# Structure Formation of Biomolecules studied with Advanced Molecular Dynamics Simulations

DISSERTATION

Manuel Patrick Luitz

TECHNISCHE UNIVERSITÄT MÜNCHEN

PHYSIK DEPARTMENT T38

typeset with XƎLATEX

Munich, 2016

PHYSIK DEPARTMENT

TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für theoretische Biophysik (T38)

# Structure Formation of Biomolecules studied with Advanced Molecular Dynamics Simulations

Manuel Patrick Luitz

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

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

| | |
|---|---|
| Vorsitzender: | Prof. Dr. Friedrich Simmel |
| Prüfer der Dissertation: | 1. Prof. Dr. Martin Zacharias |
| | 2. Prof. Dr. Ville Kaila |

Die Dissertation wurde am 15.12.2016 bei der Technischen Universität München eingereicht und durch die Fakultät der Physik am 10.03.2017 angenommen.

# Contents

**APPENDIX**

*Die Zeit, die ist ein sonderbar Ding.*
*Wenn man so hinlebt, ist sie rein gar nichts.*
*Aber dann auf einmal, da spürt man nichts als sie.*
*Sie ist um uns herum, sie ist auch in uns drinnen.*

— Hugo von Hofmannsthal, *Der Rosenkavalier*

# Acknowledgments

Before we dive into the fascinating world of biomolecular systems and, particularly their investigation with computer simulations, I want to take a deep breath to create a moment of gratitude for all the generous help that reached me from different, sometimes unexpected angles during my time as PhD student and especially with regard to this thesis. I gratefully acknowledge my wonderful supervisor Prof. Dr. Martin Zacharias who gave me the opportunity to be a member of his research team and who gave me plenty of rope to develop my own ideas but was always there when I needed somebody to discuss these ideas with or just to comfort me when things did not work out the way I expected them to do. Thank you Martin for your always open door welcoming the entering and your always open ear listening to all kinds of concerns, scientific or not. On the institutional level, I thank the Technical University of Munich (TUM), the Deutsche Forschungsgemeinschaft (DFG), and the Gauss Centre for Supercomputing e.V. for funding and providing computing time on the SuperMUC at the Leibniz Supercomputing Centre (LRZ). I deeply thank all current and former members of T38 who burn for the fascinating world of biophysics and with whom I could share this passion in a relaxed and pleasant atmosphere. A special tribute goes to Sonja who made it her personal mission to organize the glue that bonded us strongly together in terrific winter school retreats or late-night Christmas parties to name only a few of these magnificent events. I thank Rainer for being such an outstanding room mate in our homely office and for becoming a dear friend. The same goes for Fabian who became a near-room mate especially in the time when we worked on the weighted ensemble methodology and who fought with me through any weather condition on our daily cycle trips home. To Florian who generously lifted of my shoulders a lot of the heavy burden I had to carry as our computer cluster administrator, a cruel job that only appears in bright daylight when things go wrong. Thanks to Giuseppe, Alex, Mahmut, Sjoerd, Nadine, Katja, Christina, Alexei, Piotr and many more with whom it was a pleasure to enter fruitful discussions about physics or other matters. I thank our collaborators for their elementary work of providing us with the experimental fundament upon which every theoretic scaffolded needs to be build. Prof. Aymelt Itzen, Evelyn, and Rudi to work with us on Rab1b posttranslational modifications. Prof. Don Lamb, Anders, and Alvaro for their will to challenge their own field of fluorescence spectroscopy with regard to the drawbacks of labeling. Prof. Thomas Kiefhaber who initiated the RNAse-S project with his precedent studies and who critically discussed strategies to approach the topic from the theoretic standpoint. Often taken for granted, I explicitly want to mention the legions of open-source developers that provide the outstanding software with which we work from dusk till dawn and the rest of the day. Their work guarantees free access to software implementations and the unlimited freedom to tamper around with the sources, constituting the vital blood pulsing through the veins of science. Selected examples are the GNU/Linux project, the GROMACS molecular dynamics toolchain, the LaTeX project, or the vast variety of python packages such as ParmEd, NumPy, or matplotlib. My friends and family in no particular order: The EDM crew, Tobi, Konsti, Andi, Manni, Machi, Holger, and others, that you shared with and developed in me the passion for old

and rusty steel boats and that you always endure my hungry phases. My dear friends Peter and Crissi with their beloved children, and Flo, Michi, Söri, Thomas, I feel deeply touched to call you my friends. Of course, my parents, grandparents, and siblings whom I love and who continue to support me in all thinkable ways. And I thank my Sarah for her understanding, her strength, her tender solicitousness, and just for being there.

# Abstract

The cell constitutes the common structural and functional building block of all biological organisms. Enclosed within the cellular membrane are many macromolecules such as proteins and nucleic acids that function in their ensemble as the molecular machinery that maintains the metabolism and provides reproductive capabilities. Detailed knowledge about the structural and dynamic nature of these macromolecules is the key to the understanding of life and many diseases. Apart from experimental techniques, in theoretical biochemistry, valuable complementary approaches have been developed to investigate these molecular mechanisms with the help of computer simulations. An important representative of these techniques are molecular dynamics (MD) simulations, an approach that describes the atomistic interactions of biomolecules and their chemical environment with a classical Newtonian model and allows to visualize the molecular motions comparable to a fictive microscope operating at atomistic resolution. Groundbreaking discoveries have been made with the help of molecular dynamics simulations in the past. Molecular dynamics simulations are however limited by the available computational resources. This often prevents the sampling of biomolecular processes occurring on slow timescales that cannot be simulated in reasonable time. In this work, the basic theoretical concepts behind molecular dynamics simulations are introduced and modern techniques are reviewed, developed, and applied that allow the enhanced sampling of otherwise inaccessible molecular processes. In particular the weighted ensemble method is critically reviewed, a statistical approach that accelerates MD sampling along predefined reaction coordinates and rigorously reproduces both kinetic and equilibrium properties. Important convergence issues, connected to the weighted ensemble methodology, are discussed and an implementation of the algorithm is presented. An advanced sampling approach is developed in order to predict structural binding modes of small ligand molecules in the receptor binding site. The structural information about ligand/receptor complexes is frequently relevant during the process of designing new drug molecules. In another study, extensive MD simulations on an intrinsically disordered peptide are directly compared to fluorescence spectroscopy measurements. Thereby, the modifications of conformational and dynamical properties of biomolecules induced by fluorescence labeling are elucidated and useful guidelines are developed to design future experiments such that the modifications on the system remain negligible. Eventually, two biologically relevant proteins are investigated with molecular dynamics simulations and the results are compared and complemented with experimental findings. Both system have in common that they are not structurally rigid but possess highly flexible regions of structural disorder rendering them interesting for molecular dynamics simulations. The first study is targeting the signaling protein GTPase Rab1b that is involved in the intracellular vesicle transport system. Rab1b acts as a molecular switch that is characterized by the conformational transition of a functional switch region from structural disorder to order. Bacteria of the species Legionella pneumophila exploit this mechanism by covalently attaching an adenosine monophosphate to the tyrosine 77 residue of Rab1b thereby reprogramming the cellular supply system to promote their own reproduction. Umbrella sampling simulations demonstrate that the switching mechanism is hereby locked in the activate conformation. In a second study the association mechanism of S-peptide and S-protein forming RNAse-S is investigated. While S-peptide is disordered in solution, it adopts a stable helical configuration when bound to S-protein. Simulations reveal that an initial key contact between the disordered S-peptide and S-protein is sufficient to induce the folding process on the surface of S-protein. An alternative mechanism proposing that S-peptide adopts the native conformation before binding to S-protein is ruled out.

# Chapter 1

# Introduction

## 1.1 Motivation

The field I entered in the last years is strongly connected to our sheer existence as living beings. About 4 billion years ago, long before complex organisms populated the surfaces and oceans of our planet, the first self-organizing chemical units with reproductive capabilities evolved [1]. Yet before these units were organized into cells, two fundamental paradigms of life arose from purely chemical means. Information was stored in chemical molecules and together with the capability to reproduce and propagate these information storing molecules, live was born. Today, we know two kinds of biological molecules that possess the capability to store information. Both, the ribonucleic acid (RNA) and the deoxyribonucleic acid (DNA), are quite similar in their chemical composition and occupy vital and complementary positions in modern cellular live. Although initially it was unclear whether DNA preceded RNA on the evolutionary timeline or vice versa, the widely accepted *RNA world hypothesis* postulates, that the first macromolecule igniting the processes of life was indeed RNA [2–4]. Laboratory experiments, based on the finding of Miller et al. in 1955 showed that simple biochemical molecules like sugars, amino acids, or nucleotide bases are formed in prebiotic conditions from a few simple chemical ingredients that are believed to have been present in the prebiotic world. However in early experiments, the efficiency of ribonucleotides synthesis was too low to strongly support the RNA world hypothesis [5–8]. Recent studies successfully increased the efficiency of synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions [9]. Another ground breaking discovery, awarded with the Nobel price in chemistry for Altman and Cech in 1989, revealed that certain RNA sequences possess catalytic properties, for which the term *ribozym* was introduced [10–12]. Interestingly, it was found that important functional components of the ribosome, a molecular machine that translates the information stored in RNA and synthesizes the encoded proteins, are constructed from RNA themselves using proteins as scaffolding structures [13]. Additionally, a candidate for the molecular machine that catalyzes the type of polymerization required for RNA replication fully constructed from RNA was recently found [14]. Emerging from the RNA world, the next step in evolution included most probably the surrounding of self-replicating RNA units with a self-assembling bilayer membrane formed from amphiphilic fatty acids in proto-cells [15]. In modern cells with an increased level of complexity, the function of storing the genetic code was mostly shifted from the single-stranded and non-redundant RNA to the double-stranded redundant DNA molecule while RNA remained in other roles e. g. gene regulation, signal transduction and as information transporters in the process of DNA transcription and protein synthesis [16]. Proteins complement the catalytic functions of RNA by having evolved to the “workhorses” in cellular life. Having divided the tasks of information storage and catalytic function among these classes of biomolecules, complex super-cellular life evolved using cells as smallest common building block which enabled the diversification into different species. The chemical building blocks of proteins are the family of amino acids which provided a larger set of chemical diversity compared to the four nucleotides in RNA, with over 500 known members from which 20 are commonly found in proteins [17]. The amino acids are linearly linked to form a chain that folds to a functional protein after assembly. The first three dimensional (3D) structure of a protein was solved 1958 for myoglobin and marked a scientific breakthrough that established X-ray crystallography as a standard technique to solve protein structures until today [18]. During his work on RNAse-A, Christian Anfinsen discovered eventually that the amino acid sequence encodes the folded structure and thereby the function of proteins [19]. The folding process arranges functional amino acid side-chains in space such that specific chemical reactions can be catalyzed or, more generally, interactions with other molecules are enabled. In cases where the folding process ends in a non-native state, serious malfunction in the organism are often the consequence [20, 21].

## 1.2 Molecular dynamics simulations

Anfinsen's dogma, that the native structure and function of proteins is determined by the amino acid sequence, largely increased the interest in structural information about proteins and other biomolecules. It has become clear that a fundamental understanding of structural properties of biomolecules on an atomistic level is the key to the understanding of cellular life and the investigation of many diseases. Since the first protein structure was solved by X-ray crystallography, a multitude of alternative experimental techniques for structure determination has been developed [22–28]. In parallel, computational chemists developed methods to describe molecular structure and motions in proteins and other biopolymers theoretically. This thesis is focused on the application and development of methods describing molecular motions over time with the help of molecular dynamics (MD) simulations. A brief introduction to the underlying theoretical concepts behind MD simulations shall be given in this chapter. MD simulations are sometimes termed to be a *computational microscope* as they provide access to the dynamic process in biomolecules at atomistic resolution [29, 30]. The approach complements experimental techniques that are usually unable to access both atomic position and dynamics at the same time. The information how biomolecules rearrange and change their structure over time is however a crucial point for understanding the mechanism underlying cellular live. Although the theory of quantum mechanics constitutes currently the most accurate theoretic description of molecular processes, its application is limited to simple systems with a small number of atoms and to relatively short time scales due to computational constraints [31]. Therefore, molecular dynamics simulations provide an alternative pathway to simulate molecular motions of larger systems on longer timescales, embedded in a classical Newtonian corset. The concept of MD is based on the approximation of Born and Oppenheimer, that proposes the wave function in the Schrödinger equation of molecular systems to be separable in two independent electronic and nuclear contributions [32].

$$\Psi_{\text{tot}} = \Psi_{\text{electronic}} \times \Psi_{\text{nuclear}} \tag{1.1}$$

The Born–Oppenheimer approximation justifies a separation of timescales to treat molecular dynamics as the motions of the slow and heavy nuclei and independently from the fast fluctuations of the electrons. Despite this separation ansatz, the quantum mechanical treatment of large molecular systems remains still too expensive to be used for the long timescales that matter for biological macromolecules. Therefore the motions of nuclei are modeled in MD simulations with the classical description of massive particles obeying Newtons second law.

$$\vec{F} = m \cdot \ddot{\vec{x}} \tag{1.2}$$

Atoms are represented in this approach by massive point particles with Cartesian coordinates of their nuclei while the electronic contributions are accounted for with a mean field approach by assigning partial charges to the point particles. These partial charges account for the non-uniform electron distribution in molecules due to the varying electronegativity of different atoms in a covalently bonded neighborhood and are typically unphysical fractions of the elementary charge $e$ [33]. This approach however ignores the effect of electronic polarisability, the process by which the electronic distribution of the molecules spontaneously responds to the environmental electric field. As in some cases the inclusion of polarisability effects is crucial for an accurate representation of biochemical processes, efforts are undertaken to include them into modern MD approaches [34]. Having reduced the level of detail for molecular systems down to classical massive particles, the interactions between these *atoms* need to be modeled. Forces in molecular systems are represented by potential functions, representing the various interactions between covalently bonded and non-bonded atoms. The sum of bonded and non-bonded potential functions, acting in a molecular system define a so-called *force field* in MD terminology. The class of bonded or intramolecular interactions in molecular dynamics simulations involve typically at least the following three potentials [35]

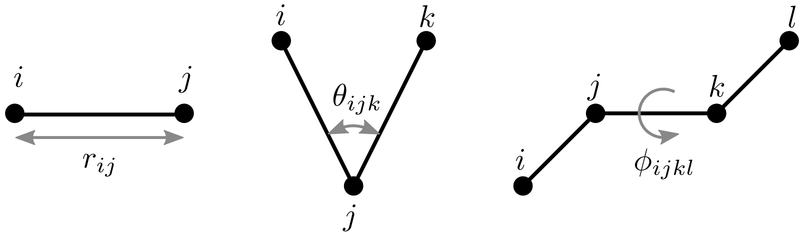

Figure 1.1: Intramolecular interactions in molecular dynamics force fields. From left to right, bond, angle and dihedral interactions are shown. Atoms are represented as black spheres with indices $i, j, k, l$ and covalent bonds are indicated as black lines. Grey arrows show the direction of the potential force.

$$V_{\text{bond}}(r_{ij}) = \frac{1}{2} k_{ij} \left( r_{ij} - r_{ij}^{(0)} \right)^2 \tag{1.3}$$

$$V_{\text{angle}}(\theta_{ijk}) = \frac{1}{2} k_{ijk} \left( \theta_{ijk} - \theta_{ijk}^{(0)} \right)^2 \tag{1.4}$$

$$V_{\text{dihedral}}(\phi_{ijkl}) = k_{ijkl} \left( 1 + \cos \left( n\phi_{ijkl} - \phi_{ijkl}^{(0)} \right) \right) . \tag{1.5}$$

Figure 1.1 depicts the motivation and application of these three potentials to covalently bound atoms. Covalent bonds, evolving from shared electrons between atoms $i, j$, are modeled as harmonic potential with a force constant $k_{ij}$ (equation 1.3). The application of bond potentials reproduces the equilibrium distance $r_{ij}^{(0)}$ between atoms, however the geometric structure of molecules is not fully described in equilibrium by only covalent bond potentials. To account for this fact further potentials based on angles between three next neighbor atoms $i, j, k$ and dihedral angles between four atoms $i, j, k, l$ are typically introduced (equations 1.4, 1.5). While the angle potential is again harmonic with force constant $k_{ijk}$ and equilibrium angle $\theta^{(0)}_{ijk}$, the dihedral potential models the periodicity of a flip around the dihedral connection in form of a periodic cosine function with $n$ minima per turn and a phase shift $\phi^{(0)}_{ijkl}$. The cosine function is offset by 1 to avoid a change of sign and a force constant $k_{ijkl}$ is defined.

The interactions between atoms that are not in a direct covalent relationship are typically modeled by two non-bonded interactions. The pairwise electrostatic interaction between atoms $i, j$ with partial charges $q_i, q_j$ is described by the Coulomb potential (equation 1.6). Furthermore, as the cores of atoms $i, j$ are positively charged and are surrounded by a negatively charged electron cloud, they interact via induced dipoles with each other, leading to a short range attractive van der Waals interaction that scales with distance $r^6_{ij}$. At close distances, on the other hand, the van der Waals attraction is replaced by a strong repulsion due to the Pauli principle of overlapping electron orbitals. This repulsion is modeled with a distance dependence $r^{12}_{ij}$. The combination of induced dipole attraction and Pauli repulsion results in the Lennard–Jones potential (equation 1.7).

$$V^{(i)}_{\text{Coulomb}}(\vec{x}_1, \ldots, \vec{x}_N) = \sum_{j, j \neq i}^{N} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{1.6}$$

$$V^{(i)}_{\text{LJ}}(\vec{x}_1, \ldots, \vec{x}_N) = \sum_{j, j \neq i}^{N} \frac{C^{(12)}_{ij}}{r^{12}_{ij}} - \frac{C^{(6)}_{ij}}{r^{6}_{ij}} \tag{1.7}$$

All these potential functions define parameters to characterize the specific interaction between different atoms in a molecular environment. For bonded potentials (1.3, 1.4, 1.5) these are the force constants $k_{ij}, k_{ijk}, k_{ijkl}$, the equilibrium distances $r^{(0)}_{ij}, \theta^{(0)}_{ijk}, \phi^{(0)}_{ijkl}$, and the number of minima $n$ per periodic flip in the dihedral potential. For non-bonded interactions (1.6, 1.7) these are the partial charges of the atoms $q_i$ and the Lennard–Jones parameters $C^{(6)}_{ij}, C^{(12)}_{ij}$. The entirety of these parameters defines a molecular dynamics force field and the process of parametrization is a non-trivial problem. It is typically achieved by fitting the parameters to experimental results or results from ab initio quantum mechanics calculations. As the complex quantum mechanic mechanisms in molecular systems are only coarsely modeled by the relatively small number of classical potential functions, it is difficult to reproduce all sorts of different molecular situations with only one set of parameters (i. e. with a single force field). Therefore the parametrization of force fields is an ongoing process and a plethora of different MD force fields has been proposed, specifically designed for the simulation of different molecular situations and environments [36–42].

Having defined the forces acting in a many particle molecular system, Newtons second law 1.2 can be iteratively solved for small time steps in order to simulate the atomic motions over time [43]. However, the algorithm would yet still only represent a $N$ particle system at constant total energy $E$ and volume $V$. In the thermodynamic world of biomolecules however, other quantities like the temperature or pressure need to be controlled by coupling the system to an external bath. The instantaneous temperature at time $t$ is defined in statistical physics over the velocity $v_i$ of the $N$ particles and their degrees of freedom $N_\mathrm{f}$

$$T(t) = \sum_{i=0}^{N} \frac{m_i v_i^2}{N_\mathrm{f} k_\mathrm{B}}. \tag{1.8}$$

The quantity in 1.8 which can be directly adjusted in MD simulations are the velocities $v_i$ (as the masses $m_i$ are constant in the classical picture). Thermostat algorithms therefore adjust the temperature to stay close to the reference temperature $T_0$ of the surrounding bath by rescaling the particles velocities on-the-fly such that $T_0 = \langle T(t) \rangle$. Several thermostat algorithms have been proposed in the past [44–47]. Another quantity that may be coupled to an external bath is the pressure $p$. The scalar pressure $p$ is defined in statistical physics as the trace of the pressure tensor $\mathbf{p}$ divided by 3. The instantaneous pressure tensor $\mathbf{p}(t)$ is defined as the sum of the kinetic energy tensor and the virial tensor at time $t$

$$\mathbf{p}(t) = \frac{1}{V} \left( \sum_{i}^{N} m_i \vec{v}_i \vec{v}_i^T + \sum_{i<j}^{N} \vec{r}_{ij} \vec{F}_{ij}^T \right) \quad \text{and} \quad p(t) = \mathrm{Tr}\left(\mathbf{p}(t)\right)/3. \tag{1.9}$$

Both, velocities $\vec{v}_i$ and positions (implicitly in the distance vector $\vec{r}_{ij}$ between atoms $i, j$) contribute to the pressure determination and can again be adjusted during the molecular dynamics simulation with the help of barostatic algorithms [44, 45, 48].

## 1.3 Outline

This thesis deals with a variety of loosely linked topics that build on common ground being situated in the field of biomolecules investigated with molecular dynamics simulations. The concept of MD simulations is applied and developed in different directions. Methods and detailed information about the used techniques are given in the respective chapters as they are required. In the first two chapters the limits of current molecular dynamics simulations are discussed with regard to the difficulties of abundant sampling and convergence. These chapters focus on the general issues of finite sampling of biomolecular systems and the different advanced sampling approaches that have been designed to bypass these problems. In chapter 2 the state of the art of advanced sampling methods is reviewed. These methods are designed to accelerate and enhance the sampling of specific molecular processes of interest that are otherwise inaccessible to molecular dynamics simulations. The discussed methods include different variations of the replica exchange methodology, simulated annealing, or meta-dynamics. In chapter 3 the weighted ensemble (WE) methodology is thoroughly reviewed and described, elucidating all advantages and drawbacks of the approach. Weighted ensemble differs from other standard advanced sampling methods by the design features to not require modifications of the Hamiltonian or a rupture in the physical continuity of trajectories. It is therefore apt to reproduce the statistically exact dynamic properties of the simulated system. A complete and efficient implementation of the WE methodology in python has been developed in the course of this thesis and is presented in appendix A. After the review of state of the art advanced sampling approaches and a technical introduction into the WE methodology, the development of a replica exchange based technique that allows the study of receptor-ligand binding processes including full flexibility of the binding partners and an explicit inclusion of solvation effects is presented in chapter 4. The approach enhances the sampling of putative ligand-receptor complexes in order to identify energetically favorable binding modes. The generic design provides a potential utility to predict binding structures during the design of drug molecules inhibiting active sites of known receptors. A connection between experimental fluorescence techniques and MD simulations is drawn in chapter 5. Fluorescence spectroscopy techniques have become important tools for the experimental investigation of conformational dynamics in biomolecules. However these techniques require the covalent attachment of labeling molecules to the target molecule. By combining MD simulations with photoinduced electron transfer fluorescence correlation spectroscopy (PET-FCS) experiments, the significant effect of fluorescence labeling on the conformational dynamics of small biomolecules is revealed. The results may be used to minimize the influence of labeling when designing new fluorescence experiments. In the chapters 6 and 7 two MD studies of biologically relevant protein systems are discussed. The first study (chapter 6) explains the effect of posttranslational modifications on the conformational switching mechanism of the cellular signaling protein Rab1b. The modification is effected by the bacterial enzyme DrrA/SidM of *Legionella pneumophila* in order to exploit the intracellular vesicle transport system of the host cell to promote the replication of the bacterium inside the host. It is found that the posttranslational modification locks the switching mechanism of Rab1b in the active signaling conformation due to the additional negative charge that is introduced by the modification. The results may also have implications for the mechanistic understanding of conformational switching in other signaling proteins. The second study (chapter 7) investigates the association mechanism of the protein complex RNAse-S. RNAse-S consists of the larger fragment S-protein and the smaller S-peptide. While intrinsically disordered in solution, S-peptide adopts a stable helical fold upon the association to S-protein. The exact mechanism, whether S-peptide adopts the native conformation before binding to S-protein or whether folding to the native structure occurs after initial key contacts are formed, is investigated.

# Chapter 2

# Exploring biomolecular dynamics and interactions using advanced sampling methods

Molecular dynamics (MD) and Monte Carlo (MC) simulations have emerged as a valuable tool to investigate statistical mechanics and kinetics of biomolecules and synthetic soft matter materials. However, major limitations for routine applications are due to the accuracy of the molecular mechanics force field and due to the maximum simulation time that can be achieved in current simulations studies. For improving the sampling, a number of advanced sampling approaches have been designed in recent years. In particular, variants of the parallel tempering replica-exchange methodology are widely used in many simulation studies. Recent methodological advancements and a discussion of specific aims and advantages are given. This includes improved free energy simulation approaches and conformational search applications.

## 2.1 Introduction

Molecular simulations are a versatile tool to study the dynamics of soft matter systems, polymeric materials, and biological macromolecules [30, 50]. Differing in their level of spatial resolution, simulations can include electrons explicitly as spatial wave function or as density function in the framework of the density functional theory. Due to the large size of biological macromolecules and many soft matter systems it is often desirable to only use the positions and momentum of whole atoms as variables. In this case the interactions are described by a classical force field based on the coordinates of atom centers. The atom-centered potentials include the average effect of electrons and are often based, at least in part, on experimental parameterization. Resolution level and associated computational demand can be even further reduced by considering whole chemical groups as single units and defining effective interactions between such coarse-grained centers of a system [51, 52]. By combining these force fields with the classical equations of motion, molecular dynamics (MD) simulations [50] can model the dynamics of soft matter systems or biological macromolecules. Alternatively, relevant conformational states of a molecular system can be generated with Monte Carlo (MC) methods [53]. The choice of the simulation ensemble depends on the physical framework. Although it is, in principle, possible to generate configurations of a system compatible with arbitrary statistical ensembles, the great majority of simulation studies are performed under conditions that are compatible with a canonical or an isobaric-isothermal ensemble. An ultimate goal is to extract realistic kinetic and thermodynamic

---

Parts of this chapter have been published in [49]

quantities from simulations and to associate it with atomic resolution structural data. To extract accurate thermodynamic and kinetic data, the sampling of relevant states during molecular simulations is of uttermost importance. Standard simulation algorithms often sample only a limited range of the relevant configurations of a given system during the available simulation time [54–56]. The improvement of simulation sampling techniques has been the major aim of advanced or enhanced sampling techniques developed in recent years [55, 56]. The number of studies in this area has increased dramatically, allowing only the discussion of a subset of important developments in the present review. It should be emphasized that many aspects of identifying and sampling thermodynamically relevant conformational states in a molecular system also play a role in other many body problems, including systems that are treated quantum mechanically (QM). Therefore, ideas and principles discussed in this chapter may also be of relevance to other fields of numerical and soft condensed matter physics.

## 2.2 Molecular mechanics force fields to study biomolecular and soft matter systems

The standard form of a classical force field was introduced in chapter 1 and shall be rewritten below,

$$\begin{aligned} V(\vec{r}_{\text{atom}}) = \sum_{\text{bonds}} \frac{1}{2} k_{\text{b}} (b - b_0)^2 + \sum_{\text{angles}} \frac{1}{2} k_\theta (\theta - \theta_0)^2 + \\ \sum_{\text{dihedrals}} \sum_{n} k_{\tau,n} (1 - \cos(n\tau + \delta_n)) + \\ \sum_i \sum_{j<i} \left( \frac{A_{ij}}{r_{ij}^{12}} + \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{r_{ij}} \right). \end{aligned} \tag{2.1}$$

In such a model, atomic interactions are approximated by atom-centered model functions that include bonded (first three summations in equation 2.1) and non-bonded contributions (last term in the aforementioned equation). To control the bond lengths ($b$) and bond angles ($\theta$) of the macromolecule quadratic energy terms with force constants ($k_b$ and $k_\theta$, respectively) matching experimental vibrational frequencies are used. A combination of periodic terms is used to control dihedral torsion angles $\tau$. The non-bonded interactions are described by van der Waals and Coulomb terms (as a double sum over all non-bonded pairs of atoms).The form of the energy function of a molecule allows a rapid calculation of the potential energy and also the calculation of forces necessary for performing MD simulations based on the numerical solution of the equations of motion. It is possible to explicitly include solvent molecules and ions around the solute molecule during the simulations. However, the explicit inclusion of a large amount of water can increase the computational demand and requires long simulations to equilibrate the solvent and ion atmosphere. Accounting implicitly for solvent effects can accelerate simulations and can also improve the convergence of calculated thermodynamic averages. Most implicit solvent models are based on macroscopic electrostatic concepts (assigning different dielectric constants to the interior and surrounding of a solute) or hydration shell models that are based on the accessible surface of a given solute. In the former case approximate solutions of the Poisson–Boltzmann equation or solutions to the Generalized Born model for a macromolecule are most frequently applied [57, 58]. However, new hardware and design of special computer architectures dramatically extended simulation time and maximum size of a simulation system including explicit solvent [30, 59–61]. Due to the availability of new hardware and design of special computational architectures [60] it is now possible to simulate systems with thousands or even millions of atoms and reaching simulation timescales in the micro-second regime (for systems with thousands of atoms). In special cases even milliseconds have been reached for small solvated protein molecules [61]. However, for routine applications the maximum simulation time is still a major limitation and is, in many cases, not sufficient to cover all relevant conformational or configurational states of a biomolecular or polymeric system. Although of major importance and coupled to the sampling problem, the question of how accurately a classical force field of the form given herein can describe a realistic molecular system is not discussed in the current review. Instead, the sampling of relevant conformational states is the main focus.

## 2.3 Sampling problem in molecular simulations

Synthetic polymeric molecules and biopolymers consist of long chain molecules typically involving rotatable chemical bonds. Molecular dynamic simulations allow only for small variations of bond lengths and bond angles of biopolymers, since large force field constants are keeping all relevant conformational states close to equilibrium geometry with respect to these variables. Conformational changes mainly arise from the bond rotation or dihedral angle change at selected positions along the polymer, possibly resulting in different conformational states that are separated by steric energy barriers. Since each building block of a polymer can contain several possible dihedral substates, the number of states for the polymeric system can grow exponentially with the polymer length. The underlying energy function is often termed the energy landscape of the polymeric system. To extract thermodynamic and kinetic quantities from simulations it is necessary to visit most or preferably all relevant states of the system. This task can be difficult or even impossible depending on the size of the system and the character of the underlying energy landscape [54]. For example, even for small systems the relevant conformational regimes with low associated conformational energy can be separated by large energy barriers (figure 2.1, left panel). On the time scale of hundreds of nanoseconds up to microseconds that can currently be covered by MD simulations, conformational transitions between stable states can still be rare events [30, 59]. In addition, even in cases with small potential energy barriers between stable states the large number of states by itself can create barriers. This is, for example, the case if the crossing of a barrier is associated with a reduction of the number of accessible states (entropic barrier). In such cases the lowering of energy barriers or increase of the temperature may not help to overcome the sampling bottleneck. In recent years, the field of biomolecular simulations has witnessed a revolution in terms of the maximum reachable simulation time scale. By means of new special purpose computer hardware [60, 61] it has become possible to run MD simulations up to the millisecond regime and beyond for not too large protein or polymer systems. Reversible folding and unfolding could be modeled for a set of 12 small proteins basing on a well parameterized molecular mechanics force field [60]. Such studies did not only offer many new insights into the atomic details of the protein folding process but also showed that molecular mechanics force fields include the essence of the important interactions realistically enough to allow the reversible folding of several different protein molecules. A drawback of brute force applications to tackle a simulation challenge is the extremely large computational demand and still limited availability. However, significant progress has been achieved not only in the design of new computer hardware but also in the development of new and smart sampling algorithms [55, 56] with the aim to sample more relevant states of a system in shorter time or with more limited computational resources, which is the focus of the topological review. In the first part of the review we give an overview on the various methods that have been proposed to overcome the conformational sampling problem during molecular simulations. For example, simulated annealing techniques open a large conformational space at high simulation temperatures to finally select low energy states by cooling down the system [62–64]. Second, potential scaling methods aim to lower barriers during energy minimization or an MD simulation [65–69] by scaling down the original potential or replacing it by a soft core potential. Third, conformational flooding [70] and meta-dynamics methods [71] specifically enhance sampling along one selected collective coordinate or a set of collective degrees of freedom of a molecular system. Alternatively, the locally enhanced sampling method makes use of multiple conformational copies of a selected region of a molecule to generate a mean field and thus to propagate the system [72]. Finally, example applications give an overview of types of problems and systems that benefit from enhanced sampling methods.

## 2.4 Advanced sampling approaches

### 2.4.1 Simulated annealing and tempering methods

The sampling problem has been recognized as a major issue since the initial developments of MC and MD simulations methods and their application to large polymeric molecule systems [54]. The simulation temperature is one commonly used classical parameter to control the ability to overcome barriers of the potential energy during simulations. In simulated annealing techniques [62–64] the system starts at a high temperature to escape from local minima in a rough energy landscape and is subsequently slowly cooled down to preferably move to the regime of the global minimum of the energy function. The technique is widely used in structural biology to generate 3D structures of biomolecules compatible with experimental data. In fact, basically every experimental biomolecular structure determination involves such a computational modeling step applied to a starting structure to maximize the agreement with experimental data [64]. It typically requires encoding the experimental data as an additional force field penalty term such that a structure with optimal agreement with respect to experiment minimizes this penalty term. For example, in the case of the structure determination using x-ray crystallography the experimental data represent the electron density of the crystal structure, and a realistic structure model should be compatible with the measured electron density or measured structure factor [64]. During the modeling process the structure factor of the structural model is estimated (by Fourier transformation of the model) and compared to the experimental structure factor. The corresponding difference is included as a force field penalty term that needs to be minimized during the structure optimization process. Another common method of structure determination uses nuclear magnetic resonance (NMR) spectroscopy and allows the derivation of short range distances and contacts in a molecular structure [73]. The experimentally determined distances can be included during a molecular simulation approach as restraints (typically as harmonic force field penalty terms with a minimum at the experimentally measured distance). Similar to structure determination with x-ray crystallography, the other force field terms limit the sampling of conformers to structures compatible with the chemical and sterical geometry of the polymer [64, 73]. In recent years many new experimental techniques for obtaining limited or low-resolution structural data on flexible molecules, intermediate states, or very large assemblies have been developed. Often no high resolution crystal structure can be determined in these cases. The most important techniques include cryo-electron-microscopy (cryoEM), which provides low to medium (near atomistic) resolution data on the electron density distribution [74]. Other examples are small angle x-ray scattering (SAXS) [75], which provides low resolution information on the shape and fluorescence energy transfer (FRET) [76, 77] that allows the estimation of distances between fluorophors in a molecule. These techniques are increasingly used not only to study biomolecules [78] but also to investigate many synthetic polymers or other soft matter systems. In general, the optimization of generated structures with respect to the experimental data may require escaping from local minima on the path to a set of conformers compatible with all experimental data. As a standard technique, simulations are initially started from a high temperature allowing the crossing of large energy barriers and are subsequently cooled down slowly in a simulated annealing protocol. A final comparison of the generated structures with the available experimental data allows retaining only those final structures fully compatible with experimental data. Starting such a simulated annealing optimization process from many different starting conformations can also be used to get an impression on how accurately a polymer structure is defined by the experimental data. Equivalent techniques are also used to generate model structures based on similarity (homology) to a known biomolecular structure (called a template structure). In such cases the experimentally derived restraints are derived from the stereo-chemistry and geometry of the structural template [79]. The realistic and optimal inclusion of such data during structure generation has emerged as an important task in structural biology. Since the often sparse experimental data are insufficient to uniquely define an atomistic structure it is necessary to combine it with the most realistic simulation conditions, including, for example, the surrounding solvent explicitly and treating molecular interactions accurately. The high initial temperatures used in simulated annealing approaches may interfere with the presence of explicit water molecules during MD simulations and also can disturb the conformation in directions not controlled by the limited experimental input data. Here, it is necessary to design combinations with other advanced sampling methods that do not disturb the simulated structure in undesirable directions. Some of the techniques discussed here could be useful to tackle this important task. Even though high temperatures in MD simulation can help to overcome energy barriers, kinetically trapped conformational states still can result from the cooling process. Hence, simulations do not guarantee the localization of a globally optimal state even if experimentally derived restraints are respected. However, in general, high simulation temperatures dramatically increase the number of relevant conformational states compared to low temperatures for a given polymer system. This is a simple consequence of the functional form of the Boltzmann factor and the density of states as a function of the energy. At high simulation temperatures the sampled states are dominated by entropy (availability of states) and less by the relative potential energy of states. In the case of the inclusion of experimental data in the form of energy penalty terms, it is possible to significantly limit the accessible conformational states even at high simulation temperatures. Without such constraints high temperature simulations sample mainly states outside the regime that are of physical importance at lower temperatures, where relative energies of states dominate. Hence, simulated annealing may not be efficient for sampling problems that do not include (experimentally derived) restraints to keep the sampled states reasonably close to a regime of interest. An interesting technique to accelerate sampling in selected collective variables during MD simulations is the temperature accelerated MD simulation [80]. In this technique a restraining potential is added to keep the sampled states close to conformations along the selected collective variable. It is possible to separate the motion using the Langevin equation into a part along the collective degrees of freedom and orthogonal coordinates. By using higher temperatures and typically larger viscosities for motions along the collective degrees of freedom, one can overcome energy barriers along the collective variables of interest more easily. However, the approach requires a selection of a collective variable of interest prior to the simulation. Switching between different simulation temperatures can alternatively be done continuously, as in the simulated tempering approach [81]. Initially, a discrete set of temperatures is chosen such that the lowest temperature represents the temperature of interest and at the highest temperature all relevant barriers can be crossed. At frequent intervals the temperature of a single simulation is switched to a higher or lower level and then continued. A switch in simulation temperature is accepted according to the following acceptance rule that preserves a canonical sampling at the selected temperatures:

$$P_{\text{accept}}(i \to j) = \min\left\{1, \frac{W_i}{W_j} e^{-\left[\beta_j V(\vec{r}) - \beta_i V(\vec{r})\right]}\right\} \tag{2.2}$$

Unfortunately, the simulated tempering method requires the estimation of appropriate weights $W_i$ (depending on the effective accessible conformational space at each temperature) for each temperature switch in advance of the production simulation. However, several methods have been recently proposed to iteratively adjust these weights on the fly during the simulation [82, 83]. Compared to simulated annealing in simulated tempering the system continuously enters high temperatures and low temperatures, increasing the chance of crossing barriers but still being able to select relevant low energy states. However, one should keep in mind that, overall, the system spends only a fraction of the total simulation time at the desired physical temperature of interest.

### 2.4.2 Scaling or deforming the force field energy function to improve sampling

Temperature is, however, not the only parameter that can be scaled to overcome energy barriers during a simulation. Scaling the original potential or replacing it with soft core potential has also proven successful to enhance conformational sampling during molecular simulations [65–69]. Alternatively, a boosting potential can be added whenever the potential energy of the sampled configuration falls below a preset energy threshold. This approach, termed accelerated MD simulation [84], also effectively lowers the difference in potential energy between low energy regimes and potential energy barriers. Any scaling of selected potential energy terms or boosting of certain energy contributions may, however, lead to a distorted energy landscape that does not preserve the structure and the distribution of minima of that of the original force field. Hence, it can lead to sampling of regions of the conformational space not relevant for the original potential energy landscape. Thus, in high dimensional coordinate space, states of little relevance for the temperature or Hamiltonian of interest may be significantly oversampled. This can also lead to a shift of the free energy difference between folded and unfolded states of a peptide or protein [66, 84]. In general, Boltzmann reweighting of the sampling on a deformed potential energy landscape can be used to recover the state distribution at the original energy function.

However, similar to the estimation of a free energy change associated with the modification of a Hamiltonian (free energy perturbation), the result depends strongly on the overlap of the sampling in the deformed potential energy landscape with respect to the sampling at the original force field (or desired temperature).

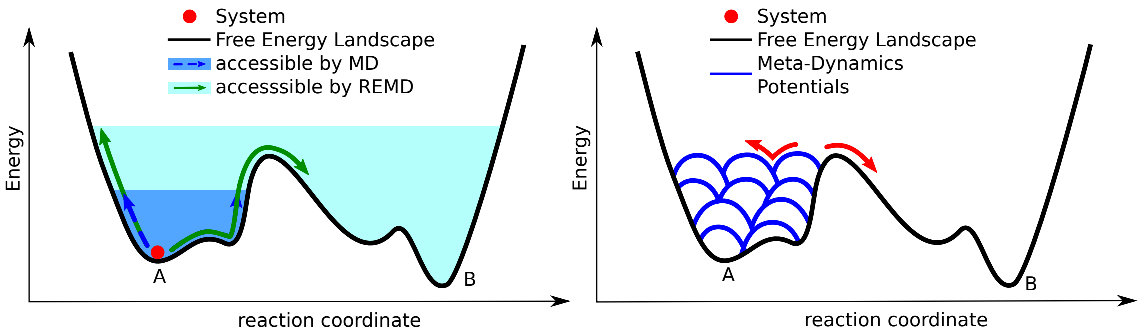

Figure 2.1: Illustration of a hypothetical energy landscape for a biopolymer along a selected reaction coordinate for conformational transitions. Starting at conformation A, the sampling is restricted to regions in the vicinity of A at low simulation temperature (marked blue), whereas crossing of large energy barriers becomes possible at higher temperatures (left panel). In meta-dynamics simulations biasing potentials (indicated in blue) are added to the force field during a simulation (right panel). The biasing potentials in the form of Gaussian functions are centered at already visited positions along the reaction coordinate and drive the simulation to explore new conformations along a reaction coordinate. Eventually, the process results in a flat free energy surface and allows extraction of the free energy function along the reaction coordinate as the sum of the Gaussian biasing potentials (with opposite sign).

### 2.4.3 Conformational flooding and meta-dynamics approaches

Although atomistic simulations provide the positions and momentum of each individual atom as a function of time, it is often desirable to focus on a more limited set of variables that are of physical interest or can be directly compared with experiments. In simple cases such variables can be (1D reaction coordinate) and may correspond to a distance between the termini of a polymer molecule or represent the radius of gyration of a flexible macromolecule. Motion in such variables requires the simultaneous and collective change of many atom positions. Along these collective coordinates the free energy landscape can contain significant free energy penalties and barriers. A classical method to guide a system along a selected collective coordinate is the umbrella sampling (US) method [85, 86] where one adds an appropriate biasing potential to the force field to attract the system toward a desired region along the collective variable (e. g. a preset distance between the ends of a polymer). Typically, one uses a simple quadratic (harmonic) biasing potential:

$$V(\vec{r}) = V_{\text{original}}(\vec{r}) + V_{\text{bias}}(D(\vec{r})) \tag{2.3}$$

with typically

$$V_{\text{bias}}(D(\vec{r})) = \frac{k}{2}\left(D(\vec{r}) - D_{\text{ref}}\right)^2 \tag{2.4}$$

Here, $D(\vec{r})$ indicates the selected collective variable that is the function of the atom coordinates $r$, and $D_{\text{ref}}$ corresponds to a preset reference value along the collective variable. The biasing potential drives the sampling of states close to regimes of interest along $D(\vec{r})$. It is also possible to extract the associated potential of mean force or change in free energy along the collective variable [87]. The US is widely used for extracting free energy changes but less as an advanced sampling method to improve the sampling in MD simulations in general. Conformational flooding [70, 88] and meta-dynamics methods [71] have been designed to specifically enhance sampling along one selected collective coordinate or a set of collective degrees of freedom of a molecular system. In contrast to the US method only a starting point for the simulations is required, and the interest is to explore the sampling along selected collective degrees without any preset interval or limit on the coordinate. Conformational flooding in its original form enhances sampling without pre-definition of a reaction coordinate of interest [70]. It makes use of a repulsive potential derived from soft principal components of motion of a system to drive the system away from the current conformational state along soft collective degrees of freedom [70, 88]. In meta-dynamics typically a 1D collective direction of interest must be provided. An additional key feature of meta-dynamics is to use the history of the simulation to flatten the free energy landscape and to guide the sampling away from already visited regimes of the conformational space. Similar in spirit and developed before meta-dynamics, in the local elevation method [89] a progressively changing biasing potential in the dihedral angles is added to the force field to drive the system away from already visited conformations. In meta-dynamics simulations potential functions of Gaussian shape are added to the force field along the collective coordinate in preset intervals [71]. The Gaussian functions are typically centered at the current sampling point along the reaction coordinate and act as a biasing potential to destabilize the conformational regime currently sampled in the simulation (illustrated in figure 2.1, right panel). This leads to a smoothing of the free energy landscape and ultimately results in a flat energy surface (along the collective coordinate) [90]. The final sum of the Gaussian functions (with a negative sign) represents the free energy function along the reaction coordinate [71, 91]. The meta-dynamics technique and several variants have been used frequently to enhance conformational sampling along a selected reaction coordinate [56]. Typically it is used in combination with 1D reaction coordinates and is less efficient in case of higher dimensions. As a recent extension of the original meta-dynamics approach, the well tempered meta-dynamics method allows a controlled inclusion of Gaussian biasing functions during the meta-dynamics process [92]. Controlled by parameter $\Delta T$, the height of deposited Gaussians decreases over sampling time, resulting in original meta-dynamics sampling for large $\Delta T$. For large $\Delta T$ the original meta-dynamics method is recovered. In addition, other variants like the multiple walker method have been combined with meta-dynamics to improve sampling along a reaction coordinate. In the multiple walker method many simulations are started in parallel at different initial conditions along the reaction coordinate and the accumulation of Gaussian functions to smooth the energy landscape is combined, which allows a faster convergence [93]. The adaptive biasing force (ABF) method aims at offsetting the mean force along a selected reaction coordinate by adding an appropriate biasing force during the simulation to overcome associated barriers [94], which can help to overcome barriers along the reaction coordinate. The iterative adaptation of the biasing force leads to an accelerated sampling and better convergence of the mean force along the reaction coordinate that, upon integration, can be used to obtain the free energy change along the collective variable. Similar to the US method the ABF and related techniques have been developed mainly to improve free energy calculations but less to improve sampling of conformations in general. Recent efforts have been directed toward simplifying the setup of biasing potentials to control and manipulate collective variables that form the basis of meta-dynamics and related simulation approaches. For example, Bonomi et al. have designed the PLUMED package [95] that can be combined with several common MD packages such as GROMACS [96], Charmm [97], or Amber [98] to provide access to several types of collective variables useful in meta-dynamics or other restraint simulation techniques. Examples of available collective variables are selected principal components of motion obtained from an unrestrained MD simulation, the relative orientation of subsets of atoms, or the number of atom-atom contacts relative to a reference set of contacts. A similar effort by Fiorin et al. [99] termed the COLVARS module is an integral component of the NAMD package [100] and also available for the LAMMBS program package [101]. The development of such modules is very helpful for the investigation of sometimes very complex collective degrees of freedom for which an enhanced statistical sampling of relevant states is desired.

### 2.4.4 Temperature for replica-exchange and parallel tempering simulations

In recent years, the replica-exchange MD (REMD) and parallel tempering methodologies have evolved to form the most widely applied and most popular advanced sampling approaches. The REMD simulation is setup by running several replicas (copies) of a simulation system parallel and independently using classical MC or MD approaches at different simulation temperatures or using force field variants (see the following paragraph on Hamiltonian replica exchange simulations). Originally, the approach was developed for simulations of spin glass systems in 1986 [102]. The applicability to improve the sampling of peptide and protein structures was proven in the late 1990s by extending it to peptide and protein simulations using MC [103] and MD approaches [104, 105]. Since then many efforts of further improving the REMD method have been aimed to enhance sampling along specific sets of conformational variables and to reduce the computational demand. In the standard application selected pairs of replicas (usually neighbors in the range of parallel running simulations) are exchanged with a specified (Metropolis) transition probability (figure 2.2). An exchange between two replicas is accepted with the probability $P$ according to a Metropolis criterion and rejected otherwise.

$$P(r_i \to r_j) = \begin{cases} 1 & \forall \Delta \leq 0 \\ e^{-\Delta} & \forall \Delta > 0 \end{cases} \quad \text{with} \quad \Delta = (\beta_i - \beta_j)\left[\mathcal{H}(\vec{r_j}) - \mathcal{H}(\vec{r_i})\right] \tag{2.5}$$

In simulated tempering the switch in temperatures of a single system requires the determination of appropriate weights for accepting or rejecting a temperature change. In replica exchange simulations this is avoided since one exchanges between two equilibrated simulations that run at different temperatures. The canonical distribution of sampled states is not disturbed by the exchanges between replicas [106]. Conformational variety arises from the random walk in the simulation temperature that allows conformations trapped in local minima to exchange with structures running in replicas at higher sampling temperature. Hence, due to the exchanges it is possible to improve sampling in each replica. However, the height and type of barriers of the molecular system have a significant influence on the efficiency of the REMD simulation [107]. If the energy barriers are not higher than the thermal energy per degree of freedom it is unlikely that T-REMD will improve sampling compared to the same number of regular cMD simulations performed at the reference temperature [107, 108]. The form of the energy landscape also influences the performance of replica-exchange MC as has been shown for two different potential landscapes. Parallel tempering was shown to be very effective in a double well potential, but the performance increased only slightly for a 'golf course' potential corresponding to a flat surface interrupted by several local deep minima [109]. When sampling systems with mostly entropic barriers, regular cMD can perform even better than T-REMD simulations [110] because high temperatures favor the high-entropy regime [111]. Exchange attempts between replicas and their acceptance frequency are of critical importance for parallel tempering simulations, since sampling benefits mostly arise from replica conformations visiting different conditions. Guaranteeing frequent exchanges is a fundamental prerequisite for enabling effective diffusion of replica conformations in the temperature range. It is typically evaluated as the round trip time for a replica to visit all temperatures of the replica ladder [112–114]. The exponential distribution of temperatures forms the standard procedure. However, other schemes taking the heat capacity of the system explicitly into account [115] or optimizing the exchanges rates on the fly have also been described [116]. It is not necessary to limit the exchanges to direct neighbors in the replica ladder. Accelerated 'diffusion' of conformers among replicas can be achieved by allowing exchanges between any pair of temperature conditions [117, 118]. To speed up the round trip rate and to prevent the formation of conformational clusters that are usually forming close to exchange bottlenecks, in the convective replica-exchange method [119, 120] one 'stick' replica is randomly selected at the beginning of an exchange cycle and exchange with its right neighbor is attempted at every updated state iteration until the exchange is accepted. The stick replica is driven in one direction until it reaches the endpoint replica, followed by reversing the exchange direction until the other endpoint is visited and reversed again to complete the round trip. Then, another replica is chosen as the new stick replica. This method alleviates the crossing of states over exchange bottlenecks and was found to accelerate mixing by a factor of two in some cases. Alternatively, the exchange mechanism can be modified by means of the Independence Sampling method [121]. In an MC-like manner the algorithm performs metropolis exchanges between randomly selected replica pairs. This process is repeated during one updated state iteration to calculate the energetically optimal state transition matrix. For $N$ replicas, this stimulated annealing of the exchange energy was found to be efficient for $N^3$ and $N^5$ repetitions. No matter how the exchanges are attempted the acceptance of exchanges between replicas requires sufficient overlap of sampled potential energies. For a given system the sampled distribution of energies varies with the temperature. The fluctuations in the sampled energies compared to the average energy decreases with the square root of the number of degrees of freedom. Relative to the mean the distribution becomes sharper with increasing system size. Hence, a given temperature range is only covered by overlapping energy distributions if the number of replicas increases with size of the system. In turn, longer total simulation times are also necessary to allow sufficient diffusion of all replicas among the increased number of replicas. This may also demand larger computational resources. Of course, by increasing the exchange attempt rate the total number of exchanges within a fixed total simulation time can be enlarged. This may indeed improve the sampling of relevant states as recent studies suggested [122–124]. It is expected to enhance the diffusion in the temperature ladder and, in turn, may improve the overall sampling (per simulation time interval). However, a small exchange interval may also lead to highly correlated exchanges, meaning that the restarted system configuration has not adapted to the new simulation conditions before a new exchange attempt. Typically, a minimum time interval of $\sim$ 1 ps is often used for allowing relaxation of the system after attempting a replica-exchange [125]. It is also important to note that the simulation thermostat (temperature control algorithm) of the replica runs can also influence the efficiency of T-REMD simulations and can create simulation artifacts [126, 127]. Some common velocity scaling schemes, in particular, do not guarantee a canonical sampling of states in each replica [127]. In addition, the order of performing exchanges between neighboring replicas can influence the efficiency of T-REMD simulations [128]. To reduce the excessive number of necessary replicas for proteins or long peptides systems, implicit solvent continuum models (e. g. Generalized Born (GB) model) [129] or a coarse-grained description of the protein [130] can be used to eliminate solvent degrees of freedom. The accuracy of current implicit solvent models may, however, fall short of providing a realistic description of the structure and dynamics of biomolecular systems [131, 132]. In hybrid explicit/implicit solvent models exchange probabilities are calculated in a continuum model while the simulation of each replica includes explicit solvent [133]. Here, the mean response of the implicit solvent to the solute conformation, which largely contributes to the exchange criterion, can differ significantly from the instantaneous potential energy of the full system with explicit water of the T-REMD simulation. It remains elusive how this potentially interferes with the conformational sampling in each replica run. Another method couples the solute to a heat bath that varies with the replicas, while the solvent is in all replica simulations coupled to the reference temperature only [134]. This concept can be extended to assign different temperatures for different degrees of freedom of the solute [135]. The number of replicas can be further decreased if only the complete solute molecule or important collective modes of motion of the solute (typically extracted from a principal component analysis) are coupled to the simulation heat bath controlling the replica runs. This is, in spirit, similar to the self-guided Langevin (SGDL) dynamics method. In this method low frequency motions of a system are accelerated (by coupling to a higher temperature bath) while keeping the high frequency motions due to the harder degrees of freedom at a lower temperature [136]. Combined with replica-exchanges this approach can accelerate the sampling in a reference simulation lacking a self-guiding contribution. One should, however, keep in mind that using heat baths of different temperatures to control different parts of a simulation system can result in non-physical (non-equilibrium) replica runs without a canonical equilibrium sampling in each replica (compared to conventional REMD). Furthermore, unpredictable artifacts within each replica simulation might be caused by the artificial temperature gradients at the solute–solvent interface (or between different degrees of freedom of the system). To avoid these problems the temperature of parts of the system can be effectively increased by means of the solute-tempering REMD method developed by Berne and coworkers [137]. Since this involves in essence the scaling of parts of the Hamiltonian, it is discussed in the paragraph on Hamiltonian replica exchange (H-REMD).

