors, which is currently accepted as the main determinant of clinical drug efficacy. In this context, the association rates calculated in chapter 7 yielded a correct ranking of known inhibitors, but showed a systematic overestimation of the absolute values. A following extensive and statistically converged kinetic analysis on adenine-derivate self aggregation in chapter 8 revealed the crucial importance of water molecule flexibility on kinetic rates. The previously (and widely used) rigid body TIP3P water model yielded accelerated dynamics also for the adenine-derivate binding rates, whereas a fully flexible water model (SPCFW) yielded kinetic rates in close agreement with experiments. In this regard, it would be promising to use the SPCFW model in the ligand association

studies for a better prediction of the ligand binding rates. In addition, changes in protonation states of receptors or inhibitors could be incorporated into the association studies by means of constant-pH simulations. These simulations allow adjusting of residue protonation states during the simulations based on a Monte Carlo scheme. However, they do not yield truly continuous trajectories. If the effects on dynamics and binding kinetics are carefully tested, the possibility of an effective on-the-fly adaption of protonation states within MD simulations could prove extremely useful. A major drawback of classical force fields could possibly be overcome. As already mentioned in chapter 7, special attention could be drawn in future studies on the influence of binding site conformational changes involved in inhibitor binding by using an equilibrated ensemble of conformations for the target protein representation.

With an adequate representation of conformational changes in the receptor molecule, also the association pathways of more complicated coupled binding processes as the dual binding of adenosine-phosphates to ADK as described in chapter 6 could be investigated. The presented study on ADK gives a systematic insight into the domain motion free energy landscapes in dependence of the binding of adenonsine-phosphates. It is not possible, however, to obtain reliable kinetic data from the equilibrium H-REMD-US simulations. An adaption of the multiscale association pathway method to such a significantly more complex coupled association process could therefore be attractive.

The methylguanine-damage recognition process of ATL, studied in chapter 9 by means of an adaption of the multiscale association pathway method, could be successfully reconstructed up to a very late stage by simulations of several µs. With increasing computational power, it might be possible to significantly extend the simulations in order to observe also the final attachment of the damaged base into the protein binding site in unbiased trajectories. Such a study could, for enzymatically active repair proteins, also be combined with quantum mechanics based QM/MM methods for the last step involving the actual chemical reaction. This would possibly yield a complete description of the recognition and repair process by means of computer simulations.

List of Publications

[1] Fabian Zeller and Martin Zacharias. "Adaptive Biasing Combined with Hamiltonian Replica Exchange to Improve Umbrella Sampling Free Energy Simulations." In: *J. Chem. Theory Comput.* 10.2 (2014), pp. 703–710.

[2] Fabian Zeller and Martin Zacharias. "Evaluation of Generalized Born Model Accuracy for Absolute Binding Free Energy Calculations." In: *J. Phys. Chem. B* 118.27 (2014), pp. 7467–7474.

[3] Fabian Zeller and Martin Zacharias. "Efficient calculation of relative binding free energies by umbrella sampling perturbation." In: *J. Comput. Chem.* 35.31 (2014), pp. 2256–2262.

[4] Fabian Zeller and Martin Zacharias. "Substrate Binding Specifically Modulates Domain Arrangements in Adenylate Kinase." In: *J. Biophys* 109.9 (2015), pp. 1978–1985.

[5] Benjamin Pelz, Gabriel Zoldak, Fabian Zeller, Martin Zacharias, and Matthias Rief. "Subnanometre enzyme mechanics probed by single-molecule force spectroscopy." In: *Nat. Commun.* 7.10848 (24, 2016).

List of Conference Contributions

[1] Fabian Zeller and Martin Zacharias. "Calculation of Binding Free Energies using PMFs and Restraining Potentials." In: *Workshop on Computer Simulation and Theory of Macromolecules, Hühnfeld* (2013). Poster Presentation.

[2] Fabian Zeller and Martin Zacharias. "Interplay between Domain Motions and Substrate Binding in Adenylate Kinase." In: *SFB863 Workshop, Schloss Hohenkammer* (2014). Poster Presentation.

[3] Fabian Zeller and Martin Zacharias. "Adaptive Biasing Combined with Hamiltonian Replica Exchange to Improve Umbrella Sampling Free Energy Simulations." In: *Biophysical Society Meeting, San Francisco* (2015). Poster Presentation.

[4] Fabian Zeller and Martin Zacharias. "Interplay of Domain Motions and Ligand Binding in Adenylate Kinase." In: *Workshop on Computer Simulation and Theory of Macromolecules, Hühnfeld* (2015). Conference Talk.

[5] Fabian Zeller, Rainer Bomblies, Manuel Patrick Luitz, and Martin Zacharias. "Influenza Neuraminidase Inhibitor Binding Studied by Molecular Dynamics Simulations." In: *SFB863 Workshop, Schloss Ringberg* (2015). Conference Talk.