Alternatively, the performance of T-REMD can be improved by coupling the highest replica to a reservoir of conformations generated independently prior to the REMD simulation and commonly via an implicit solvent model. Randomly selected conformers from this reservoir can exchange with the highest temperature replica of the T-REMD and thus promote conformational sampling. Improved sampling compared to conventional T-REMD was observed for exchanges with a Boltzmann (equilibrated) or non-Boltzmann reservoir [138–140]. For some applications covering a wide range of temperatures throughout the full simulation is unnecessary. Then, T-REMD simulations can, for example, refine the structure of a protein by starting with a small set of replicas at high temperatures, which is subsequently cooled down until the lowest temperature replica has reached the target temperature [141]. In contrast to the standard simulated annealing technique, not only a single but also a set of promising conformations are kept throughout the REMD simulation and cooled down to the low target temperature. Recently, numerical schemes have been developed that mimic a T-REMD protocol without actually performing the T-REMD simulation [108, 142, 143]. These methods are extremely valuable for investigating the efficiency of new REMD methods and variants.

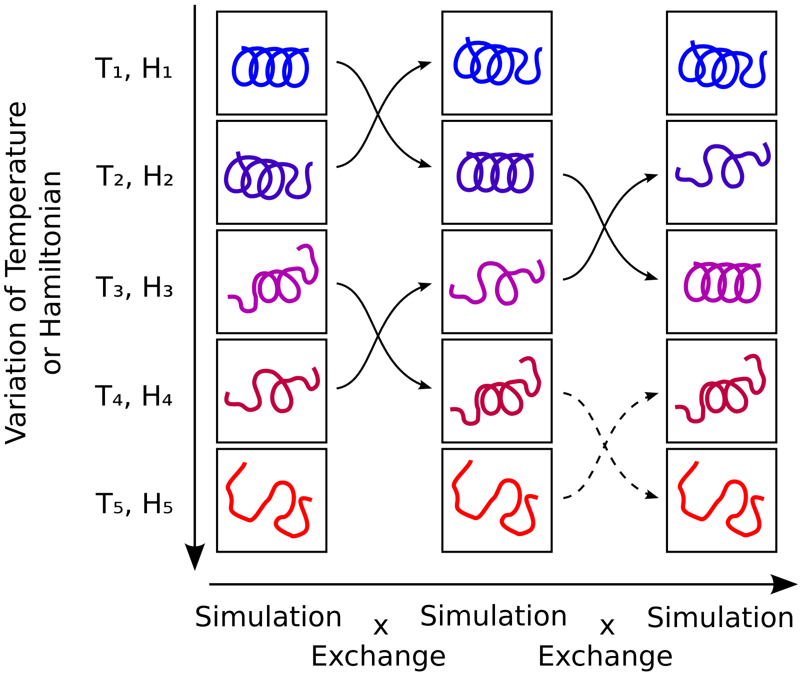

Figure 2.2: In replica-exchange MD simulations (REMD) several copies of a system (indicated as chain molecules of varying conformation) are simulated in parallel at different temperatures or under the control of different force fields (Hamiltonians, indicated by different colors). At preset intervals exchanges between neighboring replicas are attempted that are accepted or rejected according to the Metropolis criterion.

### 2.4.5 Hamiltonian replica-exchange approaches

The temperature is not the only parameter that can be varied in replica-exchange simulations: the force field or Hamiltonian of the system can also be scaled along the replica ladder. For this case the exchange criterion that preserves canonical sampling among the replicas is:

$$P(r_i \to r_j) = \begin{cases} 1 & \forall\Delta \leq 0 \\ e^{-\Delta} & \forall\Delta > 0 \end{cases} \quad \text{with} \quad \Delta = \beta \left[\mathcal{H}^j(\vec{r_i}) + \mathcal{H}^i(\vec{r_j}) - \mathcal{H}^j(\vec{r_j}) - \mathcal{H}^i(\vec{r_i})\right] \tag{2.6}$$

This Hamiltonian replica-exchange criterion can be transformed into the parallel tempering criterion if the force field is scaled linearly, $\mathcal{H}^i = \lambda_i \mathcal{H}$, along the replicas:

$$\begin{aligned} \Delta &= \beta\left[\lambda_i(\mathcal{H}(\vec{r_j}) - \mathcal{H}(\vec{r_i})) - \lambda_j(\mathcal{H}(\vec{r_j}) - \mathcal{H}(\vec{r_i}))\right] \\ &= (\beta\lambda_i - \beta\lambda_j)\left[\mathcal{H}(\vec{r_j}) - \mathcal{H}(\vec{r_i})\right] \end{aligned} \tag{2.7}$$

which is equivalent to the parallel tempering method (see section 2.4.4). In a classical simulation, however, the force field is typically a function consisting of several additive terms (equation

2.1). As a consequence the Boltzmann factor of a given system state can be split into a product of Boltzmann factors for each energy term. Only energy terms that differ in neighboring replicas affect the Boltzmann factors and enter into the Metropolis exchange criterion. Hence, Hamiltonian replica-exchange MD (H-REMD) simulations have the advantage of scaling parts of the force field among the replicas while preserving canonical sampling in each replica. This decreases the effective system size considered at every attempted exchange and can reduce the number of required replicas [55, 144]. In contrast, the temperature is a property of the whole system and the number of required replicas to cover a given temperature range increases rapidly with system size. Scaling different parts of the Hamiltonian among the replica runs has been explored in several variants of the H-REMD technique. One strategy is to identify and scale the main energy terms that are responsible for conformational energy barriers of a simulation system. For example, the non-bonded interactions of the solute–solute and solute–solvent interactions can cause such barriers for conformational transitions and have been scaled among the replicas either linearly [145] or by using a soft-core potential [146]. Although very effective in reducing sterical barriers, the approach generates many conformations with partially overlapping atoms in the higher replicas where van der Waals interactions are reduced by soft core interactions or decreased van der Waals radii. As sterical energies of these conformers may be very high in the lower replicas (closer to the original force field), the exchange rates between neighboring replica runs may decrease drastically. This in turn might be compensated with an increased number of intermediate replica conditions. Alternatively for the solute-tempering method [137, 147–149], the system energy is divided into separate interaction components corresponding to solvent–solvent, solute–solvent, and solute–solute interactions. In a typical application the solute–solvent and solute–solute interactions are then linearly reduced in the replica runs, which is equivalent to an effective simulation temperature increase in the scaled part of the system. Since it is possible to control the 'effective simulation temperature' or scaling of just the solute or of both the solute and the solute–solvent interactions, the desired temperature range can be covered by a smaller number of replicas. Promising initial results on peptide systems were not confirmed for protein folding when the method was less efficient than standard T-REMD [147]. The method was refined recently by modifying only the intra-molecular interactions in each of the replicas [148]. In further variations the solute tempering methods often differ in the type of 'tempered' or scaled interactions [149]. For example, domain motions of a protein can be studied by including domain interactions with the solvent in the tempered part but by excluding the intra-domain interactions [150]. This can lead to improvement in the sampling of domain–domain arrangements and can prevent the sampling of states that affect the individual protein domains through unfolding or refolding.

In T-REMD simulations the number of replicas and the spacing in simulation temperature can be pre-calculated or at least estimated to adjust a desired exchange rate between neighboring replicas (see last paragraph). The Boltzmann factor of every state in the system is affected by a change in temperature to the same degree (change in beta in the Boltzmann factor). In the case of H-REMD this is not necessarily the case since typically just parts of the Hamiltonian differ between replicas. For example, consider scaling of only the repulsive core of a set of atoms among the replicas in the H-REMD simulation (e. g. up to the van der Waals radius of the atoms [151]), then for many system configurations (those that do not involve atom overlaps of the scaled set of atoms) the energy difference between replicas will be zero, resulting in a high exchange/acceptance rate. For those system configurations that involve atom overlaps the exchange acceptance is, however, strongly affected by the type and degree of scaling. Even little changes in atom overlap can result in large energy differences between neighboring replicas and affect the acceptance probability. Hence, the exchange probability in case of H-REMD is determined by the fraction of system configurations that are influenced by modification of the Hamiltonian and within those configurations by the type and degree of Hamiltonian modification per replica. In the general case, it is difficult to determine these factors in advance of a simulation. In practice it is useful to perform test simulations for a given system or to adapt the scaling of the Hamiltonian among the replicas on the fly during the simulation to optimize exchange rates and the number of round trips (see below).

Several H-REMD approaches try to diminish energy barriers that occur in certain types of biomolecule, e. g. in peptides and proteins or in nucleic acids. For example, the dihedral angle biasing-potential (BP-REMD) method [152] uses a dihedral angle biasing potential specific for peptides to facilitate peptide main chain transitions (figure 2.3). Energy barriers that may prevent dihedral transitions (soft degrees of freedom) of peptides in standard MD simulations are reduced along the replica ladder. The biasing potentially is derived from MD simulations on small peptide model compounds in explicit solvent. Biasing potentials in the replica runs gradually increase among the replicas, allowing transitions at high levels of biasing. Because of the exchange criterion's independence from the solvent molecules the method requires fewer numbers of replicas than T-REMD simulations. The method was successfully applied to peptide and protein folding problems [152, 153] and also to the refinement of protein model structures [154]. Related methods for specific enhancement of dihedral transitions by scaling van der Waals interactions between one and four neighbors in peptide bonds have also been described [155]. The approach can also enhance conformational sampling of coupled nucleic acid backbone states [156, 157]. It is also possible to optimize the biasing potential applied in each replica run on the fly during the REMD simulation [156, 158]. In this case the biasing potential is adjusted during equilibration and is then used in a subsequent REMD production phase. Variants of the dihedral angle BP-REMD approach have also been developed for improved sampling of carbohydrate-based polymers [159, 160]. Conformational sampling also improves by means of the resolution exchange approach, where a biomolecule is represented as a combination of a coarse-grained force field and a fully atomistic model (possibly also including explicit solvent molecules). The degree of coarse graining increases along the replica ladder defined completely by the atomistic model in the lowest replica and by the coarse-grained representation in the highest replica run [161]. The coarse-grained model typically increases the conformational sampling in the higher replicas (because of fewer barriers). A wider sampling can also be achieved in the atomistic reference system due to the exchanges with higher replicas. The performance, however, depends sensitively on the overlap of the sampled states at the coarse-grained level with the relevant states in the all-atom representation. The flexibility of proteins can be estimated using the elastic network model (ENM). In particular for large-scale backbone motions such coarse-grained models are useful. To combine the low resolution information provided by ENM models with high-resolution atomistic MD simulations the ENM-coupled REMD approach has been developed [162, 163]. Again, a biasing potential is constructed considering the soft collective degrees of freedom (i. e. eigenmodes with small eigenvalue from the ENM calculations). This biasing potential is used in H-REMD simulations to improve the sampling of large-scale motions in multi-domain protein molecules.

As described, in meta-dynamics Gaussian potentials are added to the force field to adaptively bias MD simulations along selected reaction coordinates [71]. The reaction coordinates can correspond to different types of variables such as number of hydrogen bonds or radius of gyration or number of atomic contacts in a biomolecule. The sampling in meta-dynamics can be further accelerated by running replicas under the control of different biasing potentials and to adapt the biasing potentials on the fly. It has been demonstrated that such approaches lead to faster folding simulations using a small number of replicas [164, 165].

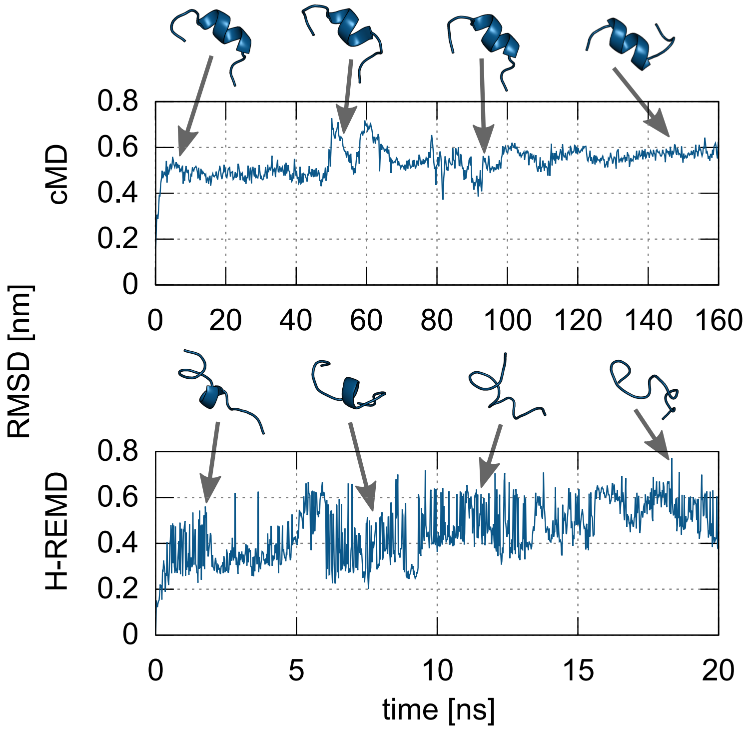

Figure 2.3: Application of the H-REMD simulation technique to enhance the sampling of RNAse S-peptide conformations. The S-peptide is a fragment of a ribonuclease enzyme that adopts an alpha-helical conformation in complex with a partner molecule (RNAse S protein) to form a functional enzyme. In the absence of a binding partner the S-peptide is unstructured. During conventional MD simulations only conformations with an intact alpha-helical core of the peptide are sampled, indicated as root-mean square deviation (RMSD) from the full alpha-helical structure and several snapshots illustrated as cartoons (upper panel). The lower panel indicates the sampling in the reference replica of the H-REMD simulation starting from the same fully helical S-peptide structure. Along the replicas a biasing potential that promoted peptide backbone dihedral transitions (biasing potential REMD) has been added [152, 153, 166]. As a result a much larger variety of peptide conformations can be sampled (indicated as snapshots) in much shorter simulation times compared to conventional MD simulations (compare upper and lower panels).

## 2.5 Application of REMD simulations to improve free energy calculations

Besides applying REMD approaches for improving conformational sampling of stably folded biopolymers or to simulate structure formation processes (described in the previous sections), it is also possible to combine it with free energy simulations. The free energy change of an alchemical transformation to annihilate or create a chemical group or to perform a conformational transition can be determined using free energy perturbation or thermodynamic integration methods [167]. For example, a chemical group can be inserted, varying the reaction coordinate $\lambda$ from 0 to 1 along the replica ladder. For $\lambda = 0$ the force field terms affecting the inserted group are turned off and are gradually, via not necessarily physical intermediates, turned on to their full strength at the replica with $\lambda = 1$. Convergence of the free energy can only be obtained if all relevant conformational states for each of these intermediates are appropriately sampled [168, 169]. Improvement of convergence can often be achieved if exchanges between intermediates are possible to prevent trapping of sampled states. Free energy simulations with exchanges of the sampled states along the reaction coordinate are equivalent to H-REMD simulations. Combining replica-exchange and alchemical free energy methods to allow exchanges between intermediate states can today be seen as a standard approach. In many cases convergence of thermodynamic integration simulations [167–170] and free energy perturbation simulations [171, 172] or US [172–177] can be improved without further computational costs. This allows for a more accurate determination of both relative free energy differences from the addition or removal of chemical groups as well as absolute binding free energies of complete ligands [178–181]. Convergence can be further improved by multi-dimensional replica-exchange (e. g. in temperature and Hamiltonian) [180, 182], yet at higher computational costs due to the increase in the number of replicas. In this case H-REMD is performed along the alchemical variable and sampling of other degrees of freedom can be improved by T-REMD in each interval of the alchemical variable.

## 2.6 Future directions

Advanced sampling techniques such as meta-dynamics and REMD simulations have found widespread applications to enhance conformational sampling and to improve the extraction of thermodynamic quantities from molecular simulations. In the current review only a fraction of the recently developed methods could be covered and the selection is necessarily biased by our own preferences. Although sampling in conventional simulations has benefited from the progress in computer hardware there is still a high demand for further development of improved algorithms to increase the sampling of relevant states in a molecular system. The aim of most recent efforts was to reduce the number of replicas for covering a desired temperature range for optimizing the performance of T-REMD simulations. The performance of the method can be enhanced by optimizing the possible exchanges at simulation intervals and allowing exchanges beyond nearest neighbors. Recent new approaches also involved the continuous adaptation and on-the-fly optimization of replica temperatures during the simulations [183, 184]. A significant advantage of the H-REMD approach is the possibility of varying only parts of the force field function among the parallel running replicas. Since the temperature is a property of the whole system such a separation of the force field into a contribution that varies among replicas and a constant part is not possible for T-REMD unless an equilibrium canonical sampling is not desired.

A major focus of new H-REMD methods is improved sampling of specific systems or of specific parameters of biomolecules including selected collective variables or the enhanced sampling of different dihedral angle states. Future development will most likely include even more structural knowledge of specific systems and may also include biasing potentials adapting to the system of interest to optimize sampling, exchanges, and round trip times during the simulation. Many recent methodological developments explore new application areas such as docking of ligands to proteins [171, 185, 186] or combining H-REMD with T-REMD to perform multi-dimensional REMD [187, 188]. The H-REMD methodology has also evolved toward a standard method for improving the convergence of free energy calculations. New developments also involve the combination with T-REMD [180] or solute tempering [189] to further improve convergence. One of the most important areas for applications is still the simulation of structure formation (folding) or proteins and other biomolecules or biopolymeric materials [54, 55, 190, 191]. The application of advanced sampling methods and of REMD methods in particular has contributed significantly to the understanding of the mechanism of structure formation and to characterizing possible intermediate states of the process. REMD simulations are an effective tool to study both biomolecules that form stable 3D structures and intrinsically disordered protein systems [192–198] that form the major part of the proteins in eukaryotic cells and that are difficult to investigate experimentally. Undoubtedly numerous new applications will give new insights into structure and function of biomolecules as well as into other soft matter systems in the years to come.

# Chapter 3

# Weighted Ensemble

## 3.1 Introduction

Energy barriers of different nature often prevent the sampling of interesting processes in biomolecular systems with MD simulations. The folding pathway of a polypeptide chain e. g. starting from its primary structure after mRNA translation over initial secondary structure formation to the functional folded protein involves several barrier transitions. The entropic search of the polypeptides backbone to discover the correct folding pathway out of a tremendous ensemble of possible backbone conformations is one fundamental barrier type. The paradox fact, that the protein can discover the correct folding configuration in finite time despite of the amount of possible conformations has been termed after its discoverer the *Levinthal paradox* [199]. Another principle barrier a protein chain has to overcome is of enthalpic nature. Functional chemical groups of sidechains or the backbone itself need to change their isomeric states or release wrong intra-chain contacts which may have formed on the way to the correct fold. As the crossing of barriers is typically not a guided or directed process but occurs randomly based on the Brownian motions of the molecule and its environment, it is a time consuming process. The waiting time for barrier crossing events is directly linked to the height of the free energy barrier (figure 3.1). A first order approximation for the rate to overcome the activation energy $\Delta G^{\ddagger}_{A\rightarrow B}$ in chemical reactions was developed by Arrhenius (equation (3.1)) [200]. The equation depicts, that the probability to transition a barrier within a certain amount of time, decreases exponentially with the free energy height, making barrier transition a rare event in some cases.

$$k_{A\rightarrow B} = Ae^{-\Delta G^{\ddagger}_{A\rightarrow B}/(k_{\mathrm{B}}T)} \tag{3.1}$$

Methods accelerating the sampling of rare events are of great interest in the field of molecular dynamics simulations, as they enable the access to timescales otherwise inaccessible with given computational capabilities. In particular the accelerated sampling of enthalpic barriers has been tackled in the past with a variety of advanced sampling techniques, to calculate thermodynamic quantities e. g. the free energy difference between two states. In order to circumvent the long waiting times of a system to transition between thermodynamic states of interest, a modified Hamiltonian is frequently used to actively drive the system along an artificial pathway connecting the states. Arbitrary pathways can be devised without the requirement to be physical meaningful creating a class of so called *alchemical* sampling techniques [201]. In order to determine the free energy difference between bound and unbound mode of a ligand which binds to a receptor protein, the ligand may either be transferred form the binding pocket to bulk solution by gradually removing all interaction terms between the ligand and the rest of the system in one state and restoring them in the other [202]. Alternatively the ligand may as well be

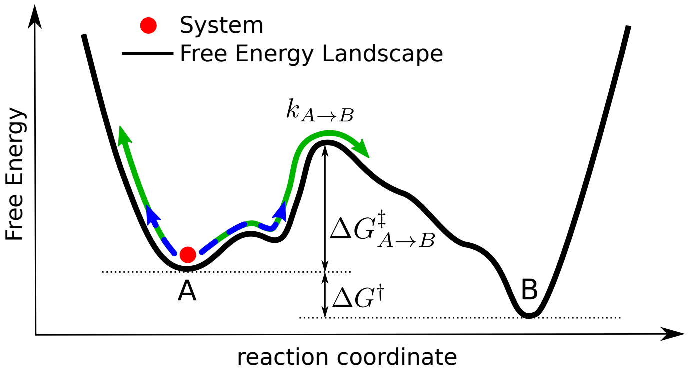

Figure 3.1: Schematic depiction of a one-dimensional free energy landscape separating two thermodynamic states $A$ and $B$. The activation free energy to transition the barrier from state $A$ to $B$ is given by $\Delta G^{\ddagger}_{A\to B}$ and determines the transition rate $k_{A\to B}$. The free energy difference between the two states is given by $\Delta G^{\dagger}$ and has no influence on the transition rate in direction $A \to B$. The activation energy of a transition in the inverse direction $\Delta G^{\ddagger}_{B\to A}$ is defined by the sum of the two energies $\Delta G^{\ddagger}_{A\to B} + \Delta G^{\dagger}$. The total free energy difference between the two states only affects the upstream rate $k_{B\to A}$.

pulled out of the binding pocket by an additional force which can later be corrected for with the Weighted Histogram Analysis Method (WHAM) [203, 204]. The class of advanced sampling techniques which require the modification of the Hamiltonian was reviewed in detail in chapter 2. Although these methods are statistically exact and yield accurate results for thermodynamic quantities when converged, the dynamic properties of the system are altered with the Hamiltonian in a way that kinetic information is lost and no longer can be rigorously extracted. This is particularly disadvantageous, when the speed of barrier transitions is of special interest for the system e. g. when determining kinetic on- and off-rates of ligand binding.

Huber and Kim developed the Weighted Ensemble (WE) method to enhance the sampling of infrequent events along certain reaction coordinates without modifying the Hamiltonian [205]. The method will be described in detail in the following section. The basic idea is, not to propagate a single long trajectory which only rarely transitions the barrier of interest, but to concurrently sample an ensemble of trajectories distributed along a reaction coordinate in such a way that barrier transition can be observed frequently. The method ensures equal sampling along the transition pathway by splitting the reaction coordinate in a number of equally sized bins and by keeping the number of trajectories residing in one bin constant. With the trajectories having assigned individual statistical weights, the unbiased dynamics can be extracted from the ensemble of trajectories. This allows in principle the determination of not only the thermodynamic stability of states, but also the associated kinetic information of transition pathways. After its invention in 1996, weighted ensemble has sunk into oblivion in the scientific community to be revived ten years later by the Zuckerman group and being continuously refined since then [206]. Several extensions have since been proposed e. g. the probability reweighting scheme to accelerate equilibration of the statistical weights in the bin space or an approach to obtain kinetic information from equilibrium simulations using history dependent probability weights [207, 208]. The method has succeeded in obtaining kinetic rates and path ensembles in several steady state or equilibrium studies on relatively simple or coarse-grained systems [206, 209–215].

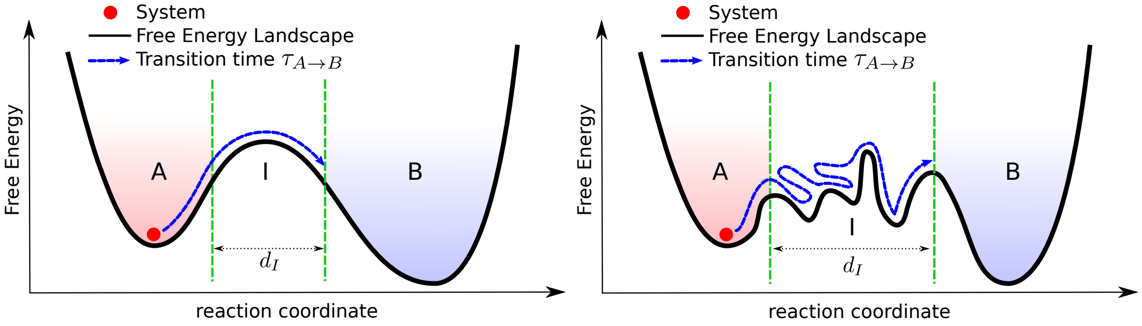

Figure 3.2: Free energy landscapes of a one-dimensional particle coupled to a thermodynamic bath. Two states $A$ and $B$ are defined with an intermediate region $I$ of width $d_I$. The two panels depict how different shapes of the intermediate barrier impact the time, the particle spends in the intermediate region $I$.

In this chapter, we aim to review the fundamental principles of the Weighted Ensemble method and to illustrate the underlying algorithms step-by-step.

## 3.2 The weighted ensemble method

### 3.2.1 Transition state theory

Before diving into the description of the WE methodology, we want to introduce some basic concepts of the transition state theory of chemical reactions which will become useful later. The left panel of figure 3.2 schematically depicts a particle in a one-dimensional free energy landscape coupled to a bath. The particle can reside in two states $A$, $B$ and the intermediate region $I$ which lies on the transition barrier between the two minima. The transition rate $k_{A\to B}$ is defined as the inverse average waiting time of the system in state $A$ to reach state $B$. When running a MD simulation on our model system, we can extract these *first passage times* as follows. Whenever the particle enters state $A$ (crosses the border to $A$) we assign a transition flag $\alpha$ to the trajectory, specifying that the system is currently in the transition $A \to B$ and start counting the time $\tau_\alpha$ it takes to reach state $B$ the next time. When the particle enters state $B$ (crosses the border to $B$) the transition flag is switched to $\beta$ the transition time $\tau_\alpha$ is stopped and stored. Having assigned the transition flag $\beta$ the trajectory is then in transition $B \to A$ and the transition time $\tau_\beta$ is started to be recorded until the trajectory enters $A$ for the next time. Note that the transition flag $\alpha$ (or $\beta$) is not changed when the trajectory exits state $A$ (or $B$) into the intermediate region $I$ and returns back into $A$ ($B$) without having visited the respective other state in the meantime. By collecting many transition events and associated first passage times $\tau$ we can then average the mean first passage time $\langle\tau_{A\to B}\rangle$. This method to extract mean first passage times from MD simulations is termed *simple counting* [216–218]. The mean first passage time leads us directly to the definition of the transition rate for both directions

$$k_{A\to B} = \frac{1}{\langle\tau_{A\to B}\rangle} = \frac{N_{A\to B}}{\sum\limits_{i=0}^{N_{A\to B}} \tau_{\alpha,i}} \tag{3.2}$$

$$k_{B\to A} = \frac{1}{\langle \tau_{B\to A}\rangle} = \frac{N_{A\to B}}{\sum\limits_{i=0}^{N_{A\to B}} \tau_{\beta,i}} \tag{3.3}$$

For a two state system like in figure 3.2 and a single continuous MD trajectory, the numbers of transitions $N_{A\to B}, N_{B\to A}$ deviates maximally by $\pm 1$ as the system propagates back and forth between the two states. At the limit of many transitions we can therefore safely assume $N_{A\to B} = N_{B\to A} = N$. It is important to note, that thermodynamic states $A$ and $B$ in figure 3.2 can be arbitrarily defined by setting the borders of the states on the reaction coordinate. Typically, thermodynamic states are not intrinsic properties of thermodynamic systems, but can be defined by any choice of connected phase space regions in a multidimensional phase space. Often, the choice is motivated by the topology of the associated phase space region e. g. a minimum (stable state) or a maximum (transition state) that entails a biological function. The accurate definition of states is de facto a problem when comparing kinetic or static properties of thermodynamic systems, obtained by different experimental or theoretical approaches as states are often only indirectly described by experimental measurements, rendering the exact translation to phase space regions difficult. Having defined the transition rates between two thermodynamic states, we can rewrite the Arrhenius equation (3.1) to yield the free energy difference between $A$ and $B$

$$\Delta G^{\dagger}_{AB} = -k_{\mathrm{B}} T \ln\left(\frac{k_{A\to B}}{k_{B\to A}}\right). \tag{3.4}$$

Although in the literature the transition rate between state $A$ and $B$ is typically defined as in equations (3.2) and (3.3), we are free to introduce a second definition of the transition rate which will become useful later when introducing transition rates in weighted ensemble. We define our transition rate $k^{\dagger}$ as

$$k^{\dagger}_{A\to B} = \frac{N_{A\to B}}{t_A} \tag{3.5}$$

$$k^{\dagger}_{B\to A} = \frac{N_{B\to A}}{t_B} \tag{3.6}$$

We use the index † in the same way as in figure 3.1 for defining the free energy difference between $A$ and $B$. The rate $k^{\dagger}$ is based again on the number of transitions $N$ and the total times $t_A$ and $t_B$ the system resides in the respective states. It only indirectly includes information about the barrier region $I$ separating the states being intrinsically included in the number of transitions $N$. In figure 3.2 we have depicted two different free energy profiles for the one-dimensional particle system. We now want to show, that our two definitions for the transition rates $k$ and $k^{\dagger}$ become equal in the limit of a vanishing intermediate region $d_I \to 0$ and full coverage of the phase space by states $A$ and $B$. By inspection of equations (3.5) and (3.2) we see that the equality of rates requires the condition

$$\sum_i^N \tau_{\alpha,i} \overset{!}{=} t_A = \sum_i^N t_{A,i} \tag{3.7}$$

to be fulfilled. We have split the total time $t_A$ the system resides in $A$ into a sum of individual residence times $t_{A,i}$ like we would inspect them when observing the system for a long time

$t_{\text{tot}}$. From the definitions of the mean passage time and the residence time follow directly that $\tau_{\alpha,i} = t_{A,i}$ in case of $d_I \to 0$ and therefore

$$\begin{aligned} \{k_{A\to B} = k^{\dagger}_{A\to B} \mid d_I = 0\} \\ \{k_{B\to A} = k^{\dagger}_{B\to A} \mid d_I = 0\} \end{aligned} \tag{3.8}$$

### 3.2.2 The weighted ensemble algorithm

In this section we want to introduce the fundamental algorithm of weighted ensemble as it was described by Huber and Kim [205]. To accelerate the frequency of barrier crossing events (figure 3.1) without modifying the Hamiltonian of the system, the phase space is discretized into $n$ bins $b_0, \ldots, b_n$ along the reaction coordinate (figure 3.4). The choice of bin size and number of bins is arbitrary but may affect the convergence as we will see later. Although we describe the algorithm on the example of an one dimensional free energy landscape and define the bin boundaries along the one dimensional reaction coordinate, any multidimensional combination of reaction coordinates may be used to define bin boundaries. It is however important that the whole phase space is covered by bins. The basic idea of the algorithm is to populate all bins with the same number of trajectories $M$ to provide equal sampling along the binned coordinate. In the one dimensional example each trajectory contains a single particle that is propagated with an MD integrator in the one dimensional free energy landscape. From time to time the WE algorithm evaluates to location of the particles on the reaction coordinate and stops the propagation of some trajectories in bins with large population while creating copies of trajectories in weakly populated bins in a process called *resampling*. Statistical weights $w_i$ are assigned to the trajectories with the sum of all weights being 1 to account for the probability of the particle to reside in a respective bins. These steps are performed in a round based algorithm with two steps per iteration. First, all trajectories are propagated with a stochastic diverging MD integrator for a time $\tau$. Secondly, the location of the particles in the trajectories are evaluated and the trajectories are assigned to the respective bins. As the particles migrate freely between the bins during the MD propagation, the trajectories need to be reassigned to bins after every propagation step (figure 3.3). After the reassignment, trajectories in bins with little population are copied to provide better sampling in these areas and deleted in bins with large trajectory population to save computational resources in areas with abundant sampling. To increase and decrease the number of trajectories within the bins, two methods are required. With the splitting routine, a trajectory within a bin is randomly selected according to its statistical weight $w$ meaning that trajectories carrying high weight have a higher chance to be selected than trajectories with low weight. The coordinates and velocity of the particle are then duplicated and a clone trajectory is forked. Both, parent and child, get assigned half of the original statistical weight of the parent $w/2$. This procedure is repeated until the desired number of trajectories $M$ is reached in every bin. Bins with no trajectories in them are unaffected by the splitting routine and remain

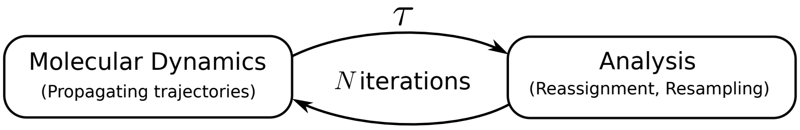

Figure 3.3: Schematic depiction of the two basic steps performed during a WE iteration.

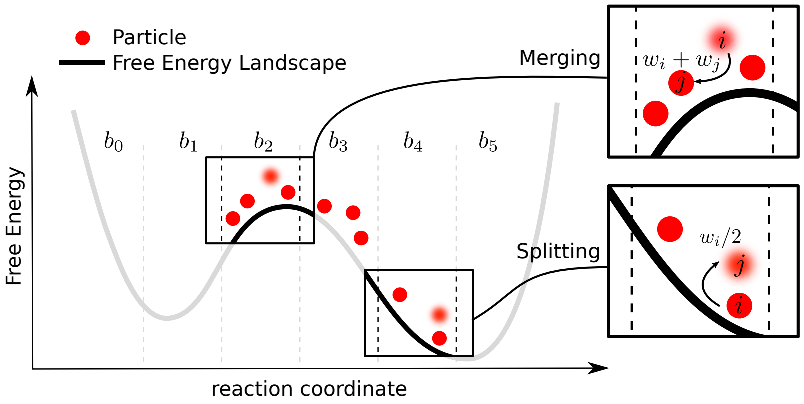

Figure 3.4: The two resampling steps in weighted ensemble simulations.

empty whereas bins with more than $M$ trajectories are submitted to a routine that decreases the number of trajectories in them. Within the validity of the diverging propagation algorithm this splitting process constitutes a valid representation of the evolution of the trajectory ensemble in phase space, as long as the total probability corresponding to the configuration in phase space remains unaffected. The detailed scheme of how many copies of the configuration are created and how the probability is distributed is in principal arbitrary and may only affect the efficiency of the sampling. At the same time, merging of trajectories in highly populated bins is necessary to prevent accumulation of a large number of trajectories in regions of low free energy. This step is crucial for the algorithm, as it avoids spending a high amount of sampling effort on regions of the reaction coordinate where sampling is abundant. Additionally it prevents an explosion of the total number of trajectories created by the splitting routine. It is realized by stopping the propagation of certain trajectories and attributing the corresponding probabilities to other trajectories within the same bin which are chosen to be continued. Mere deletion of trajectories would violate the conservation of probability within a closed thermodynamic system. So whenever a bin contains more than $M$ trajectories one of them is randomly selected for deletion according to its inverse normalized weight. A second surviving trajectory is randomly selected (according to its weight) and gets assigned the weight of the deleted trajectory plus its own weight, thereby conserving the overall probability. This process is repeated until the number of trajectories reaches eventually $M$. The resampling process of combining and splitting trajectories in the bin space is depicted in figure 3.4.

The setup of a WE simulation typically starts from one primal trajectory bearing the statistical weight $w_0 = 1$. This trajectory is the split in its starting bin until $M$ copies are generated which are then propagated for the first time step $\tau$. To ensure that trajectories containing identical particles positions and velocities diverge over time, a stochastic propagator is required as e. g. the stochastic Langevin or Brownian dynamics integrators. The particles migrate along the reaction coordinate over time and eventually populate the whole bin space while transporting probability via their statistical weights along the reaction coordinate. The equilibrium is reached when the detailed balance condition is met and the average net flux of probability over bin boundaries vanishes. From the average distribution of probabilities the free energy profile can be calculated along the bins $b_i$ according to

$$G(b_i) = -k_\mathrm{B}T \ln\langle P_{b_i}\rangle_N + c, \quad \text{with } P_{b_i} = \sum_{j=0}^{M} w_{j,b_i}. \tag{3.9}$$

The bin probability $P_{b_i}$ is calculated from the sum of weights of all trajectories residing in bin $b_0$ and with a constant $c$. The calculation of equilibrium properties like the potential of mean force (PMF) along the binning coordinate can be extracted as an intrinsic property of the weighted ensemble method. However, as mentioned before, the Hamiltonian is not modified by the approach and the kinetic properties of the system are therefore conserved. We will see in the next section how transition rates can be extracted from a weighted ensemble simulation.

### 3.2.3 History based rate calculation

The transition rate between two dense phase space regions (or in the WE terminology between two bins $i, j$) has been introduced in equations (3.5), (3.6). The rate definition has been shown in (3.8) to be equal with equations (3.2), (3.3) in case that the phase space regions $i, j$ are topologically connected. We can write the transition rate between bins $i, j$ in the weighted ensemble bin space as

$$k_{ij} = k_{ij}^{\dagger} = \frac{N_{ij}}{t_i} = \frac{N_{ij}/t_\mathrm{tot}}{t_i/t_\mathrm{tot}} = \frac{\langle F_{ij}\rangle_N}{\langle P_i\rangle_N} \tag{3.10}$$

For reasons of convenience we introduce a shorter index notation $ij$ for transition events $i \to j$. In order to translate the rate definition from the MD "simple counting" method to the probability weight based WE terminology, we introduced the total simulation time $t_\mathrm{tot}$ of a continuous MD simulation. By expanding the fraction in (3.10) with $t_\mathrm{tot}$ we directly see that the denominator becomes the fraction of residence time in state (or bin) $i$ over the total simulation time $t_\mathrm{tot}$ which represents at equilibrium the probability $\langle P_i\rangle$ to find the system in this state. The instantaneous probability $P_i$ can be extracted from the sum over all trajectories in bin $b_i$ in an equilibrated WE simulation as

$$P_i = \sum_{j\in b_i} w_j. \tag{3.11}$$

This probability might however strongly fluctuate with the particles/trajectories and their associated weights in $b_i$ among different iterations. To improve the quality of the estimate, $P_i$ needs to be averaged over a series of independent iterations $N$ yielding the equilibrium probability $\langle P_i\rangle_N$. The numerator in (3.10) on the other hand becomes the fraction of the number of transitions $i \to j$ over the total simulation time $t_\mathrm{tot}$ having the units of a rate. We introduce the average probability flux which is the probability transport per time

$$\langle F_{i\to j}\rangle = \frac{N_{ij}}{t_\mathrm{tot}} = \frac{1}{\tau}\langle P_{ij}\rangle_N. \tag{3.12}$$

Again the averaging has to be performed over multiple iterations $N$. The reason for this is, that the instantaneous flux only includes transitions times of order $\tau$. Only transition events faster than the iteration time step would be included which may be shorter than the intra bin relaxation time. The transition time distribution would be cut after longer transition times $\tau_{ij} > \tau$ and not accounted for in the average flux which would give a statistically non rigorous transition rate $k_{ij}$.

This leads directly to the problem that the rate definition (3.10) does not hold the criterion (3.8) for transition events between not connected phase space/bin regions. In order to track transition events that occur on timescales above the iteration time $\tau$ and to circumvent the problem, that the history information about trajectories that need to travel over multiple bins to reach the target state $B$ is lost after each WE iteration, a history based method to derive transition rates has been introduced [208]. We can derive the weighted ensemble definition of history based rates directly from the continuous MD definition of transition rates between states $A$ and $B$ (equation (3.2)).

$$k_{AB} = \frac{1}{\langle \tau_{AB} \rangle} = \frac{1}{\frac{1}{N_{AB}} \sum_{i=0}^{N_{AB}} \tau_{AB,i}} = \frac{N_{AB}/t_{\text{tot}}}{\sum_{i=0}^{N_{AB}} \tau_{AB,i}/t_{\text{tot}}} = \frac{\langle F(A \to B \mid \alpha) \rangle_N}{\langle P(\alpha) \rangle_N} \tag{3.13}$$

Here, we have split up the mean first passage time in a sum over the individual passage times $\tau_{AB,i}$ normalized by the number of transitions $N_{AB}$. Again we expanded the term with the total simulation time $t_{\text{tot}}$ in the picture of a MD simulation in order to translate the individual terms to probabilities and fluxes in a WE simulation. The denominator results in the probability $P(\alpha)$ to find the system in the transition event $A \to B$ averaged over multiple iterations $N$. By design the probability to find the system in either one transition direction $\{\alpha \mid A \to B\}$ or $\{\beta \mid B \to A\}$ is $P(\alpha) + P(\beta) = 1$ when considering only two states and after all trajectories have visited at least one of the two states (when starting the initial WE trajectory in the intermediate region, the direction of the transition is yet undefined). The definition of probability flux from $A$ to $B$ in the numerator has to be generalized compared to the simpler case of connected states in equation (3.10). $F(A \to B \mid \alpha)$ denotes the probability transport in $B$ per time under the condition, that the arriving trajectories have previously visited state $A$ before entering $B$.

The extraction of history based transition events from WE simulations is similar to the picture we gave for continuous MD simulations to estimate passage times from single trajectories. WE trajectories (or particles in our one dimensional picture) are tagged with a history flag $\alpha$ whenever they enter state $A$ and a flag $\beta$ whenever they enter state $B$. When trajectories get merged, the flag of the deleted trajectory is eliminated and only the weight is transferred to the surviving trajectory. When a trajectory is split on the other hand, the two children get half of the weight of their parent and inherit the history flag. It is important that the history flag is deleted upon merging, as it is intrinsically linked to the trajectories history and can not be transferred to another trajectory in contrast to the statistical weight. Contravention against this concept would introduce spurious, unphysical history transport in phase space and distort the kinetic properties of the transition rate. To evaluate eventually the average probability flux from $\langle F(A \to B \mid \alpha) \rangle_N$ in the WE simulation, the associated weights of trajectories with flag $\alpha$ entering state $B$ need to be counted and averaged over multiple iterations $N$.

### 3.2.4 Probability reweighting

The probability flow along the reaction coordinate is a time consuming process. Weighted Ensemble simulations are typically started from a single trajectory that is located in one of the bins. Returning to our one dimensional model system, let us assume, that the initial trajectory starts in bin $b_0$ which is located in state $A$ (figures 3.2, 3.4). The initial trajectory has assigned all probability weight $w_0 = 1$ and is propagated with the WE algorithm. To estimate the time dependence of the probability transport in the bin space and to reach the equilibrium probability distribution

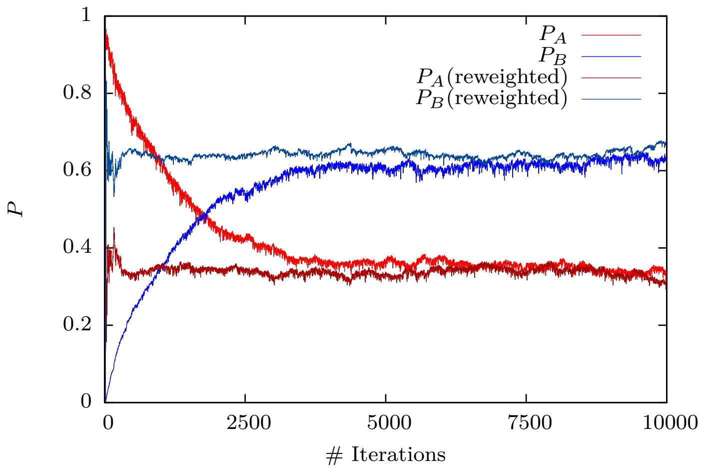

Figure 3.5: Probability equilibration in a two-state model system with and without probability reweighting approach. All weights start in state $A$ with $P_A(t = 0) = 1$.

reflecting the free energy profile (equation (3.9)), we can treat the probability transport between $A$ and $B$ as a system of coupled first order reactions

$$\begin{aligned} \frac{\mathrm{d}P_A(t)}{\mathrm{d}t} &= -k_{AB}P_A(t) + k_{BA}P_B(t) \\ \frac{\mathrm{d}P_B(t)}{\mathrm{d}t} &= -k_{BA}P_B(t) + k_{AB}P_A(t). \end{aligned} \tag{3.14}$$

These two coupled ordinary differential equations describe the evolution in time of probabilities $P_{A,B}$ in the two states. When the equilibrium is reached, the net flux of probabilities vanishes according to the detailed balance criterion ($\mathrm{d}P_{A,B}/\mathrm{d}t = 0$) [219]. If the phase space is divided in only two state $A$ and $B$ the probability of finding the system in either one state is $P_A + P_B = 1$. Using these conditions, we can then condense the two equations (3.14) to a single ordinary differential equation

$$\frac{\mathrm{d}P_A(t)}{\mathrm{d}t} = -k_{AB}P_A(t) + k_{BA}(1 - P_A(t)) \tag{3.15}$$

and give the solution

$$P_A(t) = \frac{k_{BA} + k_{AB}\exp\left(-(k_{AB} + k_{BA})t\right)}{k_{AB} + k_{BA}} P_A(t = 0). \tag{3.16}$$

We find, that the process to reach the equilibrium probability distribution in weighted ensemble simulations is subjected to an exponential decay. The relaxation time constant of the decay is given by the inverse sum of rates $t_0 = 1/(k_{AB} + k_{BA})$. Depending on the transition rates, this process typically involves many iterations and is not accelerated by the WE method itself. Figure 3.5 exemplarily depicts the temporal evolution of the probability in states $A$ and $B$ of a two state model system with a free energy landscape similar to figure 3.4 and the states stretching over the two minima. As predicted in equation (3.16) the probabilities exponentially decay over several thousands of WE iterations until reaching the flux equilibrium after about 5000 iterations in this example.

In order to accelerate the time consuming process of reaching the equilibrium probability distribution in WE simulations, the probability reweighting method has been developed by Bhatt et al. [207]. It is based on the assumption, that the steady state equilibrium between neighboring bins is only weakly coupled to the global equilibrium of the probability distribution. Therefore, the local inter-bin rates $k_{ij}$ between next neighboring bins can be estimated from an unequilibrated WE system according to equation (3.10). Although we have shown, that the given rate definition is only equivalent to the generic mean first passage time rate of next neighbor bins, it is still frequently used in the probability reweighting approach to calculate rates of non-neighboring bins under the assumption that the transition rates to farther apart bins decay quickly with the bin distance and have therefore only weak influence on the equilibrium flux equations which we will introduce now. The equilibrium detailed balance condition for a system of $n$ bins can be written as a coupled linear equation system of ordinary differential equations based on a first order rate model.

$$\begin{aligned} \frac{\mathrm{d}P_1}{\mathrm{d}t} &= -\sum_i^n k_{1,i}P_1 + \sum_{j\neq 1}^n k_{j,1}P_j \overset{!}{=} 0 \\ \vdots \quad & \qquad \vdots \qquad\qquad \vdots \\ \frac{\mathrm{d}P_n}{\mathrm{d}t} &= -\sum_i^n k_{n,i}P_n + \sum_{j\neq n}^n k_{j,n}P_j \overset{!}{=} 0. \end{aligned} \tag{3.17}$$

With the additional constraint of probability normalization

$$\sum_i^N P_i \overset{!}{=} 1. \tag{3.18}$$

In matrix representation, equation (3.17) reads

$$\mathbf{K}\cdot\mathbf{p} = \begin{pmatrix} -\sum_i^n k_{1,i} & k_{2,1} & k_{3,1} & \cdots & k_{n,1} \\ k_{1,2} & -\sum_i^n k_{2,i} & k_{3,2} & \cdots & k_{n,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ k_{1,n} & k_{2,n} & k_{3,n} & \vdots & -\sum_i^n k_{n,i} \end{pmatrix} \cdot \begin{pmatrix} P_1 \\ \vdots \\ P_n \end{pmatrix} \overset{!}{=} \mathbf{o} \tag{3.19}$$

where $\mathbf{o}$ is the $n$ dimensional null vector. In this representation it becomes obvious, that the contribution of next neighbor rates is concentrated on the main diagonal decaying quickly in the extra diagonal elements. In a WE simulation, to reach the equilibrium, the $P_i$ are the unknown variables, while under the given assumptions, the rate matrix $\mathbf{K}$ can be quickly estimated according to (3.10) over several iterations. Therefore by solving the (overdetermined) linear equation system comprised of (3.19)/(3.17) and (3.18) with a standard solver algorithms e. g. Gauss elimination or singular value decomposition [220], an on the fly estimate for the equilibrium probability distribution in the $n$ bins can be given after every iteration of the WE algorithm (the solution of the linear equation system (3.19) scales with the number of bins $n^2$ which constitutes a negligible effort compared to the MD sampling power required per bin). By adjusting the weights of trajectories in all bins accordingly, the equilibrium probability distribution is reached by the system much quicker compared to the pure diffusion process (equation (3.14)). It is important to note, that the probability reweighting algorithm yields only a rough estimate of the equilibrium probability distribution due to inaccurate evaluation of transition rates between non-neighboring bins. It is therefore typically switched off after a several equilibration iterations during the WE simulations and should be followed by a second equilibration phase with the (statistically not exact) probability reweighting method turned off (In example figure 3.5 reweighting would be typically switched off after $\sim 300$ iterations). Note further, that the probability reweighting method can in principle even be applied when only a subset $s$ of all bins $n$ is yet explored by the WE trajectories. The rate matrix $\mathbf{K}$ is then only of size $s \times s$. However, it may occur that one or multiple transition rates $k_{ij}$ among close distance bins become zero or very small numbers even after averaging them over multiple iterations, rendering the rate matrix $\mathbf{K}$ singular or its numerical solution unstable. This often indicates an inappropriate bin setup with high inter- or intra-bin free energy barriers which may be mitigated by readjusting bin boundaries, the number of bins, or the WE iteration time step $\tau$. In figure 3.5 the situation of probability equilibration in an one dimensional two-state model is depicted for a WE simulation with and without the probability reweighting method being applied. The simulation with probability reweighting reaches the probability equilibrium significantly faster than the simulation without probability reweighting.

## 3.3 Convergence bottlenecks in Weighted Ensemble

Although the Weighted Ensemble method yields statistically rigorous sampling of the phase space, the convergence of WE simulations often critically depends on various parameters [221]. In this section, we aim to critically review a number of convergence and sampling bottlenecks intrinsically linked to the WE methodology. Several examples affecting the sampling efficiency will be discussed in the following which potentially deplete or even reverse the accelerating effect of WE on sampling and convergence if not mitigated.

### 3.3.1 Intra-bin barriers

The WE method involves the discretization of one or multiple reaction coordinates into a bin structure. However the discretization itself strongly affects the convergence properties and the time scale on which the system relaxes to equilibration. A fundamental principle of the algorithm is to accelerate the crossing of major energetic barriers by discretizing the reaction pathway on the barrier in several bin areas and to provide equal sampling in each of these bins by splitting and merging trajectories, therefore increasing the probability to observe barrier transition events. Although the principle assumption is valid, the discretization itself is in general not a trivial task which involves the definition of the number of bins $M$ and the spacing between bins respectively. Especially when the free energy landscape is rugged or unknown, the accurate placement of bins can strongly affect sampling efficiency and represents therefore a critical part of the setup. Figure 3.6 depicts again our well known one dimensional two-state model system with a rugged energetic landscape involving a high barrier peak in bin $b_2$. Apparently the effect on the equilibration time scale is significant as the given bin setup fails to focus the sampling effort on the major transition barrier separating states $A$ and $B$. Assuming a total number of trajectories $N_{\mathrm{T}}$ in all bins, the WE algorithm may even perform worse than $N_{\mathrm{T}}$ individual MD simulations as in the WE approach, only the fraction of trajectories located in $b_2$ sample the

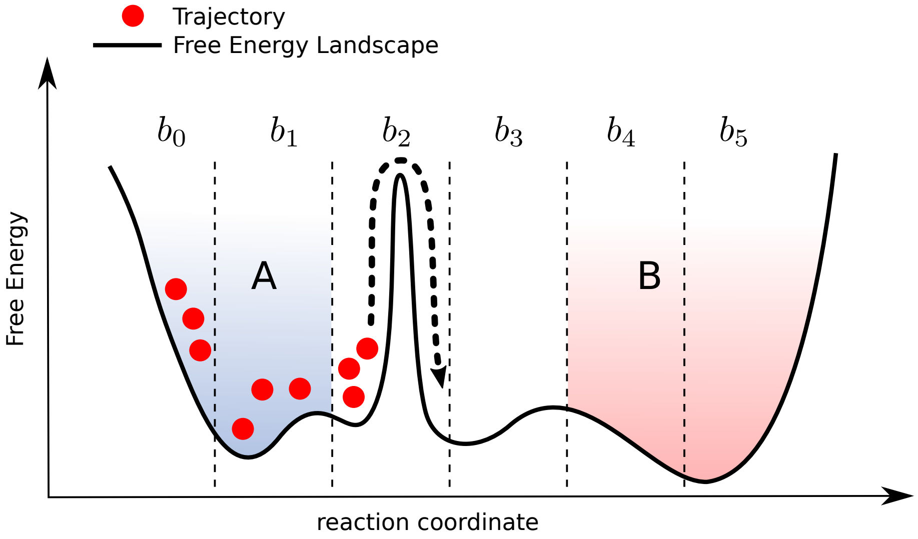

Figure 3.6: Intra-bin barriers are an equilibration bottleneck and reduce the relaxation time scale of the global probability distribution in bin space down to the slowest intra-bin probability transport timescale.

peak barrier transition while the remaining trajectories spend their sampling effort on shallower areas in phase space. It is an intrinsic problem of the WE approach, that the appropriate positioning of bin boundaries requires prior knowledge about the free energy profile along the bin coordinate. However it might be possible to conceive a binning mechanism which consumes the current knowledge about the bin space on-the-fly to adjust bin boundaries adaptively. Note in this context again that bin boundaries do not necessarily need to be isotropically spaced along the reaction coordinate. It is therefore possible to increase the number of bins in regions of steep free energy profile while covering flat free energy barriers with large bins.

### 3.3.2 Orthogonal barriers

Although inefficient binning along the reaction coordinate may produce intra-bin barriers that potentially eliminate the sampling advantage of the WE approach, possible mitigation may arise from a readjustment of bin boundaries or the increase of the number of bins in the region in question. Another less obvious class of intra-bin barriers that defies such mitigation approaches, are barriers orthogonal to the binning coordinate. Figure 3.7 depicts the top view of a two dimensional free energy landscape in contour line representation with coordinates $x, y$ and two defined states $A, B$. The bins discretize the $x$ coordinate in 5 bins and the two global minima are located in $b_1$ and $b_5$ respectively states $A$ and $B$. For a particle to overcome the predominant barrier of $> 5\ k_\mathrm{B}T$ separating the global minima, the transition pathway runs mainly along the $y$ coordinate. The binning coordinate does not reflect the predominant reaction coordinate and does therefore not improve the sampling of barrier transition events. Neither increasing the number of bins nor shifting the bin boundaries provides an effective countermeasure to circumvent this intrinsic issue. A possible approach to reinstate the advantages of WE, would be the binning along both coordinates $x, y$ or the redefinition of the binning coordinate along the predominant transition pathway. Expanding binning to multiple dimensions is however mostly limited to a

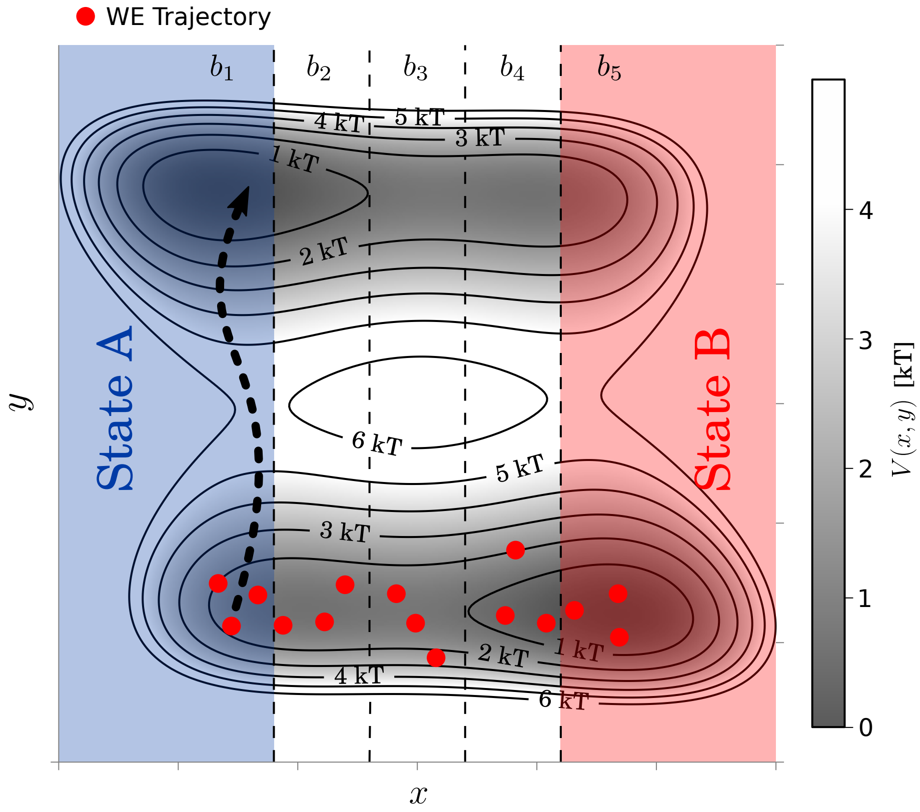

Figure 3.7: Orthogonal barrier crossing in weighted ensemble simulations does not profit from the accelerated sampling along the reaction coordinate and reduces the sampling efficiency down to the slowest orthogonal reaction barrier.

small number of additional dimensions, as the overall sampling effort scales exponentially with $\prod_n N_i$ per number of bins $N_i$ in $n$ dimensions. One further drawback of all re-binning mitigation approaches however is the fact that complete information about the free energy landscape and the association pathway is required but typically not available beforehand.

### 3.3.3 Conformational flooding

Another class of issues that negatively affects sampling efficiency is the conformational flooding of bins. This issue requires an additional degree of freedom in the system that has not been present in the one dimensional toy systems so far. It occurs in situations where a molecule can adopt alternative conformations of significant lifetime. Assuming a case where the association process of a ligand/receptor complex should be simulated with WE and the ligand can adopt two predominant conformations $c_1$ and $c_2$. While the ligand may sample both conformations in solution, only the $c_1$ conformation binds to the receptor in a conformational selection mechanism, while on the other hand the conformation $c_2$ does not fit sterically into the binding pocket. Figure 3.8 depicts the WE setup of the given system. In the optimal case the WE simulation should be started with an ensemble of ligand conformations in bin $b_0$ that reflects the equilibrium distribution of ligand conformations. Upon propagation, the unexplored bins are consecutively populated by splitting those trajectories which first arrive in unpopulated bins. Let us assume

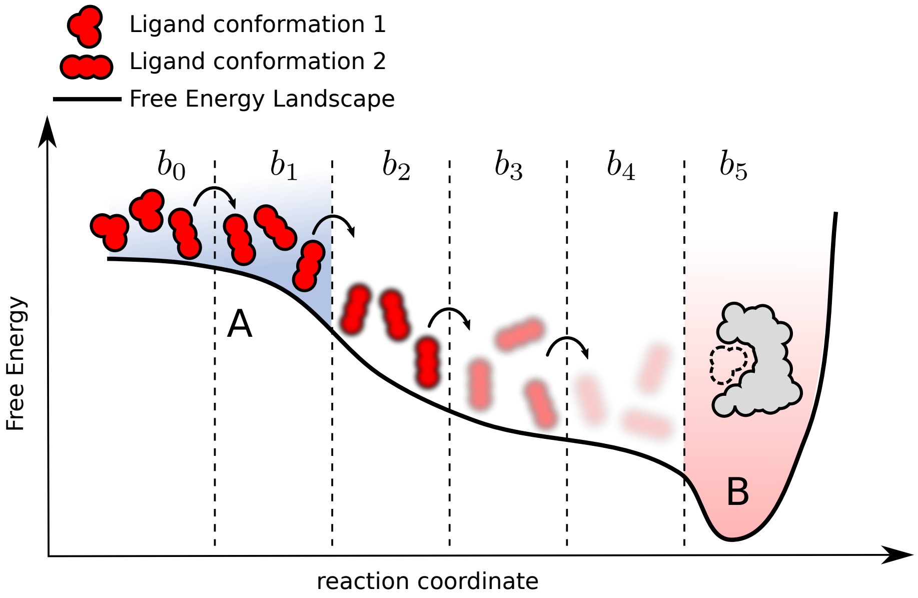

Figure 3.8: Schematic depiction of conformational flooding in WE simulations in case of the association pathway of a ligand/receptor system.

in this example that randomly conformation $c_2$ reaches $b_1$ first and is split until the number of trajectories per bin $n = 3$ is reached leading to a population consisting only of conformations $c_2$ in bin $b_1$. This over-representation of one (random) conformation at the exploration frontier in bin space effects a self-reinforcing population selection mechanism in successively populated bins at the exploration barrier caused by the splitting routine. This effect gets even worse when the transition timescale of the ligand to change its conformation is much larger than the diffusion timescale of trajectories from one end of the bin space to the other. Once all the bins have been populated by the WE algorithm, the relaxation time to reach the equilibrium population of ligand conformations slows down to the relaxation timescale of ligand conformations as these conformational transitions are not specifically accelerated by the approach. The example of conformational flooding accurately depicts the fact that even when carefully setting up the bin spacing in a weighted ensemble simulation the time scale to reach the equilibrium may strongly suffer from any orthogonality in the system. Conformational flooding can be considered as a special form of orthogonal barrier and a possible mitigation could be binning of relevant conformations along an additional reaction coordinate in conformation space. This mitigation however again increases the dimensionality of the bin space and thereby potentially reduces the overall sampling efficiency and requires prior knowledge about the conformational equilibrium ensemble of the ligand.

## 3.4 Discussion

The weighted ensemble (WE) methodology constitutes an interesting alternative to the domain of widely accepted advanced sampling approaches (see chapter 2). The elegant design allows improved sampling of kinetic and equilibrium properties along preselected reaction coordinates. The WE approach does not demand the modification of the system Hamiltonian which in return conserves kinetic properties. These can be rigorously extracted without requiring additional non-rigorous models e. g. fitting of barrier transitions kinetics with exponential rate models or the definition of Markov state models [222, 223]. Interesting proceedings have enhanced the WE approach over the last years e. g. the development of history based transition rates and the probability reweighting method to accelerate reaching the equilibrium probability distribution [207, 208]. Yet, the approach has not achieved the breakthrough to become a widely accepted and applied technique in the community of molecular dynamics scientists. We presented three scenarios of transport bottlenecks in equilibrium WE simulations that depict one fundamental drawback of WE simulations and might be a reason why mainly simple systems comprising a low number of orthogonal degrees of freedom have been published in the literature. For simple model applications, the approach yields improved sampling along the reaction coordinate, however as soon as system properties orthogonal to the binned coordinates strongly affect the equilibrium, the efficiency to reach the equilibrium drops to the slowest orthogonal relaxation timescale. The introduction of appropriate binning along the orthogonal coordinates in order to mitigate this effect is limited due to the exponential explosion of sampling effort introduced by every additional binning coordinate. The main issue of the WE approach is however not the fact that orthogonality may affect the sampling efficiency but that the impact of orthogonality is often difficult to detect in multi-dimensional biomolecular systems. Although all sorts of molecular dynamics simulations in principle suffer from this problem, WE simulations in particular may even perform worse than brute force MD approaches, when orthogonal effects drive the system to states far from the equilibrium (conformational flooding) or consume precious computational resources by sampling trivial phase space regions instead of focusing the sampling effort on important barrier transitions (orthogonal and intra bin barriers). Despite these drawbacks the WE methodology provides in principle a straight forward way to discover transition pathways in complex biomolecular systems without altering the Hamiltonian in an unphysical way. This fact has been first discovered by Dickson et al. who developed the *WEXPLORE* software which aims at exploring physical transition pathways along binning coordinates without claiming that the generated transition pathway ensemble has reached the equilibrium [214, 215]. A ready to use software framework *WESTPA* has emerged from the Zuckerman lab providing easy access to the WE methodology for a large community of simulators [213]. WESTPA interfaces with popular molecular dynamics software suites such as Amber, Gromacs, and more. For the sake of gaining a better understanding and in order to implement the basic principles of the WE algorithm combined with elegant parallelization routines, we developed in the course of this work our own WE software suite *hdWE* which is purely written in python code and therefore provides a maximum of portability to alternative operating system platforms. This is in contrast to WESTPA which contains in its code base a mixture of both Linux shell scripts and python code impeding a porting to other platforms. A plugin framework is provided with hdWE which strongly simplifies the interfacing with any MD software package (see Appendix A).

# Chapter 4

# Protein-ligand docking using Hamiltonian replica exchange simulations with soft core potentials

Molecular dynamics (MD) simulations in explicit solvent allow studying receptor-ligand binding processes including full flexibility of the binding partners and an explicit inclusion of solvation effects. However, in MD simulations the search for an optimal ligand-receptor complex geometry is frequently trapped in locally stable non-native binding geometries. A Hamiltonian replica-exchange (H-REMD) based protocol has been designed to enhance the sampling of putative ligand-receptor complexes. It is based on softening non-bonded ligand-receptor interactions along the replicas and one reference replica under the control of the original force field. The efficiency of the method has been evaluated on two receptor-ligand systems and one protein-peptide complex. Starting from misplaced initial docking geometries the H-REMD method reached in each case the known binding geometry significantly faster than a standard MD simulation. The approach could also be useful to identify and evaluate alternative binding geometries in a given binding region with small relative differences in binding free energy.

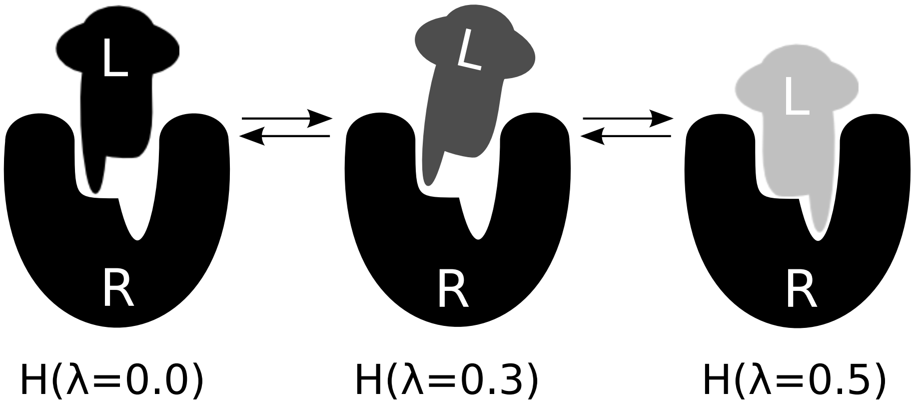

Figure 4.1: Schematics of protein-ligand docking with H-REMD based soft core scaling of non-bonded protein-ligand interactions. Ligand (L) and Receptor (R) are depicted in a key-lock representation. Non-bonded interactions of the ligand are scaled along the replicas to overcome energetic barriers of misplaced conformations in the higher replicas (represented by large $\lambda$). Energetically favorable conformations are swapped back to the reference replica at $\lambda = 0$.

---

Parts of this chapter have been published in [224]

## 4.1 Introduction

The design and identification of putative drug molecules can greatly benefit from the prediction of the structure of ligand-receptor complexes [225, 226]. It involves the computational docking of possible ligand molecules to cavities or ligand binding pockets on the surface of a target biomolecule. The docking search is typically performed without explicit inclusion of the surrounding solvent and no or only limited inclusion of receptor flexibility [162, 227]. Predicted complexes are evaluated by simple scoring functions that include geometrical fit and different energetic contributions [162, 225, 227]. It is frequently possible to correctly predict the recognition site on the receptor protein but the placement of the ligand can significantly deviate from the experimentally observed geometry [228, 229]. The approximations inherent to the scoring function may not allow precise and realistic distinction between different sterically possible complex structures [230]. Furthermore, explicit solvent molecules are mostly neglected in docking approaches to accelerate the computations. However, water molecules can play a significant role in ligand–receptor recognition.

Molecular Dynamics (MD) simulations in explicit solvent offer a possible route for the refinement of incorrectly placed ligands in the binding site. In an MD simulation full atomistic flexibility of both partners (ligand and receptor) can be considered and solvent effects can be included more realistically by explicit waters compared to the approximate scoring functions used in typical docking searches [231, 232]. Indeed, in ultra-long MD simulations it has been demonstrated that the complete process of ligand binding to a receptor cavity can be explored [233, 234]. However, such simulations require many microseconds of simulation time and the ligand can still be trapped for long times in metastable states separated by high-energy barriers from other realistic binding geometries. For practical applications it would be desirable to accelerate the MD search without loss in the accuracy of representing the system and the aqueous environment. In Hamiltonian replica exchange MD (H-REMD) simulations the force field of the system is modified in parallel replica runs that exchange with a reference simulation under the control of the original force field.

An advantage of H-REMD compared to the more common variation of the temperature among replicas (T-REMD) is the possibility to vary only part of the Hamiltonian of the system among replicas. This allows acceleration of transitions along selected degrees of freedom of the system typically requiring fewer replicas compared to T-REMD [166]. A variaty of H-REMD approaches have been described scaling different parts of the force field depending on the specific sampling task [145, 154, 164, 235–240]. Due to the variety of choices, it is not always clear which part of the Hamiltonian to modify among replicas in order to best tackle a given sampling problem. However, in case of ligand-receptor interactions the trapping of the ligand on the surface of a receptor molecule at a non-native locally stable placement is due to non-bonded interactions of the ligand and receptor (or with the solvent). In order to improve the sampling of relevant ligand-receptor structures in MD-refinement simulations, we have applied H-REMD based on softening the ligand-receptor interactions along the replicas. Such softening of non-bonded interactions has been used already to improve the conformational sampling of molecules in solution [145, 235] but not for ligand-receptor docking in explicit solvent. It allows for a broader sampling and accelerated barrier crossing in the replicas which in turn improves the sampling of relevant states in the reference replica. The technique is able to generate a pool of possible ligand orientations in a known binding pocket and to discriminate among those which are energetically favorable.

Significantly better sampling of relevant states compared to standard MD simulations was observed. It is further demonstrated that the method is applicable to the refinement of peptide-protein complexes where side chain placement is uncertain and only backbone conformations are known approximately.

## 4.2 Methods

### 4.2.1 H-REMD Docking

Energetic and entropic driving forces of ligand-receptor complex formation are closely linked to the representation of solvent molecules and ions and often demand a more accurate treatment than given by simplified docking algorithms. MD simulations have the potential to become a valuable tool for elucidating the association of ligand-receptor complexes. Seminal studies of ultralong MD simulations on special purpose machines reproduced in a proof of principle manner the binding event of small ligands to receptor molecules [233, 234]. Recent speedup of MD algorithms with graphics processing units (GPUs) allowed the production of $\mu$s binding trajectories on readily affordable hardware [241]. The computational effort, however, is still tremendous. The major drawback of MD simulations is the incapability of scanning the relevant phase space region in reasonable computing time due to energy barriers. In order to accelerate the generation of conformations and to rapidly scan relevant areas in the ligand-receptor free energy landscape a Hamiltonian replica exchange (H-REMD) method was employed.

The method accelerates the searching of possible ligand binding modes in a known binding site by scaling the non-bonded interactions of the ligand with the rest of the system. A coupling coordinate $\lambda$ is introduced to connect the unmodified Hamiltonian $\mathcal{H}_\mathrm{A}$ of the system with a Hamiltonian $\mathcal{H}_\mathrm{B}$ where all non-bonded interactions of the ligand with the rest of the system are turned off (dummy ligand). The transition from the $\mathcal{H}_\mathrm{A}$ to the $\mathcal{H}_\mathrm{B}$ regime is mapped to parallel replica simulations at different $\lambda_i$ values. The reference simulation at $\lambda = 0$ is connected to replicas at higher $\lambda$ values by the H-REMD technique [167] as depicted in Figure 4.1. The increment of $\lambda$ between the replica has to be chosen such that sufficient phase space overlap ensures reasonable exchange rates. To avoid numerical instabilities at $\lambda$ values close to 1 and 0, the Hamiltonians are connected with soft core scaling potentials [242, 243].

### 4.2.2 Simulation Setup

Simulations were performed with an in house implementation of H-REMD in GROMACS release 4.5 and with release 4.6 of GROMACS which includes the implementation of H-REMD and is significantly faster than release 4.5. Structural data was taken from PDB files and processed with the GROMACS toolchain [96, 244] to solvate the proteins in a truncated octahedron water box with periodic boundary conditions. Crystal water and ions were removed from the PDB file in order to not overlap with manually generated initial ligand placements for FKBP-52 but crystal water was kept for FKBP-12 and MHC class I. FKBP-52 was cleaved after residue Gly139 to obtain the ligand binding subdomain solely. Hydrogen atoms provided by the PDB files were discarded and automatically assigned by the gromacs tool pdb2gmx which derived protonation states for all residues. No possible differences in protonation states between bound and unbound receptor protein were considered. The Amber parm99SB forcefield [245] was used in conjunction with the TIP3P water model [246]. Forcefield parameters for the ligand molecules were parametrized in the general Amber forcefield GAFF [247] with the Antechamber software package [248]. Counterions were added to render the system electrostatically neutral. The box dimensions were adjusted with an initial minimum distance of 1 nm of the solute to the box boundaries. Energy minimization via steepest decent was performed with a convergence criterion of 20 k steps or 100 kJ mol$^{-1}$ nm$^{-1}$ followed by an equilibration simulation in the NVT ensemble for 250 ps and a time step size of 1 fs. A second equilibration phase of the same duration and step size adapted the system to 1.01 bar in constant pressure NPT ensemble. The equations of motion during equilibration were integrated with the leap-frog integrator (MD). Temperature was coupled to the system with the velocity rescale thermostat [47] at 298 K and the Berendsen barostat [45] was used for NPT. During energy minimization and equilibration simulations non-hydrogen atoms of the protein were restraint in space with a harmonic potential at force constant of 1000 kJ mol$^{-1}$ nm$^{-2}$. The bond lengths of H atoms were constrained with the Linear Constraint Solver [249] with a coupling matrix extension order of 12 (4 in production run) as proposed in Ref. [47] Electrostatic interactions were calculated with the particle–mesh Ewald algorithm [250] with a grid spacing of 0.12 nm and cubic B-spline interpolation. The Lennard-Jones interactions were switched to zero after 0.8 nm and a cutoff at 0.9 nm. Long-range dispersion correction resulting from the truncated Lennard-Jones interactions was applied to pressure and energy.

### 4.2.3 Hamiltonian Replica Exchanges

The replica exchange docking simulations were performed with the stochastic velocity Verlet [251] integrator which handles the temperature coupling implicitly (298 K). The Parrinello-Rahman barostat [48] (1.01 bar) was used to generate an NPT ensemble. A 2 fs time step was used. The non-bonded ligand interactions (with receptor and solvent) were scaled in each replica using the soft core scaling method [242, 243] with $\alpha = 0.3$ the soft core power $p = 1.0$ and an interaction radius of $\sigma = (C_{12}/C_6)^{1/6}$ or $\sigma = 0.25$ when Van der Waals parameters $C_{12}$ or $C_6$ were zero.

$$\begin{aligned} V_{\text{softcore}}(r) &= (1-\lambda)V_{\text{A}}(r_{\text{A}}) + \lambda V_{\text{B}}(r_{\text{B}}) \\ r_{\text{A}} &= \left(\alpha\sigma_{\text{A}}^6\lambda^p + r^6\right)^{\frac{1}{6}} \\ r_{\text{B}} &= \left(\alpha\sigma_{\text{B}}^6(1-\lambda)^p + r^6\right)^{\frac{1}{6}} \end{aligned} \tag{4.1}$$

The parameter $\lambda$ allows a smooth scaling of the non-bonded interactions with $\lambda$ = 0 representing the reference states with all interactions switched on and $\lambda$ = 1 the state with all ligand atoms represented as (noninteracting) dummy atoms. To avoid sampling of completely unrealistic ligand placements (with large possible overlap with receptor atoms) a maximum $\lambda$ = 0.54 was used in the highest of 10 replicas and $\lambda$ increased in each replica by $\Delta\lambda = 0.06$. Test simulations indicated that this step size resulted in a reasonable exchange acceptance rate of about 20%. Replica exchanges were attempted every 1000 steps among neighboring replicas ($\lambda_i$ and $\lambda_{i+1}$ where $i$ was alternating between odd or even replica numbers). A further increase of the maximum $\lambda > 0.54$ resulted in a rapid drop of the exchange acceptance rates because many conformations with close ligand-receptor atom-atom distance (and large non-bonded overlap energies in neighboring replicas with smaller $\lambda$ values) were sampled. Note, that the soft core scaling of the electrostatic interactions was only performed for the real space interaction. Due to the shift factor $\alpha$ the non-bonded potential is shifted in realspace which can lead to slight discontinuities of the electrostatic interactions at the cutoff radius (in the replicas but not in the reference replica with the original force field that was used for all the analysis of the simulations).

Starting positions for protein-ligand compounds were generated either by manually rotating and translating the ligand in its binding pocket or by running a continuous MD simulation at $\lambda = 0.54$ and subsequent energy minimization using the full force field. The down scaling of ligand-receptor interactions can cause dissociation of the ligand from the receptor surface. In order to prevent the dissociation of the ligand from the receptor especially in the high $\lambda$ replica windows, a position restraint based on the center of mass (COM) pull code was implemented in GROMACS. A harmonic potential $f(r)$ with force constant $k_{\text{com}} = 1000\,\text{kJ mol}^{-1}\,\text{nm}^{-2}$ along the connection vector of the COMs of ligand and protein was applied when the actual COM distance $r$ moved farther than a penalty value $d$ away from the equilibrium distance $r_{\text{eq}}$.

$$f(r) = \begin{cases} k_{\text{com}}(r - r_{\text{eq}} + d)^2, & r < r_{\text{eq}} - d \\ k_{\text{com}}(r - r_{\text{eq}} - d)^2, & r > r_{\text{eq}} + d \\ 0, & r_{\text{eq}} - d \leq r \leq r_{\text{eq}} + d \end{cases} \tag{4.2}$$

We adjusted the penalty value to $d = 0.7\,\text{nm}$.

### 4.2.4 Test systems

The human FKBP protein (FKBP-52) in complex with the high-affinity ligand FK506 (PDB code 4LAX) and a complex with the lower affinity ligand SB3 [252] (in complex with FKBP-12, PDB code 1FKG) served as test systems for the H-REMD docking simulations. The dissociation constant of SB3 (1,3-Diphenyl-1-Propyl-1-(3,3-Dimethyl-1,2-Dioxyphenyl)-2-Piperidine Carboxylate) and the FKBP domain is $K_{\text{d}} = 10\,\text{nM}$ compared to $K_{\text{d}} = 0.4\,\text{nM}$ of the native FKBP inhibitor FK506. Refinement of docked peptides was tested on the peptide binding domain of a murine MHC class I molecule in complex with a viral antigen (PDB code 2VAB). The antigen sequence is FAPGNYPAL and is derived from the Sendai virus nucleoprotein (324-332), SEV-9 [253].

## 4.3 Results and Discussion

### 4.3.1 FKBP ligand-receptor complexes

MD simulations have been used in the past to refine docked ligand-receptor complexes [254–256]. In order to evaluate the capability of continuous MD (cMD) simulations for refining docked ligand-receptor complexes simulations were started from two initial ligand placements denoted as W1 and W2 that deviated by 0.5 and 0.4 nm of $\text{RMSD}_{\text{ligand}}$ from the native FKBP–FK506 complex (Figure 4.2). The root mean-square deviation ($\text{RMSD}_{\text{ligand}}$) corresponds to the deviation of the ligand atoms from the native structure after best superposition of the receptor structure onto the receptor of the native complex. An $\text{RMSD}_{\text{ligand}}$ of less than 0.3 nm was considered as a near-native binding mode. In order to provide a fair comparison with the H-REMD approach, 10 independent simulations of 20 ns duration for each starting structure W1 and W2 were performed. Starting velocities for each run were assigned randomly according to the Maxwell distribution at temperature 298 K. The cMD simulation setup was chosen such that approximately the same computational resources were used for the comparative H-REMD run with 10 replicas.

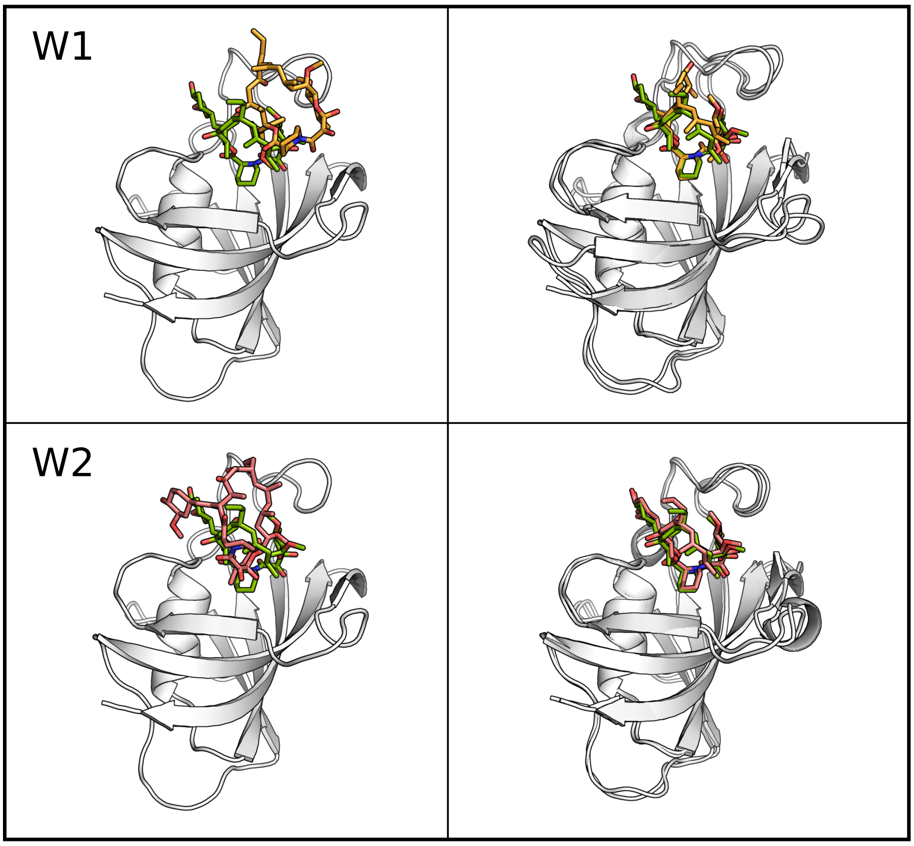

Figure 4.2: Starting (left panels) and final (right panels) W1 and W2 structures for H-REMD simulations of FKBP-52 protein in complex with ligand FK506 Tacrolimus (sticks). Protein backbone structures of FKBP-52 are aligned with the crystal structure (PDB Code 4LAX) and the native ligand binding mode is colored green.

For the first starting structure W1, all cMD simulations were unsuccessful to move the ligand closer to the known binding configuration within 20 ns simulation time (Figure 4.3A). A cluster analysis revealed that simulations were kinetically trapped in the initial ligand configuration or moved away from the native binding mode (Figure 4.3B).

For the second structure W2, two cMD simulations explored the native binding mode after 5 ns and 10 ns respectively, however the majority of trajectories moved away from the native ligand conformation and were trapped in metastable non-native binding modes (Figure 4.3A). The cluster representing the native binding mode covers approximately 10% of the joined 200 ns W2 trajectories but is not the largest cluster.

In contrast to the cMD simulations, the H-REMD simulations resulted in both cases in a rapid drop of the $\mathrm{RMSD}_{\mathrm{ligand}}$ with respect to the native complex structure within less than 1 ns in case of initial structure W2 and 1.3 ns simulation time in case of start structure W1 (Figure 4.3). A cluster analysis of the sampled states in the replica under the control of the original force field identified in both cases ligand placements with an $\mathrm{RMSD}_{\mathrm{ligand}} < 0.3$ nm as most populated states within 15 ns H-REMD simulations (Figure 4.3B).

The ligand binding modes sampled at the final stage of the two H-REMD simulations starting

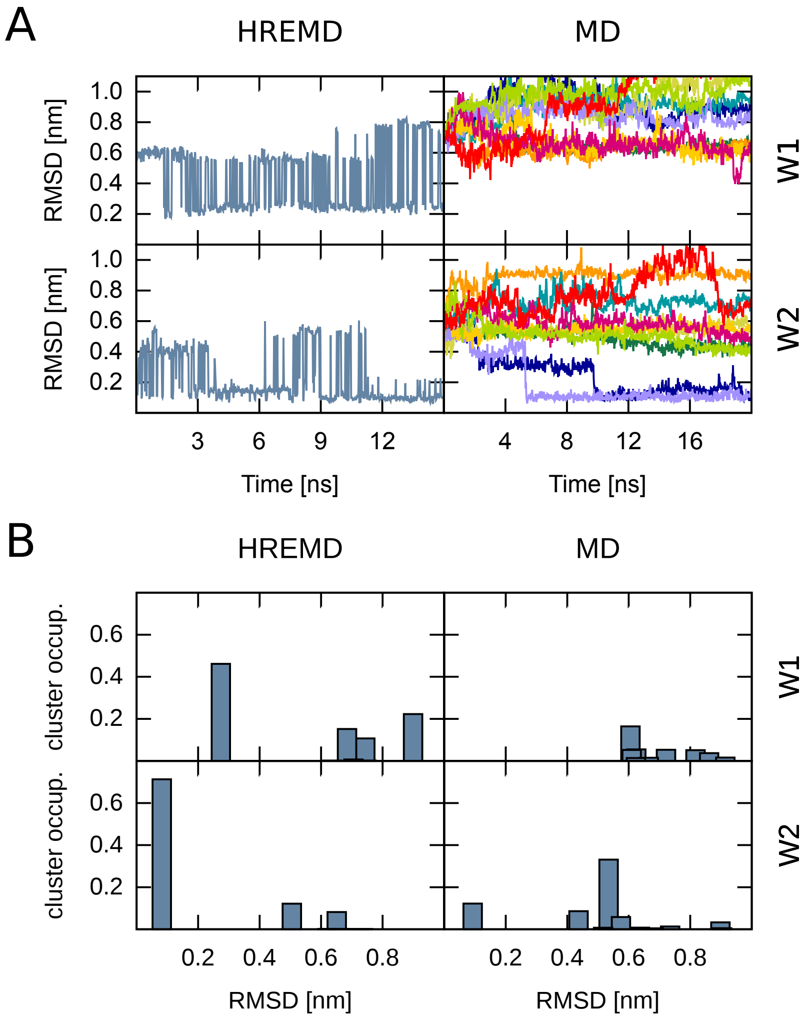

Figure 4.3: H-REMD simulations of FK506 binding to the FKBP-52 protein. Two starting conformations W1 (upper row) and W2 (lower row) of FK506 ligand were used to initially feed H-REMD and control MD simulations (start structures are shown in Figure 2). (**A**) $\text{RMSD}_{\text{ligand}}$ of FK506 versus crystal structure for the H-REMD reference replica (left panel) and 10 continuous MD runs (right panel). The FKBP-52 backbone was aligned with respect to the crystal structure prior to $\text{RMSD}_{\text{ligand}}$ calculations. (**B**) Relative population of the seven largest clusters for H-REMD and MD simulations, respectively, vs. $\text{RMSD}_{\text{ligand}}$. The cluster occupancy is the fraction of cluster members over all frames and gives a measure of the relative cluster size. Binding modes are clustered with the single-linkage method and the cluster $\text{RMSD}_{\text{ligand}}$ to reference crystal structure was calculated for the cluster member with smallest average $\text{RMSD}_{\text{ligand}}$ to all other structures of the cluster. MD runs have been concatenated to one 200 ns trajectory for W1 and W2, respectively, before clustering.

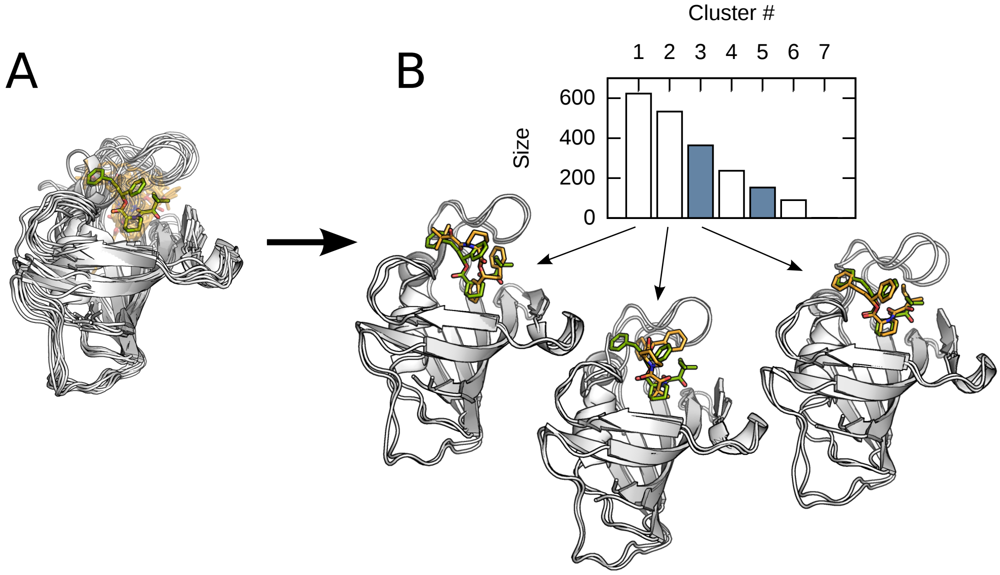

Figure 4.4: H-REMD simulations of FKBP-12 in complex with ligand SB3. (**A**) Pool of initial SB3 conformations for H-REMD (shaded orange) superimposed onto the crystal structure (green). (**B**) Cluster analysis of SB3 binding modes observed in the reference H-REMD window (blue boxes correspond to near native solutions). Average structures of the three most populated clusters are illustrated (predicted ligand as orange vs. native ligand in green stick representation).

from different initial placements differed in $\mathrm{RMSD}_{\mathrm{ligand}}$. The placement that was reached from the second start structure W2 reached a considerably lower $\mathrm{RMSD}_{\mathrm{ligand}}$ than the simulation starting from W1 (Figure 4.3). Inspection of the binding pocket indicated that differences in the side chain conformations of the receptor in the binding pocket are a likely reason for the alternative ligand placements. The Tyr92 side chain hindered the (1S,2S)-2-Methoxycyclohexanol motive of FK506 from moving in the correct configuration (see W1 right panel in Figure 4.2). It should be noted, that during H-REMD only the receptor-ligand interactions were scaled along the replicas but not the side chain–side chain interactions in the binding region. This indicates that the H-REMD approach rapidly identifies near-native binding modes of the FKBP-52–FK506 complex. However, a full equilibration could not be achieved since alternative binding modes still competed during the 15 ns in the reference replica and for both start structure cases slightly different receptor conformations evolved during the search.

For the second test system FKBP-12 with ligand SB3 an alternative strategy was evaluated. Instead of starting from only one ligand-receptor complex different start structures (one in each replica) were used. Each start structure had an $\mathrm{RMSD}_{\mathrm{ligand}} > 0.4$ nm from the native binding mode. Application of 10 independent cMD simulations (each 20 ns) resulted in kinetically trapped complex structures and none of the simulations reached any near-native complex geometry (supp. Material, Figure S1). In contrast, cluster analysis of the H-REMD application resulted in several notable near-native binding modes in the reference replica (Figure 4.4). Cluster 3 was close to the crystal structure binding mode with $\mathrm{RMSD}_{\mathrm{ligand}} < 0.2$ nm. Noteworthy, cluster 5 which showed a quasi symmetric flip of the two phenyl rings in the SB3 1,3-diphenylprobyl subunit compared to crystal structure was otherwise very similar to the experimental geometry (see Figure 4.4).

A 20 ns H-REMD simulation on the crystal structure swapped after 15 ns an alternative binding mode with a rearranged loop region into the reference replica. This binding mode survived for the last 5 ns in the lower $\lambda$ replica and was present in the reference replica after its first occurrence with a 5% probability. This mode was not found in a standard MD simulation starting from the FKBP-12–SB3 crystal structure (data not shown). However, the observation of alternative binding modes for the SB3 ligand with lower affinity compared to the first ligand-receptor system in line with previous ultra-long MD simulations on the same system [257]. It indicates that the present approach could be especially useful to identify putative alternative binding modes for a ligand in shorter simulation time than standard MD simulations.

For a fully equilibrated simulation the relative populations of conformational clusters should directly reflect the relative free energies of the clusters. Hence, the accumulation of ligand placements in the most populated cluster can be used to identify the most likely binding mode in the typical case when the native complex is not known. Indeed, in case of the high-affinity FKBP–FK506 complex the most populated clusters corresponded to the near-native complex (although not with a large gap relative to alternative placements) and its population was larger for the second half compared to the first half of the H-REMD simulation. However, for the second system the near native geometry was sampled but not as most populated cluster (Figure 4.4). In practical applications one is interested to identify putative binding modes as quickly as possible and will not be able to run simulations to full equilibration. It is therefore useful to look into alternative strategies to identify the most likely near-native geometry, eg. by re-scoring of sampled ligand-receptor geometries. Interestingly, for the present systems the average ligand-receptor interaction energy showed a correlation with the deviation of the ligand from the near-native placement (see supp. Material Figure S2). Near-native sampled complexes of FKBP–FK506 showed on average a more favorable interaction energy compared to complexes with larger deviation from the native placement. Presumably this is due to an optimal complementarity of interactions in the native binding placement. For the FKBP–SB3 system the near native placement and an alternative configuration at $\mathrm{RMSD}_{\mathrm{ligand}}\approx 0.5$ nm showed similar interaction energies (supp. Material Figure S2). Hence, for the present cases the interaction energy could be used (besides of cluster population) as a criterion to select possible near-native binding modes. Note, however, that the binding affinity is in general not only determined by direct ligand-receptor interactions but also by other energetic and entropic contributions.

### 4.3.2 Refinement of a MHC class I peptide-protein complexes

MHC class I proteins play a central role in the recognition of antigenic peptides and their representation to the immune system. Class I proteins contain a narrow cleft between two $\alpha$-helices to bind antigenic peptides of 8-10 residues with an extended backbone structure [258]. The receptor conformation and bound peptide backbone conformation is similar for most antigenic peptides of the same length which is confirmed by over 170 crystallized structures in the Protein Data Bank. It offers the possibility to test the H-REMD method on predicting the side chain structure for an approximately known peptide backbone conformation. A peptide start structure with extended backbone and randomized side chain conformations was generated and energy minimized in the binding groove to remove possible sterical clashes (Figure 4.5). The ligand backbone conformation was weakly restrained to the reference coordinates of the crystal structure with a force constant of $1000\,\mathrm{kJ\,mol^{-1}\,nm^{-2}}$ which allowed backbone conformational fluctuations of

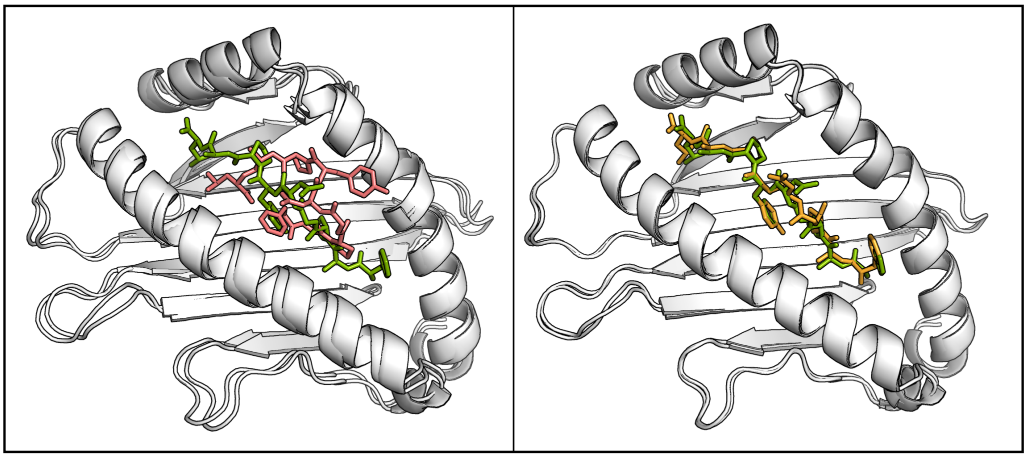

Figure 4.5: MHC class I protein in complex with viral antigen SEV-9. The MHC class I protein backbone structure is aligned with the crystal structure (PDB Code 2VAB). The native antigen peptide configuration is colored green. On the left side the starting structure for MD and H-REMD simulations is indicated in red. The right panel shows the average structure representing the second largest cluster (orange) obtained after H-REMD simulation (green: native ligand conformation). The $\mathrm{RMSD}_{\mathrm{ligand}}$ of the predicted structure is below 0.2 nm with respect to the native ligand placement.

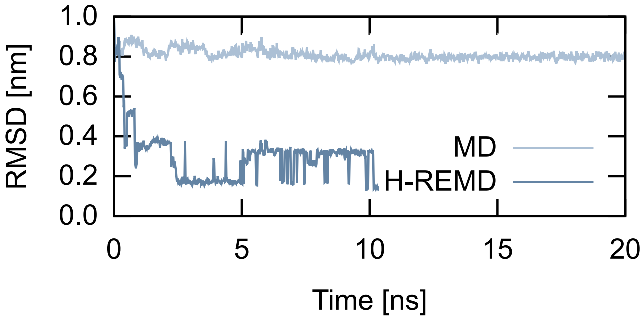

Figure 4.6: $\mathrm{RMSD}_{\mathrm{ligand}}$ vs. simulation time for MD simulation and reference replica of the H-REMD simulation of MHC class I protein in complex with viral protein fragment SEV-9. Simulations were started from a starting structure shown in Figure 4.5. The ligand conformation is trapped near its starting conformation by sterical barriers during the whole MD simulation. In the H-REMD reference window a lower RMSD configuration is already found within $< 0.5$ ns equilibration. Only two main binding modes survive in the reference replica after 0.5 ns which differ mainly in the side chain rotamer of Phe1.

±0.05 nm. In a standard MD simulation of 20 ns the peptide was trapped over the whole simulation time close to the initial conformation due to sterical barriers in the narrow binding cleft of MHC class I protein (Figure 4.6). The cluster analysis for ligand conformations found only one cluster with more than one member. The average structure was similar to the starting structure (Figure 4.5). In contrast, the H-REMD simulations sampled the correct binding mode with most side chains in near-native rotameric states in less than 3 ns. Already during the equilibration phase a conformation with lower ligand RMSD than the starting structure was swapped in the reference replica and only two relevant clusters survived for the rest of the simulation. The two clusters differed mainly in the side chain rotamer of the Phe1 peptide residue. Similar to the FKBP systems (previous paragraphs) a noteworthy correlation of the peptide-receptor interaction energy with respect to deviation from the native complex structure was found (supp. Material Figur S2) indicating that also in this case the interaction energy could serve as an additional criterion to select realistic docking structures.

### 4.3.3 Additional Information

$\text{RMSD}_{\text{ligand}}$ plots of 10 independent MD simulations on the FKBP-12–SB3 complex are depicted in figure 4.7. The $\text{RMSD}_{\text{ligand}}$ was calculated with respect to the native complex. In addition, the correlation of $\text{RMSD}_{\text{ligand}}$ with interaction energy between ligand and receptor for all three test systems is shown 4.8. The correlation was calculated from the reference replica of the H-REMD simulations.

## 4.4 Conclusions

The realistic prediction of ligand-receptor binding geometries is an important goal of in silico drug discovery. Molecular docking methods are widely used to generate and evaluate putative binding geometries mostly employing simple scoring functions that include flexibility of

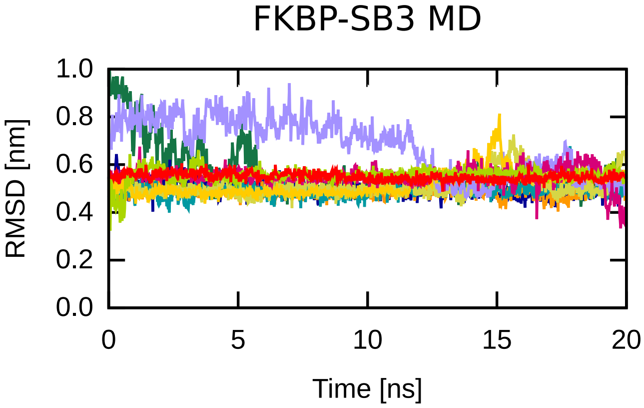

Figure 4.7: Deviation of the SB3 ligand ($\text{RMSD}_{\text{ligand}}$) with respect to the placement in the native complex in 10 independent continous MD simulations of 20 ns duration (indicated by different line colors) on the FKBP-12–SB3 complex starting from different starting structures. Each simulation was trapped in non-native placements and no near native binding geometry (with $\text{RMSD}_{\text{ligand}} <$ 0.5 nm) was sampled.

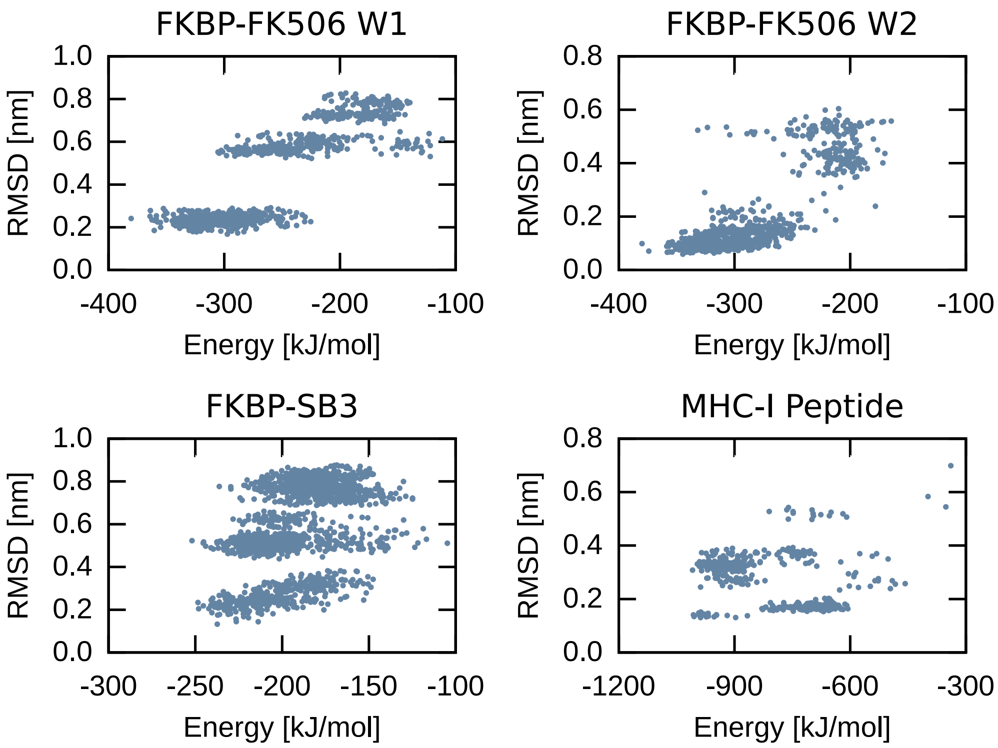

Figure 4.8: Correlation of the deviation of the ligand from the native placement (in terms of $\mathrm{RMSD}_{\mathrm{ligand}}$) and interaction energy between ligand and receptor molecules for conformations sampled in the reference replica during H-REMD simulations. Plots are shown for all three systems investigated by the H-REMD docking approach.

the binding partners and solvation of the ligand and receptor binding site only approximately. MD simulations on the other hand include both full flexibility of binding partners and explicit solvation of partner molecules and are increasingly being used for studying ligand-receptor association events. The major drawback of standard MD simulations, however, is the large computational demand coupled with sampling mostly irrelevant states before reaching a native binding mode. High energy barriers separating favorable putative binding geometries prevent the identification of near-native binding modes in reasonable sampling times. In order to accelerate the search we tested an H-REMD approach based on soft core scaling non-bonded interactions between partners along the replicas. Similar approaches have already been used to enhance conformational sampling of isolated molecules [145, 235] in explicit solvent but not to refine putative ligand-receptor complex structures. The approach requires a smaller number of replicas compared to standard temperature (T-)REMD because the scaling affects only a small fraction of the system. For the present test cases ten replicas were sufficient to significantly reduce barriers, enhance sampling and still allow reasonable exchange rates between neighboring replicas. The method was tested on three systems including a peptide-protein complex. In all cases significant improvement of initially misplaced ligand-receptor complexes and side chain placements in case of the peptide-protein complex was obtained. In contrast, standard MD simulations largely failed on the test cases to refine the initial conformations. The approach is especially powerful to sample a broader range of possible ligand-receptor complex structures compared to standard MD simulation. A full equilibration of sampled states was still not achieved for the ligand-receptor complexes since different distributions of sampled states were obtained depending on the start structure (e. g. in the FKBP–FK506 system). This can be attributed in part to limited sampling but may also be due to alternative receptor conformations evolving during the simulations starting from different initial ligand placements.

A possible extension of the approach is to not only scale ligand-receptor interactions among replicas but also the interactions between side chains of the receptor in the vicinity of the ligand binding site which could result in a simultaneous enhanced sampling of alternative receptor conformations. It should be noted that the computational demand of the approach is much higher than of standard docking tools. However, it should also be emphasized that the approach is not intended for systematic evaluation of thousands of putative ligands or binding modes but to refine a small fraction of preselected putative docking geometries (e. g. obtained form a systematic docking run). With increasing efficiency of MD simulations such approach can become a valuable alternative to simple scoring methods of docking geometries that largely neglect the dynamics of partners and do not include explicit solvent molecules.

Another interesting possible application is the generation of alternative binding geometries for a given ligand in a binding site that are of similar binding free energy. We identified an alternative binding mode of the FKBP-12–SB3 system in a H-REMD simulation that started from crystal structure configuration. The finding agrees with previous MD studies which identified alternative binding modes for the same complex [257]. The population of different binding modes in a given binding site reflects the relative stability of these modes and may give hints which interactions (chemical groups on the ligand or possible chemical ligand modifications) may stabilize or destabilize a given binding geometry.

# Chapter 5

# Covalent dye attachment influences the dynamics and conformational properties of flexible peptides

Fluorescence spectroscopy techniques like Förster resonance energy transfer (FRET) and fluorescence correlation spectroscopy (FCS) have become important tools for the *in vitro* and *in vivo* investigation of conformational dynamics in biomolecules. These methods rely on the distance-dependent quenching of the fluorescence signal of a donor fluorophore either by a fluorescent acceptor fluorophore (FRET) or a non-fluorescent quencher, as used in FCS with photoinduced electron transfer (PET). The attachment of fluorophores to the molecule of interest alters the molecular properties and may affect the relevant conformational states and dynamics especially of flexible biomolecules like intrinsically disordered proteins (IDP). Using the intrinsically disordered S-peptide as a model system, we investigate the impact of terminal fluorescence labeling on the molecular properties. We perform extensive molecular dynamics simulations on the labeled and unlabeled peptide and compare the results with *in vitro* PET-FCS measurements. Experimental and simulated timescales of end-to-end fluctuations were found in excellent agreement. Comparison between simulations with and without labels reveal that the $\pi$-stacking interaction between the fluorophore labels traps the conformation of S-peptide in a single dominant state, while the unlabeled peptide undergoes continuous conformational rearrangements. Furthermore, we find that the refolding rate of S-peptide is decreased by at least one order of magnitude by the fluorophore attachment. Our approach combining experimental and *in silico* methods provides a benchmark for the simulations and reveals the significant effect that fluorescence labeling can have on the conformational dynamics of small biomolecules. The presented protocol is not only useful to compare PET-FCS experiments with simulation results but provides a strategy to minimize the influence on molecular properties when designing fluorescence labeling.

## 5.1 Introduction

Changes in the three dimensional structure on various timescales are an omnipresent feature of proteins and other biopolymers. Fluorescence spectroscopy techniques provide a useful toolset to investigate the dynamics and extent of these structural rearrangements *in vitro* and *in vivo* [259–263]. One of the most common approaches is the use of Förster resonance energy transfer

---

This study was performed in collaboration with the research group of Prof. Don C. Lamb, Department Chemie, Physikalische Chemie, Ludwig-Maximilians-Universität München. PET-FCS and CD spectrum measurements were performed by Anders Barth and Alvaro H. Crevenna.

(FRET) between two fluorophores attached to the molecule of interest [264]. The radiationless transfer of energy from the excited donor dye to the red-shifted acceptor dye depends on the relative orientation and distance between the fluorophores. Its high sensitivity in the range of 20–100 Å renders the effect interesting for experimental determination of distances at the molecular scale. Another example of radiation-less energy transfer is photoinduced electron transfer (PET) [259, 260, 265–268]. The excited-state energy may dissipate via electron transfer from the fluorophore to the quencher or vice versa, depending on the redox potentials of the excited state fluorophore and quencher. Relaxation to the ground state then occurs non-radiatively by charge recombination of the radical donor/acceptor ion pair. The timescale of the PET reaction resides in the range of femtoseconds to picoseconds [269, 270] which is significantly faster than the fluorescence lifetime of the fluorophores of typically a few nanoseconds [271, 272]. The efficiency of PET decays exponentially with distance on the length scale of a few Å, showing effectively an all-or-nothing quenching behaviour. PET enables the *in vitro* time resolved detection of closed and open contacts between fluorophore and quencher in proteins and other biomolecules. Due to the on-off characteristics of PET, it is commonly used with fluorescence correlation spectroscopy (FCS) [273] to study the timescale of the dynamic changes of the fluorescence signal. FCS is based on the analysis of the time correlation of the detected signal and is thus sensitive to all processes that affect the fluorescence signal. Most commonly FCS is being used to study the diffusion properties of molecules [274], but it is also a powerful tool to study conformational dynamics when combined with FRET or PET [275, 276].

In proteins, tryptophan is the only naturally occurring amino acid which possesses photophysical properties that enable fluorescence quenching through PET [277]. If tryptophan is not part of the protein sequence, a PET experiment requires the artificial inclusion of a Trp residue in the region of interest by either mutation or the adherence of an additional residue. Trp residues which might interfere in an undesirable way with the fluorophore need to be deleted from the protein. The labeling of proteins with fluorophores usually requires modification of the protein sequence. Typically, labeling is performed by reacting a maleimide derivative of the fluorescent dye with cystein residues in the protein. To achieve specificity, this approach requires the removal of natural cystein residues or the introduction of additional cystein residues by mutation. Other labeling approaches target amino groups or rely on bioorthogonal labeling strategies based on the addition of unnatural amino acids [278]. After attaching the fluorphore to the molecule of interest, careful control measurements have to be performed to ensure that the photophysical properties are not altered by the local environment, and that no sticking interactions occur which would impair the rotational freedom of the fluorophore.

Although fluorescence techniques have been applied successfully on a variety of systems [262, 279, 280], usually little information is available to what extent the structural or dynamical properties of the studied system are modified by the attachment of the fluorescence labels. Molecular dynamics studies can help in the interpretation of experimental results and detection of potential artifacts introduced by the dye label [281–283]. As the fluorophores typically exceed the size of naturally occurring amino acids, it is expected that at least the local diffusivity is modified. Furthermore, many readily available fluorophores contain rigid ring systems which function as light absorbing centers (e. g. oxazine derivatives MR121, Atto655 and Atto Oxa11) and potentially facilitate hydrophobic or $\pi$-stacking interactions with other aromatic ring structures especially of the quencher. A significant influence of the fluorescent label on the local structure, the conformational dynamics and the overall functionality of the protein can thus usually not be excluded, and careful controls have to be performed to ensure the validity of the experimental results.

In the present study we conducted a comparative *in vitro* and *in silico* study on the 14 amino acid long truncated S-peptide, which historically served as a model system for intrinsically disordered peptides [284–288]. We N-terminally attached fluorophore Atto655 and added a tryptophan residue to the C-terminus serving as a fluorescence quencher [259]. For the labeled peptide the fluorescence quenching autocorrelation was measured and analyzed with respect to the dynamic contribution, which is a measure for the end-to-end dynamics of the peptide chain. Additionally, we performed extensive, continuous molecular dynamics (MD) simulations on S-peptide with and without the fluorescence labels, to provide atomistic insight into the dynamic and conformational regime of the peptides.

Quantitative agreement of the experimental quenching autocorrelation of labeled S-peptide and *in silico* results was obtained. Comparison of the simulations reveals, that the dynamical and conformational regime of S-peptide was significantly altered by the attachment of Atto655 and Trp15. This study sheds light on systematic modifications of macromolecular properties introduced by fluorescence labeling and provides valuable insight for the design of future fluorescence spectroscopy experiments.

## 5.2 Results

### 5.2.1 MD simulations

Labeled and unlabeled versions of S-peptide were extensively simulated with molecular dynamics simulations (MD) for 30 µs. To give qualitative insight in the refolding dynamics of intrinsically disordered S-peptide, the evolution of the RMSD with respect to the starting structure was calculated (Figure 5.1) and trajectories were visually inspected. To allow direct comparison between RMSD regimes of both systems, the RMSD was only calculated for residues 1 to 14 (without Atto655 and Trp15).

Unlabeled S-peptide rapidly fluctuated between conformational modes on the timescale of several nanoseconds as expected for an intrinsically disordered peptide. Fluctuation of the RMSD indicated no stable conformation surviving in the microsecond time regime throughout the whole simulation. Labeled S-peptide, however, showed a significantly reduced bandwidth of RMSD fluctuations with several plateaus in the RMSD evolution. Visual inspection confirmed metastable states surviving for several microseconds during the trajectory. Many configurations revealed close contacts between the two ring systems of terminal Atto655 and Trp15 indicating a strong stacking interaction that traps the system in a quenched state. After about 12 µs the backbone locked into a stable $\beta$-sheet like configuration and remained in this state until finally folding to a $\beta$-sheet structure after 29 µs (see Figures 5.1,5.2 and conformational regime clusters #1 and #4).

### 5.2.2 Conformational regime

Conformations from MD trajectories were clustered for both systems separately to quantify the impact of labeling on the conformational regime. Clustering along the RMSD was performed with the single-linkage algorithm using a RMSD cutoff of 0.25 nm and $10^5$ frames from each

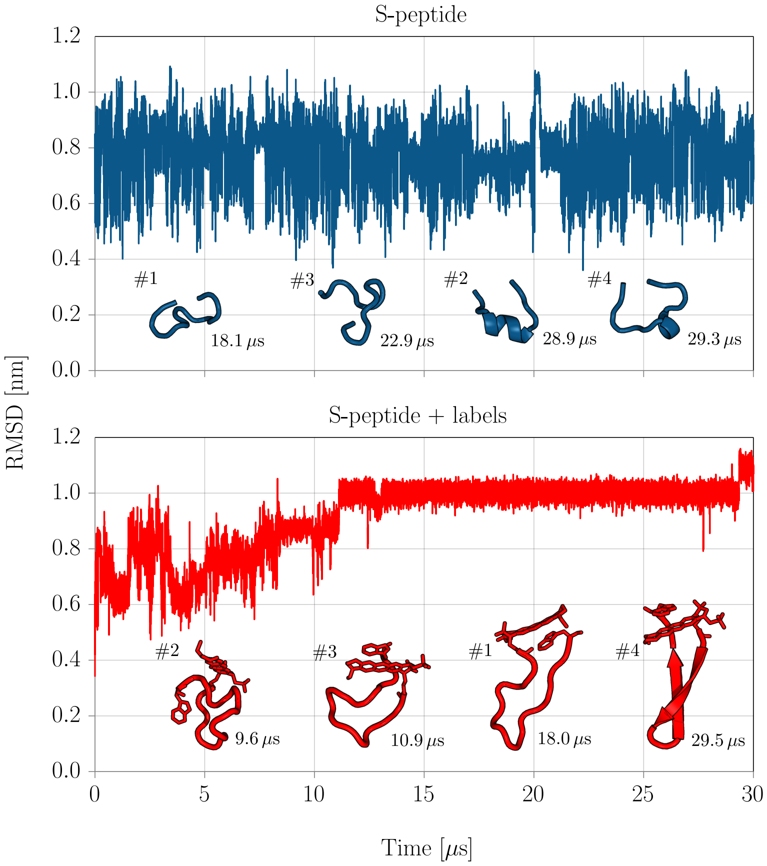

Figure 5.1: The RMSD of heavy atoms of residues 1-14 with respect to the unfolded starting structure for simulations with (lower panel) and without (upper panel) labels. The mean structures of the respectively four largest clusters are shown and their cluster index is indicated (#). Additionally the timestamp of the clusters mean structures occurrence during the simulation is indicated at the bottom right of each cluster structure.

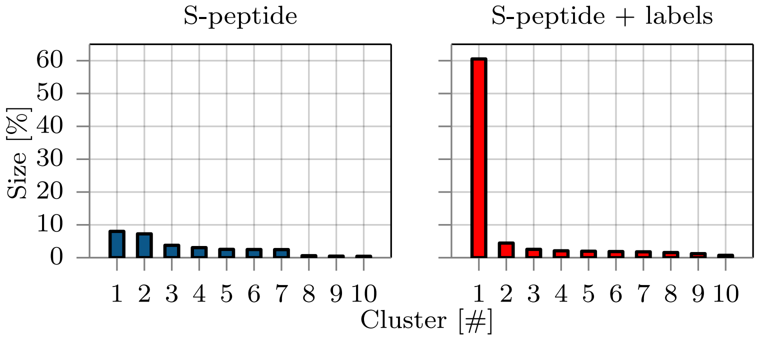

Figure 5.2: The population size of the ten largest clusters in percentage of livetime compared to the whole trajectory from simulations with (right panel) and without (left panel) fluorescence labels. Clustering was based on the RMSD and the single linkage algorithm [289] with a 0.25 nm cutoff was used.

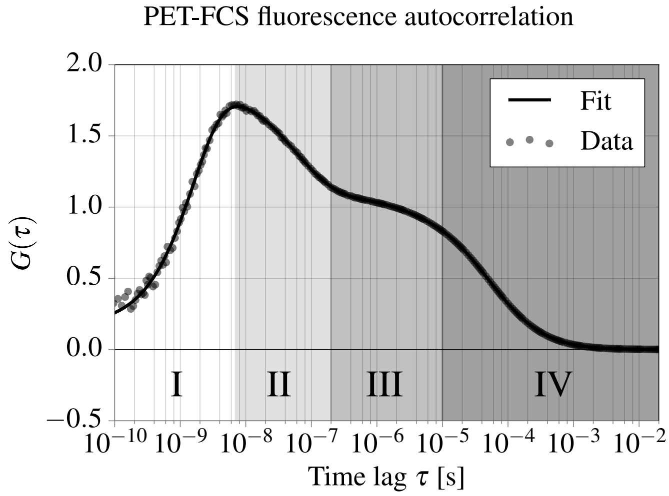

Figure 5.3: Experimentally obtained FCS curve and model fit function for labeled S-peptide. Indicated are the four main time regimes of the relevant processes. I: Photon antibunching, II: Chain dynamics, III: Photophysics, IV: Diffusion.

trajectory [289]. Clusters were sorted and numbered by their frequency of occurrence and the distribution of the ten biggest clusters is shown in Figure 5.2.

As expected, cluster sizes of unmodified S-peptide reflect the typical conformational behavior of an intrinsically disordered peptide. The decrease in probability with increasing cluster index is relatively moderate suggesting low free energy differences between neighboring clusters. A total of about 4700 clusters was found where S-peptide adopted even the configuration of the largest cluster only during 6% of the total simulation time. The mean structures of the first four clusters are depicted in figure 5.1 and give insight into the variability of conformations.

With the attachment of Atto655 and Trp15 to the termini of S-peptide, the conformational behavior however changed significantly. The variability of clusters narrowed down to about 1000 different clusters, with many showing stacked Atto655/Trp15 configurations. The largest cluster, found between 12–29 µs, dominates the conformational regime with a probability of over 60% and indicates a shift from intrinsic disorder to a meta stably folded peptide, reducing the conformational variability significantly. Three out of four mean structures of the largest clusters show strong stacking interaction between terminal labels (figure 5.1).

### 5.2.3 PET-FCS measurements

The dye-quencher dynamics of labeled S-peptide were measured with a PET-FCS setup experimentally. The full PET-FCS correlation function is shown in Figure 5.3. Since FCS is sensitive to all processes that affect the fluorescence signal, multiple phenomena are observed (Regimes I-IV). Photon antibunching is observed on the timescale of the fluorescence lifetime of $\sim$ 2 ns, a typical property of quantum emitters [290]. Diffusion through the confocal volume occurs on timescales between several tenths of µs to ms, depending on the size of the observation volume and the diffusion coefficient of the molecule. Most fluorescent dyes can undergo intersystem crossing from the excited singlet state into a dark triplet state with lifetimes in the range of several µs. Any conformational dynamics are superimposed onto these processes. The fast

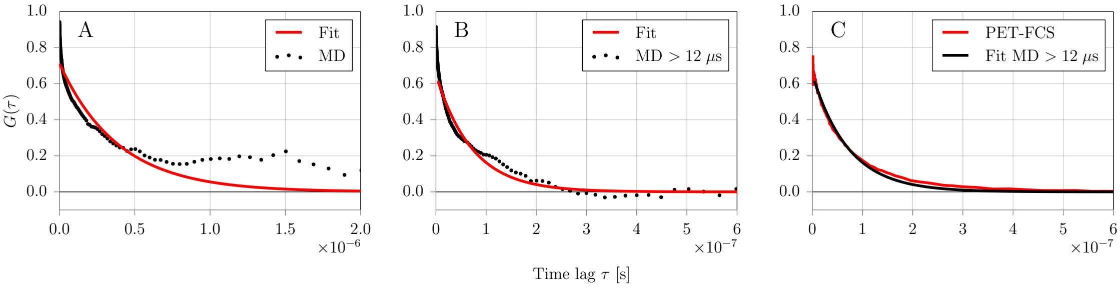

Figure 5.4: Atto655/Trp15 fluorescence quenching autocorrelation data fitted with a two-state exponential model function. Data and fits are shown for MD simulations (A, B) and experimental PET-FCS measurement (C). (A) Data from MD calculated over the whole simulation time (30 µs). (B) Initial 12 µs of MD data omitted. (C) Dynamic part of the correlation curve from experimental PET-FCS measurement (red) overlayed with the fitted MD data collected after 12 µs (black).

chain dynamics of intrinsically disordered peptides or unfolded proteins usually occur on the submicrosecond timescale [291], while slower conformational dynamics involving large conformational rearrangements usually take place in the range of ms to s [292].

The quenching contribution to the correlation function is indicated in regime II (Figure 5.3). To determine the timescales of the chain dynamics, we fit the correlation function with a model accounting for the listed contributions (Equation (5.3)). By careful inspection of the weighted residuals of our correlation fit function, we find, that the addition of a second dynamic contribution to the model function improves the quality of the fit significantly (Figure 5.7). It remains however unclear, whether the second component arises from an alternative conformational regime of S-peptide (see discussion). Therefore, we limit our discussion to the average timescale and overall amplitude of the two components, because the length of the MD simulation is not sufficient to address the existence of two dynamic contributions. To directly compare the measured dynamics to the correlation functions obtained from MD, we isolate the dynamic contribution by dividing the correlation function by the contributions of diffusion, photophysics and antibunching (Figure 5.4 B). We converted the observed amplitude and relaxation time to off- and on-rates by equations (5.4) and (5.5) yielding an off-rate $k_{\text{off}} = 5.0\,\mu\text{s}^{-1}$ and an on-rate $k_{\text{on}} = 7.4\,\mu\text{s}^{-1}$.

### 5.2.4 Fluorescence quenching

Configurations of Atto655/Trp15 from simulation were classified as "dark" state when the distance between the geometric centers of their ring compounds was below a quenching distance of $r^* < 0.55$ nm or as fluorescent otherwise [293, 294]. The quenching autocorrelation function was fit to a two-state kinetic model (Equation (5.1)). Due to the global conformational rearrangements of S-peptide during the initial 12 µs and the associated metastable states with lifetimes in the microsecond regime, the convergence of quenching autocorrelation data was insufficient. Especially large correlation times $\tau > 500$ ns suffered from the lack of sampling (Figure 5.4, A). The long-lived metastable states however dominated the quenching dynamics resulting in relaxation timescales of $\tau_r = 391$ ns (Table 5.1).

Because S-peptide locked in a quasi-stable folded $\beta$–sheet like configuration after 12 µs simulation time, we decided to treat the initial 12 µs as equilibration time and recalculate the quenching

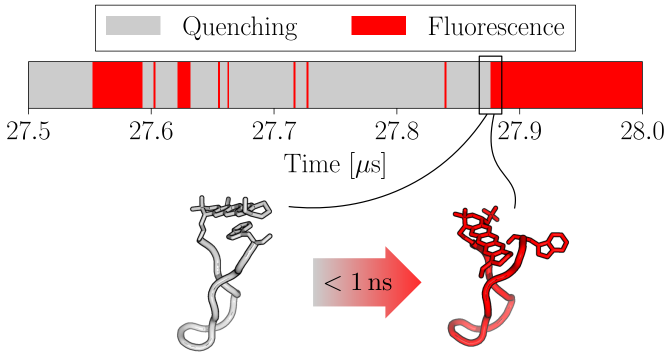

Figure 5.5: Quenched and fluorescent conformations in the timeframe between 27.5–28.0 µs are shown over time. Although the backbone conformation was locked in cluster #1 during this timeslot, spontaneous unstacking of Atto655/Trp15 was observed. Two exemplary structures shortly before and after an unstacking event are shown below. The stacked (grey) configuration quenches the Atto655 fluorescence, while the unstacked (red) configuration allows fluorescence. Unstacking was observed to occure on sub nanosecond timescales.

autocorrelation for only the second part of the simulation with subsequently refitting the two-state model (Figure 5.4, B). The resulting quenching relaxation timescale of $\tau_r = 72$ ns was about 4-5 times faster as dye and quencher could not diffuse far away from each other by the confined $\beta$-sheet like backbone structure. In the investigated simulation time window (12–30 µs) no global backbone rearrangements of S-peptide were observed and lifetimes of quenching states of the fluorescence labels were found in the range of hundreds of nanoseconds (see Figure 5.5). Omitting the initial 12 µs as equilibration also led to a significant improvement of coincidation between the two-state exponential fit model and the simulation data especially for long relaxation times $\tau_r$. Similar to PET-FCS data treatment we also calculated on and off rates by equations (5.4) and (5.5) from fluorescence amplitude and relaxation time. Comparative data between experiment and simulation for fitted relaxation parameters and rates are shown in Table 5.1. We find that the quenching dynamics for simulation data after 12 µs are in good agreement with the experimental results.

| | $a_r$ | $\tau_r$ [ns] | $k_{\text{on}}$ [$\mu$s$^{-1}$] | $k_{\text{off}}$ [$\mu$s$^{-1}$] |
|---|---|---|---|---|
| PET-FCS | 0.67 | 80.4 | 7.4 | 5.0 |
| MD > 12 µs | 0.65 | 72.1 | 8.4 | 5.5 |
| MD all | 0.71 | 391.3 | 1.5 | 1.1 |

Table 5.1: Relaxation time scales ($\tau_r$) and amplitudes ($a_r$) of the fluorescence autocorrelation fitted with an exponential two-state model. Data is shown for experimental PET-FCS measurements and MD simulations. For PET-FCS data, the sum of amplitudes and the average relaxation time of the two dynamic components are shown. Additionally relaxation times and amplitudes have been converted to microscopic on and off rates of the related quenching process with formulas (5.4), (5.5). The rates accord with the average opening and closing frequency of quenching contact formation between dye and quencher in labeled S-peptide.

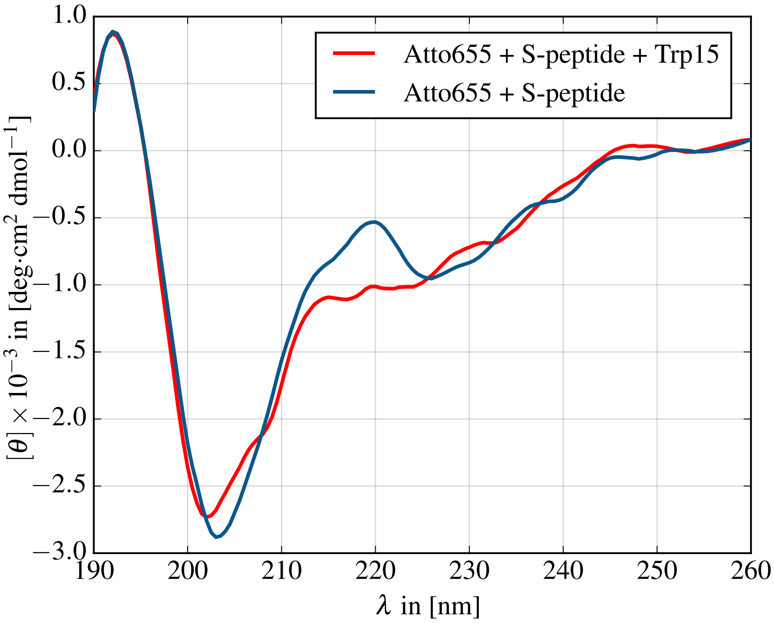

Figure 5.6: Mean residue elliptisity from circular dichroism spectra of labeled S-peptide with and without Atto655 stacking partner Trp15. The peak at 220nm indicates residual $\alpha$-helix formation in labeled S-peptide without Trp15.

### 5.2.5 Refolding dynamics

To characterize the effect of the fluorescent labels on the dynamics of S-peptide, we analyzed the end-to-end distance $d$ of the peptide as a measure for the refolding dynamics. The distance $d$ was calculated between C-$\beta$ atoms of residues Lys1 and Asp14 again for both systems with and without labels. We split the distance ensemble in two regimes, to characterize the switching dynamics between a folded and unfolded regimes of S-peptide. Distances $d < 1.3$ nm at immediate contact were assigned to a “close” regime while distances $d > 2.5$ nm were assigned to an “open” regime. By counting the number of transitions from one regime to the other and dividing it by the total simulation time, a mean refolding rate of opening and closing events of the peptide was calculated. For S-peptide without labels a refolding rate of 34.9 µs$^{-1}$ was found while refolding dynamics for labeled S-peptide were slowed by more then one order of magnitude to 2.7 µs$^{-1}$. Note, that this refolding rate definition only roughly correlates with the $k_{\text{on}}$ and $k_{\text{off}}$ rates of quenching contact formation which are related to fluorescence autocorrelation model via formulas (5.4), (5.5). The quenching rates describe the frequency of opening and closing quenching contact between fluorophore Atto655 and quencher Trp15. This rates can be extracted from both experiment and simulations and serve to compare the agreement between those. However, to compare the dynamics between simulations with and without fluorescence labels, we defined a two-state model (open/close) based on a feature that both systems have in common.

### 5.2.6 Circular dichroism spectra

To identify to what extend the $\beta$-sheet conformation that labeled S-peptide adopts in simulation is part of the equilibrium regime and to measure the modification in the conformational regime by dye-quencher interactions, we measured circular dichroism spectra of labeled S-peptide and compared it with the spectrum of S-peptide labeled with Atto655 but without Trp15 (Figure 5.6). When Trp15 is added to the peptide the peak at 220 nm vanishes, indicating an increase in $\beta$-sheet structure induced by the $\pi$-stacking interaction between the termini. The CD spectrum reveals that the conformational regime is not dominated by only $\beta$-sheet structures but includes contributions of helical and random coil conformations. This supports the results from the 30 µs simulation where labeled S-peptide adopted $\beta$-sheet conformation only during a fraction of the simulation time and exposed random coil backbone configurations otherwise (Figure 5.1, cluster #1).

## 5.3 Discussion

The interpretation of fluorescence spectroscopy measurements depends on the assumption that the artificial attachment of fluorophores does not alter the conformations and dynamics of the target molecule itself. We performed comparative MD simulations and PET-FCS measurements on fluorescently labeled 14 residue S-peptide, serving as a model system of an intrinsically disordered peptide. Our results reveal that the labeling strongly affects both the conformational and dynamical properties of S-peptide, shifting it from the disordered conformational regime to a semi-stable fold with $\beta$-sheet content. The reason is the strong $\pi$-stacking interaction between the rigid ring systems of Atto655 and Trp15, which traps the termini of the peptide to remain at close distance. This effect may occur with many widely used fluorophores containing rigid ring systems as their light absorbing centers that facilitate stacking interactions with other aromatic ring structures.

The agreement between quenching correlation functions from MD simulations and PET-FCS measurement is surprisingly good, although only when skipping the initial 12 µs of the MD trajectory as equilibration time (Table 5.1). As the starting structure of labeled S-peptide was chosen from the regime of the unlabeled S-peptide (helical structure), we assumed that the system requires some time to reach equilibrium and treated the initial 12 µs as equilibration time, thereby dividing the dynamic S-peptide regime in two characteristic regimes. The first part of the simulation was dominated by global backbone rearrangments and slow quenching dynamics as the peptide backbone continously stretched and refolded. During the second part the S-peptide backbone locked into a meta-stable fold with only the label side chains stacking interaction opening and closing from time to time. Simulation time, however, is finite, therefore it remains unclear to what degrees these two regimes contribute to the equilibrium regime. As we reproduce good agreement between simulation and experiment for the second regime with faster quenching dynamcis, we assume that S-peptide preferably resides in this regime. However, the fit quality of experimental PET-FCS data can be improved by adding an additional dynamic term with a slower relaxation coefficient. We speculate, that this second dynamic term might correspond to the regime of global rearrangments seen in the initial part of MD simulations.

The good agreement between experiment and simulation is even more surprising as the quenching state is judged only by a simple distance criterion which was proposed earlier by Vaiana et al. [293] and was based on a fit of fluorophore MD trajectories to correlation spectroscopy data. However, previous studies successfully demonstrated that the connection between atomistic MD simulations and experimental PET-FCS measurements can be drawn with such a simple criterion [294]. Gaining a more detailed understanding of the relation between dye-quencher orientation and fluorescence quenching would require to derive redox potentials for single MD frames via quantum chemical calculations. However we use the experimental PET-FCS measurements only for the validation of our labeled S-peptide simulations, whereas the comparison between simulations with and without labels reveal the stark influence of fluorescence labeling on the peptide dynamic and conformational regime. Our circular dichroism measurements on labeled and unlabeled S-peptide strongly support these findings.

The influence of the local environment on the photophysical properties of fluorophores is a well-known artifact and can be addressed experimentally. Changes in the absorption and emission spectra or quantum yield (and related, lifetime) all affect the Förster radius in FRET experiments, requiring careful control measurements. It is additionally crucial that the rotational freedom of the fluorophore is not compromised by geometric hindrance, specific interactions with aromatic rings, or electrostatic or hydrophobic interactions, as the FRET efficiency critically depends on the relative orientation of donor and acceptor fluorophores. The rotational freedom of the fluorophore can be addressed by means of the fluorescence anisotropy [295]. Our study, on the other hand, sheds light on the effect of the chemical modification on the target biomolecule. Certain considerations need to be taken into account to prevent that the fluorescence label affects the conformational and dynamical properties of the target biomolecule in an undesirable way. In particular PET-FCS experiments, which intrinsically require that fluorophore and quencher can come at close proximity, suffer from possible interactions between them. The influence of specific dye-quencher or dye-dye interactions on conformational dynamics is especially pronounced when unstructured systems like unfolded or unstructured proteins as presented here are studied, for which PET and FRET are often the method of choice. In general, it should be good advise to use fluorescence labeling only in systems where the size of the dye is significantly smaller than the target molecule to avoid that a single dye interaction dominates the conformational properties. The position for the fluorescent label should be chosen in a well-structured and solvent-exposed part of the protein sequence to minimize the influence on the local structure and dynamics, especially when partly disordered proteins are studied. We demonstrate that MD is a valuable tool to quantitatively analyze PET-FCS experiments and relate the measurements to atomistic details. Furthermore, simulations provide a useful strategy to predict the influence of labeling configurations on systematic properties when designing fluorescence experiments.

## 5.4 Methods

### 5.4.1 Molecular Dynamics

Two simulations of S-peptide with and without labels were started from extended peptide conformations. The periodic box boundaries were set at minimal distance of 1 nm to the peptide. The peptide was parametrized with the Amber99sb-ILDN forcefield [42]. Parameters for the Atto655 fluorophore were generated as follows. First the electron density was calculated using Gaussian09 [296] with B3LYP [297, 298] at 6-31G* basis set level [299], second the partial charges were generated with the restraint electrostatic potential protocol (RESP) [300]. Finally, atom types and bonded/nonbonded parameters were assigned from the general amber forcefield GAFF [247]. Both systems were energy minimized and equilibrated in NPT ensemble (298 K, 1.01 bar) after the addition of explicit solvent Tip3P water molecules [246]. Respectively four positive and negative counter ions were added to the solution keeping the box charge neutral. The production run was performed at timestep of 3 fs with the GROMACS 5.0 [96, 244] MD software suite using the velocity rescaling thermostat in NVT ensemble (298 K) [301].

### 5.4.2 Peptide synthesis

The 14 amino acid truncated S-peptide (KETAA AKFER QHMD) was used as a model system for an intrinsically disordered peptide [284–287]. To fluorescently label the S-peptide for PET-FCS measurements, the fluorophore Atto655 was attached to the N terminus and an additional Tryptophan was added to the C terminus serving as a fluorescence quencher [260]. The resulting sequence of labeled S-peptide was Atto655 KETAA AKFER QHMDW. Peptide synthesis, labeling and purification was performed as described previously [302].

### 5.4.3 PET-FCS Measurements

PET-FCS measurements were performed on a custom-built confocal single-molecule fluorescence microscope. The sample was excited with a diode laser (LDH-D-C-640, PicoQuant) operated in continuous wave mode at an average laser power of 280 μW as measured before the aperture of the objective. The fluorescence signal was passed through a pinhole and split on two avalanche photodiodes (SPCM-AQR-14, Perkin Elmer Optoelectronics) by a 50:50 beam splitter to avoid detector dead time in a Hanbury Brown-Twiss-type detector arrangement [303]. Fluorescence signal was passed through an emission filter (ET670/30, AHF Analysentechnik). The range of the emission filter was chosen as to avoid detector crosstalk due to the breakdown flash of the APDs [304]. Individual photon arrival times were recorded with 16 ps resolution using time-correlated single photon counting (TCSPC) hardware (HydraHarp400, PicoQuant). S-peptide was dissolved in standard PBS buffer with 0.005% Tween-20 to prevent sticking to the cover slide surface, and diluted to a final concentration of $\sim$ 1 nM. FCS data was collected over a time of 16 h at room temperature.

### 5.4.4 Quenching Autocorrelation

Configurations of Atto655/Trp15 from simulation were classified as “dark” state when the distance between the geometric centers of their ring compounds was below the quenching distance of $r^* <$ 0.55 nm or as fluorescent otherwise [293, 294]. To calculate the quenching autocorrelation $10^6$ frames from the 30 μs MD trajectory were analyzed. The quenching signal autocorrelation data was then fitted with a two-state model of single exponential signal decay

$$G_{\text{dyn}}(\tau) = a_r e^{-\tau/\tau_r} \tag{5.1}$$

where $a_r$ is the amplitude and $\tau_r$ the relaxation time constant of the quenching relaxation autocorrelation.

### 5.4.5 FCS Data Analysis

The second order intensity cross-correlation function $G_{ij}(\tau)$ between two channels $i$ and $j$ is defined by:

$$G_{ij}(\tau) = \frac{\langle I_i(t) I_j(t+\tau)\rangle}{\langle I_i(t)\rangle\langle I_j(t)\rangle} \tag{5.2}$$

where $I_i(t)$ is the intensity in channel $i$ at time $t$, $\tau$ is the time lag and $\langle\rangle$ denotes time averaging. Cross-correlation functions between the two detectors were computed using custom-written software based on a multiple-tau correlation algorithm [305]. Error bars are determined by splitting the measurement into ten segments of equal length and computing the standard error of mean of the correlation functions. Fitting of the correlation function was performed in MATLAB (The Mathworks, Inc.) using the non-linear least squares fit routine by minimizing the weighted residuals. Confidence intervals (95%) of determined parameters are computed from the covariance matrix obtained from the fit procedure. FCS curves are fit using a model accounting for photon antibunching, triplet kinetics and diffusion, as well as one or two additional bunching terms for the observed kinetics:

$$\begin{aligned} G(\tau) = \frac{\gamma}{N}\left(1+\frac{\tau}{\tau_D}\right)^{-1}\left(1+\frac{\tau}{p^2\tau_D}\right)^{-1/2} \times \\ \left(1+\frac{T}{1-T}e^{\frac{\tau}{\tau_T}}\right)\left(1+a_{r,1}e^{-\tau/\tau_{r,1}}+a_{r,2}e^{-\tau/\tau_{r,2}}\right) \end{aligned} \tag{5.3}$$

where $a_{r,1/2}$ are the amplitudes and $\tau_{r,1/2}$ the relaxation time constants of the quenching relaxation autocorrelation [277], $\tau_D$ is the diffusion time constant, $N$ is the average number of particles in the focus, $\gamma = 2^{-3/2}$ is a correction factor accounting for the shape of the confocal volume, $p$ is the ratio of axial to lateral size of the confocal volume, $\tau_T$ is the triplet time constant and $T$ is the triplet fraction. The parameters of the dynamic quenching term are related to the off- and on-rates of the quenching process by [306]:

$$\tau_r = \frac{1}{k_{\text{on}} + k_{\text{off}}} \tag{5.4}$$

$$a_r = \frac{k_{\text{off}}}{k_{\text{on}}} \tag{5.5}$$

In terms of the system at hand, $k_{\text{on}}$ and $k_{\text{off}}$ can be interpreted as the microscopic rate constants of end-to-end contact formation and dissociation.

## 5.5 Supplementary experimental results

### 5.5.1 Control measurements

We characterized the labeled S-peptide in terms of fluorescence lifetime and anisotropy with respect to a control construct without tryptophan (Figure 5.8 A and B). The fluorescence lifetime of Atto655 changed slightly upon attachment to the peptide from 1.8 ns for the free dye (data not shown) to 2.18 ns. Upon addition of the tryptophan, the lifetime decreased slightly to 2.05 ns,

| | | | | | | |
|---|---|---|---|---|---|---|
| $\tau_1$ | $A_1$ | $\tau_2$ | $A_2$ | | | |
| $37 \pm 5$ ns | $0.32 \pm 0.05$ | $120 \pm 11$ ns | $0.35 \pm 0.05$ | | | |
| $N$ | $\tau_D$ | $p$ | $\tau_{AB}$ | $A_{AB}$ | $\tau_T$ | $T$ |
| $0.411 \pm 0.001$ | $51.4 \pm 0.2\ \mu s$ | $4.28 \pm 0.05$ | $1.67 \pm 0.05$ ns | $0.91 \pm 0.02$ | $2.4 \pm 0.2\ \mu s$ | $0.085 \pm 0.003$ |

Table 5.2: Fit results for PET-FCS measurement. Parameters as described in the main text. Errors are given as 95% confidence intervals.

which is due to dynamic quenching on the timescale of the fluorescence lifetime [260]. Using the Stern-Volmer equation [307], the dynamic quenching rate $k_q$ can be calculated from the quenched and unquenched lifetime $\tau_q$ and $\tau_u$ as:

$$k_q = \tau_q^{-1} - \tau_u^{-1} \tag{5.6}$$

This yields a relaxation time for the dynamic quenching process of ~35 ns, in very good agreement with the observed fast quenching constant. No significant change in the fluorescence anisotropy was observed upon addition of the tryptophan, thus the rotational movement of the dye is not affected by the quenching interaction. Upon removal of the tryptophan, the dynamic quenching term disappears (Figure 5.8 C). Still, a small triplet amplitude is detected even in the absence of tryptophan. To confirm that the observed amplitude is indeed caused by photophysics, we performed a series of measurements at different excitation powers (Figure 5.8 D-F). The only significant change in the obtained correlation functions was the triplet amplitude (see Figure 5.8 F and inset in Figure 5.8 D), whereas the other kinetic amplitudes showed no power dependence (Figure 5.8 E).

### 5.5.2 Circular dichroism

CD measurements were performed on a Jasco J-715 spectrophotometer at 25 ℃. Labeled S-peptide with and without Trp15 was solvated at 1 mg ml$^{-1}$ concentration in DPBS buffer (Sigma-Aldrich). The samples were measured in a quartz cell with 0.2 mm path lenght. Data is expressed in terms of mean residue molecular elliptisity $[\theta]_\lambda$. CD data was smoothed using a Savitzky-Golay-Filter with order 3 and a window of 10 nm [308].

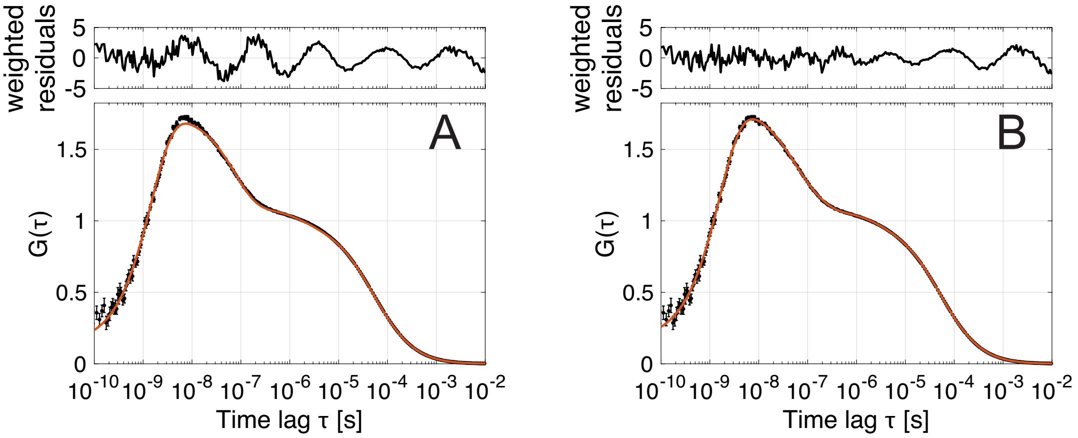

Figure 5.7: Fit of the experimental correlation function using a model accounting for one kinetic component (A) and two kinetic components (B). The weighted residuals show that one kinetic component is not sufficient to describe the observed dynamics. The $\chi^2_{red.}$ goodness-of-fit measure changes from 2.80 to 1.09 upon inclusion of a second kinetic component. The Bayesian information criterion (BIC) is significantly lowered from 765 to 340, justifying the inclusion of the second kinetic component.

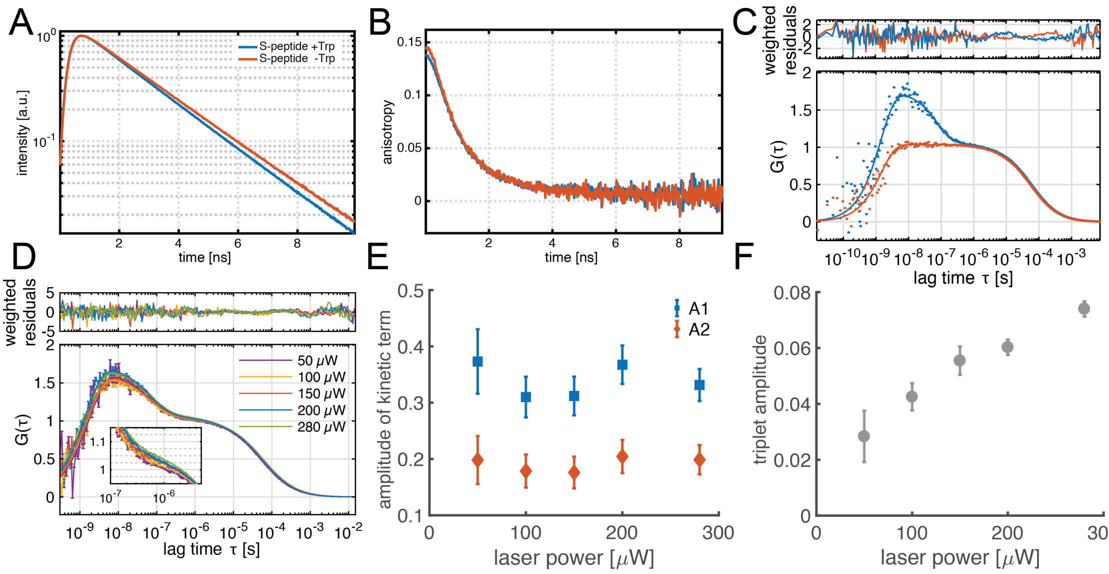

Figure 5.8: (A) Normalized intensity decays for S-peptide. The lifetime changes from 2.2 ns to 2.1 ns upon addition of the tryptophan residue. (The lifetime of free Atto655 dye is 1.8 ns) (B) The time-resolved fluorescence anisotropy shows no significant change. (C) PET-FCS curves reveal no dynamic term in the control construct, although a small triplet fraction is visible. Measurement performed at 100 μW laser power. (D-E) Testing for power dependence of observed bunching terms reveals that only the triplet term at ∼ 3 μs depends on laser power, while the kinetic terms are not affected. (A1) Amplitude for ∼ 45 ns component. (A2) Amplitude for ∼ 130 ns component.

# Chapter 6

# Adenylylation of Tyr77 stabilizes Rab1b GTPase in an active state: A molecular dynamics simulation analysis

The pathogenic pathway of *Legionella pneumophila* exploits the intracellular vesicle transport system via the posttranslational attachment of adenosine monophosphate (AMP) to the Tyr77 sidechain of human Ras like GTPase Rab1b. The modification, termed adenylylation, is performed by the bacterial enzyme DrrA/SidM, however the effect on conformational properties of the molecular switch mechanism of Rab1b remained unresolved. In this study we find that the adenylylation of Tyr77 stabilizes the active Rab1b state by locking the switch in the active signaling conformation independent of bound GTP or GDP and that electrostatic interactions due to the additional negative charge in the switch region make significant contributions. The stacking interaction between adenine and Phe45 however, seems to have only minor influence on this stabilisation. The results may also have implications for the mechanistic understanding of conformational switching in other signaling proteins.

## 6.1 Introduction

The orchestration of intracellular protein interaction networks requires tight temporal and spatial regulation in order to maintain homeostasis and to react to changing environmental conditions. Small Ras-like GTPases (guanosine triphosphate phosphohydrolases), also referred to as G-proteins, play a pivotal role in the coordination of intracellular signaling by acting as binary molecular switches [310]. The discrimination between the active and inactive state is achieved by differential co-factor binding: GTPases are in the active ("on") state when bound to guanosine triphosphate (GTP), but inactive ("off") when complexed with guanosine diphosphate (GDP). Signaling is promoted by the binding and recruitment of effector proteins that specifically interact with the active state of the GTPase. Due to their fundamental role in coordinating signaling it is not very surprising that many intracellularly replicating bacterial pathogens have evolved with mechanisms to interfere with GTPase activities and thereby promote their survival [311]. Among several activity modulating strategies, the covalent attachment of additional functional groups (also referred to as posttranslational modifications (PTMs)) appears to be particularly prominent and interesting. We and others have previously observed that the pathogenic bacterium *Legionella pneumophila* covalently modifies the human G-protein Rab1b via the posttranslational attachment of adenosine monophosphate (AMP) from adenosine triphosphate (ATP) to the side

---

Parts of this chapter have been published in [309]

chain of residue Tyr77 with the help of the bacterial protein DrrA (also known as SidM). This enzymatic adenylylation (also termed AMPylation) exploits the intracellular vesicle transport system of its host cell [312]. It has been observed that adenylylation of small GTPases can abrogate the interaction with GAPs [312] or downstream effectors [313, 314]. In particular, adenylylation of Rab1 blocks the access of the human GAP TBC1D20 and the Legionella GAP LepB in vitro [312, 315, 316]. Thus, Rab1 adenylylation appears to stabilize the G-protein in the active GTP-state by inhibiting GAP-mediated GTP-hydrolysis. It is, however, less obvious whether the adenylylated Rab1 protein also maintains an active conformation that is in principle capable of interacting with GTP-state specific cellular factors. The activity state of a GTPase is communicated to interaction partners mainly via two highly important regulatory regions that are referred to as switch I and switch II. In the inactive GDP-bound state, these regions are structurally disordered but they become highly conformationally restrained in the active GTP-bound form. Interacting molecules very sensitively probe the switch conformations and thus can bind specifically either the GDP- or the GTP-state. Interestingly, the adenylylated residue Tyr77 of Rab1 is located in the switch II region and consequently the question arises as to how this modification may affect the configurational ensemble of the switch II and/or switch I regions. The X-ray structure of the AMP-Rab1:GTP complex revealed a stacking interaction of the adenine base of the adenylylated Tyr77 of switch II with a highly conserved phenylalanine (Phe45) side chain [312]. This interaction may suggest that Tyr77-adenylylation fixes switch II in a defined and active–state like conformation. Furthermore, we have observed recently that the deadenylylation reaction of AMP-Rab1 by SidD is independent of the nucleotide-state of the G-protein, i. e. AMP-Rab1:GDP and AMP-Rab1:GTP show no difference in their substrate properties [317].

This observation is a very astonishing finding since most GTPase interacting molecules very sensitively discriminate between the GDP- and GTP-states by binding only to the inactive or the active switch conformations, respectively. The lack of discrimination of SidD between the GDP- and GTP-states of AMP-Rab1 may therefore suggest that the switch regions are locked in identical conformations. Consequently, adenylylation of Rab1b could force the switch regions into the active conformation even if the protein is actually in the GDP-state possibly due to the stacking interaction with the Phe45 residue observed in the crystal structure.

In order to elucidate this mechanism we performed extensive continuous Molecular Dynamics (MD) and Umbrella Sampling (US) based free energy simulations to compare the influence of adenylylation on Rab1b conformational states bound to either GTP or GDP. The simulations indicate a stabilizing effect of the Tyr77 adenylylation on the active form of Rab1b even in the presence of GDP. In addition, electrostatic energy analysis of conformational ensembles close to the active states vs. states representing the inactive form reveals that electrostatic interactions make the major favorable contribution to the active state stabilization in the presence of the Tyr77 adenylylation. The simulation study indicates that stabilizing effects of side chain modifications in GTPases (not necessarily close to the GTP/GDP binding site) might not only be mediated by contacts but also indirectly e. g. by electrostatic interactions. The result may also have important implications for understanding the influence of other modifications on signaling proteins.

## 6.2 Results

### 6.2.1 Molecular Dynamics simulations and in vitro deadenylylation assay on Rab1b

In order to elucidate the influence of the bound nucleotide and adenylylation of Tyr77 on the conformational flexibility and stability of Rab1b we first performed a series of continuous (c)MD-simulations in explicit solvent. The simulations were started from the crystal structure of the GTP bound form with native or adenylylated residue Tyr77 (i. e. AMP covalently attached to the OZ atom of Tyr77). Starting structures in complex with GDP with or without Tyr77 modification were generated in silico by removing the corresponding atoms from the crystal structure. In all simulations the protein structure remained overall close to the starting conformation with an overall backbone root-mean-square deviation (RMSD) of $< 0.2$ nm with respect to the crystal structure (Figure 6.1). The calculated root-mean-square-fluctuation (RMSF) of the switch I and II regions showed larger fluctuations in the GDP vs GTP complexes but no sign of unfolding, e. g. towards the inactive form in the presence of GDP (Figure 6.2).

The $Mg^{2+}$ ion stayed close to the initial placement in simulations with either GTP or GDP bound to Rab1b. In simulations with GTP two water molecules persisted in the close vicinity of $Mg^{2+}$ exposing the oxygen atom which carries the negative partial charge to the Mg ion (Figure 6.3). This induced a local shielding effect of the two positive charges by the dipole field of the water molecules. These water molecules were also found in the crystal structure of Rab1b [318]. The cavity arising from the hydrolysis of GTP to GDP (i.e removal of the $\gamma$-phosphate from GTP) was filled with a third water molecule which persisted throughout all simulations with GDP bound to Rab1b. Residues Ser39, Thr40 and Tyr37 located in the switch I region formed persistent H-bonds with the $\gamma$-phosphate during the simulation in line with the X-ray structure [318]. Also in agreement with experiment, one persistent H-bond between $\gamma$-phosphate and the backbone of switch II residue Gly66 was observed.

The X-ray structure of the AMP-Rab1b:GppNHp complex revealed a planar packing of the

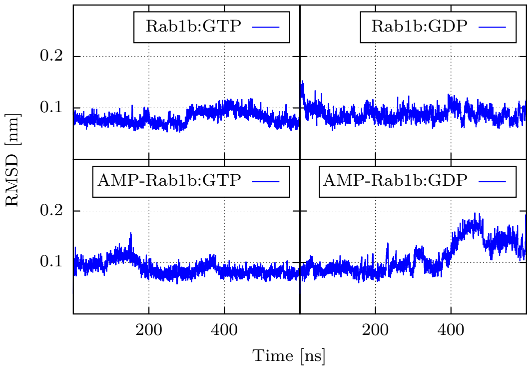

Figure 6.1: Root-mean-square deviation (RMSD) of protein backbone with respect to the AMP-Rab1b:GppNHp X-ray structures of four cMD simulations for different Rab1b systems. The flexible five N- and C-terminal residues were excluded from the RMSD calculations, respectively. The initial conformations were well conserved throughout the 600 ns simulations which is reflected in a RMSD below 0.2 nm for all simulations.

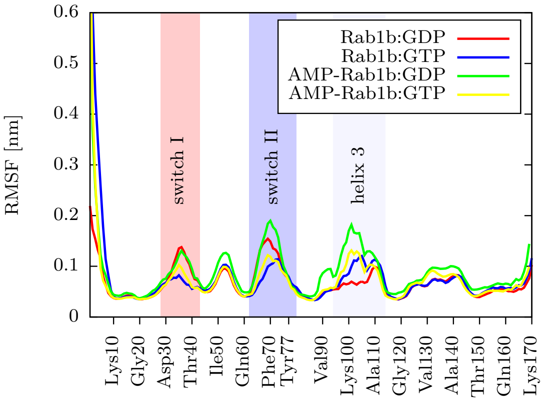

Figure 6.2: Average root mean square fluctuations (RMSF) of Rab1b C-$\alpha$ atoms in GTP and GDP bound form taken from 600 ns MD simulations. The difference of both curves highlights the change in flexibility depending on the present nucleotide. The removal of the $\gamma$-phosphate destabilizes switch I and II regions due to the loss of hydrogen bonds network and the shift in the electrostatic field.

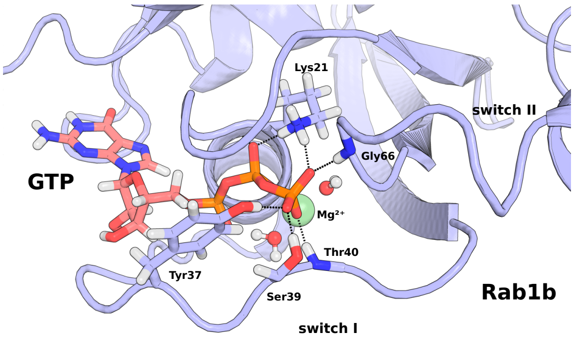

Figure 6.3: Noncovalent interaction network of GTP bound to Rab1b observed during Molecular Dynamics simulations. GTP is depicted as atom-color-coded sticks while Rab1b is indicated as blue cartoon. The bound magnesium ion is shown as a green sphere. Rab1b residues forming bonds with GTP are shown as sticks. The $\gamma$-phosphate group of GTP forms 3 hydrogen bonds with switch I residues Tyr37, Ser39, and Thr40 of Rab1b. One additional bond is formed between switch II residue Gly66 and GTP. The nucleotide is further stabilized in the binding pocket by two salt bridges formed with core residue Lys21. Two water molecules build a charge shielding shell around $Mg^{2+}$ ion and are tightly bound during all the simulations.

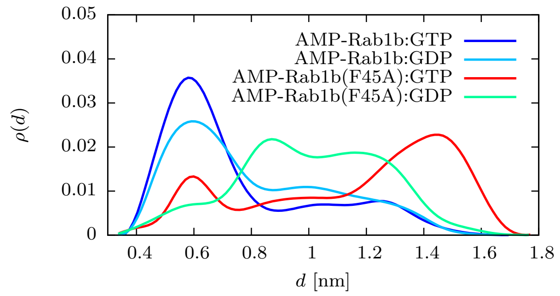

Figure 6.4: Probability distribution of sampled distances between the C$\beta$-atom of residues 45 and the adenine double ring structure of AMP-Tyr77 during cMD simulations. A short distance (below 0.7 nm) indicates a stacking interaction between the adenine and residue 45. Sampling of larger distances corresponds to non-contacting states.

highly conserved Phe45 phenyl ring with the adenine double ring of adenylylated Tyr77 which suggests a functionally relevant $\pi$-stacking interaction [312]. Based on this suggestion additional 600 ns simulations of the adenylylated Rab1b mutant F45A (Rab1b(F45A)) in the presence of either GDP or GTP were performed and compared with simulation results of wild type AMP-Rab1b. Again, neither the GTP nor the GDP bound case resulted in significant structural changes in the switch regions during the simulation time (see Figure 6.1). However, the conformational sampling of the adenine base attached to Tyr77 is altered by the F45A mutation: A histogram plot of distances between the C$\beta$-atom of residue 45 and the adenine double ring of AMP-Tyr77 reveals a significant loss of the stacking interaction between the Ala45 side chain and the adenine ring in the Rab1b(F45A) mutant compared to wild-type (both in the GTP or GDP bound case, Figure 6.4). Cluster analysis indicated a dominating cluster for AMP-Rab1b with about 74% of all frames representing a stacking interaction between Phe45 and AMP when GTP was bound and slightly less with bound GDP (see Figure 6.4, 6.5A). For the F45A mutant the clustering revealed that the AMP modification gains conformational freedom visiting a broader range of configurations (Figure 6.5, C,B, and D: three of the five largest clusters). Importantly, the average distance between AMP group and residue 45 is considerably larger in case of the F45A mutation compared to the wild type (Figure 6.4) incompatible with an effective stacking interaction.

In order to probe its relevance for the switch mechanism we performed enzymatic deadenylylation reactions of preparatively modified AMP-Rab1b. We have observed previously that the deadenylylation enzyme SidD does not discriminate between GDP- and GTP-bound AMP-Rab1b [317]. However, SidD shows much weaker activity toward the synthetic adenylylated switch II peptide TITYAMPYRGAHGC in comparison to AMP-Rab1b:GTP ($k_{\text{cat}}/K_{\text{M}}(\text{peptide}) = 150\,\text{M}^{-1}\,\text{s}^{-1}$ vs. $k_{\text{cat}}/K_{\text{M}}(\text{AMP-Rab1b:GTP}) = 5 \times 10^5\,\text{M}^{-1}\,\text{s}^{-1}$) (Figure 6.7) [319]. This observation is a strong indication that SidD requires the recognition of specific structural elements in Rab1b rather than merely binding to the adenylylated tyrosine. The structural specificity of SidD in particular and since GTPase-binding proteins and enzymes usually discriminate very sensitively activity states by probing the conformations of the switch regions, the hypothesis was derived that adenylylation locks both GDP- and GTP-forms in the same conformation. If the Phe45-adenine interaction was relevant for this locking effect, a F45A substitution would be expected to promote conformational segregation and thus affect SidD catalytic rates (resulting in preference of AMP-Rab1(F45A):GTP over AMP-Rab1(F45A):GDP). However, the differences

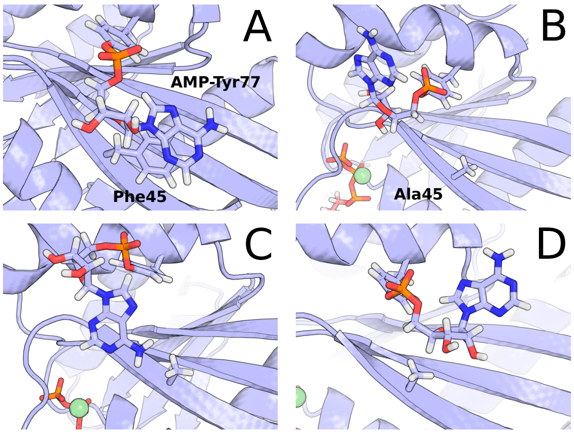

Figure 6.5: Stacking interactions between the adenine double ring of adenylylated Tyr77 with the phenyl ring of Phe45 observed during 600 ns continuous MD simulation of AMP-Rab1b:GTP. Clustering of AMP-Tyr77 orientations indicated stacked conformations (largest conformational cluster) occurring in 74% of the whole simulation trajectory (A). The stacking interaction, however, is greatly reduced in the simulation of the F45A mutation. The decoupled AMP sidechain showed high flexibility and visited various states at increased distance to Ala45 during a 600 ns simulation of the mutant. The three largest clusters are depicted (B,C,D).

in catalytic deadenylylation of AMP-Rab1(F45A):GppNHp vs AMP-Rab1(F45A):GDP were negligible (Figure 6.7) and thus similar (active) conformations of the switch regions in the AMP-Rab1(F45A) proteins are expected. Consequently, the Phe45 may not contribute significantly to rigidifying the conformations of adenylylated Rab1 in both activity states, suggesting that other molecular effects are predominant in locking the conformational states.

### 6.2.2 Free energy calculation of switch region unfolding reveals stabilization by adenylylation

In order to directly probe the effect of chemical modification and/or mutation of Rab1b on the transition between active and inactive states (conformational ensembles), we performed Umbrella Sampling (US) free energy simulations. This approach permits to induce conformational transitions associated with the active and inactive states using a penalty potential to unfold the switch region during simulations. It also allowed us to calculate the associated change in free energy (also termed potential of mean force: PMF) for the transition and how it depends on adenylylation and on bound nucleotide. As a reaction coordinate for the US simulations the mean deviation of a set of distances within the switch II region from the active state was employed (dRMSD coordinate, illustrated in Figure 6.6 and explained in detail in the Methods section). A small reference dRMSD results in sampling of conformations close to the active GTPase conformation whereas unfolding of the switch region is induced with increasing reference dRMSD (modified in 14 umbrella windows from $\mathcal{D}_0^A = 0\,\text{nm}$ to $\mathcal{D}_0^B = 0.4\,\text{nm}$). In order to improve the convergence of the US simulations frequent replica exchanges between neighboring US windows were allowed (H-REUS technique, see Methods). The calculated PMF showed reasonable convergence after 80 ns of data gathering time in each US window (Figure 6.9). The active state ensemble is represented by conformations close to the minimum of the calculated free energy curve (Figure

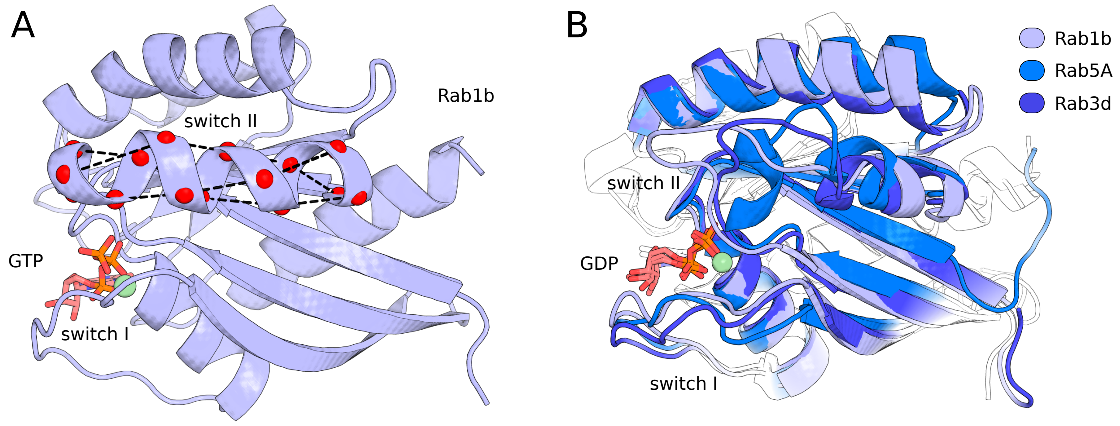

Figure 6.6: (A) Illustration of the set of distances which contributed to the dRMSD bond network in order to bias the unfolding of the switch II region. Contributing atoms are highlighted as red spheres, the distances are shown as dashed lines. Rab1b is shown in its active conformation taken from X-ray structure (PDB code 3NKV) but without adenylylation. (B) Superimposed structures of Rab1b homologs human RAS-related proteins Rab3d (PDB code 2GF9) and Rab5a (PDB code 3CLV) in GDP bound form with a representative inactive state snapshot from dRMSD US simulations. The snapshot agrees qualitatively with unfolded features of switch I and switch II regions observed in the Rab1b GDP bound homologs.

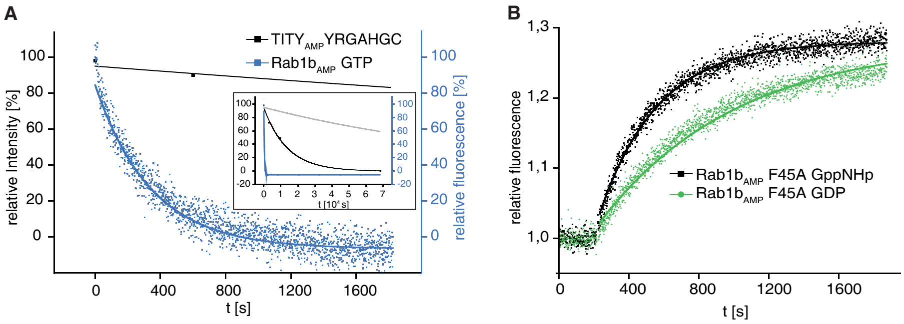

Figure 6.7: Kinetics of deadenylation (A) Deadenylation of AMP-Rab1b:GTP and the adenylylated peptide TITY$_{AMP}$YRGAHGC by SidD revealing a significant preference for the adenylylate protein. Demodification of AMP-Rab1b:GTP (5 μM) or peptide (50 μM) were initiated with catalytic amounts of SidD (50 nM or 500 nM, respectively). Deadenylation was monitored using the change in intrinsic tryptophan fluorescence (AMP-Rab1b:GTP) or by quantifying reaction products on reversed phase chromatography (peptides). Data were fitted to a single exponential function. Inset: Comparison of reaction progress on long time scales. For comparison, the reaction progress curve for 5 μM peptide-AMP with 50 nM SidD (grey) was simulated as described in methods. (B) Deadenylation of Rab1b AMP-Rab1b(F45A):GDP (green) and AMP-Rab1b(F45A):GppNHp (black) by SidD. Deadenylation of 1 μM Rab1b was initiated by addition of 100 nM SidD and monitored via intrinsic tryptophan fluorescence.

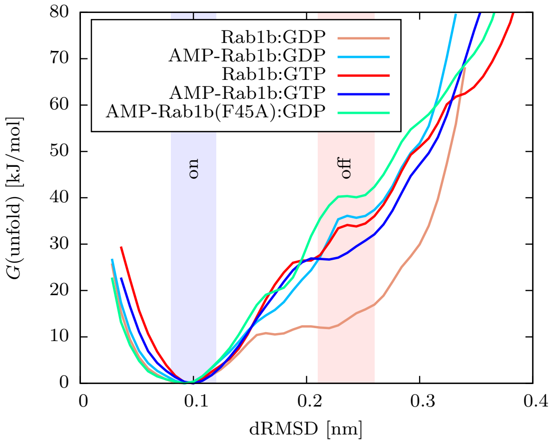

Figure 6.8: The effect of different modifications in Rab1b on the unfolding free energy of switch II along the dRMSD coordinate. Only the inactive Rab1b:GDP form has a significantly lower unfolding free energy of the switch II region compared to the other Rab1b modifications. The removal of the Phe45 stacking interaction with adenylylated Tyr77 by F45A mutation shows no notable difference in the unfolding free energy profile compared to the AMP-Rab1b:GDP version.

6.8) at small dRMSD (below 0.15 nm). Already at dRMSD > 0.15 nm the switch region starts to unfold. The definition of a dRMSD range for the inactive state was based on the comparison with GDP bound GTPase X-ray structures and their associated dRMSD values (0.2 nm for the set of distances, which is also close to the plateau regime observed in the calculated free energy curves, Figure 6.8). Note, that during the US simulations along the dRMSD coordinate an ensemble of conformations at the regime of the inactive state was sampled (Figure S4). Hence, the simulations support the view that the inactive state is not represented by a single stable conformation but compatible with various unfolded conformations. In all systems the unfolding process resulted in an increase in free energy along the reaction coordinate (Figure 6.8). However, in case of an unmodified Rab1b:GDP the free energy increase was significantly (about 14 kJ/mol) smaller than in case of a bound GTP. The calculated free energy changes were also tested with respect to changes in the force constants used to control the dRMSD deviation from a reference during the US simulations (supp. Information, Figure S5). For increasing or lowering the force constant by a factor of 2 or 4, respectively, almost the same free energy change (within $\sim 3\,\mathrm{kJ\,mol^{-1}}$, supp. Information, Figure S5) was obtained indicating the calculated PMFs are robust with respect to changes in the force constants used in the US simulations. In addition, the presence of the adenylylated Tyr77 appears to stabilize the active state of the GTPase for both the GTP and GDP bound cases (Figure 6.8). The calculated free energy change for the AMP-Rab1b:GDP case is similar to the Rab1b:GTP curve. For Rab1b:GDP a consecutive unfolding was found starting with the switch II helix unfolding in lower dRMSD replica and followed by switch I at higher dRMSD indicating a coupling of both conformational regimes. Unfolding simulations of switch II revealed that the major conformational rearrangement occurred in the N-terminal part of the helix between residues Gly66 and Tyr77 (see MD-snapshot in Figure 6.6). In addition to the wild type Rab1b protein, free energy simulations were also performed on the F45A mutation of AMP-Rab1b:GDP. Surprisingly, the unfolding characteristics of the F45A mutant were similar to the AMP-Rab1b:GDP simulation indicating that the observed transient stacking between AMP and

| Difference electrostatic unfolding energy [a] | $\Delta\Delta E_{\text{coulomb}}$ | $\Delta\Delta E_{\text{reaction-field}}$ | $\Delta\Delta E_{\text{total}}$ |
|---|---|---|---|
| AMP-Rab1b:GDP − Rab1b:GDP | $-2.6 \pm 1.8$ | $22.4 \pm 10.4$ | $19.8 \pm 12.2$ |
| Rab1b:GTP − Rab1b:GDP | $-1.5 \pm 1.4$ | $30.5 \pm 10.4$ | $29.0 \pm 11.8$ |

[a] values are given in [kJ/mol]

Table 6.1: Electrostatic contributions to the mean energy difference of inactive vs. active conformational ensembles of Rab1b:GDP in the presence or absence of the adenylylation at Tyr77 based on FDPB calculations (see Methods for details). A positive $\Delta\Delta E$ indicates a stronger favorisation of the active form (relative to inactive) for the adenylylated variant (equivalent to a relative stabilization of the inactive ensemble vs active ensemble in the absence of adenylylation). The energetic contributions are split into two contributions for direct Coulomb interactions and the electrostatic solvation (reaction field) term, respectively. The top row shows the relative electrostatic stabilization of the wildtype vs. the adenylylated Rab1b (both with GDP bound). The second row indicates the corresponding electrostatic energy differences for active vs inactive conformations in case of GTP vs. GDP bound to Rab1b (indicating an electrostatic stabilization of the active conformational ensemble by the presence of GTP compared to GDP).

Phe45 (observed in the cMD simulations) may not be of dominant importance for stabilizing the active state (Figure 6.8).

### 6.2.3 Electrostatic effects of adenylylation affect Rab1b conformations

Since the proposed stacking interaction between Phe45 and the adenine base appeared not to be of major significance for stabilizing the switch region, we aimed at analyzing the electrostatic effects resulting from the presence of AMP on Rab1b conformations. The adenylylated Tyr77 residue is located in the switch II region relatively far from the GTP binding site. In order to investigate the influence of long-range electrostatic interactions we compared the electrostatic energy of Rab1b in the active and inactive states. The electrostatic energies were calculated for an ensemble of conformations extracted from the AMP-Rab1b:GDP umbrella sampling simulations with dRMSD values of around 0.1 nm representing the active state. A conformational ensemble of unfolded switch II region with dRMSD around 0.25 nm represented the inactive conformational ensemble. Snapshots were taken from the trajectories every 0.6 ns skipping the initial 12 ns to account for equilibration. Electrostatic contributions can be split into Coulomb interactions between atoms belonging to Rab1b in the different ensembles (termed Coulomb contributions) and secondary interactions of the protein atoms with the surrounding solvent (termed electrostatic solvation or reaction field contribution). Both average contributions can differ in the active versus inactive ensembles depending on distances between charges (Coulomb contribution) and accessibility to solvent (solvation or reaction field contribution). The total electrostatic energy is given as a sum of these two contributions.

The Coulomb contributions and solvent induced reaction field contributions were calculated with the finite-difference Poisson-Boltzmann (FDPB) approach and a continuum solvent representation (see Methods for details). In order to focus on the influence of adenylylation and to minimize the numerical error of the FDBP solutions, the calculations were repeated after replacing the adenylylated residue 77 with an unmodified Tyr residue in the active and inactive ensembles. This procedure does not account for possible differences in the sampled ensembles

| System | $\Delta\Delta G$ | $\Delta G$ |
|---|---|---|
| AMP-Rab1b:GDP | 20.1 | 32.1 |
| AMP-Rab1b(F45A):GDP | 25.6 | 37.6 |
| AMP-Rab1b:GTP | 14.7 | 26.7 |
| Rab1b:GTP | 18.8 | 30.8 |
| Rab1b:GDP | – | 12.0 |

Table 6.2: Calculated relative free energy differences of unfolding the switch II region with respect to Rab1b:GDP ($\Delta\Delta G$) and absolute unfolding free energy ($\Delta G$) in [kJ/mol] from US simulations. The absolute unfolding free energies $\Delta G$ were calculated by integrating the PMF for the active regime (corresponding dRMSD range 0.08–0.12 nm) and the inactive regime (dRMSD range 0.21–0.26 nm) and substracting the values. The free energy of unfolding for the Rab1b:GDP case was again substracted from $\Delta G$ values to get relative $\Delta\Delta G$.

between adenylylated and unmodified Rab1b but allows to focus directly on the adenylylation effect.

While the direct Coulomb interactions favor the unfolded structure of AMP-Tyr77 compared to unmodified Tyr77 by about $-2.6\,\mathrm{kJ\,mol^{-1}}$, the inverse effect was observed for the reaction field contribution with about $22.4\,\mathrm{kJ\,mol^{-1}}$ (Table 6.1). In total, the wildtype Rab1b:GDP inactive ensemble of switch II is favored electrostatically by $19.8\,\mathrm{kJ\,mol^{-1}}$ compared to AMP-Rab1b:GDP. The calculated difference in electrostatic energy is consistent with the PMF unfolding curves (suggesting a free energy difference of about $20\,\mathrm{kJ\,mol^{-1}}$, Table 6.2) and provides a possible energetic explanation for the stabilizing effect of AMP on the switch regions. Although not at the focus of the present study we also compared the mean electrostatic energy of Rab1b:GTP versus Rab1b:GDP in active versus inactive ensembles using the same procedure. In this case the calculations predicted a strong favorisation of the inactive ensemble vs. active ensemble of the Rab1b:GDP complex relative to the Rab1b:GTP complex by $29.0\,\mathrm{kJ\,mol^{-1}}$ (Table 6.1). This result predicts that electrostatic interactions stabilize an active state ensemble (relative to the inactive state ensemble) considerably more in the presence of GTP versus GDP (in agreement with the experimental observation). Note, that the same trend was also found for calculations using an internal dielectric constant of 2 or 4 for the protein (see supp. Material Table S1) which in part accounts for the possible relaxation of the protein charge distribution upon removal of the Tyr77 modification or switch from GTP to a bound GDP in the trajectory analysis. It is also consistent with the calculated trend from the free energy simulations which predicted a relative stabilization free energy of $18.8\,\mathrm{kJ\,mol^{-1}}$ (Table 6.2). Note, that an agreement is not necessarily expected because other energetic and entropic contributions also contribute to the relative stabilities of inactive and active conformational ensembles. These contributions are accounted for the in free energy simulations but not in the electrostatic energy calculations.

## 6.3 Discussion

The molecular mechanism how adenylylation of residue Tyr77 within the switch II region affects the active and inactive conformations of Rab1b has not been addressed previously. Here, we employed MD and free energy simulations to characterize the conformational flexibility and stability of Rab1b in different GDP/GTP states in dependence of Tyr77 adenylylation. On the time scale of 600 ns of our continuous MD simulations similar flexibility patterns of the different Rab1b nucleotide complexes were found. The observed hydrogen bonding pattern and location of water molecules was compatible with available experimental structures [312]. In case of Tyr77 adenylylation, the adenine group of AMP stacked on the neighboring Phe45 for a significant fraction of the simulation time but also other states with fully solvent exposed AMP group were sampled. A F45A substitution in silico resulted in the elimination of this $\pi$-stacking interaction and increased the conformational flexibility of Tyr77-AMP. The modified and unmodified Rab1b:GDP complexes showed generally larger fluctuations in the switch I and switch II regions compared to GTP-bound complexes. However, no spontaneous unfolding of the switch regions characteristic for an inactive conformation in the presence of GDP was observed on the time scale of the cMD simulations. This is not surprising since the estimated time scale of such changes is in the range of tenth of seconds [320–322].

In order to still calculate the free energy change associated with a transition to the inactive state we employed US based on a dRMSD coordinate that allows gradual unfolding of the switch region and transition to conformations representing the inactive state. The sampled inactive conformations agreed qualitatively well with switch II helix and switch I loop regions found in related Ras-GTPases (X-ray structures of the GDP bound form of Rab proteins , PDB codes 2GF9, 3CLV and 4Q21, Figure 6.6) [323]. The simulations predicted in all cases an increase of the free energy towards unfolding of the switch region. Although this may support a view that Rab1b behaves as a non-classical GTPase there is are many experimental studies that demonstrate that Rab1b indeed operates by a classical GTPase-mode. This is supported by the crystal structure of a homologous Rab1a GTPase (more than 95% sequence identity to Rab1b) in the presence of GDP (PDB code 2FOL) which indicates an unfolded and disordered structure of the switch regions. In addition, it has been shown that several Rab1 effector proteins (such as Mical-1, Mical-2C, Mical-3C, Golgin84, GM130, p115, Rabaptin5) favor the GTP-bound state over the GDP state [324–331]. In addition, the GDP-specific Rab-binding protein GDI has been shown in several instances that it does not bind to the GTP-state but only interacts with the inactive GDP Rab1b form [332, 333]. These results indicate that Rab1b significantly differs in its conformations between the GDP- and GTP-states since effectors (GTP-specific) and GDI (GDP-specific) discriminate unambiguously between the activity states. Our US simulations along the dRMSD coordinate did not reveal a single compact conformation representing the inactive state but suggest a largely disordered more solvent exposed ensemble of conformations. It is well known that current molecular mechanics force fields overestimate the stability of compact states even of disordered proteins [334, 335]. In turn this might cause an artificial over-stabilization of the folded compact states of the switch regions. Nevertheless, the calculated free energy change associated with a transition from an active to an inactive state was significantly smaller in the presence of GDP compared to GTP confirming that the relative change for the transition is in qualitative agreement with the experimental observation.

Furthermore, the US simulations showed that the adenylylation of Tyr77 stabilizes the active state relative to the inactive state regardless of GDP or GTP binding. The free energy costs of unfolding the switch region (AMP-Rab1:GDP to Rab1:GDP) were similar to the level obtained for Rab1b:GTP (Rab1:GTP to Rab1:GDP). This result agrees with an interpretation that Tyr77 adenylylation could lock Rab1b in an active conformation in the AMP-modified form [312, 317]. Hypothetically, the stacking interaction of the AMP group with the Phe45 observed in the experimental X-ray structure [312] and during the cMD simulations may be involved in this stabilization. However, a similar calculated free energy change associated with the unfolding process was obtained for a F45A mutation in silico, suggesting that base stacking is not the major contribution to conformational stabilization of AMP-Tyr77. An analysis of the electrostatic energy of the active and inactive states revealed that the presence of the adenylylation destabilizes the ensemble of the inactive conformations mainly due to a reduced electrostatic solvation (solvent reaction field) contribution. In combination with our observation, that the geometric coordination of the adenine group of AMP does not impair the stabilization of the active state ensemble we assume, that mainly the inclusion of an additional negative charge in switch II located at the phosphate group of AMP modifies the electrostatic field such that the reaction of surrounding dielectric solvent molecules inhibits a transition to the inactive ensemble.

The combination of MD simulations and in vitro enzymatic activity assays suggest that adenylylation of Rab1 at switch II tyrosine 77 significantly changes conformational transitions of the GTPase domain. As a consequence, the G-protein surprisingly adopts an active-like conformation in both nucleotide states and thus AMP-modification may in this instance uncouple Rab-activation from the conventional cyclic GDP-GTP-binding. Generally, the effects of PTMs on the conformations of small GTPases using structural or computational methods are only poorly characterized. In one study, Kalbitzer and coworkers used 1H and 31P nuclear magnetic resonance (NMR) spectroscopy to investigate the consequences of Thr35-glycosylation of H-Ras to gain insight into structural consequences of this particular PTM [336]. Their work supported an earlier model from crystallographic studies of glycosylated H-Ras that suggested a disordering of switch I due to interference with $Mg^{2+}$ coordination and Tyr32-nucleotide interactions [337]. In this respect, glycosylation impaired effector protein binding by promoting a disordered switch I state and by sterically interfering with protein-protein interactions. Rab1 adenylylation appears to have the opposite effect and actually promotes the active conformations of the switch region. It will be interesting to see whether other modifications of small GTPases (i. e. adenylylation of Cdc42 on Thr35 or Tyr32, phosphocholination of Rab1b on Ser76 [313, 314, 338]) will have similar conformational consequences.

Our study indicates that chemical modifications located at a distance from the nucleotide binding site of a Rab1b protein may considerably affect the conformational equilibrium of active and inactive states. In addition to contacts in the neighborhood of the modification, long range electrostatic interactions contribute significantly to stabilizing an active signaling state which may also be relevant for understanding the influence of other modifications on the signaling mechanism in other signaling proteins.

## 6.4 Methods

### 6.4.1 Molecular Dynamics Simulation Setup

The crystal structure of the Rab1b protein with an Adenosinmonophosphate (AMP) moiety covalently attached to OZ atom of Tyr77 in complex with the GTP analog Phosphoaminophosphonic Acid Guanylate Ester (GNP) (PDB code 3NKV) served as start structure for the simulations. The nitrogen atom in GNP bridging the beta to the gamma phosphate group was replaced by an oxygen atom which resulted in the AMP-Rab1b:GTP complex model structure. For simulations on unmodified Rab1b or bound to GDP the AMP modification and/or the $\gamma$-phosphate group were removed to create start structures of the Rab1b:GTP, AMP-Rab1b:GDP, and Rab1b:GDP complexes, respectively. All simulations were performed with the GROMACS molecular dynamics software suite release version v4.6 [96, 244]. The Amber ff99sb-ILDN forcefield [42] was used for the protein and parameters for guanosine nucleotides GDP and GTP were taken from Carlson et al. [339]. Partial charges for the adenylylated Tyr77 residue were calculated using Gaussian03 [296] with B3LYP [297, 298] with the 6-31G* basis set level [299] and a total charge of $-1\,e$ following the restraint electrostatic potential (RESP) protocol [300]. The antechamber program [248] of the AmberTools13 package [98] in combination with the general amber forcefield GAFF [247] was used to assign atom types and bonded parameters. The system topologies were finally prepared with the tleap tool from the AmberTools13 software package and then translated to Gromacs topology files via acype [340]. To render the systems charge neutral sodium ions were added and the protein was solvated with the TIP3P water model [246] in a truncated octahedron box with periodic boundary conditions and a minimum distance of 1 mm from solute to the box boundaries.

### 6.4.2 Simulation protocol

After the setup procedure, energy minimization with the steepest decent algorithm was performed until one convergence criterion of either 20 k steps or a maximum force below $100\,\mathrm{kJ\,mol^{-1}\,nm^{-1}}$ was reached. The minimization was followed by two short equilibration runs of each 150 ps duration at a time step of 1 fs, primarily in the NVT and then in the NPT ensemble. All heavy backbone atoms were restraint in space with a harmonic potential at force constant of $1000\,\mathrm{kJ\,mol^{-1}\,nm^{-2}}$ to avoid conformational rearrangements whilst equilibration. The equations of motions were solved according to the leap-frog integrator (MD) and the Particle Mesh Ewald (PME) algorithm was used to calculate long range electrostatics [250] with a grid interpolation up to the order of 6 (4 in NPT and production run) and FFT grid spacing of 0.12 nm. The Lennard–Jones interactions were switched to zero after 1.0 nm with a cutoff value of 1.1 nm for both Lennard–Jones and real space electrostatic interactions. The temperature was adapted to a reference of 298 K with the velocity rescale [47] algorithm and pressure in the NPT equilibration phase was controlled with the Berendsen barostat [45] to equal 1.01 bar. For production runs the time step size was increased to 2 fs and the Parinello–Rahman barostat was applied [48]. A long range dispersion correction for energy and pressure was applied to account for the error introduced by truncated Lennard-Jones interactions. The LINear Constraint Solver [249] with a coupling matrix extension order of 12 (4 in production run) constrained the bond lengths involving H atoms.

### 6.4.3 Stacking interaction between Phe45 and AMP-Tyr77 sidechains

Sidechain conformations of AMP-Tyr77 were clustered for the 600 ns trajectories of systems Rab1b:GTP, Rab1b:GDP, Rab1b(F45A):GTP, and Rab1b(F45A):GDP in order to characterise the prevalence of the $\pi$-stacking interaction between adenine double ring of AMP-Tyr77 with the phenyl ring of Phe45. The protein trajectories were aligned before clustering of the AMP-Tyr77 sidechain was performed using the single linkage method with a RMSD cutoff of 0.07 nm as implemented in g_cluster of the GROMACS toolchain. Additionally the distance between C$\beta$ atom of residue 45 and the center of mass of the adenine double ring has been monitored over time and a histogram was calculated by splitting the observed distance range in 100 bins (Figure 6.4).

### 6.4.4 Free energy simulations based on the root mean square deviation of a set of intramolecular distances

The root mean square deviation of a set of distances (dRMSD) obtained during the simulation with respect to the same set of distances in a reference structure was used as reaction coordinate. In contrast to Cartesian RMSD this reaction coordinate does not require a superposition step and is invariant under rotation. An additional advantage is that it is possible to include only subsets of distances that are spatially separated in the dRMSD coordinate (e. g. only local short range distances) which would require several superposition operations if using the Cartesian RMSD as reaction coordinate. The collective variable is defined as the RMS sum over a set of $N$ interatomic distances $d_i$ with a reference distance $d_{i0}$ (equation 6.1).

$$\mathcal{D}(d_i) = \sqrt{\frac{1}{N}\sum_{i}^{N}\Big(d_i - d_{i0}\Big)^2} \tag{6.1}$$

The spring like biasing potential which emerges from the dRMSD (equation 6.2) is expanded to two reference dRMSD values $\mathcal{D}_0^A$ and $\mathcal{D}_0^B$ which are linearly connected over a coordinate $\lambda$ and allows for the calculation of a Potential of mean Force (PMF) curve along this coupling coordinate.

$$V(d_i, \lambda) = \frac{k_0}{2}\Big(\mathcal{D}(d_i) - (1-\lambda)\mathcal{D}_0^A - \lambda\mathcal{D}_0^B\Big)^2 \tag{6.2}$$

The biasing potential was used to perform umbrella sampling (US) free energy simulations along the reaction coordinate $\lambda$. The simulations were carried out with an in-house implementation of the dRMSD potential in the GROMACS v4.6 software. In order to improve the sampling of relevant conformational states along the dRMSD coordinate during US simulations replica exchanges between simulations of neighboring $\lambda$ values were introduced (REUS-technique). Replica exchanges were attempted every 1000 steps alternating between odd and even indexed neighboring replica pairs to avoid that one system state could swap with more than one replica window per exchange step. The force constant for the harmonic dRMSD potential function was set to a value of 1000 kJ mol$^{-1}$ nm$^{-2}$ and the reference dRMSD was changed in equidistant steps from $\mathcal{D}_0^A = 0.0$ nm to $\mathcal{D}_0^B = 0.40$ nm in the replica windows. The number of 14 replica windows was adjusted in test simulations such that exchange rates higher than 30 % were recorded. In order to seed the reference distances for the distance pairs which accounted to the dRMSD, average distances were calculated from the 600 ns cMD simulations. The PMF along the dRMSD coordinate was calculated with the WHAM [341] algorithm implemented by Grossfield et al. [342], using 100 bins and a tolerance of $1 \times 10^{-5}$ kJ.

### 6.4.5 Electrostatic energy calculations

The electrostatic energy of Rab1b conformations in different modification and in the presence of different nucleotides was calculated using the finite-difference Poisson Boltzmann (FDPB) approach implemented in the APBS software package [343]. Structures representing the active state (represented by replica window with average dRMSD $\approx 0.1$ nm with respect to the crystal start structure) and the inactive state (replica with dRMSD $\approx 0.25$ nm) were obtained as snapshots from the trajectories every 600 ps skipping the initial 12 ns to account for equilibration. All water molecules and surrounding ions were removed and the linearized FDPB was solved for each snapshot with the APBS software package [343] using a two-stage focusing (initial boundary conditions were calculated with the multiple Debye-Hückel option) and a final grid spacing of 0.3 Å. The divalent $Mg^{2+}$ was included in the continuum electrostatics calculation as it has a significant effect on the local electrostatic field of the nucleotide binding site and the nearby switch regions and was found to be conserved in its binding site throughout all simulations. A dielectric constant of $\epsilon = 80$ for solvent was used [344] for protein in water was used. A permittivity of $\epsilon = 1$ in the protein was used to directly compare with the explicit solvent simulations which employ a vacuum permittivity and implicitly include orientational polarisation effects since a trajectory of conformations was analysed. Electrostatic energy calculations for the AMP-Rab1b:GDP ensemble in inactive and active states were repeated after replacing the adenylylated residue 77 with tyrosine (resulting in Rab1b:GDP structures) in order to estimate the electrostatic contribution of the AMP modification to the switch opening. This procedure minimizes all errors of electrostatic energy calculations which arise from the grid representation of the molecules because the investigated structures differ only in the presence or absence of the AMP group but not in the coordinates of all other atoms (placement relative to the grid). The same protocol was applied to inactive and active conformations of the Rab1b:GTP ensembles and subsequent in silico mutation to Rab1b:GDP by deletion of the $\gamma$-phosphate in GTP, in order to give an estimate for the electrostatic contribution of GTP vs. GDP to the switch I/II stability.

### 6.4.6 Protein Expression and Purification

Rab1b proteins, $SidD_{37-350}$ and $DrrA_{8-533}$ were produced as described previously [312, 317, 332]. In brief, $SidD_{37-350}$ and $DrrA_{8\text{-}533}$ were cloned in a pET19 vector with N-terminal Hexahistidine-tag ($His_6$-tag) and a tobacco etch virus cleavage site. Protein production in *E. coli* BL21 CodonPlus (DE3)-RIL cells was induced by addition of 0.5 mM IPTG overnight at 20 °C. Purification was achieved by Nickel affinity chromatography including cleavage of the $His_6$-tag and final gel filtration in 20 mM HEPES pH 7.5; 100 mM NaCl; 2 mM DTE; 1 mM $MgCl_2$. $Rab1b_{3\text{-}174}$ proteins were produced in a pMAL vector with N-terminal $His_6$-MBP tag and a tobacco etch virus cleavage site. Rab1b mutant proteins were generated by site-directed mutagenesis. Protein production in *E. coli* BL21 CodonPlus (DE3) cells was induced by addition of 0.5 mM IPTG overnight at 20 °C. Purification was achieved by Nickel affinity chromatography including cleavage of the $His_6$-MBP-tag and final gel filtration in 20 mM HEPES pH 8.0; 50 mM NaCl; 2 mM DTE; 1 mM $MgCl_2$; 10 µM GDP.

### 6.4.7 Preparative nucleotide exchange

Nucleotide exchange of Rab1b proteins was performed as described earlier [312]. In brief, Rab proteins were incubated with 5 mM ethylendiamintetraacetid acid (EDTA) and a 20 times molar excess of nucleotide at 25 °C for at least 2 hours in exchange buffer (20 mM HEPES pH 8.0; 50 mM NaCl, 2 mM DTE). Excess nucleotide was removed by using a PD-10 column (GE Healthcare) in storage buffer (20 mM HEPES pH 8.0; 50 mM NaCl, 2 mM DTE; 1 mM $MgCl_2$; 10 µM nucleotide). Completeness of the exchange was verified by reversed phase HPLC analysis.

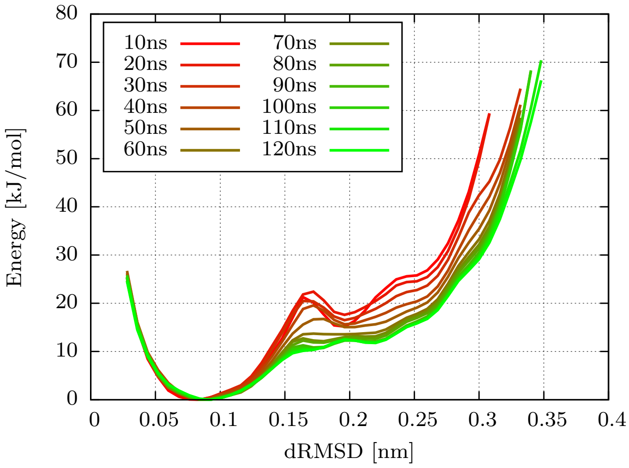

Figure 6.9: Convergence of PMF along the global dRMSD coordinate for AMP-Rab1b:GDP system over time. The cumulative PMF is plotted every 10 ns and the initial 5 ns are excluded from the calculation for equilibration. PMF does not change by more than 0.1 kJ mol⁻¹ ns⁻¹ after approximately 80 ns.

### 6.4.8 Preparative Adenylation

Preparative adenylation of Rab1b was performed as described previously [312]. In brief, Rab1b was incubated with a 2.5 molar excess of ATP and an 0.01 molar ratio of DrrA$_{8\text{-}533}$ at room temperature. Completeness of the reaction was verified by mass spectrometry. The modified protein was purified by size exclusion chromatography (20 mM HEPES pH 8.0; 50 mM NaCl, 2 mM DTE; 1 mM $MgCl_2$; 10 µM GDP).

### 6.4.9 Deadenylation of Peptide-AMP by SidD

Deadenylation of the switch II peptide TITY$_{AMP}$YRGAHGC by SidD$_{37-350}$ was analyzed in a time-dependent manner by reversed phase chromatography using an Aeris C4 widepore column (Phenomenex) on a Shimadzu HPLC system. Peptide (50 µM) was incubated with 0.5 µM SidD at 25 °C and subjected to reversed phase HPLC analysis at indicated time points. A binary gradient of 100% $H_2O$ containing 0.01% trifluoracetic acid (TFA) and 100% acetonitrile containing 0.01% TFA from 5 to 25% acetonitrile at 1 ml min$^{-1}$ flow rate was used to elute peptides. Peptides were detected by measuring tyrosine absorption at 274 nm. The progress curve of 5 µM peptide-AMP by 50 nM SidD was calculated in OriginPro v8.6G (OriginLab, Northampton, MA) using a 10 times lower rate constant.

### 6.4.10 Deadenylation assay

Deadenylation by SidD$_{37\text{-}350}$ was measured using the change in tryptophan fluorescence as reported previously [345] in a Fluoromax-3 spectrophotometer (HORIBA Jobin Yvon) (excitation at 297 nm; emission at 340 nm). Start of the deadenylation reaction was induced by adding 50 nM and 100 nM of SidD$_{37\text{-}350}$ to 5 µM Rab1b$_{3\text{-}174}$ Q67A and 1 µM Rab1b$_{3\text{-}174}$ F45A in 20 mM HEPES pH 7.5, respectively; 50 mM NaCl; 5 mM $MgCl_2$; 2 mM DTE at 25 °C.

### 6.4.11 Convergence of dRMSD simulations

In order to estimate the required simulation time to obtain well converged potential-of-mean-force (PMF) curves for simulations with a dRMSD biasing potential on the switch II region, we tested different simulation times of up to 120 ns per dRMSD interval (window) with 14 replica windows on the unmodified Rab1b:GDP system biasing the unfolding of switch II along the dRMSD coordinate. The cumulative PMF along the dRMSD coordinate was plotted every 10 ns (figure 6.9). We found that the PMF change was smaller than 1 kJ mol$^{-1}$ per 10 ns or 0.1 kJ mol$^{-1}$ ns$^{-1}$ after approximately 80 ns. Accordingly, following these benchmark results we chose a simulation time of 80 ns for all dRMSD H-REUS simulations. The convergence criterion of 0.1 kJ mol$^{-1}$ ns$^{-1}$ for the change in PMF per simulation time was met by all dRMSD PMF simulations within the 80 ns .

| System differences | $\epsilon_{\text{protein}}$ | $\Delta\Delta E_{\text{coulomb}}$ | $\Delta\Delta E_{\text{reaction-field}}$ | $\Delta\Delta E_{\text{total}}$ |
|---|---|---|---|---|
| AMP-Rab1b:GDP − Rab1b-GDP | 1 | $-2.6 \pm 1.8$ | $22.4 \pm 10.4$ | $19.8 \pm 12.2$ |
| AMP-Rab1b:GDP − Rab1b-GDP | 2 | $-1.3 \pm 0.9$ | $16.2 \pm 9.4$ | $14.9 \pm 10.3$ |
| AMP-Rab1b:GDP − Rab1b-GDP | 4 | $-0.7 \pm 0.5$ | $9.9 \pm 7.6$ | $9.2 \pm 8.1$ |
| Rab1b:GTP − Rab1b-GDP | 1 | $-1.5 \pm 1.4$ | $30.5 \pm 10.4$ | $29.0 \pm 11.8$ |
| Rab1b:GTP − Rab1b-GDP | 2 | $-0.8 \pm 0.7$ | $18.9 \pm 13.5$ | $18.1 \pm 13.2$ |
| Rab1b:GTP − Rab1b-GDP | 4 | $-0.4 \pm 0.4$ | $11.6 \pm 7.7$ | $11.2 \pm 8.1$ |

Table 6.3: Electrostatic contributions to the mean energy difference of inactive vs. active conformational ensembles of Rab1b:GDP in the presence or absence of the adenylylation at Tyr77 based on FDPB calculations (see Methods for details). Energies have been calculated for three different dielectric constants in protein ($\epsilon = 1, 2, 4$). The reaction field stabilizes the active conformation more when AMP or GTP are present and the effect is reproducible for different protein permittivities $\epsilon_{\text{protein}}$. The reduction in calculated energy differences with increasing $\epsilon$ results the increasing similarity of solvent and solute permittivity. The conformations that were used for FDPB calculations were extracted from simulations in explicit solvent using a dielectric permittivity of $\epsilon = 1$ which therefore is the correct permittivity for comparison with MD results. $\Delta\Delta E$ values are given in [kJ/mol].

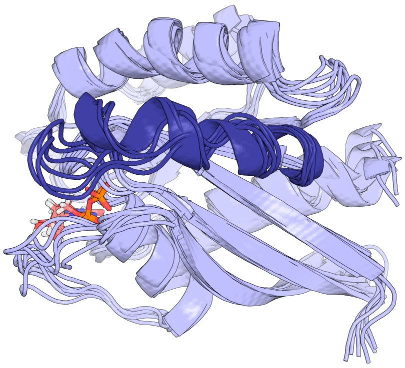

Figure 6.10: An ensemble of conformations representing the transition from active to inactive state. Switch II is colored in dark blue, the rest of Rab1b is colored light blue. Snapshots are taken from different replicas of the dRMSD H-REUS simulations.

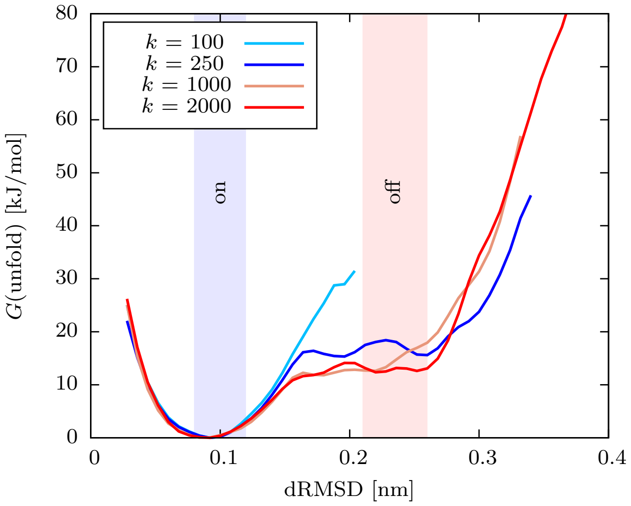

Figure 6.11: Different force constants for the dRMSD biasing potential (equation 2) have been tested for unfolding of switch II region exemplarily on the Rab1b:GDP system. Force constants $k$ are given in [kJ mol$^{-1}$ nm$^{-2}$]. The biasing potential was increased in 14 steps from a reference value of $R_0^A = 0.0$ to $R_0^B = 0.65$ (in production runs $R_0^B = 0.40$ was used). Note that the effective force constant acting on a dRMSD atom pair is $k_{\text{eff}} = k/N$ with the number of atom pairs $N = 11$ resulting from the derivative of the biasing potential after $d_i$. The resulting PMFs converge for force constants over 250 kJ mol$^{-1}$ nm$^{-2}$. At even lower force constant of 100 kJ mol$^{-1}$ nm$^{-2}$ the biasing potential is unable to unfold switch II region resulting in no sampling of the inactive configurations.

# Chapter 7

# From Chaos to Order: The association process of RNAse-S studied by Molecular Dynamics Simulations

S-peptide undergoes a transitions from intrinsic disorder to an ordered helical state upon binding to its partner S-protein when forming the complex RNAse-S. Although RNAse-S is one of the most studied protein systems, the driving principles of the association mechanism are still not understood in full detail [346]. We use Molecular dynamics (MD) and advanced sampling approaches to investigate the binding process of the 14 residue S-peptide at atomic resolution. Our results complement the picture drawn by published literature. In agreement with experiment we find significant conformational fluctuations of the isolated S-peptide compatible with a disordered regime and only little residual helical structure. Based on this solution regime and with the help of two simplified diffusive models we rebut that S-peptide binds to S-protein via a conformational selection mechanism. With systematic in silico Alanine scanning we uncover that the major contribution to complex stability emerges from the C-terminal helical turn consisting of residues 8-13 while the N-terminal residues 1-7 contribute only little. Comparative simulations of S-protein in presence and absence of S-peptide reveal that isolated S-protein undergoes a global pincer like conformational change that narrows the S-peptide binding cleft. This conformational change is reversed by S-peptide association which also stabilizes conformational fluctuations in S-protein. In complex, the N-terminal helix of S-peptide unfolds and refolds repeatedly revealing that helix is only part of the equilibrium regime for these residues while the C-terminal residues are tighly confined the conformation that is found in the X-ray structure. This is in line with speculations from previous studies that helix formation constitutes the final step of the association process. We ultimately simulate 100 trajectories of S-peptide fragment consisting of residues 8-14 positioned in the close vicinity of the S-protein binding site. While in the majority of simulations S-peptide diffuses away from the binding site, many of those trajectories that attach to the S-protein surface, form initial native contacts with residues Phe8 or Met13. We assume that the initial anchoring of these residues is followed by helix formation of S-peptide on the surface which agrees with experimental findings that at least Phe8 plays an important role for the transition state.

## 7.1 Introduction

The initial stage in proteolysis of bovine pancreatic ribonuclease (RNAse-A) is a hydrolytic cleavage of the protein backbone between residues 20 and 21 by the serine protease subtilisin that conserves catalytic activity in the resulting complex RNAse-S [347, 348]. RNAse-S consists of the smaller N-terminal fragment S-peptide$_{1\text{-}20}$ (residues 1-20) and the larger C-terminal S-protein (residues 21-124). In complex, S-peptide residues 15-21 are coordinated in a disordered flexible loop region while the N-terminal residues 1-14 adopt an ordered helical conformation similar to the conformation in RNAse-A [349, 350]. However when unbound in solution the conformational regime of S-peptide is governed by intrinsic disorder [284–286, 351]. How S-peptide transitions from the disordered regime to the ordered helical state upon association to S-protein, is still not fully understood, despite the plethora of studies RNAse-S has been subjected to in the past [346]. Essentially two distinct association mechanisms have been proposed [352]. The *induced fit* hypothesis assumes that binding and folding of S-peptide to its native structure is coupled such that folding occurs simultaneously or quickly after initial contacts with the S-protein surface are formed. On the other hand the *conformational selection* hypothesis expects S-peptide to recurrently adopt the native conformation in solution and to bind only then to S-protein. The question has been targeted within a variety of studies to investigate first principles of disorder to order transitions induced by peptide binding.

Systematic truncation studies of the S-peptide sequence by Finn and Hofmann have shown that the first 14 residues of S-peptide$_{1\text{-}14}$ are sufficient to fully restore enzymatic activity [353]. Their study further revealed that the smallest possible motive to recover RNAse-S activity is S-peptide$_{8\text{-}12}$, however only at high molar surplus. Later a study with Alanine mutations of residues which did not contact S-protein led to a simplified analog of S-peptide$_{1\text{-}15}$ forming a native like complex with S-protein at slightly reduced activity and ten fold higher dissociation constant (AEAAAAKFARAHMAA compared to the native KETAAAKFERQHMDS) [354, 355]. These findings together with the X-ray structures of RNAse-S complex emphasized the importance of residues His12, Met13, and Phe8 functioning as scaffolding residues for S-peptide where additionally His12 participates in the enzymatic breakdown mechanism of RNA [349, 350, 356]. Their side chains represent 64% to the S-peptide surface area that becomes buried in complex and strongly interact with the hydrophobic patch at the S-protein binding surface [323, 357]. N-terminal residues (1-7) on the other hand only weakly affect the specific binding mechanism and complex stability as was shown by their replacement with Alanine. Tritium exchange measurements further suggested that helix unzipping without unbinding of S-peptide occurs as a native state fluctuation in RNAse-S, an effect which is enhanced at both low pH and salt concentration [358]. Removal of Lys1 has no large effect on activity and dissociation but removal of Glu2, which forms a salt bridge with Arg10 in the crystal that potentially stabilizes the helix, increases dissociation by a factor of 4 [357].

A sidechain replacement study measuring the effect on binding kinetics by Goldberg and Baldwin later identified Phe8 as key residue to stabilize the transition state [352]. While Phe8Ala mutation strongly affected both binding kinetics and complex stability the mutation of Met13 and His12, two other complex stabilizing residues, had only little effect on $k_{\text{on}}$ leading to the assumption that native interaction between these residues and S-protein are not present in the transition state. The authors however were not decisive whether the transition state requires residual helix content from disordered S-peptide prior to association. To illuminate this question, they proposed an experiment that should locally destabilizes S-peptide helix properties and measure the effect on association kinetics [359]. Such an experiment was performed by Bachmann et al. by replacing amino bonds in the S-peptide backbone with thioxo bonds. The local helix destabilization revealed that the effect on association rates $k_{\text{on}}$ is weak for most backbone positions except for Phe8 which reduced wildtype association rates by a factor of two. Helix destabilization however affected complex stability at least when included in residues Lys7 to His12 [288]. They concluded that the initial recognition does not require helical packing in S-peptide and that folding occurred on the surface of S-protein.

In this work we carefully review the literature on RNAse-S with regard to complex formation and complement the general picture with *in silico* experiments. We use Molecular Dynamics (MD) simulations to quantify the disordered conformational ensemble of unbound S-peptide $_{1\text{-}14}$ and draw connections to other studies. Based on these findings we take an alternative approach of two simple diffusion models to discuss and rule out the conformational selection hypothesis. We perform computational Alanine scanning on the S-peptide $_{1\text{-}14}$ sequence and identify Met13, His12, and Phe8 as important complex stabilizers, in good agreement with experiment. Surprisingly we find that Arg10 contributes also significantly to complex stability, a residue which is located in the minimal binding motive of S-peptide $_{8\text{-}12}$ required for RNAse-S activity. A comparative simulation of S-protein with and without S-peptide bound, reveals that the S-protein structure is stabilized by S-peptide binding and the binding site is narrowed in absence of S-peptide. In agreement with predictions on helix unzipping we find that the salt bridge between Arg10 and Glu2 which stabilizes the N-terminal helix opens and closes in the equilibrium leading to unfolding and refolding of the native helix in residues 1-7 while in the meantime residues 8-14 remain firmly bound to their native sites. Although the time scale of association hinders the simulation of the full pathways of S-peptide $_{1\text{-}14}$ association, we perform 100 independent simulations of the shorter fragment S-peptide $_{8\text{-}14}$ starting in the proximity of the S-protein binding site. We find that Phe8 or Met13 bind first to the hydrophobic pocket in the S-protein binding site at their native position serving as the earlier predicted anchor prior to S-peptide folding.

## 7.2 Results and Discussion

### 7.2.1 Intrinsic disorder of S-peptide

The key characteristic in the association mechanism of RNAse-S is the intrinsic disorder of unbound S-peptide in solution. To investigate the conformational regime and to answer the question how often S-peptide adopts the bound helix conformation yet before interacting with S-protein we performed an extensive 10 µs continuous MD simulation on unbound S-peptide $_{1\text{-}14}$ at 298 K (see Figure 7.1). During the whole simulation time S-peptide underwent a continuous refolding process, thereby sampling a variety of diverse conformations. The metastable conformations possessed lifetimes of at most several hundreds of nanoseconds before undergoing the next global structural transformation. An average fraction of helicity of 17% with most of the helical content being confined to residues Lys7 to Arg10 was found (figure 7.1). Measuring the root mean square deviation (RMSD) to the native bound S-peptide conformation over time revealed that less than 0.5% of the configurations were closer than 0.5 nm. The before mentioned conformational selection mechanism would require that conformations close to the bound one are sampled which was the case in our simulations only during 0.5% of the time. Otherwise a coupled binding and folding mechanism would pose no such restrictions to the conformational regime of unbound S-peptide.

Altough sampling of the dynamic regime was certainly not exhaustive during the 10 µs simulation, the simulation draws a picture of the intrinsically disordered S-peptide in solution that was expected and outlined previously in experiments: The increased helicity between residues Lys7

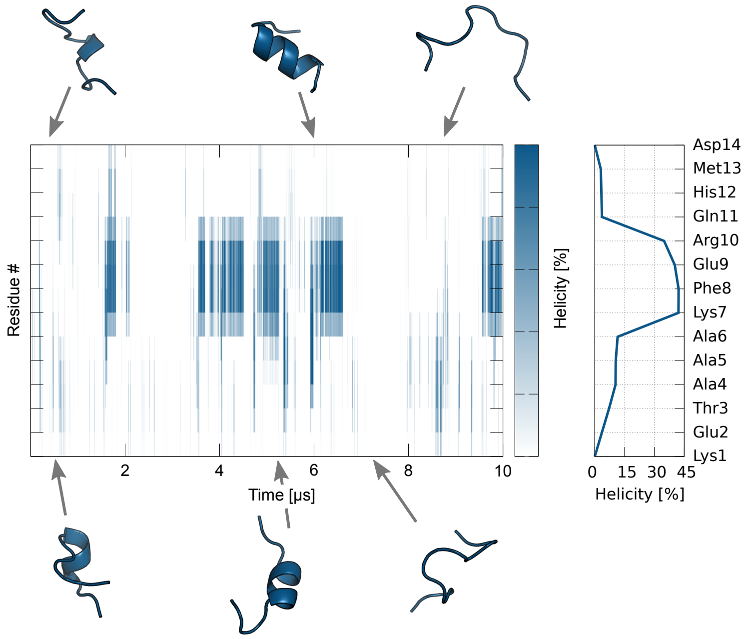

Figure 7.1: The helicity per residue over time of freely diffusing S-peptide $_{1\text{-}14}$ calculated from a 10 µs continuous MD simulation. Several structural snapshots of S-peptide $_{1\text{-}14}$ are depicted above and below in blue cartoon representation. The average helicity per residue is shown in the left panel.

and Arg10 of S-peptide is in good agreement with $^{15}$N relaxation parameter measurements on the 22 residue recombinant variant of S-peptide by Alexandrescu et al. which resulted in a comparable helicity profile [360]. Interestingly, early measurements on the 20 residues S-peptide $_{1\text{-}20}$ revealed, that even at low temperatures helix formation stops before Thr17 possibly near Met13, the residue where helix propagation stops in RNAse-S [361]. In our simulations we used truncated S-peptide $_{1\text{-}14}$ which could explain that helix propagation stops already at residue Arg10 as for short peptides, the terminal residues have increased entrophic freedom resulting in a decreased propensity to be ordered. S-peptide is known to have overall helicity values of 22–30% at low temperatures 0 ℃ down to only 5% at 25 ℃ [351, 360–362]. Furthermore, Kim and Baldwin reported for isolated S-peptide $_{1\text{-}20}$ that the helix propagation is terminated closely after residue Met13 and their data suggested helix formation in isolated S-peptide $_{1\text{-}20}$ to be limited to the same regions as in complex with S-protein [361]. Experimental evidence that the helix is stabilized at low pH and high salt (1 M) and model building studies suggested that a salt bridge formation between Glu9- and His12+ might be the reason [362, 363] which has been refuted later and was also not found in our simulation [364]. Our finding that the region of increased helicity largely overlaps with the minimal required motive S-peptide $_{8\text{-}12}$ to restore RNAse-S activity, leads to the speculation that increased packing in the region where transition state formation is expected to happen might be a requirement for complex formation. Note, that the transmutation among conformational states of S-peptide occurred on a timescale of several hundreds of nanoseconds in our simulation. This might be a hint that folding timescales of S-peptide on the S-protein surface reside in the same order of magnitude or are possibly even slower which would in turn render the full simulation of S-peptide association events difficult with the currently available simulation capabilities.

### 7.2.2 Diffusion controlled models for the S-peptide association

Two distinct mechanisms have been proposed for the binding mechanism of RNAse-S namely *conformational selection* and *induced fit* [288]. While for the induced fit mechanism, the conformation of S-peptide is not relevant for the initial association event, the conformational selection mechanism requires, that the disordered S-peptide adopts the bound conformation in solution previous to binding. Our previous results, however, showed that only less than 1% of the conformations sampled by S-peptide in solution are close to the bound conformation which drastically reduces the probability to form an encounter complex in case of a conformational selection mechanism. With the help of two diffusion based models, we estimate an upper limit for the association rate of S-peptide binding to S-protein under the fictitious assumption that S-peptide always adopts the bound conformation in solution. Our approach thereby neglects first the intrinsic disorder of S-peptide and simulates the optimal conditions of a conformational selection mechanism. The estimated rates are then compared to the experimentally derived value of $k_{\mathrm{on}} = 4.4 \times 10^5\,\mathrm{M}^{-1}\,\mathrm{s}^{-1}$ [288]. Under the assumption that the association mechanism is a conformational selection, we reduce the resulting rates by two orders of magnitude to include the effect of intrinsic disorder of S-peptide in solution. As in both models boundary conditions are used which define a very generous reaction condition and include only the diffusion controlled part of the association pathway, the real association rates are expected to be even further reduced by the subsequent binding process. By comparison with the experimental rate we then find, whether or not a conformational selection mechanism can be ruled out based on our simple diffusive models.

We first apply an analytic diffusion model which was originally invented by Schlosshauer et al. to predict association rates for protein–protein complexes in cases where binding partners are rigid and the binding mechanism is mainly diffusion controlled [365, 366]. Long-range electrostatic interactions between the binding partners are neglected in the model although S-peptide has a net charge of $+1$ at physiologic pH that can be assumed to decelerate the association kinetics with the positively charged binding region on S-protein due to electrostatic repulsion. Experimental results have confirmed this speculation as the $k_{\mathrm{on}}$ rate of S-peptide slows down with decreasing salt concentration [288]. The model treats the binding partners as rigid spheres which can react when they contact with the reactive regions of their surface. The reactive regions are defined by polar angles of the spherical representation of the binding partners ($\theta_{\mathrm{A,B}}$, see figure 7.2). In addition to contact, the reactive patches have to be orientated axially within certain tolerances ($\delta\phi_0, \delta\chi_0$). The diffusive properties of the binding partners are treated via their translational and rotational diffusion constants.

We used the sum of the radii of gyration as contact distance and for the calculation of rotational diffusion constants with the Stokes-Einstein formula. Translational diffusion constants were extracted from continuous MD simulations by fitting the root mean square displacement (see methods). The most critical and difficult to justify parameters of the model are the size of the reactive patches of the binding partners and the axial orientation tolerances. Based on a set of parameters fitted by Schlosshauer et al. on a variety of protein-protein complexes we calculated $k_{\mathrm{on}}$ rates for several sizes of reactive patches ($\theta_{\mathrm{A,B}}$) and tolerances ($\delta\phi_0, \delta\chi_0$, see table 7.3). The

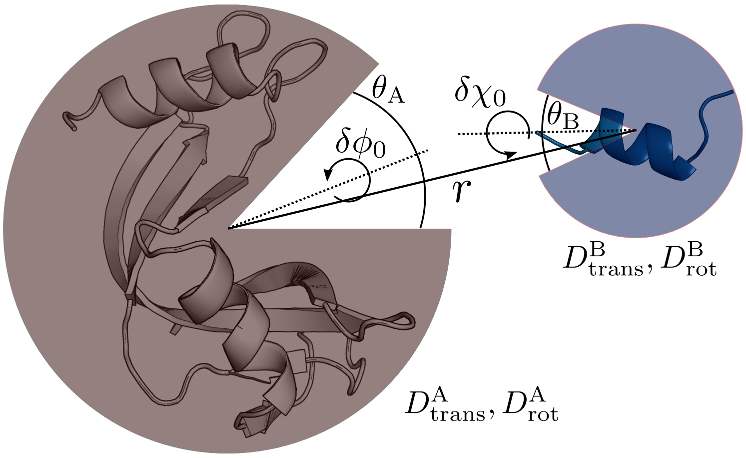

Figure 7.2: Coordinate definition for the analytic diffusion model of Schlosshauer et al. [365]. S-protein and S-peptide are represented as spheres of radii $R_A, R_B$ for which the radii of gyration were used. Binding occurs when $r \leq R_A + R_B$ and the reactive patches (defined by polar angles $\theta_{A,B}$) align within certain axial tolerances ($\delta\phi_0, \delta\chi_0$).

resulting association rates greatly depend on the choice of reaction parameters. For extremely large reactive patch size and tolerance (e. g. $\theta_{A,B}, \delta\phi_0, \delta\chi_0 = 60°$) the rate can be boosted to arbitrarily fast association, however, parameters close to the maximum found by Schlosshauer et al. [366] in a set of protein–protein complexes ($\theta_{A,B}, \delta\phi_0, \delta\chi_0 = 20°$) give a reasonable upper diffusive limit for the association rate of RNAse-S $k_{on}(\text{ana}) = 5.5 \times 10^6\,\text{M}^{-1}\,\text{s}^{-1}$. This is about one order of magnitude faster than measured in experiment, but would rule out the conformational selection mechanism because less than 1% of the encounter complexes match the reaction criterion and retard therefore the association kinetics estimate by at least two orders of magnitude being then one order of magnitude slower than the experimental value. A coupled folding and binding mechanism on the other hand involves a subsequent folding step to the native structure which can be expected to slow down the estimated association rate thereby potentially matching the experiment. The model provides a reasonable upper limit for the diffusion controlled association rate to form the encounter complex, that we can safely assume to be further slowed down to the total association rate when encounter complex subsequently folds to the native mode.

Although the analytic model given by Schlosshauer is inherently correct, its validity strongly depends on the accurate choice of reaction parameters. We therefore applied a second numeric model based on Brownian Dynamics (BD) simulations in order to estimate association rates as presented by Northrup et al. [367, 368]. In this model, BD simulations were initiated with a starting distance between the binding partners at which pairwise interactions are negligible. The simulations are terminated when the binding partners diffuse either away to a larger distance threshold or satisfy the reaction criterion for which a distance root mean square deviation of dRMSD $< 5\,\text{Å}$ from the native X-ray structure involving four interface residues was chosen (see table 7.2). Electrostatic interactions between the binding partners were accounted for by solving the Poisson-Boltzmann equation. The final $k_{on}$ rate was then calculated from a set of $2 \times 10^5$ trajectories resulting in $k_{on}(\text{BD}) = 1.0 \times 10^7\,\text{M}^{-1}\,\text{s}^{-1}$ (100 mM NaCl). The reaction criterion was weak enough that encounter complexes did not form van der Waals contacts between S-protein and S-peptide but ensured a reasonable orientation of the binding partners. The resulting rate should therefore be interpreted as an upper limit estimate for the diffusion from bulk into a pre-bound encounter complex requiring subsequent binding steps to form the native complex. The real $k_{\mathrm{on}}$ rate to the native complex can be safely assumed to be lower than this estimate. Again the result rules out the conformational selection model as a reduction of at least two orders of magnitude (due to subsequent binding steps and S-peptide conformations) leads to binding kinetics slower than those measured in experiment. In order to investigate the influence of salt concentration on the association kinetics, we performed a second BD run with electrostatic interactions at 10 mM NaCl. Similar to findings of Bachmann et al. we find an association rate $k_{\mathrm{on}}(\mathrm{BD}) = 4.4 \times 10^{6}\,\mathrm{M^{-1}\,s^{-1}}$ at 10 mM NaCl that slows down with decreasing salt concentration (compared to $k_{\mathrm{on}}(\mathrm{BD}) = 1.0 \times 10^{7}\,\mathrm{M^{-1}\,s^{-1}}$ at 100 mM NaCl). The effect arises from the positive net charges of both S-peptide $_{1\text{-}14}$ (+1) and the S-protein binding site. The slight electrostatic repulsion gets damped at high salt concentration allowing S-peptide to form the encounter complex more often.

### 7.2.3 Free energy calculations on S-peptide Alanine mutants

Residues Phe8, His12, and Met13 of S-peptide are known to contribute strongly to RNAse-S complex stability [369]. Folding studies with a fluorescently labeled S-peptide $_{1\text{-}15}$ further revealed that Phe8 has a predominant effect on the transition state stability and was identified as a key residue to form specific contact between S-peptide and S-protein during the transition state [352]. To investigate the effect of specific sidechain interactions between S-peptide and the hydrophobic binding site of S-protein on the complex stability, we performed a systematic computational Alanine scan on the complete sequence of S-peptide and measured the contribution to the binding free energy.

The results of the free energy calculations are shown in table 7.1 and are in good agreement with available data from experimental binding studies [288, 352]. The systematic underestimation of the difference binding affinities of about 3–4 kJ mol$^{-1}$ may be partly attributed to a residual effect of the helical starting structures especially for the free S-peptide mutations which would require more conformational equilibration in order to reflect the intrinsic behavior in solution. However also other free energy studies using the same Amber-99sb forcefield found that the calculated binding affinities were slightly below the experimental results which indicates a systematic cause by the forcefield [370]. As expected the largest sidechain contribution to complex stability is found for the Alanine substitution of scaffolding residue Phe8 in good agreement with previous findings. To our surprise Arg10 adds the second largest contribution (at significant errors tough) to the binding free energy only then followed by the documented stabilizers Met13 and His12. This is surprising as it was not expected after experimental substitution of Arg10Phe by Bachmann et al. [288] resulted in only 2.6 kJ, however this measurement was made in 1% DMSO compared to the other substitutions as the authors claimed this would increase S-peptide solubility however DMSO is known to modify binding properties in proteins [371]. Additionally the Arg10Phe mutation may result in completely different binding properties compared to Arg10Ala. Mutation of Glu2 reveals only weak contribution to the binding affinity although the salt bridge formation with Arg10 in the bound state would have been expected to have higher impact on the binding affinity. One reason might be that mutation simulations in unbound S-peptide were started from native X-ray structure biasing the helical state of the otherwise intrinsically disordered S-peptide. Salt bridge formation between Glu2-Arg10 was expected

| *Mutation* | $\Delta\Delta G^{\text{fec}}$ [kJ/mol] | $\Delta\Delta G^{\text{exp}}$ [kJ/mol] |
|---|---|---|
| Asp14Ala | $-1.07 \pm 3.26$ | – |
| Met13Ala | $10.32 \pm 0.35$ | $15.58 \pm 0.19$ |
| His12Ala | $7.87 \pm 0.90$ | $11.08 \pm 0.34$ |
| Gln11Ala | $1.86 \pm 0.44$ | – |
| Arg10Ala | $11.62 \pm 2.71$ | – |
| Glu9Ala | $-3.65 \pm 0.89$ | – |
| Phe8Ala | $25.87 \pm 0.64$ | $29.06 \pm 0.50$ |
| Lys7Ala | $-0.66 \pm 0.59$ | – |
| Thr3Ala | $0.80 \pm 0.31$ | – |
| Glu2Ala | $2.42 \pm 2.72$ | – |

Table 7.1: Difference free energy of binding of S-peptide Alanine mutations. Experimental data from Baldwin et al. [352] measured at pH 6.7, 10 mM Mops, 9.9 ℃.

to increase complex stability by stabilizing the N-terminal turns of S-peptide helix but our results indicate a minor importance to binding affinity. The major contribution to binding affinity arises from residues between Phe8 and Met13. In particular N-terminal residues before Phe8 have only weak influence on complex stability. This agrees with previous findings that the reduction in binding free energy by N-terminal truncation amounts only 6.3 kJ mol$^{-1}$ for S-peptide$_{4\text{-}14}$ and 10.5 kJ mol$^{-1}$ for S-peptide$_{8\text{-}14}$ compared to a total binding energy of $\sim$ 37 kJ mol$^{-1}$ for wildtype S-peptide$_{1\text{-}14}$ [288, 372].

There is an ongoing debate about the role of His12 protonation state in the association process which we want to briefly review. The acid-base properties of the imidazole ring in His12 play a critical role in the catalytic mechanism of RNAse-S [356]. Its protonation state is affected by substrate binding and respectively two p$K_\text{a}$ values were measured with and without substrate presence in the active site [373]. While the p$K_\text{a} = 5.75$ without substrate is close to the p$K_\text{a}$ value of the His sidechain in solution [374], the presence of the substrate shifts the tendency of His12 to be protonated to p$K_\text{a} = 7.0$ [375]. When no substrate is bound the His12-$\delta$ tautomer was speculated to be hydrogen bonded to the hydroxyl oxygen of Thr45 [376]. Studies on the C-peptide lactone (CPL), the N-terminal residues 1-13 of RNAse-A, revealed that the helix is unstable at standard conditions (26 ℃, ionic strength 100 µM) partial helix formation is conserved at low temperatures (1 ℃, ionic strength 100 µM) [363, 377]. At these low temperatures the helix stability of S-peptide was found to strongly correlate with the pH but melts out rapidly with increasing temperature being independent of temperature above 30 ℃ even at low pH being [363]. It has been speculated that the residual helix content may be caused by a salt bridge formation between residues Glu9- and His12+ at low pH, a salt bridge that is not present in the crystal structure [363]. This hypothesis has been proven wrong by experiments with C-peptide analogs and our simulations on free S-peptide$_{1\text{-}14}$ [364]. An interaction between His12 and Phe8 has been identified to increase the helix stability at low temperature while stabilizing effect from the protonation of His12 imidazole ring at low pH arose mainly from favorable interactions with the helix backbone [378]. Later studies postulated that a ring interaction between Phe8 and His12+ is the primary mechanism by which His12 stabilizes the C-peptide helix [379]. This led to the hypothesis that the C-peptide sequence might contain sufficient information to act as an

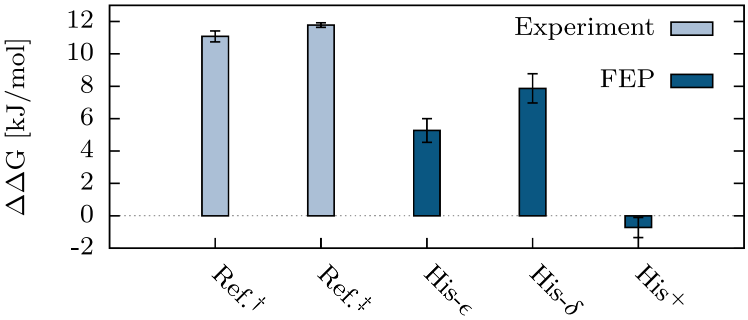

Figure 7.3: Calculated difference free energy of binding for Alanine mutants of the three His12 tautomers compared with experimental results. Experimental values are Ref.$^\dagger$ from Baldwin et al. (pH 6.8) [352] and Ref.$^\ddagger$ from Bachmann et al. (pH 6) [288]. Simulated values are derived from free energy perturbation (FEP) combined with H-REMD.

autonomous folding unit [379]. The effect of His12 protonation on the association transition state and complex stability remained unresolved. We performed free energy calculations to determine to what degree complex stability is affected by Ala substitutes of His12 tautomers and His12+ mimicking low pH. Comparing the difference binding affinities of our simulations with experimental His12Ala mutations studies reveals that the His12-$\delta$ tautomer is the predominant protonation state during S-peptide/S-protein complex formation (figure 7.3). Our results show, that the protonation of His12+ at low pH is unfavorable for the complex stability although the helix stability of S-peptide in solution was previously found to be increased. This is in good agreement with the experimental finding, that association kinetics are slowed down at low pH confirming the hypothesis that little helical structure is present in the transition state [288] and impedes previous assumptions that S-peptide acts as autonomous folding unit prior to association [379]. Additionally, the deceleration of association kinetics at low pH is amplified by the increased electrostatic repulsion between S-peptide (net charge increased to 2+ due to charged His12+) and the positively charged S-protein binding site.

### 7.2.4 Conformational regimes of S-protein and S-peptide

Circular dichroism specra of RNAse-S and S-protein indicated the same composition of secondary structure elements of S-protein and RNAse-S (pH 6.8, 10 °C) and a similar spectrum of S-protein at temperature below 30° with slight modifications at increased temperature [380, 381]. While the predicted difference in secondary structure of S-protein in absence of S-peptide is only small, we investigate possible changes in the tertiary structure by performing 1 µs MD simulation of S-protein in presence and absence of bound S-peptide. The conformational regime found for S-protein bound to S-peptide is very similar to the X-ray structure of RNAse-S with a RMSD of S-protein mostly below 0.2 nm (figure 7.4, B) which is in good agreement with experimental observations [380]. In the absence of S-peptide however, S-protein underwent a structural rearrangement of its global conformation. The hydrophobic S-peptide binding site narrowed down by a pincer-like closing motion of helix I and $\beta$-sheet II with a concurrent distortion of $\beta$-sheet I serving as a rotational axis (figure 7.4, A). The transition occurred rapidly after S-peptide was removed within the first 200 ns of the simulation. S-protein also showed increased flexibility and

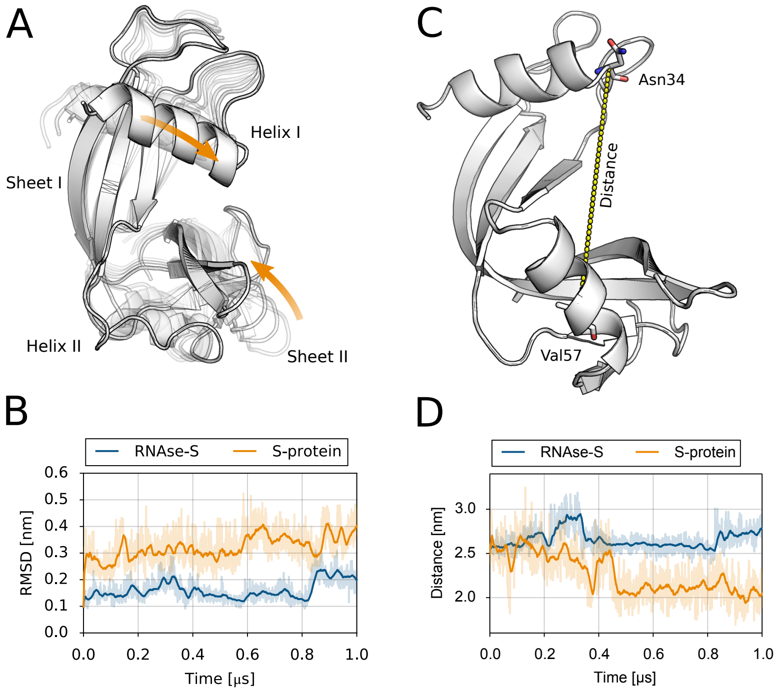

Figure 7.4: Cartoon illustration of the pincer-like conformational transition of S-protein in absence of S-peptide (A). The hydrophobic binding site of S-peptide is narrowed by the upward moving $\beta$-sheet II and the downward motion of Helix I. Helix II is stretched during this process and unfolds. The RMSD of the S-protein backbone with respect to X-ray structure is shown in (B) both for simulations in presence and absence of S-peptide bound to S-protein. The N-terminal five residues (21-25) of S-protein were removed from the RMSD calculation to mitigate noise from terminal fluctuations. The conformational transition occurs within the first 50 ns of the simulation without S-peptide bound. The distance between S-protein residues Asn34 in Helix I and Val57 in the hydrophobic pocket (C) is measured in both simulations with and without S-peptide present (D). The compression of the binding pocket in S-protein is characterized by a shortening of this distance by more then 0.5 nm when S-peptide is absent. Running averages are depicted in both RMSD and distance plots as solid lines while the actual data points are shown as light shades.

was conformationally less confined than RNAse-S which led to higher fluctuations in RMSD. The compression of the binding site was visualized over the distance between Helix I residue Asn34 and S-peptide binding site residue Val57 (figure 7.4, C). When S-peptide was removed, the distance decreased by more than 0.5 nm (figure 7.4, D). In presence of S-peptide on the other hand, the S-protein structure and the whole complex became inherently stabilized. We speculate, that the narrowed binding region potentially decelerates the mean first passage time of binding but not necessarily affects the second order kinetics of S-peptide association. The finding is surprising as to our knowledge a conformational change of S-protein coupled to S-peptide association has not been reported in the literature so far in particular as no X-ray structure of S-protein alone is currently available.

In our simulation of RNAse-S on the other hand, the salt bridge between sidechains Arg10 and Glu2 in S-peptide opened after about 50 ns and unlocked a flexible regime for the N-terminal

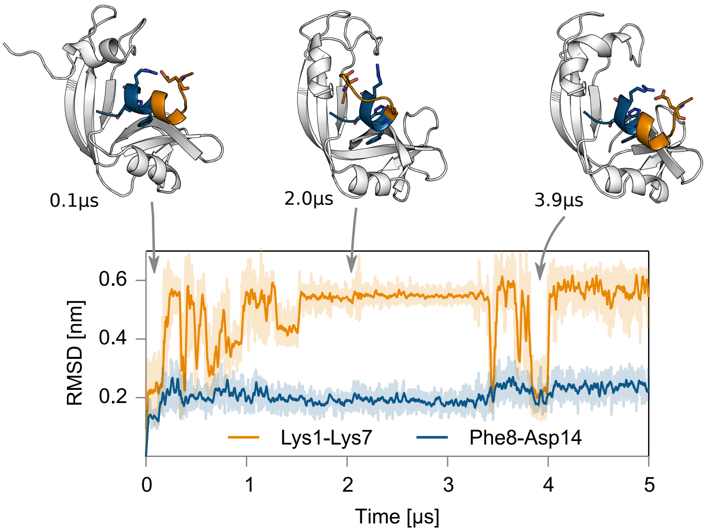

Figure 7.5: RMSD of S-peptide during a 5 µs simulations of RNAse-S. Two RMSDs are calculated with respect to X-ray structure for S-peptides residues Lys1-Lys7 (orange) and residues Phe8-Asp14 (blue). Snapshots of the RNAse-S are shown above to visualize the partial unzipping of N-terminal S-peptide helix. S-peptide residues 1-7 are shown in orange cartoon, residues 8-14 are shown as blue cartoon and S-protein is shown in white cartoon representation. Running averages are depicted as solid lines while actual data points are visualized as light shades.

residues 1-7 including uncoiling of the involved helical turn. Partial helix unzipping fluctuations of bound S-peptide were previously detected by tritium exchange experiments in the native state of RNAse-S without delivering an atomistic explanation and strongly support our findings. The effect was found to be enhanced at both low pH and low salt [358]. It was further speculated, that helix formation represents the final step of the association event [288]. Based on these findings, we performed another 5 µs simulation of RNAse-S to investigate to what extent unzipping fluctuations of S-peptide helix are part of the equilibrium ensemble in solution (figure 7.5). The C-terminal residues 8-14 remained close to the crystal structure conformation including RNAse-S catalytic residues His12 during the whole simulation. The stability was provided by Phe8 sidechain as the N-terminal residue of this stable motive which remained firmly attached to the hydrophobic S-protein binding pocket formed by Val54, Val108, Phe120 and Pro117 serving as an stabilizing anchor for the following residues 9-14. Residues 1-7 however, that feature little contact with S-protein, rearranged their conformation by unzipping and refolding the N-terminal helix (figure 7.5). The helix stabilizing salt bridge formed between Glu2 and Arg10, which is present in crystal structure, repeatedly opened and closed during the simulation and revealing alternative conformations of the N-terminal tail of S-peptide where in particular the positively charged tail interacted with charged alternative sites on S-protein and S-peptide. After about 4 µs the S-peptide formed again a helical structure with closed Glu2-Arg10 salt bridge and all native contacts restored very similar to the X-ray structure. Our results provide evidence that in equilibrium, S-peptide continuously unfolds and refolds the helical part between residues 1-7 while the C-terminal residues 8-14 are tightly confined to the X-ray conformation. The salt bridge between Arg10 and Glu2 that was expected to increase helix stability is only a temporary feature in the RNAse-S complex. This finding is confirmed by studies on a simplified S-peptide analog which formed a catalytically active complex with S-protein but lacked the salt bridge between Arg10 and Glu2 which indicates that the salt bridge formation is of minor importance for the RNAse-S complex and activity [355]. Note, that catalytic activity is even conserved after removal of Lys1 and Glu2 in S-peptide $_{3\text{-}14}$ [353].

### 7.2.5 S-peptide $_{8\text{-}14}$ association simulations

Based on experiments with truncated S-peptide, which revealed that S-peptide $_{8\text{-}12}$ constitutes the minimal required motive to restore catalytic activity with S-protein, however at strongly reduced activity of only 11% compared to S-peptide $_{1\text{-}14}$ and at a surplus of 8000:1 molar ratio to S-protein, we designed a setup to simulate the initial step in S-peptide/S-protein recognition with atomistic MD simulations [353]. Due to the weak activity of S-peptide $_{8\text{-}12}$ we included the complex stabilizing residue Met13 and the additional negatively charged Asp14 to overcome the slight electrostatic repulsion which is found unfavorable for the transition state formation. This is in line with the fact that already the addition of Met13 in S-peptide $_{8\text{-}13}$ strongly increased the activity in the truncation studies [353]. We assume that S-peptide $_{8\text{-}14}$ includes all essential binding features and removal of the N-terminal residues is justified as they form an ensemble of metastable states in equilibrium including N-terminal helix unzipping. We assume therefore that these residues are not involved in the primal step of the association pathway. To test whether the complex of S-protein and S-peptide $_{8\text{-}14}$ exhibits similar stability as with S-peptide $_{1\text{-}14}$ we performed a 1 µs simulation starting from truncated X-ray structure. Indeed, S-peptide $_{8\text{-}14}$ remained firmly bound without notable rearrangements of the complex structure (data not shown). In particular the catalytic sidechain of His12 and the scaffolding residue Phe8 remained in an X-ray structure like configuration, supporting the finding that enough structural information is conserved in the S-peptide $_{8\text{-}14}$ motive to restore RNAse-S activity.

We simulated 100 trajectories with S-peptide $_{8\text{-}14}$ positioned close to the binding site but without direct van der Waals contacts between S-peptide and S-protein to identify potential common features of the binding pathway and investigate the recognition step (see methods). Of the initial 100 trajectories, 79 were abandoned when S-peptide diffused away from the proximity of the binding site. The remaining 21 trajectories were terminated after 1 µs of simulation time. Although only in one trajectory S-peptide $_{8\text{-}14}$ formed an X-ray structure like complex with S-protein most of the 21 trajectories revealed a common binding feature. In 12 out of the 21 simulations the scaffolding residue Phe8 bound to the hydrophobic pocket in the S-protein binding site formed by residues Val47, Val54, Ile106, Val108, Pro117, and Phe120 (See figure 7.6) and in further 5 trajectories complex stabilizer Met13 bound to this pocket instead. This initial anchoring to the binding site successfully hindered S-peptide to diffuse away from the binding site in those cases. This finding confirms earlier speculations that at least hydrophobic residue Phe8 serves as an initial anchor for the subsequent folding of S-peptide on the surface of S-protein [288, 359]. In one trajectory S-peptide adopted a helical conformation and subsequently bound to the X-ray like pose forming crystal structure like RNAse-S in an conformational selection manner (See figures 7.6, 7.7, trajectory 10). We assume however, that neither the sampling time nor the number of simulated trajectories is sufficient to allow the drawing of a concluding picture of the binding mechanism. Especially as structural rearrangements of S-peptide on the S-protein surface occurred on the timescale of several hundreds of nanoseconds. While looking at the pincer like structural change of S-protein that narrows the S-peptide binding site, we found several

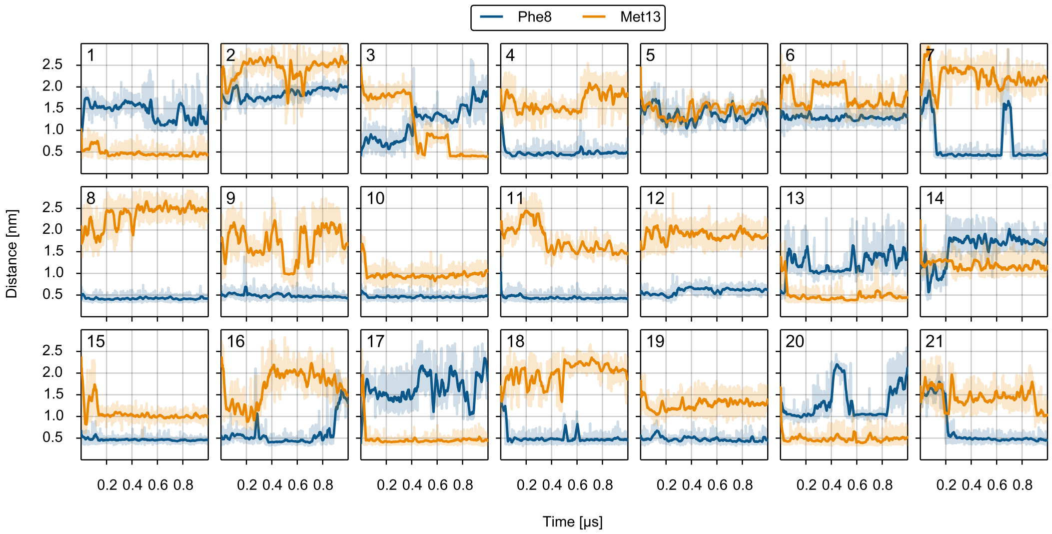

Figure 7.6: The distance of residues Phe8 (blue) and Met13 (orange) to the center of the hydrophobic pocket in S-protein formed by residues Val47, Val54, Ile106, Val108, Pro117, and Phe120. The distance is measured for those 21 out of 100 simulations of S-peptide$_{8\text{-}14}$ association where S-peptide$_{8\text{-}14}$ remained closer than 1.8 nm to the binding site during the 1 µs simulation time. The running average over 10 ns is depicted as solid line, actual data is shown transparently. A distance of about 0.5 nm corresponds to the native side chain pose of Phe8 in the RNAse-S complex.

trajectories where in particular the flexibility of helix I hindered some S-peptide residues to lock into their native binding site (figure 7.7 depicts the S-protein/S-peptide $_{8\text{-}14}$ conformation after 1 µs simulation time). We speculate that this global conformational change in S-protein limits association kinetics by impeding the hydrophobic collapse of S-peptide on the binding surface to form the helical turn in residues 8-14.

## 7.3 Conclusion

This study aims to complement and complete the picture of the mechanism underlying complex formation of S-peptide with S-protein and the associated transition from disorder to order of S-peptide conformations. Combining the present knowledge and our findings this picture can be summarized as follows: In the absence of S-peptide, S-protein undergoes a pincer like conformational rearrangement that narrows the cleft where S-peptide binds. The conformational rearrangement comes at increased flexibility of the whole protein. This behavior is reversed upon S-peptide association and helix formation. While intrinsically disordered in solution S-peptide folds on the surface of S-protein after formation of initial key contacts of residues Phe8 or Met13 with the hydrophobic pocket in the S-protein binding site. The hypothesis that conformational selection is the predominant association pathway can safely be ruled out by considerations based on two diffusion models. It is however still unclear whether increased helicity in the region of residues 7-10 as found in experiment and our simulations is a requirement for complex formation. At least when helix propensity at Phe8 is decreased in experiment, the association rate is halved [288] which might arise from both retarded folding on the surface or conformational

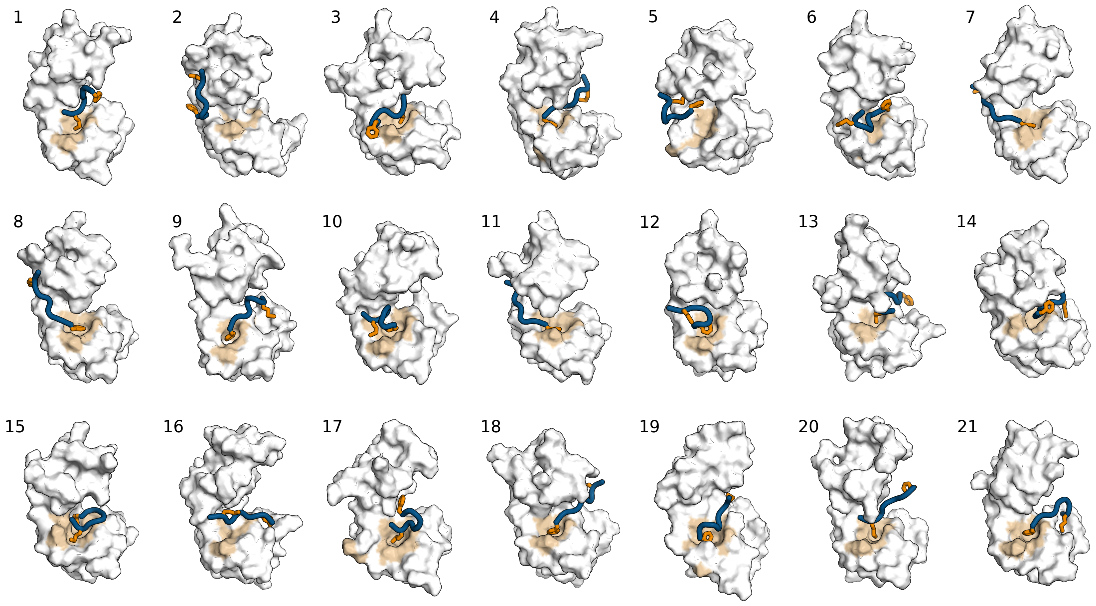

Figure 7.7: Final snapshots of S-peptide$_{8\text{-}14}$ association simulations of those 21 out of 100 simulations where S-peptide$_{8\text{-}14}$ remained close to the binding site of S-protein. S-peptide$_{8\text{-}14}$ is shown in blue cartoon with residues Met13 and Phe8 highlighted as orange stick. S-protein is depicted as white surface with the hydrophobic patch in the S-peptide binding site colored in light orange.

selection of the C-terminal helix turn of S-peptide (residues 8-12). The helix formation of the N-terminal residues 1-7 on the other hand seems to be the final step in the association process. This helical turn can be stabilized by a salt bridge formation between residues Arg10 and Glu2 but in equilibrium the N-terminal tail of S-peptide undergoes continuous conformational rearrangements with the helix being only one out of many allowed conformations in the native state ensemble. This behavior contrasts with the C-terminal residues 8-13 which remain tightly confined to their positions in complex and contribute most to binding affinity. By comparing free energy simulations of His12 mutation to Alanine with experimental free energy differences we find that His12+ protonation slightly defavours complex stability and association kinetics and that the predominant protonation state during complex formation must be the N-$\delta$ tautomer. Interestingly we find, in agreement with experiment, that diffusion of S-peptide into the vicinity of the S-protein binding site is slightly slower at low salt concentration, an effect which evolves presumably from the damping of repulsive electrostatics between positively charged S-protein and S-peptide. At high S-peptide concentrations, the reaction becomes concentration independent which shows that the concentration dependent encounter complex formation is followed by a concentration independent structural rearrangement [352, 358].

The question remains why S-peptide requires the N-terminal residues although they do not remain in a stable helical fold in RNAse-S and their mutation to Alanine has only little or no impact on complex stability. Possibly, the N-terminal residues increase the specificity of S-peptide and prevent unspecific sticking to the S-protein surface or formation of inactive complex configurations as found in our association simulations of S-peptide$_{8\text{-}14}$. Definitely, their presence strongly improves the catalytic activity of RNAse-S.

## 7.4 Methods

### 7.4.1 Preparations

We used the crystal structure of wild-type bovine pancreatic RNAse A as starting conformation for our simulations (PDB Code 1FS3). Residues 15 to 20 were removed to produce the complex of RNAse-S as those residues were found to be unimportant for both binding and enzymatic activity [357]. Superimposition of crystal structures of RNase-S (PDB Code 2RNS) with RNase A showed identical conformations within the estimated errors which of 0.05 Å [323] which renders any of the X-ray structures a valid starting point for the simulations. X-ray water molecules were kept and the protein was solvated in a rhombic dodecahedron box with periodic boundary conditions and rendered electrostatically neutral by the inclusion of chloride ions. Atomic interactions were parametrized with the Amber ff99sb-ILDN forcefield [42] and the TIP3P [246] explicit solvent model. To account for the correct ionisations states of acidic and basic residues, we carefully performed a $pK_A$ computation with the Karlsberg++ web server [373, 382] and included the resulting protonation states at $pH = 7.0$ in the setup. Simulations were performed with the GROMACS 4.6 molecular dynamics software package [96, 244]. Simulations of unbound S-peptide were always started from the bound configuration (residues KETAAAKFERQHMD in PDB 1FS3) with stripped of S-protein and crystal waters if not indicated differently.

### 7.4.2 Simulation protocol

Following the setup procedure, energy minimization with the steepest decent algorithm was performed to remove steric clashes until one of the two convergence criteria was reached (20 k steps or the maximal force below 100 kJ mol$^{-1}$ nm$^{-1}$). The system was then heated to 303 K in the NVT ensemble for 150 ps with the velocity rescaling thermostat [47] and equilibrated at standard condition pressure 1.01 bar in a NPT ensemble using the Berendsen barostat [45] for another 150 ps. During equilibration the time step for the leap-frog integrator was set to 1 fs and increased to 2 fs for the production runs. Production runs were performed with the Parinello-Rahman barostat [48]. Longrange electrostatic interactions were treated with the Particle Mesh Ewald (PME) algorithm [250] with grid interpolation of 6th order (4 during production runs) and a fast Fourier transform spacing of 0.12 nm. Short range interactions were switched off with a cutoff distance of 1.1 nm and the error introduced by the truncation of Lennard–Jones interactions to pressure and energy was mitigated with a long-range dispersion correction. LINCs at a coupling matrix extension order of 12 (4 in production) was applied to covalent bonds comprising hydrogen atoms [249]. To prevent undesired conformational rearrangements of the protein during the equilibration, backbone atoms were restraint to their X-ray structure position with a force constant of 100 kJ mol$^{-1}$ nm$^{-2}$.

### 7.4.3 Free energy calculations for Alanine mutations

The difference free energy of binding associated with the mutation of S-peptide residues to Alanine were simulated with a H-REMD free energy protocol as described previously [370]. In brief, interactions of the mutated sidechain with the rest of the system were gradually switched off while interactions of the Alanine methyl group were simultaneously turned on via a coupling coordinate $\lambda$ based on a dual topology approach [383]. Both non-bonded van der Waals and

| Distance | S-peptide | S-protein |
|---|---|---|
| $d_1$ | Phe8 | Leu51 |
| $d_2$ | Phe8 | Phe120 |
| $d_3$ | Asp14 | Tyr26 |
| $d_4$ | Ala5 | Val116 |

Table 7.2: Residue pairs that defined the distances $d_i$ in the dRMSD calculation. The dRMSD was used as the reaction criterion between S-peptide and S-protein in the BD based association model.

electrostatic interactions were represented by softcore potentials ($\alpha_{sc} = 0.3, \sigma_{sc} = 0.25$ and a softcore power of 6) [242, 243]. The transformation was performed along 21 equidistant $\lambda$ steps and replica exchange was attempted every 1000 time steps. In order to calculate difference free energies of binding, the mutation was simulated twice, once in S-peptide/S-protein complex and once in unbound S-peptide. Each $\lambda$ window was simulated for 20 ns and to account for conformational rearrangements especially in unbound S-peptide, the initial 5 ns were skipped for the free energy calculations. The free energy difference between wild type and Alanine mutant was then estimated with the Bennett Acceptance Ratio method [384]. All windows were started from the X-ray configuration of the RNAse-S complex with removed loop residues 15 to 20 (PDB code 1FS3) and removed S-protein for mutations in unbound S-peptide. Errors were estimated by splitting the data in five blocks assuming uncorrelated sampling amongst those and calculating the mean variance. The corresponding change in binding affinities for the mutations was then calculated as the difference free energy of mutating the residue once in RNAse-S complex and once in unbound S-peptide.

### 7.4.4 Association simulations

To generate an ensemble of conformations a 1 µs simulation was performed of S-peptide$_{8\text{-}14}$ in water. S-peptide structures from this ensemble were then isotropically positioned on a 10 nm shell around S-protein and Brownian Dynamics simulations were performed. The BD simulations were terminated either when S-peptide diffused away to a 12 nm radius or came close to the binding site of S-protein. The criterion for close configurations was measured as the distance root mean square displacement (dRMSD) of pairs of residue distances listed in table 7.2 and reference distances of 0.4 nm. Whenever the dRMSD was below 0.9 nm, the BD simulation was terminated. BD simulations were performed until a set of 100 starting structures with close S-peptide/S-protein separation was generated. These configurations were the solvated in TIP3P water [246] at 100 mM NaCl and minimized and equilibrated for 100 ps at 298 K and 1.01 bar in an NPT simulation. Hydrogen mass repartitioning was used [385] to support a simulation time step of 0.4 fs. The 100 equilibrated conformations were subsequently simulated for either 1 µs in an NVT ensemble or terminated when the center of mass (COM) of S-peptide diffused away from the binding site of S-protein (COM of residue Val47) by more than 2 nm. Trajectories where S-peptide remained close to the binding site were further analyzed.

### 7.4.5 Diffusion controlled association rate calculation

We wrote an in house software implementation for the protein-protein association rate calculation based on a purely diffusion controlled model of two spheres with orientation restraints as it was presented by Schlosshauer et al. [365]. Translational diffusion constants for both S-protein and S-peptide were estimated from continuous MD simulations of 0.4 µs and 1.3 µs duration by measuring the mean square displacement over time and fitting the Einstein relation [386]. The resulting values $D^{\text{trans}}_{\text{S-protein}} = 2.75(696)\,\text{Å}^2\,\text{ns}^{-1}$ and $D^{\text{trans}}_{\text{S-peptide}} = 67.24(788)\,\text{Å}^2\,\text{ns}^{-1}$ were found. Rotational diffusion constants were estimated from the Stokes–Einstein relation $D^{\text{rot}} = k_{\text{B}}T/6\pi\eta R$ with a viscosity of $\eta = 8.9\,\text{N}\,\text{s}\,\text{m}^{-2}$ and $T = 300\,\text{K}$ ($D^{\text{rot}}_{\text{S-protein}} = 3.3\,^{\circ}\,\text{ns}^{-1}$ and $D^{\text{rot}}_{\text{S-peptide}} = 24.0\,^{\circ}\,\text{ns}^{-1}$). For the contact radii $R$ we calculated the average radius of gyration over the continuous MD trajectories and plugged them into the relation for the rotational diffusion constant ($R_{\text{S-protein}} = 14.7\,\text{Å}$ and $R_{\text{S-peptide}} = 7.6\,\text{Å}$). $k_{\text{on}}$ values for different angular restraints $\delta\phi_0, \delta\chi_0, \theta^0_{\text{A}}$, and $\theta^0_{\text{B}}$ were calculated. It should be noted, that even the upper limit of 20° resulted in an increase of the association rate of only about one order of magnitude ($k_{\text{on}} = 5.5 \times 10^6\,\text{M}^{-1}\,\text{s}^{-1}$). The absolute theoretical diffusion limited association rate for two uniformly reactive spheres is given by Smoluchovskis formula $k_{\text{D}} = 4\pi DR$ with the relative diffusion constant $D = D_{\text{A}} + D_{\text{B}}$ and the reaction contact distance of the two spheres $R = R_A + R_B$ [387]. The infinite series of rotation matrix integrations was truncated after $N = 40$ (Equation 7.1).

### 7.4.6 Analytic Schlosshauer model

For the calculation of association rates we applied a model system consisting of two spheres in full diffusion controlled limit. The model assumes that reaction occurs when the spheres are at contact distance $R = R_{\text{A}} + R_{\text{B}}$. To account for the fact that for proteins, the reaction can only take place if the binding partners are facing each other with their binding site various constraints can be applied. A often used assumption describes the reactive part of the surface with a polar angle $\theta_{\text{A,B}}$ a so called reactive patch which has to be aligned before reaction [387, 388]. Schlosshauer derived a even more general expression for the $k_{\text{on}}$ rate which includes also the requirement for an axial alignment of the two reactive patches defined by angle restraints $\delta\phi_0, \delta\chi_0$ (see figure 7.2) [365].

$$k_{\text{on}} = D(Ra_0/8\pi^2)^2 \times \left[ \frac{D}{\kappa} a_0 - R \sum_{l,l_1,l_2}^{\infty} \frac{K_{l+1/2}(\xi^*)}{l K_{l+1/2}(\xi^*) - \xi^* K_{l+3/2}(\xi^*)} \times q_{ll_1l_2} \sum_{n=-l_1}^{l_1} \left( \sum_{m=-l_1}^{l_1} \hat{C}^{mn}_{l_1l_2} \begin{pmatrix} l & l_1 & l_2 \\ 0 & m & -m \end{pmatrix} \right)^2 \right]^{-1} \tag{7.1}$$

with $q_{ll_1l_2} = 16\pi(2l+1)(2l_1+1)(2l_2+1)$ and $\kappa$ quantifying the extent of diffusion control in the reaction where $\kappa \to \infty$ in the full diffusion controlled limit. The $K_{l+1/2}$ are the modified bessel functions of the second kind. $\begin{pmatrix} l & l_1 & l_2 \\ 0 & m & -m \end{pmatrix}$ is the Wigner $3j$ symbol and

$$\hat{C}^{mn}_{l_1 l_2} = \frac{\sin(m\delta\phi_0)}{m}\frac{\sin(n\delta\chi_0)}{n} \int_0^{\theta^0_{\mathrm{A}}} \sin(\theta_{\mathrm{A}}) d^{l_1}_{mn}(\theta_{\mathrm{A}}) \mathrm{d}\theta_{\mathrm{A}} \int_0^{\theta^0_{\mathrm{B}}} \sin(\theta_{\mathrm{B}}) d^{l_2}_{-m-n}(\theta_{\mathrm{B}}) \mathrm{d}\theta_{\mathrm{B}}. \tag{7.2}$$

The relative translational diffusion constant is defined as the sum of the two respective diffusion constants $D = D_{\mathrm{A}} + D_{\mathrm{B}}$. The diffusion dependent variable $\xi^*$ is given as

$$\xi^* = R \left[ l_1(l_1+1)\frac{D^{\mathrm{rot}}_{\mathrm{A}}}{D} + l_2(l_2+1)\frac{D^{\mathrm{rot}}_{\mathrm{B}}}{D} \right]^{1/2}. \tag{7.3}$$

The integral over Euler angles $a_0$ is given as

$$a_0 = (4\pi)^3 \delta\phi_0 \delta\chi_0 (1 - \cos\theta^0_{\mathrm{A}})(1 - \cos\theta^0_{\mathrm{B}}) \tag{7.4}$$

with Wigner's small $d^l_{mn}(\theta)$ rotation matrix. We truncated the infinite sum over indices $l, l_1, l_2$ after $N = 40$ as the relative change of $k_{\mathrm{on}}$ per additional iteration dropped under 1%. The nomenclature in the literature is ambiguous as the $K_{l+1/2}$ are also sometimes termed modified Bessel functions of the *third* kind [365, 388]. To avoid ambiguity the modified Bessel functions of the second kind used here are defined according to Solc and Stockmeyer [389] as:

$$K_{l+1/2}(x) = K_{1/2}(x) \sum_{m=0}^{l} \kappa_{m,l}\, x^{-m} \tag{7.5}$$

with the Bessel function of order $1/2$,

$$K_{1/2}(x) = (\pi/2x)^{1/2}\, \mathrm{e}^{-x} \tag{7.6}$$

and the recurrence relation for the coefficients $\kappa_{m,l}$

$$\kappa_{m,l+1} = \kappa_{m,l-1} + (2l+1)\kappa_{m-1,l} \quad \text{with} \quad 0 < m < l. \tag{7.7}$$

Initial values are given by

$$\kappa_{m,l} = 0, \qquad m > l \tag{7.8}$$

$$\kappa_{0,l} = 1, \qquad \kappa_{l,l} = (2l-1)!!. \tag{7.9}$$

Equation (7.1) was numerically solved with an in house software implementation written in python.

### 7.4.7 Brownian Dynamics Simulations

Brownian Dynamics (BD) simulations [368] were used to estimate the rate of S-peptide to form an encounter complex for the scenario of a conformational selection mechanism. The electrostatic potential around S-protein was first calculated by solving the Poisson Boltzmann equation at 298 K and a relative permittivity of water $\epsilon = 78.5$ with the ABPS software package [390]. The

| $\theta_A^0$ | $\theta_B^0$ | $\delta\phi_0$ | $\delta\chi_0$ | $k_{\text{on}}$ [$\text{M}^{-1}\text{s}^{-1}$] |
|---|---|---|---|---|
| 10 | 10 | 10 | 10 | $4.1 \times 10^5$ |
| 15 | 15 | 15 | 15 | $1.7 \times 10^6$ |
| 10 | 30 | 15 | 15 | $3.0 \times 10^6$ |
| 20 | 20 | 20 | 20 | $5.5 \times 10^6$ |
| 60 | 60 | 60 | 60 | $6.8 \times 10^8$ |

Table 7.3: Association rates of S-peptide to S-protein for several reactive patches sizes and axial orientation tolerances. Angular constraints $\theta_A^0, \theta_B^0, \delta\phi_0, \delta\chi_0$ are denoted in degrees.

electrostatic potential was saved on a grid around the centered S-protein to calculate electrostatic interactions during the BD simulations. Effects of S-peptide on the electrostatic field were neglected. To smooth the surface of the binding partners, hydrogen atoms were removed during the BD simulation and their charge was added to the connected heavy atoms. Spacial exclusion between S-peptide and S-protein was set to an atomic collision distance of 2 Å. BD simulations were started with randomly oriented S-peptide at a distance of $b = 100$ Å from the center of mass (COM) of S-protein where interactions were isotropic and negligible and terminated when either S-peptide diffused away from S-protein by more than $q = 120$ Å or when the encounter complex was formed. When the distance root mean square deviation (dRMSD) of four residue distances between S-peptide and S-protein (table 7.2) became smaller than dRMSD $< 5$ Å the configuration was counted as encounter complex. The dRMSD (Equation (7.10)) was calculated from $N = 4$ atomic distances $d_i$ between the binding interface of S-protein and S-peptide thereby simultaneously ensuring reasonable orientation and separation of the encounter complex.

$$\text{dRMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(d_i - d_{i0})^2} \tag{7.10}$$

The reference distances $d_{i0}$ were taken from the X-ray structure of RNAse-S. The rate of forming the encounter complex was then calculated with the method proposed by Northrup et al. [367]

$$k_{\text{on}}^{\text{ES}} = 4\pi D_0 b \beta_\infty \tag{7.11}$$

where $D_0$ is the relative diffusion constant of the binding partners which was calculated from Stokes–Einstein relation from the radii of gyration and

$$\beta_\infty = \frac{\beta_{\text{BD}}}{1 - (1 - \beta_{\text{BD}})\Delta} \tag{7.12}$$

with $\beta_{\text{BD}}$ being the fraction of encounter complexes formed during all started BD trajectories and

$$\Delta = \frac{b}{q}. \tag{7.13}$$

# Conclusion and Outlook

When Karplus, Levitt, and Warshel received the Nobel price in chemistry in 2013 for their contributions in the field of molecular dynamics simulations and particularly the combination of different levels of theory in multiscale approaches, the field of computational biochemistry in general crossed an important border on its way to becoming a widely accepted theory outside the boundaries of a specialized scientific community. When asked for a simple description of his work, Nobel price laureate Martin Karplus answered after winning the price: “If you like to know how a machine works, you take it apart, we do that for molecules.” He inherently included in his statement the picture of molecular dynamics being a computational microscope that enables scientists to investigate the structural and dynamical processes of biological molecules at atomistic resolution. The track of success of molecular dynamics simulations that recently climaxed in the Nobel price award started back in the late 1950’s when the age of computers was making its first steps. When the first stereochemical structures of larger biomolecules were solved, experimentalists used models of plastic balls and sticks to visualize the spatial arrangement of atoms and bonds between them. Electron density maps calculated from X-ray diffraction data were frequently visualized with manual drawings. A publication from the early years of protein structure determination, investigating the structure of the protein–protein complex RNAse-S by Wyckhoff et al. in 1967 gives a good impression how tedious this process must have been at the time before computers came into play [391]. Nowadays, computer programs are not only performing the calculations to determine molecular structures or serve as a visualization platform but even allow the simulation of molecular structure formation directly *in silico*. Currently the simulation of complete folding pathways of proteins is limited on standard hardware but the development of *Anton*, a special purpose molecular dynamics computer, permitted already in 2011 a glimpse into this future when successfully simulating the folding pathways of several fast-folding proteins on the millisecond timescale [60]. This might soon be possible on readily available standard hardware if the industrial ambition of developing faster computer hardware continues with the current speed, a progress which just recently enabled simulations of reasonable protein sizes for several microseconds on graphics processing units (GPUs). In principle, the structure prediction of proteins simply starting from the knowledge of the genetic information that encodes the protein sequence would then be possible by the abundant simulation of folding pathway ensembles, thereby potentially replacing the complex and time consuming processes involved in the experimental structure determination by a generic computer simulation. On other scientific frontiers like the field of drug development, the simulation of binding affinities and association pathways for a large number of potential drug candidates could replace cost intensive high-throughput screening techniques in the laboratory. The simulation techniques for this are already nowadays available and have proven their predictive capabilities on the small scale.

In that sense, the future sounds like a world where simulators might one day completely replace experimentalists in a lab coat. Although this is certainly thinkable for specialized applications, theory and experiment have ever been bonded together in the progress of science and certainly will be in the future. The reason is, that physics, chemistry, and biology are empiric sciences that develop theories to describe the world based on experiences and measurements which are obtained from experiments. The predictions of theoretic models therefore require the normative power of experimental measurements to improve their quality in a most probably never ending feedback loop. Although current models already have proven their predictive quality in many situations they are still incomplete. That counts for the description of molecules with quantum mechanics calculations (the best theoretic description of molecules currently available) and obviously even more for the classical modeling of atoms as balls connected by springs in molecular dynamics simulations. Systematic errors are inherent in these models and are difficult to quantify, may they evolve from incomplete description of physical processes or just be a matter of numeric discretization. The dependence on experiments might first seem a drawback of theoretic models in general but it constitutes the most powerful paradigm in empirical sciences. It is based on the recognition that the provability of theories is logically excluded as their models are solely based on a finite series of limited empirical observations. Theories therefore can either be falsified by rebutting their predictions with an experiment or the predictions are confirmed by the experiment but the theory itself will never be proven correct. A theory remains valid as long as its predictions are not contradicted by experimental results. The quality of our current models and computational limitations resulting in insufficient sampling are typical reasons why simulators have often difficulties with their results being considered equal by experimental colleagues. This is well reflected by the fact that predictive simulation studies are rarely found in high impact peer reviewed journals. If they make their way into these journals, they are typically strongly connected to experimental findings or have reproductive character. This sometimes seduces experimentalists to sit on a high horse claiming that their results represent the *real* world a believe that is certainly often justified (they may be interpreted incorrectly though), however thereby ignoring the predictive power inherent in simulations. Simulations can provide a comfortable shortcut compared to experiment that ideally sharpens the eye of the experimentalists by telling them where to look at and by delivering alternative insight. On the other hand, simulators sometimes lift off the ground of empiricism and neglect the fact that their results remain predictions and require experimental confirmation. For the sake of an improved scientific progress and in order to close the confidence gap between these two species of scientists, we need strong communication and tight collaboration between both worlds.

In this thesis a series of studies is presented showing, that simulations and experiments are at their best when tightly coupled together. Inspired by experimental work of the group of Thomas Kiefhaber we approached the association mechanism of RNAse-S with molecular dynamics simulations. RNAse-S is a product of the proteolytic cleavage of RNAse-A which is according to a review by Ronald T. Raines from 1998 the "most extensively studied enzyme of the 20th century" [346]. RNAse-S is formed when the smaller S-peptide binds to its larger partner S-protein. S-peptide is an intrinsically disordered peptide in solution but adopts an helical conformation when bound to S-protein. Despite of the groundbreaking experimental work of the Kiefhaber group and others, the complete association process is still not fully understood. By carefully comparing our simulation results with the vast amount of experimental studies around RNAse-S we validated every step of our proceeding. With a simple diffusion model calculation we could rule out one of the two proposed binding mechanisms and confirm the interpretation of the Kiefhaber study. But our simulations also indicated additional, so far unknown effects. Not only S-peptide seems to undergo a conformational transition upon binding, but also its larger partner S-protein which has been expected to be conformationally stable undergoes a global conformational rearrangement. Discussions with our collaborators in the Kiefhaber group resulted in the idea of measuring this prediction *in vitro* with the help of an small angle X-ray scattering experiment. Theoretically predicted scattering curves seem promising that the effect could be visible in such an experiment. Also our simulations indicate that S-peptide forms an initial key contact with S-protein with specific anchoring residues first before folding on the surface to the native helical conformation. This mechanism has been suspected in the past and could be produced under the "computational microscope" but would be difficult to observe with experimental techniques otherwise. However the sampling capabilities of molecular dynamics simulations performed on GPUs of the latest generation are still not sufficient to sample complete association pathways of S-peptide to S-protein in an unmodified forcefield as full association occurs on timescales above tens of microseconds.

Our study on the switching mechanism of Rab1b is based on an experimental finding, that bacteria of the species *Legionella pneumophila* express the enzyme DrrA/SidM to introduce a posttranslational modification in Rab1b by adenylylating the Tyrosine residue 77. Aymelt Itzen and his group challenged us whether we could come up with a reasonable model about what this modification effects in Rab1b and in particular how the switching mechanism is affected. Based on umbrella sampling simulations and Poisson–Boltzmann calculations we found that the switching mechanism of Rab1b is locked in its active configuration by the posttranslational modification, providing a mechanism for the bacterium to interact with the vesicle trafficking control system of the host cell. We find that the effect is mainly caused by the negative charge of the modification which is introduced into the functional switch region of Rab1b. Interestingly, another bacterial enzyme SidD has been identified that removes the adenylylation from Rab1b and thereby reinstates the original cellular control mechanism. In the meantime our explanation of the modified switching mechanism could be fully confirmed by preliminary nuclear magnetics resonance spectra of adenylylated and wild type Rab1b in their GDP (inactive) and GTP (active) form. Interestingly, further preliminary results from simulations and experiments indicate that another posttranslational modification, the posphocholination of neighboring residue Serine 76 which is net charge neutral, has no impact on the structural coordination of Rab1b.

An impressive example how simulations may help experimentalists to figure out the issues inherently drowsing in experimental techniques is given in our study on fluorescence labeling of S-peptide. While discussing about the possibilities of a fluorescence spectroscopy study of the S-peptide/S-protein association process with collaborators of the Don Lamb group, we realized, that S-peptide might be too small to be fluorescently labeled without significantly influencing the conformations and dynamics. We asked ourselves whether it might be possible to identify such a modification in silico and whether such simulations could be confirmed by connecting them to experimental fluorescence spectroscopy measurements. We used continuous molecular dynamics simulations to investigate the conformational and dynamical regime of S-peptide with and without the fluorescence labels being attached. We validated the results of the simulation of labeled S-peptide by comparing them with the experimental PET-FCS measurements. The successful validation allowed us to compare the simulated conformational and dynamical regimes between labeled and unlabeled S-peptide. The impact of the labeling was found to be surprisingly strong. Not only the diffusive properties of S-peptide were heavily affected but the peptide was also locked in a metastable $\beta$-sheet configuration during the simulations caused by a short-range stacking interaction between the fluorescence labels. In order to confirm the findings from the simulations, we measured the circular dichroism spectrum of both the labeled and unlabeled S-peptide and found accordance with our predictions from the simulations. This study highlights the potential of molecular dynamics simulations to help experimentalists to identify whether a measurement technique itself modifies the behavior of the measured system, an effect which is often difficult to access by solely experimental means.

In a further study, we developed a protocol to identify ligand binding modes in the receptor binding site with simulations. With regard to the field of drug design it is quite possible, that parts of the work which is currently performed in the laboratory might soon be replaced by such simulation protocols. With high-throughput screening techniques, large libraries of thousands of drug candidates are frequently checked in a trial and error approach for their binding capabilities to the drug target. This process is both expensive and time consuming. The theoretic concepts and protocols to calculate binding affinities of drug candidates on the computer are already available but they are computationally still too expensive to be applied on the large scale. However these computational approaches promise more than to be one day just a cheaper replacement for the lab techniques. They give access to detailed atomistic insight in the dynamic processes occurring in the binding pocket. Such information paves the way to the sophisticated and specific design of drug molecules instead of the brute force trial and error testing of thousands ligands.

The progress in the understanding of life will in the future even stronger depend on the interaction between different groups of scientists be they experimentalists or theoreticians. Tools and simulation protocols need to be developed in order to shortcut often time consuming, expensive, and error prone laboratory techniques. Yet, simulation techniques are capable to provide valuable insight into biomolecular processes that is otherwise difficult to obtain in experiment. However molecular dynamics simulations still require improvements especially in terms of force-field parametrization and the inclusion of additional physical effects in order to strengthen their predictive power and their accurate representation of molecular processes. Experiments are the ultimate instance to check these predictions for their validity in the real world.

# Appendix A

# hdWE: A hyper-dimensional Weighted Ensemble implementation

In the course of this thesis, we developed from scratch an implementation of the weighted ensemble methodology. The original idea was to allow the simultaneous simulation of multiple binning coordinates in one WE setup, therefore we came up with the name *hyper-dimensional Weighted Ensemble* or short hdWE for the implementation, a name which is admittedly somewhat extravagant. The code is purely written in Python 2 to guarantee a maximum of platform independence and provide an easy start for coding newcomers who want to understand and expand the implementation. The WE algorithm is subdivided in two basic parts: While the bookkeeping of bin structures, WE iterations, analysis of bin coordinates, and stochastic weights $w_i$ of the trajectories is handled by the hdWE program itself, the molecular dynamics simulations are outsourced to an external MD software suite (e. g. AMBER, Gromacs). This subdivision allowed us to focus our efforts on the design and performance improvement of the WE implementation itself while taking advantage of state of the art MD implementations at the same time. Two key benefits of this approach should be noted. First, as the weighted ensemble does not require modifications of the Hamiltonian, we rely on the fastest available implementations of pure MD algorithms which are nowadays typically accelerated with the help of General Purpose graphics processing Units (GPUs) and allow us to access timescale of several µs for a typical biomolecular system of $\sim 50000$ atoms. Second, during the propagation step of a WE iteration, the trajectories are completely independent from each other and can therefore be propagated in multiple parallel MD threads for the next WE time step $\tau$ (figure 3.3). This form of trivial parallelizability is an intrinsic advantage of the WE algorithm. After the trajectories have been propagated, their end structures are evaluated with respect to their new position on the binning coordinates and resorted into the bins. Then the trajectories are split and merged according to the resampling mechanism to ensure that the bins are evenly filled. Eventually the algorithm reenters the MD propagation step.

## A.1 Implementation details

In this section we want to briefly outline the important data structures and the hierarchy among them to provide insight into the design of the hdWE program. At the lowest hierarchical level, the WE algorithm organizes a number of trajectories which are evenly distributed along the binning coordinate. Trajectories are represented in hdWE as an instance of the class `Segment()`. For

---

The full source code of hdWE was released under the license terms of the GPLv3 and can be obtained from https://github.com/enzyx/hdWE the better understanding, we want to look at a pseudo code representation of this class, which is reduced to the principle features.

```
class Segment(object):
    def __init__(self, probability, parent_iteration_id,
                       parent_bin_id, parent_segment_id,
                       iteration_id, bin_id, segment_id):
        self.probability          = probability          # float
        self.parent_iteration_id  = parent_iteration_id  # int
        self.parent_bin_id        = parent_bin_id        # int
        self.parent_segment_id    = parent_segment_id    # int
        self.bin_id               = bin_id               # int
        self.segment_id           = segment_id           # int
        self.iteration_id         = iteration_id         # int
        self.coordinates          = None                 # list of floats
```

A `Segment()` object is located and identifiable in time and bin space with three ids defining the iteration (`iteration_id`), bin (`bin_id`) and position in the bin (`segment_id`). In addition to that, the segment stores information about its recent history in the variables prefixed with `parent_`. The ids define the position of a segment in the bin data structure above and are directly linked to a trajectory coordinate file. The statistical weight $w_i$ of the trajectory assigned by the WE algorithm is stored in `probability` and the coordinate(s) of the trajectory on the binning coordinate(s) are stored in as an array in `coordinates`, e. g. in case of a single binning coordinate, the array contains only one element. The segments are organized in the `Bin()` class, containing an array `segments` of segment instances.

```
class Bin(object):
    def __init__(self, iteration_id, bin_id, target_number_of_segments,
                       coordinate_ids):
        self.iteration_id               = iteration_id    # int
        self.bin_id                     = bin_id          # int
        self.target_number_of_segments  = \
                            target_number_of_segments     # int
        self.coordinate_ids             = coordinate_ids
        self.segments                   = []
        self.initial_segments           = []
        self.resampling_history         = []
```

The `Bin()` is again identifiable via `iteration_id` and `bin_id` which are redundantly stored in the segment instances for simple access to identifiers. The bin class carries information about its position along the binning coordinate(s) in the `coordinate_ids` array and stores the target number of segments. In order to preserve the full history about segments being merged and split during the resampling procedure, a copy of the `segments` array is stored in `initial_segments` before resampling. Every step of the resampling process is stored in the `resampling_history` array. The resampling elements in this array are instances of the classes `Split()` or `Merge()` containing mainly `segment_ids` of the merged and split segments.

```
class Merge(object):
    """
    surviving_segment: The index of the surviving segment
    deleted_segments: List of segment_ids which are
                      deleted probability
    """
    def __init__(self, surviving_segment_id, deleted_segments_ids):
        self.surviving_segment_id  = surviving_segment_id
        self.deleted_segments_ids  = deleted_segments_ids

    def getType(self):
        return type(self).__name__

class Split(object):
    """
    parent_segment: segment_id of the split segment
    m: Number of segments which result from splitting
    """
    def __init__(self, parent_segment_id, m):
        self.parent_segment_id  = parent_segment_id
        self.m                  = m

    def getType(self):
        return type(self).__name__
```

The WE algorithm is round based and the underlying data structure representing one WE round is the class `Iteration()`. The iteration class stores information about the binning coordinates (`boundaries`) which are referenced by the bin instances in `bin.coordinate_ids`. The iteration counter is stored in `iteration_id` and the list of bin instances in the `bins` array.

```
class Iteration(object):
    def __init__(self, iteration_id, boundaries, n_starting_structures):
        self.iteration_id  = iteration_id    # int
        self.boundaries    = boundaries      # array
        self.bins          = []
```

Having introduced the data structure hierarchy, we now can analyze the main program loop in hdWE. For educational reasons, we skip the setup routines and jump directly into a pseudo-code representation. The code has been slightly simplified in terms of call parameters and additional function calls, compared to the actual hdWE implementation, to keep the level of detail at a minimum and focus on the essentials of the WE routine.

```
for iteration_counter in range(MAX_ITERATIONS):
    # 1. Intialize iteration and sort segments into bins
    iterations.append(Iteration(iteration_counter,
                                iterations[-1]))
    resorting.copyBinStructureToLastIteration(iterations)
```

```
    resorting.resort(iterations)

    # 2. Backup the segments lists of all bins
    for this_bin in iterations[-1]:
        this_bin.backupInitialSegments()

    # 3. Resampling (split/merge trajectories)
    resampler.resample(iterations[-1])

    # 4. Reweighting
    reweighter.doProbabilityReweighting(iterations)

    # 5. Run MDs
    md_module.runMDs(iterations[-1])

    # 6. Calculate Segment Coordinates
    md_module.calcCoordinates(iterations[-1])
```

The loop runs over a predefined number of WE iterations `MAX_ITERATIONS` and the instances of the `Iteration()` class are stored in the `iterations` array. First, a new instance of `Iteration()` is created and appended to the `iterations` list. The `resorting` instance provides functions to copy the bins and segments of the previous iteration to our new iteration object. Because segments have been propagated with MD at the end of the previous iteration, the trajectories (segments) have new coordinates in the bin spaces and need to be assigned to the bins with the `resorting.resort()` routine. If a trajectory accessed a previously empty bin region during the last iteration, a new bin instance is generated on-the-fly. In a second step, the `segments` list in the bins is backed up to the `initial_segments` list before the resampling routine merges and splits segments. The reweighting routine can (optionally) be applied, readjusting the probability weights of the segments. Internally, the reweighting routine calculates the averaged rate matrix over a given number of iterations. Then the algorithm enters the propagation step which is handled by the `md_module` which provides an interface to the external Molecular Dynamics program. The `MD_module()` class functions as a wrapper to different MD programs on a plugin basis. The specific call syntax of the individual MD programs is hidden behind general function calls (e. g. `runMDs()`, `calcCoordinates`) which have to be implemented when aiming to support an alternative MD software package in hdWE. The ultimate step in the main loop is the calculation of the new bin coordinates (`md_module.calcCoordinates()`) for the propagated trajectories to allow resorting them into bins in the subsequent iteration.

## A.2 Usage of hdWE

In this section, we briefly present the individual steps required to setup and perform a WE simulation with the hdWE program and the Amber molecular dynamics software [98]. hdWE uses a global configuration file (`hdWE.conf`) to define the WE simulation parameters. Its syntax is based on the python module `ConfigParser`. The following example shows a configuration file to simulate the binding process of a ligand to a protein with WE using the distance between protein and ligand as binning coordinate.

```
[hdWE]
workdir                  = /work/dir/path/
starting-structures      = ./pro-lig.rst7
jobname                  = pro-lig
segments-per-bin         = 10
max-iterations           = 10000
boundaries               = 10.0 12.0 14.0 16.0 18.0 20.0
sample-region            = -99999 99999
keep-coords-frequency    = 100
keep-trajectory-files    = False
resampling-mode          = westpa
merge-threshold          = 0.01
compress-iteration       = False
number-of-threads        = 8

[amber]
topology-path            = pro-lig.top
infile-path              = pro-lig.in
coordinate-masks         = pro-lig.mask
binary                   = pmemd.cuda
cuda_visible_devices     = 0,1,2,3
parallelization-mode     = mpi
mpirun                   = /usr/bin/mpirun -np 4
```

The workflow initially requires the setup of topology and coordinate files of the protein and ligand system with the AMBER package. Instructions on the setup protocol can be found in the AMBER manual. The resulting topology (`pro-lig.top`), coordinate (`pro-lig.top`), and Amber run file `pro-lig.in` have to be specified in the WE configuration file. The `workdir` parameter points to the directory where `hdWE` searches for files and writes the output of the WE simulation. In order to calculate the binning coordinate(s) from the MD trajectories, hdWE relies on the Amber tool `cpptraj`. The syntax for `cpptraj` input files is documented in the Amber manual and can be any type of coordinate. To specify one ore multiple binning coordinates, an Amber mask has to be written to the `pro-lig.mask` input file in one line per coordinate.

```
distance d :1-367 :LIG
```

In this example, the protein is defined by residues 1–367 while the ligand molecule is identified with the residue name `LIG`. The `out` directive for the definition of the `cpptraj` output file will automatically be added by `hdWE`. The bin boundaries are then defined with the `boundaries` parameter. In our example we have 7 bins covering distances $b_0 = \{\, d \mid 0 < d \leq 10 \,\}, b_1 = \{\, d \mid 10 < d \leq 12 \,\}, \ldots, b_6 = \{\, d \mid 20 < d \leq \infty \,\}$. The bins are open at their endpoints covering the whole phase space up to the region specified in `sampling-region`. In our case we just specified a numerically large range compared to the system size. The units of the distance are taken from `cpptraj` which uses Å for distances. If multiple binning coordinates are used, additional coordinate boundaries are specified in a comma separated list format. The target number of trajectories per bin is given in `segments-per-bin` and is required by the resampling routine to define, how often trajectories are to be split or merged. `max-iterations` gives the number of WE iterations and the additional flags `keep-trajectory-files` and `keep-coords-frequency` specify, whether the trajectory output files are stored and at how often the coordinate files resulting from each WE iteration are stored. The amber binary `pmemd.cuda` should be in the `$PATH` variable of the shell, alternatively a full path can be specified. `hdWE` starts multiple `pmemd` programs in parallel via a `mpi` pipeline, therefore the user can fine tune the parallelization via the `mpirun` command which may also run on multiple nodes via the `--host` or `--hostfile` flag of `mpirun`. Parallelization is switched on with the `parallelization-mode` parameter and the available GPUs are defined over the `cuda_visible_devices` parameter which is translated internally to the `CUDA_VISIBLE_DEVICES` shell variable. After having setup the configuration file, the simulation is ready to be started with `hdwE`. When calling `hdWE` with the flag `-h` from a terminal, it outputs an overview of the available command line options.

```
usage: hdWE.py [-h] [-c FILE] [-d] [-a | -o] [-n]

  _         ___          ________
 | |       | \ \        / /  ____|
 | |__    __| |\ \  /\  / /| |__
 | '_ \ / _` | \ \/  \/ / |  __|
 | | | | (_| |  \  /\  /  | |____
 |_| |_|\__,_|   \/  \/   |______|

A hyperdimensional weighted ensemble simulation implementation.

optional arguments:
  -h, --help            show this help message and exit
  -c FILE, --conf FILE  The hdWE configuration file
  -d, --debug           Turn debugging on
  -a, --append          continue previous iterations (use parameters from conf
                        file when --read is given)
  -o, --overwrite       overwrite previous simulation data
  -n, --new-conf        Read new boundaries from config file when --append is
                        used
```

In our example of a protein–ligand complex formation, the command to initiate the simulation would be

```
hdwe.py -c hdWE.conf
```

The program creates two output directories `pro-lig-log` and `pro-lig-run` containing log files of every iteration in the first and the trajectory/coordinate files in the second folder. The iteration log filenames are the zero padded iteration indices with the filename extension `.iter` (e. g. `00000000.iter, 00000001.iter`). The naming scheme for the coordinate and trajectory output files is assembled again from zero padded iteration, bin, and trajectory index (e. g. file `00000034_00002_00013.rst7` identifies as segment 13 of bin 2 in iteration 34). `hdWE` provides a set of tools to analyze the results of a WE simulation. In the given example the transition rates between arbitrary states can be calculated with the `ana_TraceFlux` program.

```
ana_TraceFlux -l pro-lig-log --state-A 0 10 --state-B 18 99999  -B 1000 -bs 500
```

The program requires the log file directory of the WE simulation and the definition of state intervals along the binning coordinate. The `-B` flag effects that the initial 1000 WE iterations are skipped as equilibration for the analysis. `ana_TraceFlux` performs error analysis according to the statistical bootstrapping method, and therefore splits the data in blocks of size `-bs 500` [392]. The `ana_BinPMF` tool calculates the PMF along the binning coordinate.

```
ana_BinPMF.py -l pro-lig-log -o pmf.dat
```

It again requires the path of the log file directory and an optional output filename. The PMF data is stored as an ASCII file which can be conveniently visualized with standard plotting software.

# Appendix B

# Distance based RMSD potential in GROMACS

In chapter 6 the influence of Tyr77 adenylylation on the conformational regime of Rab1b was investigated with the help of umbrella sampling (US) molecular dynamics simulations combined with the replica exchange (RE) technique. As reaction coordinate to unfold the switch regions in Rab1b a *distance based root mean square deviation* (dRMSD) potential was chosen and implemented natively in GROMACS version 4.6. The sampling along the dRMSD coordinate provides a powerful tool to investigate conformational changes of biomolecules and we want to briefly introduce the fundamental concepts as well as an application note of our dRMSD implementation in GROMACS.

## B.1 Theory

The dRMSD coordinate is defined as the mean deviation of a set of distances $d_i$ from their references $d_{i0}$. The distances $d_i$ are typically defined between pairs of atoms $a_{i1}, a_{i2}$ and the reference distances correspond to a target molecular structure. We write the definition of the dRMSD ($\mathcal{D}$) as

$$\mathcal{D}(d_1, \ldots, d_N) = \sqrt{\frac{1}{N}\sum_{i}^{N}(d_i - d_{i0})^2} \tag{B.1}$$

where the index $i$ runs over all distances between $N$ atom pairs which contribute to $\mathcal{D}$. The dRMSD is similar to the root mean square deviation (RMSD), which describes the deviation of a set of coordinates from their reference position and is often applied in the field of molecular dynamics to analyze the conformational evolution of molecules in simulations. Despite their similarity in definition and applicability, the dRMSD coordinate offers significant advantages over the RMSD with respect to software implementation and algorithmic performance. The main reason is, that distances between atom pairs can be trivially extracted during a MD simulation, as they are invariant from the rotation and translation of the molecule. The RMSD on the other hand, requires the molecular structure to be superimposed with its reference structure to eliminate the rotational and translational contributions and to yield only the internal conformational deviation from the reference. The superimposition step requires the determination of a translation vector to align the centroids of the molecules current and reference structures and an optimal rotation matrix for rotational alignment. The Kabsch algorithm provides a fully analytic solution

---

The full source code can be obtained from https://github.com/enzyx/gromacs-4.6-drmsd for the alignment problem [393, 394]. However, its application involves the numerical solution of a linear equation system, which is not always numerically stable. Furthermore, when used as a reaction coordinate in MD simulations, the superimposition has to be performed during every step of the MD algorithm, constituting additional computational cost. Therefore the dRMSD coordinate constitutes typically the safer to implement choice which delivers better computational performance over RMSD. In order to perform an US simulation along the dRMSD coordinate, a potential has to be defined, which traps the simulation at a specific point along the reaction coordinate. Frequently, a spring like potential is used in the form

$$V(d_1,\dots,d_N) = \frac{1}{2}k_0\big(\mathcal{D}(d_1,\dots,d_N) - \mathcal{D}_0\big)^2. \tag{B.2}$$

with a reference dRMSD $\mathcal{D}_0$ and a force constant $k_0$. In order to translate the dRMSD potential to forces which can be applied to the atoms in a MD simulation, we keep in mind that the distances $d_i$ are functions of the atom pair coordinates $\vec{r}_{i1}, \vec{r}_{i2}$.

$$d_i \equiv d_i(\vec{r}_{i1}, \vec{r}_{i2}) = |\vec{r}_{i1} - \vec{r}_{i2}| = \sqrt{(\vec{r}_{i1} - \vec{r}_{i2})^2} \tag{B.3}$$

The forces on the Cartesian atom coordinates $x_{i1}, y_{i1}, z_{i1}$ and $x_{i2}, y_{i2}, z_{i2}$ of atoms $a_{i1}, a_{i2}$ involved in distance $d_i$ are taken from the total derivative of $\mathcal{D}$ with respect to the atom coordinate e. g. $x_{1i}$

$$\begin{aligned} F_{x_{i1}} &= -\frac{\mathrm{d}V(d_1,\dots,d_N)}{\mathrm{d}x_{i1}} \\ &= -\frac{k_0}{N}\cdot\frac{\mathcal{D}(d_1,\dots,d_N) - \mathcal{D}_0}{\mathcal{D}(d_1,\dots,d_N)}\cdot(d_i - d_0)\cdot\frac{x_{i1} - x_{i2}}{d_i} \\ &= -F_{x_{i2}} \end{aligned} \tag{B.4}$$

According to Newtons 3rd law, the equal force on $x_{1i}$ acts on $x_{2i}$ in opposite direction. In vectorial form, equation B.4 may be rewritten as

$$\vec{F}_{a_{i1}} = -\frac{k_0}{N}\cdot\frac{\mathcal{D}(d_i) - \mathcal{D}_0}{\mathcal{D}(d_i)}\cdot(d_i - d_0)\cdot\frac{\vec{d}_i}{d_i} = -\vec{F}_{a_{i2}} \tag{B.5}$$

Having extracted the forces, the dRMSD potential can then be applied on a MD simulation. In order to drive a system between two states $A$ and $B$ however, we need to extend $\mathcal{D}$ with information about the endpoints $A, B$. Endpoints in terms of dRMSD can be defined by different sets of reference distances $d_{i0}^A, d_{i0}^B$ for the two states. The sets of reference distances may represent two different conformations of the molecule of interest for example. To control the potential between the endpoints, a scaling parameter $\lambda \in [0, 1]$ is introduced. The $\lambda$ dependent dRMSD coordinate is then defined as

$$\mathcal{D}(d_i, \lambda) = \sqrt{\frac{1}{N}\sum_i^N \Big(d_i - (1-\lambda)d_{i0}^A - \lambda d_{i0}^B\Big)^2} \tag{B.6}$$

The system is driven to the endpoint $A$ if $\lambda = 0$ and to $B$ if $\lambda = 1$ by linearly switching between the reference distance sets. The $\lambda$ dependent dRMSD potential is further extended by two reference dRMSDs $D_0^A, D_0^B$ and takes the form

$$V(d_i, \lambda) = \frac{k_0}{2}\Big(\mathcal{D}(d_i, \lambda) - (1-\lambda)\mathcal{D}_0^A - \lambda\mathcal{D}_0^B\Big)^2 \tag{B.7}$$

Again the forces on Cartesian coordinates of atoms $a_{i1}, a_{i2}$ have the form e. g. for $x_{i1}, x_{i2}$

$$\begin{aligned} F_{x_{i1}} &= -\frac{\mathrm{d}V(d_i,\lambda)}{\mathrm{d}x_{i1}} \\ &= -\frac{k_0}{N} \cdot \frac{\mathcal{D}(d_i,\lambda) - (1-\lambda)\mathcal{D}_0^A - \lambda\mathcal{D}_0^B}{\mathcal{D}(d_i,\lambda)} \cdot \\ &\quad \left(d_i - (1-\lambda)d_{i0}^A - \lambda d_{i0}^B\right) \cdot \frac{x_{i1} - x_{i2}}{d_i} \\ &= -F_{x_{i2}} \end{aligned} \tag{B.8}$$

When performing US simulations with several $\lambda$ windows along the reaction coordinate, the Replica Exchange technique may be applied in addition to enhance the diffusion of conformations between the end points and accelerate the discovery of alternative structures. This may be necessary especially as the conformational space that “fulfills” a certain dRMSD condition may vary at different $\lambda$ values.

## B.2 Application notes

The input files of GROMACS have been extended by several parameters to setup a Hamiltonian Replica Exchange Umbrella Sampling simulation (H-REUS). The simulation parameters are adjusted in the configuration file syntax (`.mdp`) and the following additional parameters are available.

**drmsd-pot:**

**yes**

Enable distance based RMSD potential.

**no**

Disable distance based RMSD potential.

**drmsd-ref: (0.0) [nm]**

Reference distance RMSD

**drmsd-refB: (drmsd-ref) [nm]**

Reference distance RMSD for state B

**drmsd-k0: (1000.0) [kJ mol$^{-1}$ nm$^{-2}$]**

Force konstant of the distance RMSD potential

**nstdrmsdpout: (100)**

Frequency of writing distance RMSD potential output (unimplemented yet)

The dRMSD potential can be turned on and off with the `drmsd-pot` switch. The parameters `drmsd-ref`, `drmsd-refB` and `drmsd-k0` correspond to the reference dRMSD values $\mathcal{D}_0^A, \mathcal{D}_0^B$ and the force constant $k_0$ in equation B.7. The standard units of GROMACS have to be used for the parameters as indicated. Atom pairs and their reference distances are defined in a dedicated section in the system topology file (`.top`). The position of the additional section `[ drmsd_restraints ]` in the topology file is not relevant. An exemplary definition of three distances has the following representation

```
[ drmsd_restraints ]
;    ai       aj       type     d0       d0B
     9        39       1        0.5      0.6
     19       49       1        0.5      0.6
     29       59       1        0.5      0.6
```

The first line represents the section header, while the second line is a comment (prefixed with the escape character " ; " in GROMACS topology files). The subsequent lines are organized in white space separated columns. The first and second columns indicate the atom indices of $a_{i1}, a_{i2}$ defining the distance $d_i$. The third column is internally used by GROMACS to identify the bond type and should always be 1. The reference distances $d_{i0}^A, d_{i0}^B$ are given in the last two columns. The different $\lambda$ windows in a replica exchange simulation are controlled with the `.mdp` file parameter

**bonded-lambdas:**

Zero, one or more lambda values for which Delta H values will be determined and written to dhdl.xvg every nstdhdl steps. Values must be between 0 and 1. Only the bonded interactions are controlled with this component of the lambda vector.

E. g. for a setup with 6 equidistant $\lambda$ windows, the input line would be

```
bonded-lambdas          = 0.00 0.20 0.40 0.60 0.80 1.00
```

To enable the $\lambda$ control, the `.mdp` file parameter **free-energy** must be set to **yes**. The rest of the system setup and the production run follows the standard GROMACS guidelines. As the dRMSD implementation is natively integrated in the replica exchange framework of GROMACS, it can be combined with other techniques requiring modifications to the Hamiltonian e. g. free energy calculations or temperature replica exchange [370].

The dRMSD implementation makes full use of the MPI and OpenMP paradigms in GROMACS. However, the long-range interaction between far distant atom pairs may break the domain decomposition mechanism used in GROMACS to parallelize large systems [395, 396]. Therefore the user can either opt for the particle domain decomposition scheme with the `mdrun` flag `-pd` or decide to use only OpenMP to parallelize the propagation within a single $\lambda$ window while using MPI to manage parallelization on the higher replica exchange level. An error message is shown to the user when the setup violates these requirements.

The output files, the dRMSD results are directed to, are defined with the `-drmsd` flag. A typical run command of `mdrun` using 6 MPI hosts for the $\lambda$ windows with each 16 OpenMP threads to parallelize the internal force calculation and propagation would be:

```
mpirun -n 6 mdrun -multi 6 -ntomp 16 -replex 1000  \
-deffnm protein_ -s protein_.tpr -drmsd drmsd_.xvg
```

The dRMSD output is directed to the files (`drmsd_0.xvg, drmsd_1.xvg, ...`) and stored in the standard ASCII based GROMACS `.xvg` file format. The current $\lambda$ value, the reference dRMSD values and the force constants are stored in the file header. The simulation data is written in three columns, saving the simulation time stamp $t$, the dRMSD value $\mathcal{D}(t)$, and the associated dRMSD potential $V(\mathcal{D}(t))$. In order to calculate a potential of mean force along the dRMSD coordinate, these files can be conveniently analyzed with the Weighted Histogram Analysis method (WHAM) [341].

# Appendix C

# Symbols and Abbreviations

### Symbols and constants

| | |
|---|---|
| $S$ | Entropy |
| $F$ | Helmholtz free energy |
| $G$ | Gibbs free energy |
| $T$ | Temperature |
| $p$ | Pressure |
| $V$ | Volume |
| $k_\mathrm{B}$ | Boltzmann constant |
| $\beta$ | $1/(k_\mathrm{B}T)$ |

### Constants

| | | |
|---|---|---|
| $\epsilon_0$ | Vacuum Permittivity | $8.8542 \times 10^{-12}\,\mathrm{F/m}$ |
| $\mu_0$ | Vacuum Permeability | $4\pi \times 10^{-7}\,\mathrm{N/A^2}$ |
| $e$ | Electron Charge | $1.6022 \times 10^{-19}\,\mathrm{C}$ |
| $k_\mathrm{B}$ | Boltzmann constant | $1.3807 \times 10^{-23}\,\mathrm{J/K}$ |
| $\pi$ | Pi | $16\arctan(\frac{1}{5}) - 4\arctan(\frac{1}{239})$ |

## Abbreviations

| | |
|---|---|
| ABF | Adaptive Biasing Force |
| AMBER | Assisted Model Building with Energy Refinement (Software) [98] |
| APBS | Adaptive Poisson–Boltzmann Solver (Software) [343] |
| BD | Brownian Dynamics |
| cMD | continous Molecular Dynamics |
| dRMSD | distance Root Mean Square Deviation |
| FDPB | Finite Difference Poisson–Boltzmann |
| GDP | Guanosine Diphosphate |
| GROMACS | GROningen Machine for Chemical Simulations (Software) [96, 244] |
| GTP | Guanosine Triphosphate |
| H-REMD | Hamiltonian Replica Exchange Molecular Dynamics |
| H-REUS | Hamiltonian Replica Exchange Umbralla Sampling |
| MC | Monte Carlo |
| MD | Molecular Dynamics |
| NPT | Constant number of particles (N), pressure (P), temperature (T) |
| NVT | Constant number of particles (N), volume (V), temperature (T) |
| PDB | Protein Data Bank |
| PMF | Potential of Mean Force |
| PTM | Post Translational Modification |
| QM | Quantum Mechanics |
| REMD | Replica Exchange Molecular Dynamics |
| RMSD | Root Mean Square Displacement |
| T-REMD | Temperature Replica Exchange Molecular Dynamics |
| US | Umbrella Sampling |
| WE | Weighted Ensemble |
| WHAM | Weighted Histogram Analysis Method |

# List of Figures

# List of Tables

# List of Publications

## Peer-Reviewed Publications

[1] R. Bomblies, **M. P. Luitz**, and M. Zacharias. “Molecular Dynamics Analysis of 4E-BP2 Protein Fold Stabilization Induced by Phosphorylation”. *Journal of Physical Chemistry B* (2016), in press.

[2] R. Bomblies, **M. P. Luitz**, and M. Zacharias. “Mechanism of pKID/KIX Association Studied by Molecular Dynamics Free Energy Simulations”. *Journal of Physical Chemistry B* 120.33 (Apr. 2016), pp. 8186–8192.

[3] **M. P. Luitz**, R. Bomblies, E. Ramcke, A. Itzen, and M. Zacharias. “Adenylylation of Tyr77 stabilizes Rab1b GTPase in an active state: A molecular dynamics simulation analysis”. *Scientific reports* 6 (Jan. 2016), p. 19896.

[4] **M. Luitz**, R. Bomblies, K. Ostermeir, and M. Zacharias. “Exploring biomolecular dynamics and interactions using advanced sampling methods”. *Journal of Physics: Condensed Matter* 27.32 (July 2015), p. 323101.

[5] **M. P. Luitz** and M. Zacharias. “Protein–Ligand Docking Using Hamiltonian Replica Exchange Simulations with Soft Core Potentials”. *Journal of Chemical Information and Modeling* 54.6 (May 2014), pp. 1669–1675.

[6] **M. P. Luitz** and M. Zacharias. “Role of tyrosine hot-spot residues at the interface of colicin E9 and immunity protein 9: A comparative free energy simulation study”. *Proteins: Structure, Function, and Bioinformatics* 81.3 (Mar. 2013), pp. 461–468.

## Book Chapters

[1] C. N. Cavasotto, R. Bomblies, **M. Luitz**, and M. Zacharias. “Free Energy Calculations of Ligand–Protein Binding”. In: *In Silico Drug Discovery and Design: Theory, Methods, Challenges, and Applications.* CRC Press, 2015, pp. 313–335.

## Publications in Preparation

[1] F. Zeller, **M. P. Luitz**, R. Bomblies, and M. Zacharias. “Accurate Multi-scale Simulation of Receptor-drug Association Kinetics: Application to Neuramidase Inhibitors”. *under review at Angewandte Chemie International Edition* (2017).

[2] R. Bomblies, **M. P. Luitz**, S. Scanu, T. Madl, and M. Zacharias. “Transient Helicity in Intrinsically Disordered Axin-1 Depends on Force Field”. *in preparation* (2017).

[3] **M. P. Luitz**, A. Barth, A. H. Crevenna, R. Bomblies, D. Lamb, and M. Zacharias. “Covalent Dye Attachment Influences the Dynamics and Conformational Properties of Flexible Peptides”. *Submitted to Plos One* (2017).

[4] **M. P. Luitz**, R. Bomblies, and M. Zacharias. “From Chaos to Order: The Association Process of RNAse-S Studied by Molecular Dynamics Simulations”. *in preparation* (2017).

## Conferences

[1] F. Zeller, R. Bomblies, **M. P. Luitz**, and M. Zacharias. “Influenza Neuraminidase Inhibitor Binding Studied by Molecular Dynamics Simulations”. *SFB863 Meeting*. Forces in Biomolecular Systems (Nov. 2015). Kreuth, Germany, Conference Talk.

[2] **M. P. Luitz**, R. Bomblies, and M. Zacharias. “From Chaos to Order: The Association Process of RNAse-S studied by MD Simulations”. *SFB1035 Annual Meeting*. Conformational Switches (July 2015). Garching, Germany, Poster Presentation.

[3] R. Bomblies, **M. P. Luitz**, and M. Zacharias. “The Mechanism of pKID-KIX Complex Formation studied by Molecular Dynamics Simulations”. *SFB1035 Annual Meeting*. Conformational Switches (July 2015). Garching, Germany, Poster Presentation.

[4] R. Bomblies, **M. P. Luitz**, and M. Zacharias. “The Mechanism of pKID-KIX Complex Formation studied by Molecular Dynamics Simulations”. *SFB1035 International Conference*. Conformational Switches (May 2015). Venice, Italy, Poster Presentation.

[5] **M. P. Luitz**, R. Bomblies, and M. Zacharias. “From Chaos to Order: The Association Process of RNAse-S studied by MD Simulations”. *SFB1035 International Conference*. Conformational Switches (May 2015). Venice, Italy, Poster Presentation.

[6] **M. P. Luitz**, R. Bomblies, M. Zacharias, E. Bender, and A. Itzen. “Conformational Transitions in Switch Regions of the Ras-Like GTPase Rab1B Studied by Free Energy Simulations”. *Biophysical Society Meeting* (Jan. 2015). Baltimore, USA, Conference Talk.

[7] R. Bomblies, **M. P. Luitz**, and M. Zacharias. “Characterization of Transiently Stable Structural Motifs in Intrinsically Disordered Proteins using Free Energy Simulations”. *Biophysical Society Meeting* (Jan. 2015). Baltimore, USA, Poster Presentation.

[8] **M. P. Luitz**, R. Bomblies, and M. Zacharias. “From Chaos to Order: The Association Pathway of RNAse-S”. *Biophysical Society Thematic Meeting*. Disordered Motifs and Domains in Cell Control (Oct. 2014). Dublin, Ireland, Poster Presentation.

[9] R. Bomblies, **M. P. Luitz**, and M. Zacharias. “The Mechanism of pKID-KIX Complex Formation Studied by Molecular Dynamics Simulations”. *Biophysical Society Thematic Meeting*. Disordered Motifs and Domains in Cell Control (Oct. 2014). Dublin, Ireland, Poster Presentation.

[10] **M. P. Luitz**, R. Bomblies, and M. Zacharias. “In silico investigation of posttranslational Tyr77 modification on Rab1b switching mechanism”. *SFB1035 Annual Meeting*. Conformational Switches (July 2014). Garching, Germany, Poster Presentation.

[11] R. Bomblies, **M. P. Luitz**, and M. Zacharias. “Coupled Association and Structure Formation studied by Molecular Dynamics Simulations”. *SFB1035 Annual Meeting*. Conformational Switches (July 2014). Garching, Germany, Poster Presentation.

[12] R. Bomblies, **M. P. Luitz**, and M. Zacharias. “MD Studies on Intrinsically Disordered Proteins”. *SFB1035 PhD retreat*. Conformational Switches (Apr. 2014). Spitzingsee, Germany, Conference Talk.

[13] **M. P. Luitz**, R. Bomblies, and M. Zacharias. “Coupled Folding and Association: RNAse-S and pKID-KIX studied with MD Simulations”. *Faltertage* (Oct. 2013). Regensburg, Germany, Poster Presentation.

[14] **M. P. Luitz**, R. Bomblies, and M. Zacharias. “Coupled Folding and Association: RNAse-S and pKID-KIX studied with MD Simulations”. *SFB1035 Annual Meeting*. Conformational Switches (July 2013). Garching, Germany, Poster Presentation.

[15] **M. P. Luitz**, R. Bomblies, and M. Zacharias. “Association of Intrinsically Disordered Proteins: A Case Study of RNAse-S”. *SFB1035 PhD retreat*. Conformational Switches (Mar. 2013). Aschau, Germany, Conference Talk.

# Bibliography

[1] E. A. Bell, P. Boehnke, T. M. Harrison, and W. L. Mao. “Potentially Biogenic Carbon Preserved in a 4.1 Billion-Year-Old Zircon”. *Proceedings of the National Academy of Sciences* 112.47 (Oct. 2015), pp. 14518–14521.

[2] L. Orgel. “Evolution of the Genetic Apparatus”. *Journal of Molecular Biology* 38.3 (Dec. 1968), pp. 381–393.

[3] F. Crick. “The Origin of the Genetic Code”. *Journal of Molecular Biology* 38.3 (Dec. 1968), pp. 367–379.

[4] L. E. Orgel. “Some Consequences of the RNA World Hypothesis”. *Origins of life and evolution of the biosphere* 33.2 (2003), pp. 211–218.

[5] S. L. Miller. “Production of Some Organic Compounds under Possible Primitive Earth Conditions1”. *Journal of the American Chemical Society* 77.9 (1955), pp. 2351–2361.

[6] C. Ponnamperuma and N. W. Gabel. “Current Status of Chemical Studies on the Origin of Life”. *Space life sciences* 1.1 (1968), pp. 64–96.

[7] Y. Wolman, W. J. Haverland, and S. L. Miller. “Nonprotein Amino Acids From Spark Discharges and Their Comparison With the Murchison Meteorite Amino Acids”. *Proceedings of the National Academy of Sciences* 69.4 (1972), pp. 809–811.

[8] J. L. Bada. “New insights into prebiotic chemistry from Stanley Miller’s Spark Discharge Experiments”. *Chemical Society Reviews* 42 (5 2013), pp. 2186–2196.

[9] M. W. Powner, B. Gerland, and J. D. Sutherland. “Synthesis of Activated Pyrimidine Ribonucleotides in Prebiotically Plausible Conditions”. *Nature* 459.7244 (May 2009), pp. 239–242.

[10] K. Kruger, P. J. Grabowski, A. J. Zaug, J. Sands, D. E. Gottschling, and T. R. Cech. “Self-Splicing RNA: Autoexcision and Autocyclization of the Ribosomal RNA Intervening Sequence of Tetrahymena”. *Cell* 31.1 (1982), pp. 147–157.

[11] S. Altman. “A View of RNase P”. *Molecular Biosystems* 3 (9 2007), pp. 604–607.

[12] P. G. Higgs. “RNA Secondary Structure: Physical and Computational Aspects”. *Quarterly Reviews of Biophysics* 33.3 (Aug. 1, 2000), pp. 199–253.

[13] V. Ramakrishnan, B. T. Wimberly, D. E. Brodersen, W. M. Clemons, R. J. Morgan-Warren, A. P. Carter, C. Vonrhein, and T. Hartsch. *Nature* 407.6802 (Sept. 2000), pp. 327–339.

[14] W. K. Johnston, P. J. Unrau, M. S. Lawrence, M. E. Glasner, and D. P. Bartel. “RNA-Catalyzed RNA Polymerization: Accurate and General RNA-Templated Primer Extension”. *Science* 292.5520 (2001), pp. 1319–1325.

[15] I. A. Chen. “GE Prize-Winning Essay: The Emergence of Cells During the Origin of Life”. *Science* 314.5805 (Dec. 2006), pp. 1558–1559.

[16] P. Forterre. “The two ages of the {RNA} world, and the transition to the {DNA} World: A Story of Viruses and Cells”. *Biochimie* 87.9–10 (2005). Facets of the {RNA} world, pp. 793–803.

[17] I. Wagner and H. Musso. “New Naturally Occurring Amino Acids”. *Angewandte Chemie International Edition in English* 22.11 (Nov. 1983), pp. 816–828.

[18] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips. “A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis”. *Nature* 181.4610 (Mar. 1958), pp. 662–666.

[19] C. B. Anfinsen. “Principles that Govern the Folding of Protein Chains”. *Science* 181.4096 (1973), pp. 223–230.

[20] F. Chiti and C. M. Dobson. “Protein Misfolding, Functional Amyloid, and Human Disease”. *Annu. Rev. Biochem.* 75.1 (June 2006), pp. 333–366.

[21] M. Bucciantini, E. Giannoni, F. Chiti, F. Baroni, L. Formigli, J. Zurdo, N. Taddei, G. Ramponi, C. M. Dobson, and M. Stefani. “Inherent Toxicity of Aggregates Implies a Common Mechanism for Protein Misfolding Diseases”. *Nature* 416.6880 (Apr. 2002), pp. 507–511.

[22] J. Lipfert and S. Doniach. “Small-Angle X-Ray Scattering from RNA, Proteins, and Protein Complexes”. *Annual Review of Biophysics and Biomolecular Structure* 36.1 (June 2007), pp. 307–327.

[23] L. M. Jackman and S. Sternhell. *Application of Nuclear Magnetic Resonance Spectroscopy in Organic Chemistry: International Series in Organic Chemistry*. Elsevier, 2013.

[24] V. V. Volkov and D. I. Svergun. “Uniqueness of ab Initio Shape Determination in Small-Angle Scattering”. *Journal of Applied Crystallography* 36.3 (Apr. 2003), pp. 860–864.

[25] P. Bernadó, E. Mylonas, M. V. Petoukhov, M. Blackledge, and D. I. Svergun. “Structural Characterization of Flexible Proteins Using Small-Angle X-ray Scattering”. *Journal of the American Chemical Society* 129.17 (2007). PMID: 17411046, pp. 5656–5664.

[26] S. J. Opella and F. M. Marassi. “Structure Determination of Membrane Proteins by NMR Spectroscopy”. *Chemical Reviews* 104.8 (Aug. 2004), pp. 3587–3606.

[27] J. Frank. “Single-Particle Imaging of Macromolecules by Cryo-Electron Microscopy”. *Annual Review of Biophysics and Biomolecular Structure* 31.1 (June 2002), pp. 303–319.

[28] P. Unwin and R. Henderson. “Molecular Structure Determination by Electron Microscopy of Unstained Crystalline Specimens”. *Journal of Molecular Biology* 94.3 (May 1975), pp. 425–440.

[29] E. H. Lee, J. Hsin, M. Sotomayor, G. Comellas, and K. Schulten. “Discovery Through the Computational Microscope”. *Structure* 17.10 (Oct. 2009), pp. 1295–1306.

[30] R. O. Dror, R. M. Dirks, J. Grossman, H. Xu, and D. E. Shaw. “Biomolecular Simulation: A Computational Microscope for Molecular Biology”. *Annu. Rev. Biophys.* 41.1 (June 2012), pp. 429–452.

[31] H. M. Senn and W. Thiel. “QM/MM Methods for Biomolecular Systems”. *Angewandte Chemie International Edition* 48.7 (2009), pp. 1198–1229.

[32] M. Born and R. Oppenheimer. “Zur Quantentheorie der Molekeln”. *Annalen der Physik* 389.20 (1927), pp. 457–484.

[33] L. Pauling. “The Nature of the Chemical Bond. IV. The Energy of Single Bonds and the Relative Electronegativity of Atoms”. *Journal of the American Chemical Society* 54.9 (1932), pp. 3570–3582.

[34] C. M. Baker. “Polarizable Force Fields for Molecular Dynamics Simulations of Biomolecules”. *Wiley Interdisciplinary Reviews: Computational Molecular Science* 5.2 (2015), pp. 241–254.

[35] D. Frenkel and B. Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Vol. 1. Elsevier (formerly published by Academic Press), 2002, pp. 1–638. ISBN: 0-12-267351-4.

[36] S. L. Mayo, B. D. Olafson, and W. A. Goddard. “DREIDING: A Generic Force Field for Molecular Simulations”. *The Journal of Physical Chemistry* 94.26 (1990), pp. 8897–8909.

[37] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. “A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules”. *Journal of the American Chemical Society* 117.19 (1995), pp. 5179–5197.

[38] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. “Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids”. *Journal of the American Chemical Society* 118.45 (1996), pp. 11225–11236.

[39] C. Oostenbrink, A. Villa, A. E. Mark, and W. F. Van Gunsteren. “A Biomolecular Force Field Based on the Free Enthalpy of Hydration and Solvation: The GROMOS Force-Field Parameter Sets 53A5 and 53A6”. *Journal of Computational Chemistry* 25.13 (2004), pp. 1656–1676.

[40] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case. “Development and Testing of a General Amber Force Field”. *Journal of Computational Chemistry* 25.9 (2004), pp. 1157–1174.

[41] A. Pérez, I. Marchán, D. Svozil, J. Sponer, T. E. C. III, C. A. Laughton, and M. Orozco. “Refinement of the {AMBER} Force Field for Nucleic Acids: Improving the Description of α/γ Conformers”. *Biophysical Journal* 92.11 (2007), pp. 3817–3829.

[42] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw. “Improved Side-Chain Torsion Potentials for the Amber Ff99SB Protein Force Field”. *Proteins: Structure, Function, and Bioinformatics* 78.8 (2010), pp. 1950–1958.

[43] L. Verlet. “Computer ”Experiments” on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules”. *Physical Review* 159.1 (1967), pp. 98–103.

[44] H. C. Andersen. “Molecular Dynamics Simulations at Constant Pressure and/or Temperature”. *The Journal of Chemical Physics* 72.4 (1980), p. 2384.

[45] 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”. *The Journal of Chemical Physics* 81 (1984), pp. 3684–3690.

[46] S. Nosé. "A Unified Formulation of the Constant Temperature Molecular Dynamics Methods". *The Journal of Chemical Physics* 81.1 (1984), pp. 511–519.

[47] G. Bussi, D. Donadio, and M. Parrinello. "Canonical Sampling Through Velocity Rescaling." *The Journal of Chemical Physics* 126.1 (Jan. 2007), pp. 014101–7.

[48] M. Parrinello and A. Rahman. "Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method." *Journal of Applied Physics* 52 (1981), pp. 7182–7190.

[49] **M. Luitz**, R. Bomblies, K. Ostermeir, and M. Zacharias. "Exploring biomolecular dynamics and interactions using advanced sampling methods". *Journal of Physics: Condensed Matter* 27.32 (July 2015), p. 323101.

[50] M. Karplus and J. A. McCammon. "Molecular Dynamics Simulations of Biomolecules". *Nature Structural Biology* 9.9 (Sept. 2002), pp. 646–652.

[51] S. Riniker, J. R. Allison, and W. F. van Gunsteren. "On Developing Coarse-Grained Models for Biomolecular Simulation: A Review". *Physical Chemistry Chemical Physics* 14.36 (2012), p. 12423.

[52] M. G. Saunders and G. A. Voth. "Coarse-Graining Methods for Computational Biology". *Annual Review of Biophysics* 42.1 (May 2013), pp. 73–93.

[53] S. Keskin. "Gas Adsorption and Diffusion in a Highly CO2 Selective Metal–Organic Framework: Molecular Simulations". *Molecular Simulation* 39.1 (Jan. 2013), pp. 14–24.

[54] K. Tai. "Conformational Sampling for the Impatient". *Biophysical Chemistry* 107.3 (Feb. 2004), pp. 213–220.

[55] K. Ostermeir and M. Zacharias. "Advanced Replica-Exchange Sampling to Study the Flexibility and Plasticity of Peptides and Proteins". *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1834.5 (May 2013), pp. 847–853.

[56] C. Abrams and G. Bussi. "Enhanced Sampling in Molecular Dynamics Using Metadynamics, Replica-Exchange, and Temperature-Acceleration". *Entropy* 16.1 (Dec. 2013), pp. 163–199.

[57] F. Fogolari, A. Brigo, and H. Molinari. "The Poisson-Boltzmann Equation for Biomolecular Electrostatics: A Tool for Structural Biology". *Journal of Molecular Recognition* 15.6 (2002), pp. 377–392.

[58] D. Bashford and D. A. Case. "Generalized Born Models of Macromolecular Solvation Effects". *Annual Review of Physical Chemistry* 51.1 (Oct. 2000), pp. 129–152.

[59] X. Daura, B. Jaun, D. Seebach, W. F. van Gunsteren, and A. E. Mark. "Reversible Peptide Folding in Solution by Molecular Dynamics Simulation". *Journal of Molecular Biology* 280.5 (July 1998), pp. 925–932.

[60] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw. "How Fast-Folding Proteins Fold". *Science* 334.6055 (Oct. 2011), pp. 517–520.

[61] S. Piana, K. Lindorff-Larsen, and D. E. Shaw. "Protein Folding Kinetics and Thermodynamics From Atomistic Simulation". *Proceedings of the National Academy of Sciences* 109.44 (July 2012), pp. 17845–17850.

[62] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. “Optimization by Simulated Annealing”. *Science* 220.4598 (May 1983), pp. 671–680.

[63] P. J. M. van Laarhoven and E. H. L. Aarts. “Simulated Annealing”. *Simulated Annealing: Theory and Applications* (1987), pp. 7–15.

[64] A. T. Brunger, P. D. Adams, and L. M. Rice. “Annealing in Crystallography: A Powerful Optimization Tool”. *Progress in Biophysics and Molecular Biology* 72.2 (Aug. 1999), pp. 135–155.

[65] J. Kostrowicki and H. A. Scheraga. “Application of the Diffusion Equation Method for Global Optimization to Oligopeptides”. *Journal of Physical Chemistry* 96.18 (Sept. 1992), pp. 7442–7449.

[66] T. Huber, A. E. Torda, and W. F. van Gunsteren. “Structure Optimization Combining Soft-Core Interaction Functions, The Diffusion Equation Method, and Molecular Dynamics”. *The Journal of Physical Chemistry A* 101.33 (Aug. 1997), pp. 5926–5930.

[67] K. Tappura, M. Lahtela–Kakkonen, and O. Teleman. “A New Soft-Core Potential Function for Molecular Dynamics Applied to the Prediction of Protein Loop Conformations”. *Journal of Computational Chemistry* 21.5 (2000), pp. 388–397.

[68] R. Riemann and M. Zacharias. “Reversible Scaling of Dihedral Angle Barriers During Molecular Dynamics to Improve Structure Prediction of Cyclic Peptides”. *Journal of Peptide Research* 63.4 (Apr. 2004), pp. 354–364.

[69] R. N. Riemann and M. Zacharias. “Refinement of Protein Cores and Protein-Peptide Interfaces Using a Potential Scaling Approach”. *Protein Engineering Design and Selection* 18.10 (Aug. 2005), pp. 465–476.

[70] O. F. Lange, L. V. Schäfer, and H. Grubmüller. “Flooding in GROMACS: Accelerated Barrier Crossings in Molecular Dynamics”. *Journal of Computational Chemistry* 27.14 (2006), pp. 1693–1702.

[71] A. Laio and M. Parrinello. “Escaping Free-Energy Minima”. *Proceedings of the National Academy of Sciences* 99.20 (Sept. 2002), pp. 12562–12566.

[72] C. Simmerling, J. L. Miller, and P. A. Kollman. “Combined Locally Enhanced Sampling and Particle Mesh Ewald as a Strategy To Locate the Experimental Structure of a Nonhelical Nucleic Acid”. *Journal of the American Chemical Society* 120.29 (July 1998), pp. 7149–7155.

[73] H.-L. Liu and J.-P. Hsu. “Recent Developments in Structural Proteomics for Protein Structure Determination”. *Proteomics* 5.8 (May 2005), pp. 2056–2068.

[74] J. L. S. Milne, M. J. Borgnia, A. Bartesaghi, E. E. H. Tran, L. A. Earl, D. M. Schauder, J. Lengyel, J. Pierson, A. Patwardhan, and S. Subramaniam. “Cryo-Electron Microscopy - A Primer for the Non-Microscopist”. *FEBS Journal* 280.1 (Dec. 2012), pp. 28–45.

[75] C. E. Blanchet and D. I. Svergun. “Small-Angle X-Ray Scattering on Biological Macromolecules and Nanocomposites in Solution”. *Annual Review of Physical Chemistry* 64.1 (Apr. 2013), pp. 37–54.

[76] R. Roy, S. Hohng, and T. Ha. “A Practical Guide to Single-Molecule FRET”. *Nature Methods* 5.6 (June 2008), pp. 507–516.

[77] E. Sisamakis, A. Valeri, S. Kalinin, P. J. Rothwell, and C. A. Seidel. "Accurate Single-Molecule FRET Studies Using Multiparameter Fluorescence Detection". *Single Molecule Tools, Part B:Super-Resolution, Particle Tracking, Multiparameter, and Force Based Methods* (2010), pp. 455–514.

[78] B. John. "Comparative Protein Structure Modeling by Iterative Alignment, Model Building and Model Assessment". *Nucleic Acids Research* 31.14 (July 2003), pp. 3982–3992.

[79] A. Fiser. "Template-Based Protein Structure Modeling". *Computational Biology* (2010), pp. 73–94.

[80] L. Maragliano and E. Vanden-Eijnden. "A Temperature Accelerated Method for Sampling Free Energy and Determining Reaction Pathways in Rare Events Simulations". *Chemical Physics Letters* 426.1-3 (July 2006), pp. 168–175.

[81] E. Marinari and G. Parisi. "Simulated Tempering: A New Monte Carlo Scheme". *Europhysics Letters (EPL)* 19.6 (July 1992), pp. 451–458.

[82] S. Park and V. S. Pande. "Choosing Weights for Simulated Tempering". *Physical Review E* 76 (1 July 2007), p. 016703.

[83] P. H. Nguyen, Y. Okamoto, and P. Derreumaux. "Communication: Simulated Tempering With Fast on-the-fly Weight Determination". *Journal of Chemical Physics* 138.6 (2013), p. 061102.

[84] D. Hamelberg, J. Mongan, and J. A. McCammon. "Accelerated Molecular Dynamics: A Promising and Efficient Simulation Method for Biomolecules". *Journal of Chemical Physics* 120.24 (2004), p. 11919.

[85] G. Torrie and J. Valleau. "Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling". *Journal of Computational Physics* 23.2 (Feb. 1977), pp. 187–199.

[86] J. Kästner. "Umbrella Sampling". *WIREs Computational Molecular Science* 1.6 (May 2011), pp. 932–942.

[87] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman. "The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method". *Journal of Computational Chemistry* 13.8 (Oct. 1992), pp. 1011–1021.

[88] H. Grubmüller. "Predicting Slow Structural Transitions in Macromolecular Systems: Conformational Flooding". *Physical Review E* 52.3 (Sept. 1995), pp. 2893–2906.

[89] T. Huber, A. E. Torda, and W. F. van Gunsteren. "Local Elevation: A Method for Improving the Searching Properties of Molecular Dynamics Simulation". *J Computer-Aided Mol Des* 8.6 (Dec. 1994), pp. 695–708.

[90] A. Laio, A. Rodriguez-Fortea, F. L. Gervasio, M. Ceccarelli, and M. Parrinello. "Assessing the Accuracy of Metadynamics". *Journal of Physical Chemistry B* 109.14 (Apr. 2005), pp. 6714–6721.

[91] G. Bussi, A. Laio, and M. Parrinello. "Equilibrium Free Energies from Nonequilibrium Metadynamics". *Physical Review Letters* 96.9 (Mar. 2006).

[92] A. Barducci, G. Bussi, and M. Parrinello. "Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method". *Physical Review Letters* 100.2 (Jan. 2008).

[93] P. Raiteri, A. Laio, F. L. Gervasio, C. Micheletti, and M. Parrinello. “Efficient Reconstruction of Complex Free Energy Landscapes by Multiple Walkers Metadynamics”. *Journal of Physical Chemistry B* 110.8 (Mar. 2006), pp. 3533–3539.

[94] E. Darve, D. Rodríguez-Gómez, and A. Pohorille. “Adaptive Biasing Force Method for Scalar and Vector Free Energy Calculations”. *Journal of Chemical Physics* 128.14 (2008), p. 144120.

[95] M. Bonomi, D. Branduardi, G. Bussi, C. Camilloni, D. Provasi, P. Raiteri, D. Donadio, F. Marinelli, F. Pietrucci, R. A. Broglia, and et al. “PLUMED: A Portable Plugin for Free-Energy Calculations With Molecular Dynamics”. *Computer Physics Communications* 180.10 (Oct. 2009), pp. 1961–1972.

[96] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl. “GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation”. *Journal of Chemical Theory and Computation* 4.3 (Mar. 2008), pp. 435–447.

[97] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. “CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations”. *Journal of Computational Chemistry* 4.2 (1983), pp. 187–217.

[98] D. Case et al. *Amber 13*. University of California, San Francisco, 2012.

[99] G. Fiorin, M. L. Klein, and J. Hénin. “Using Collective Variables to Drive Molecular Dynamics Simulations”. *Molecular Physics* 111.22-23 (Dec. 2013), pp. 3345–3362.

[100] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten. “Scalable Molecular Dynamics With NAMD”. *Journal of Computational Chemistry* 26.16 (2005), pp. 1781–1802.

[101] S. Plimpton. “Fast Parallel Algorithms for Short-Range Molecular Dynamics”. *Journal of Computational Physics* 117.1 (Mar. 1995), pp. 1–19.

[102] R. H. Swendsen and J.-S. Wang. “Replica Monte Carlo Simulation of Spin-Glasses”. *Physical Review Letters* 57.21 (Nov. 1986), pp. 2607–2609.

[103] T. Okabe, M. Kawata, Y. Okamoto, and M. Mikami. “Replica-Exchange Monte Carlo Method for the Isobaric–Isothermal Ensemble”. *Chemical Physics Letters* 335.5-6 (Mar. 2001), pp. 435–439.

[104] U. H. Hansmann. “Parallel Tempering Algorithm for Conformational Studies of Biological Molecules”. *Chemical Physics Letters* 281.1-3 (Dec. 1997), pp. 140–150.

[105] Y. Sugita and Y. Okamoto. “Replica-Exchange Molecular Dynamics Method for Protein Folding”. *Chemical Physics Letters* 314.1-2 (Nov. 1999), pp. 141–151.

[106] Y. Okamoto. “Generalized-Ensemble Algorithms: Enhanced Sampling Techniques for Monte Carlo and Molecular Dynamics Simulations”. *Journal of Molecular Graphics and Modelling* 22.5 (May 2004), pp. 425–439.

[107] C. Predescu, M. Predescu, and C. V. Ciobanu. “On the Efficiency of Exchange in Parallel Tempering Monte Carlo Simulations”. *The Journal of Physical Chemistry B* 109.9 (2005). PMID: 16851481, pp. 4189–4196.

[108] D. M. Zuckerman and E. Lyman. “A Second Look at Canonical Sampling of Biomolecules Using Replica Exchange Simulation”. *Journal of Chemical Theory and Computation* 2.4 (July 2006), pp. 1200–1202.

[109] J. Machta. “Strengths and Weaknesses of Parallel Tempering”. *Physical Review E* 80.5 (Nov. 2009).

[110] H. Nymeyer. “How Efficient Is Replica Exchange Molecular Dynamics? An Analytic Approach”. *Journal of Chemical Theory and Computation* 4.4 (Apr. 2008), pp. 626–636.

[111] R. Denschlag, M. Lingenheil, and P. Tavan. “Efficiency Reduction and Pseudo-Convergence in Replica Exchange Sampling of Peptide Folding–Unfolding Equilibria”. *Chemical Physics Letters* 458.1-3 (June 2008), pp. 244–248.

[112] W. Nadler, J. H. Meinke, and U. H. E. Hansmann. “Folding Proteins by First-Passage-Times-Optimized Replica Exchange”. *Physical Review E* 78.6 (Dec. 2008).

[113] D. Gront and A. Kolinski. “Efficient Scheme for Optimization of Parallel Tempering Monte Carlo Method”. *Journal of Physics: Condensed Matter* 19.3 (Jan. 2007), p. 036225.

[114] M. J. Abraham and J. E. Gready. “Ensuring Mixing Efficiency of Replica-Exchange Molecular Dynamics Simulations”. *Journal of Chemical Theory and Computation* 4.7 (July 2008), pp. 1119–1128.

[115] S. Trebst, M. Troyer, and U. H. E. Hansmann. “Optimized Parallel Tempering Simulations of Proteins”. *Journal of Chemical Physics* 124.17 (2006), p. 174903.

[116] W. Nadler and U. H. E. Hansmann. “Optimized Explicit-Solvent Replica Exchange Molecular Dynamics from Scratch”. *Journal of Physical Chemistry B* 112.34 (Aug. 2008), pp. 10386–10387.

[117] F. Calvo. “All-Exchanges Parallel Tempering”. *Journal of Chemical Physics* 123.12 (2005), p. 124106.

[118] P. Brenner, C. R. Sweet, D. VonHandorf, and J. A. Izaguirre. “Accelerating the Replica Exchange Method Through an Efficient All-Pairs Exchange”. *Journal of Chemical Physics* 126.7 (2007), p. 074103.

[119] G. F. Signorini, E. Giovannelli, Y. G. Spill, M. Nilges, and R. Chelli. “Convective Replica-Exchange in Ergodic Regimes”. *Journal of Chemical Theory and Computation* 10.3 (Mar. 2014), pp. 953–958.

[120] Y. G. Spill, G. Bouvier, and M. Nilges. “A Convective Replica-Exchange Method for Sampling New Energy Basins”. *Journal of Computational Chemistry* 34.2 (Sept. 2012), pp. 132–140.

[121] J. D. Chodera and M. R. Shirts. “Replica Exchange and Expanded Ensemble Simulations as Gibbs Sampling: Simple Improvements for Enhanced Mixing”. *Journal of Chemical Physics* 135.19 (2011), p. 194110.

[122] D. Sindhikara, Y. Meng, and A. E. Roitberg. “Exchange Frequency in Replica Exchange Molecular Dynamics”. *Journal of Chemical Physics* 128.2 (2008), p. 024103.

[123] D. J. Sindhikara, D. J. Emerson, and A. E. Roitberg. “Exchange Often and Properly in Replica Exchange Molecular Dynamics”. *Journal of Chemical Theory and Computation* 6.9 (Sept. 2010), pp. 2804–2808.

[124] N.-V. Buchete and G. Hummer. “Peptide Folding Kinetics From Replica Exchange Molecular Dynamics”. *Physical Review E* 77.3 (Mar. 2008).

[125] X. Periole and A. E. Mark. “Convergence and Sampling Efficiency in Replica Exchange Simulations of Peptide Folding in Explicit Solvent”. *Journal of Chemical Physics* 126.1 (2007), p. 014903.

[126] E. Rosta and G. Hummer. “Error and Efficiency of Simulated Tempering Simulations”. *Journal of Chemical Physics* 132.3 (2010), p. 034102.

[127] E. Rosta, N.-V. Buchete, and G. Hummer. “Thermostat Artifacts in Replica Exchange Molecular Dynamics Simulations”. *Journal of Chemical Theory and Computation* 5.5 (May 2009), pp. 1393–1399.

[128] M. Lingenheil, R. Denschlag, G. Mathias, and P. Tavan. “Efficiency of Exchange Schemes in Replica Exchange”. *Chemical Physics Letters* 478.1-3 (Aug. 2009), pp. 80–84.

[129] A. Onufriev, D. Bashford, and D. A. Case. “Exploring Protein Native States and Large-Scale Conformational Changes With a Modified Generalized Born Model”. *Proteins: Structure, Function, and Bioinformatics* 55.2 (Mar. 2004), pp. 383–394.

[130] Y. Chebaro, X. Dong, R. Laghaei, P. Derreumaux, and N. Mousseau. “Replica Exchange Molecular Dynamics Simulations of Coarse-Grained Proteins in Implicit Solvent”. *Journal of Physical Chemistry B* 113.1 (Jan. 2009), pp. 267–274.

[131] H. Nymeyer and A. E. Garcia. “Simulation of the Folding Equilibrium of -Helical Peptides: A Comparison of the Generalized Born Approximation With Explicit Solvent”. *Proceedings of the National Academy of Sciences* 100.24 (Nov. 2003), pp. 13934–13939.

[132] R. Zhou. “Free Energy Landscape of Protein Folding in Water: Explicit vs. Implicit Solvent”. *Proteins* 53.2 (Sept. 2003), pp. 148–161.

[133] W. Xu, T. Lai, Y. Yang, and Y. Mu. “Reversible Folding Simulation by Hybrid Hamiltonian Replica Exchange”. *Journal of Chemical Physics* 128.17 (2008), p. 175105.

[134] X. Cheng, G. Cui, V. Hornak, and C. Simmerling. “Modified Replica Exchange Simulation Methods for Local Structure Refinement”. *Journal of Physical Chemistry B* 109.16 (Apr. 2005), pp. 8220–8230.

[135] M. B. Kubitzki and B. L. de Groot. “Molecular Dynamics Simulations Using Temperature-Enhanced Essential Dynamics Replica Exchange”. *Biophysical Journal* 92.12 (June 2007), pp. 4262–4270.

[136] X. Wu, M. Hodoscek, and B. R. Brooks. “Replica Exchanging Self-Guided Langevin Dynamics for Efficient and Accurate Conformational Sampling”. *Journal of Chemical Physics* 137.4 (2012), p. 044106.

[137] P. Liu, B. Kim, R. A. Friesner, and B. J. Berne. “Replica Exchange With Solute Tempering: A Method for Sampling Biological Systems in Explicit Water”. *Proceedings of the National Academy of Sciences* 102.39 (Sept. 2005), pp. 13749–13754.

[138] A. Okur, D. R. Roe, G. Cui, V. Hornak, and C. Simmerling. “Improving Convergence of Replica-Exchange Simulations through Coupling to a High-Temperature Structure Reservoir”. *Journal of Chemical Theory and Computation* 3.2 (Mar. 2007), pp. 557–568.

[139] J. Z. Ruscio, N. L. Fawzi, and T. Head-Gordon. "How Hot? Systematic Convergence of the Replica Exchange Method Using Multiple Reservoirs". *Journal of Computational Chemistry* (2009), NA–NA.

[140] A. E. Roitberg, A. Okur, and C. Simmerling. "Coupling of Replica Exchange Simulations to a Non-Boltzmann Structure Reservoir". *Journal of Physical Chemistry B* 111.10 (Mar. 2007), pp. 2415–2418.

[141] S. Kannan and M. Zacharias. "Simulated Annealing Coupled Replica Exchange Molecular Dynamics – An Efficient Conformational Sampling Method". *Journal of Structural Biology* 166.3 (June 2009), pp. 288–294.

[142] W. Zheng, M. Andrec, E. Gallicchio, and R. M. Levy. "Simple Continuous and Discrete Models for Simulating Replica Exchange Simulations of Protein Folding". *The Journal of Physical Chemistry B* 112.19 (2008). PMID: 18251533, pp. 6083–6093.

[143] W. Zheng, M. Andrec, E. Gallicchio, and R. M. Levy. "Simulating Replica Exchange Simulations of Protein Folding With a Kinetic Network Model". *Proceedings of the National Academy of Sciences* 104.39 (Sept. 2007), pp. 15340–15345.

[144] H. Fukunishi, O. Watanabe, and S. Takada. "On the Hamiltonian Replica Exchange Method for Efficient Sampling of Biomolecular Systems: Application to Protein Structure Prediction". *Journal of Chemical Physics* 116.20 (2002), p. 9058.

[145] R. Affentranger, I. Tavernelli, and E. E. Di Iorio. "A Novel Hamiltonian Replica Exchange MD Protocol to Enhance Protein Conformational Space Sampling". *Journal of Chemical Theory and Computation* 2.2 (Mar. 2006), pp. 217–228.

[146] J. Hritz and C. Oostenbrink. "Hamiltonian Replica Exchange Molecular Dynamics Using Soft-Core Interactions". *Journal of Chemical Physics* 128.14 (2008), p. 144121.

[147] X. Huang, M. Hagen, B. Kim, R. A. Friesner, R. Zhou, and B. J. Berne. "Replica Exchange with Solute Tempering: Efficiency in Large Scale Systems". *Journal of Physical Chemistry B* 111.19 (May 2007), pp. 5405–5410.

[148] L. Wang, R. A. Friesner, and B. J. Berne. "Replica Exchange with Solute Scaling: A More Efficient Version of Replica Exchange with Solute Tempering (REST2)". *Journal of Physical Chemistry B* 115.30 (Aug. 2011), pp. 9431–9438.

[149] P. Liu, X. Huang, R. Zhou, and B. J. Berne. "Hydrophobic Aided Replica Exchange: an Efficient Algorithm for Protein Folding in Explicit Solvent". *Journal of Physical Chemistry B* 110.38 (Sept. 2006), pp. 19018–19022.

[150] S. L. C. Moors, S. Michielssens, and A. Ceulemans. "Improved Replica Exchange Method for Native-State Protein Sampling". *Journal of Chemical Theory and Computation* 7.1 (Jan. 2011), pp. 231–237.

[151] S. G. Itoh, i. Okumura, and Y. Okamoto. "Replica-Exchange Method in Van Der Waals Radius Space: Overcoming Steric Restrictions for Biomolecules". *Journal of Chemical Physics* 132.13 (2010), p. 134105.

[152] S. Kannan and M. Zacharias. "Enhanced Sampling of Peptide and Protein Conformations Using Replica Exchange Simulations With a Peptide Backbone Biasing-Potential". *Proteins: Structure, Function, and Bioinformatics* 66.3 (Nov. 2006), pp. 697–706.

[153] S. Kannan and M. Zacharias. “Folding Simulations of Trp-Cage Mini Protein in Explicit Solvent Using Biasing Potential Replica-Exchange Molecular Dynamics Simulations”. *Proteins: Structure, Function, and Bioinformatics* 76.2 (Aug. 2009), pp. 448–460.

[154] S. Kannan and M. Zacharias. “Application of Biasing-Potential Replica-Exchange Simulations for Loop Modeling and Refinement of Proteins in Explicit Solvent”. *Proteins: Structure, Function, and Bioinformatics* 78.13 (Aug. 2010), pp. 2809–2819.

[155] Y. Mu. “Dissociation Aided and Side Chain Sampling Enhanced Hamiltonian Replica Exchange”. *Journal of Chemical Physics* 130.16 (2009), p. 164107.

[156] J. Curuksu and M. Zacharias. “Enhanced Conformational Sampling of Nucleic Acids by a New Hamiltonian Replica Exchange Molecular Dynamics Approach”. *Journal of Chemical Physics* 130.10 (2009), p. 104110.

[157] M. Kara and M. Zacharias. “Influence of 8-Oxoguanosine on the Fine Structure of DNA Studied with Biasing-Potential Replica Exchange Simulations”. *Biophysical Journal* 104.5 (Mar. 2013), pp. 1089–1097.

[158] K. Ostermeir and M. Zacharias. “Hamiltonian Replica-Exchange Simulations With Adaptive Biasing of Peptide Backbone and Side Chain Dihedral Angles”. *Journal of Computational Chemistry* 35.2 (Oct. 2013), pp. 150–158.

[159] S. K. Mishra, M. Kara, M. Zacharias, and J. Koca. “Enhanced Conformational Sampling of Carbohydrates by Hamiltonian Replica-Exchange Simulation”. *Glycobiology* 24.1 (Oct. 2013), pp. 70–84.

[160] D. S. Patel, R. Pendrill, S. S. Mallajosyula, G. Widmalm, and A. D. MacKerell. “Conformational Properties of α- or β-(1⟶6)-Linked Oligosaccharides: Hamiltonian Replica Exchange MD Simulations and NMR Experiments”. *Journal of Physical Chemistry B* 118.11 (Mar. 2014), pp. 2851–2871.

[161] E. Lyman and D. M. Zuckerman. “Resolution Exchange Simulation with Incremental Coarsening”. *Journal of Chemical Theory and Computation* 2.3 (May 2006), pp. 656–666.

[162] M. Zacharias. “Combining Elastic Network Analysis and Molecular Dynamics Simulations by Hamiltonian Replica Exchange”. *Journal of Chemical Theory and Computation* 4.3 (Mar. 2008), pp. 477–487.

[163] K. Ostermeir and M. Zacharias. “Hamiltonian Replica Exchange Combined With Elastic Network Analysis to Enhance Global Domain Motions in Atomistic Molecular Dynamics Simulations”. *Proteins: Structure, Function, and Bioinformatics* 82.12 (Oct. 2014), pp. 3410–3419.

[164] S. Piana and A. Laio. “A Bias-Exchange Approach to Protein Folding”. *Journal of Physical Chemistry B* 111.17 (May 2007), pp. 4553–4559.

[165] P. Cossio, F. Marinelli, A. Laio, and F. Pietrucci. “Optimizing the Performance of Bias-Exchange Metadynamics: Folding a 48-Residue LysM Domain Using a Coarse-Grained Model”. *Journal of Physical Chemistry B* 114.9 (Mar. 2010), pp. 3259–3265.

[166] K. Ostermeir and M. Zacharias. “Advanced Replica-Exchange Sampling to Study the Flexibility and Plasticity of Peptides and Proteins”. *Biochimica et Biophysica Acta* 1834.5 (2013), pp. 847–853.

[167] C. J. Woods, J. W. Essex, and M. A. King. “The Development of Replica-Exchange-Based Free-Energy Methods”. *Journal of Physical Chemistry B* 107.49 (Dec. 2003), pp. 13703–13710.

[168] I. V. Khavrutskii and A. Wallqvist. “Improved Binding Free Energy Predictions from Single-Reference Thermodynamic Integration Augmented with Hamiltonian Replica Exchange”. *Journal of Chemical Theory and Computation* 7.9 (Sept. 2011), pp. 3001–3011.

[169] D. K. Shenfeld, H. Xu, M. P. Eastwood, R. O. Dror, and D. E. Shaw. “Minimizing Thermodynamic Length to Select Intermediate States for Free-Energy Calculations and Replica-Exchange Simulations”. *Physical Review E* 80.4 (Oct. 2009).

[170] M. P. Luitz and M. Zacharias. “Protein–Ligand Docking Using Hamiltonian Replica Exchange Simulations with Soft Core Potentials”. *Journal of Chemical Information and Modeling* 54.6 (June 2014), pp. 1669–1675.

[171] Y. Meng, D. Sabri Dashti, and A. E. Roitberg. “Computing Alchemical Free Energy Differences with Hamiltonian Replica Exchange Molecular Dynamics (H-REMD) Simulations”. *Journal of Chemical Theory and Computation* 7.9 (Sept. 2011), pp. 2721–2727.

[172] K. Ostermeir and M. Zacharias. “Rapid Alchemical Free Energy Calculation Employing a Generalized Born Implicit Solvent Model”. *Journal of Physical Chemistry B* 119.3 (Jan. 2015), pp. 968–975.

[173] J. Curuksu, J. Sponer, and M. Zacharias. “Elbow Flexibility of the kt38 RNA Kink-Turn Motif Investigated by Free-Energy Molecular Dynamics Simulations”. *Biophysical Journal* 97.7 (Oct. 2009), pp. 2004–2013.

[174] J. C. Gumbart, B. Roux, and C. Chipot. “Standard Binding Free Energies from Computer Simulations: What Is the Best Strategy?” *Journal of Chemical Theory and Computation* 9.1 (Jan. 2013), pp. 794–802.

[175] J. C. Gumbart, B. Roux, and C. Chipot. “Efficient Determination of Protein–Protein Standard Binding Free Energies from First Principles”. *Journal of Chemical Theory and Computation* 9.8 (Aug. 2013), pp. 3789–3798.

[176] F. Zeller and M. Zacharias. “Adaptive Biasing Combined with Hamiltonian Replica Exchange to Improve Umbrella Sampling Free Energy Simulations”. *Journal of Chemical Theory and Computation* 10.2 (Feb. 2014), pp. 703–710.

[177] F. Zeller and M. Zacharias. “Efficient Calculation of Relative Binding Free Energies by Umbrella Sampling Perturbation”. *Journal of Computational Chemistry* 35.31 (Sept. 2014), pp. 2256–2262.

[178] T. Rodinger, P. L. Howell, and R. Pomès. “Calculation of Absolute Protein-Ligand Binding Free Energy Using Distributed Replica Sampling”. *Journal of Chemical Physics* 129.15 (2008), p. 155102.

[179] W. Jiang, M. Hodoscek, and B. Roux. “Computation of Absolute Hydration and Binding Free Energy with Free Energy Perturbation Distributed Replica-Exchange Molecular Dynamics”. *Journal of Chemical Theory and Computation* 5.10 (Oct. 2009), pp. 2583–2588.

[180] W. Jiang and B. Roux. “Free Energy Perturbation Hamiltonian Replica-Exchange Molecular Dynamics (FEP/H-REMD) for Absolute Ligand Binding Free Energy Calculations”. *Journal of Chemical Theory and Computation* 6.9 (Sept. 2010), pp. 2559–2565.

[181] M. Lapelosa, E. Gallicchio, and R. M. Levy. “Conformational Transitions and Convergence of Absolute Binding Free Energy Calculations”. *Journal of Chemical Theory and Computation* 8.1 (Jan. 2012), pp. 47–60.

[182] Y. Sugita, A. Kitao, and Y. Okamoto. “Multidimensional Replica-Exchange Method for Free-Energy Calculations”. *Journal of Chemical Physics* 113.15 (2000), p. 6042.

[183] M. S. Lee and M. A. Olson. “Comparison of two Adaptive Temperature-Based Replica Exchange Methods Applied to a Sharp Phase Transition of Protein Unfolding-Folding”. *Journal of Chemical Physics* 134.24 (2011), p. 244111.

[184] M. A. Olson and M. S. Lee. “Evaluation of Unrestrained Replica-Exchange Simulations Using Dynamic Walkers in Temperature Space for Protein Structure Refinement”. *PLOS ONE* 9.5 (May 2014). Ed. by Y. Zhang, e96638.

[185] K. Wang, J. D. Chodera, Y. Yang, and M. R. Shirts. “Identifying Ligand Binding Sites and Poses Using GPU-accelerated Hamiltonian Replica Exchange Molecular Dynamics”. *Journal of Computer Aided Molecular Design* 27.12 (Dec. 2013), pp. 989–1007.

[186] A. de Ruiter and C. Oostenbrink. “Protein–Ligand Binding from Distancefield Distances and Hamiltonian Replica Exchange Simulations”. *Journal of Chemical Theory and Computation* 9.2 (Feb. 2013), pp. 883–892.

[187] N. M. Henriksen, D. R. Roe, and T. E. Cheatham. “Reliable Oligonucleotide Conformational Ensemble Generation in Explicit Solvent for Force Field Assessment Using Reservoir Replica Exchange Molecular Dynamics Simulations”. *Journal of Physical Chemistry B* 117.15 (Apr. 2013), pp. 4014–4027.

[188] C. Bergonzo, N. M. Henriksen, D. R. Roe, J. M. Swails, A. E. Roitberg, and T. E. Cheatham. “Multidimensional Replica Exchange Molecular Dynamics Yields a Converged Ensemble of an RNA Tetranucleotide”. *Journal of Chemical Theory and Computation* 10.1 (Jan. 2014), pp. 492–499.

[189] L. Wang, B. J. Berne, and R. A. Friesner. “On Achieving High Accuracy and Reliability in the Calculation of Relative Protein-Ligand Binding Affinities”. *Proceedings of the National Academy of Sciences* 109.6 (Jan. 2012), pp. 1937–1942.

[190] K. Lindorff-Larsen, N. Trbovic, P. Maragakis, S. Piana, and D. E. Shaw. “Structure and Dynamics of an Unfolded Protein Examined by Molecular Dynamics Simulation”. *Journal of the American Chemical Society* 134.8 (Feb. 2012), pp. 3787–3791.

[191] F. Jiang and Y.-D. Wu. “Folding of Fourteen Small Proteins with a Residue-Specific Force Field and Replica-Exchange Molecular Dynamics”. *Journal of the American Chemical Society* 136.27 (July 2014), pp. 9536–9539.

[192] C. K. Fisher and C. M. Stultz. “Constructing Ensembles for Intrinsically Disordered Proteins”. *Current Opinion in Structural Biology* 21.3 (June 2011), pp. 426–431.

[193] M. Cecchini, F. Rao, M. Seeber, and A. Caflisch. “Replica Exchange Molecular Dynamics Simulations of Amyloid Peptide Aggregation”. *Journal of Chemical Physics* 121.21 (2004), p. 10748.

[194] M. Han and U. H. E. Hansmann. “Replica Exchange Molecular Dynamics of the Thermodynamics of Fibril Growth of Alzheimer’s Aβ42 Peptide”. *Journal of Chemical Physics* 135.6 (2011), p. 065101.

[195] P. Anand, F. S. Nandel, and U. H. E. Hansmann. "The Alzheimer's β Amyloid (Aβ[sub 1–39]) Monomer in an Implicit Solvent". *Journal of Chemical Physics* 128.16 (2008), p. 165102.

[196] P. H. Nguyen, M. S. Li, and P. Derreumaux. "Effects of All-Atom Force Fields on Amyloid Oligomerization: Replica Exchange Molecular Dynamics Simulations of the Aβ16–22 Dimer and Trimer". *Physical Chemistry Chemical Physics* 13.20 (2011), p. 9778.

[197] Y. Chebaro, N. Mousseau, and P. Derreumaux. "Structures and Thermodynamics of Alzheimer's Amyloid-β Aβ(16–35) Monomer and Dimer by Replica Exchange Molecular Dynamics Simulations: Implication for Full-Length Aβ Fibrillation". *Journal of Physical Chemistry B* 113.21 (May 2009), pp. 7668–7675.

[198] H.-L. Chiang, C.-J. Chen, H. Okumura, and C.-K. Hu. "Transformation Between α-Helix and β-Sheet Structures of One and Two Polyglutamine Peptides in Explicit Water Molecules by Replica-Exchange Molecular Dynamics Simulations". *Journal of Computational Chemistry* 35.19 (May 2014), pp. 1430–1437.

[199] R. Zwanzig, A. Szabo, and B. Bagchi. "Levinthal's Paradox." *Proceedings of the National Academy of Sciences* 89.1 (Jan. 1992), pp. 20–22.

[200] H. Eyring. "The Activated Complex in Chemical Reactions". *The Journal of Chemical Physics* 3.2 (1935), p. 107.

[201] M. R. Shirts, D. L. Mobley, and J. D. Chodera. "Chapter 4 Alchemical Free Energy Calculations: Ready for Prime Time?" In: ed. by D. Spellmeyer and R. Wheeler. Vol. 3. Annual Reports in Computational Chemistry. Elsevier, 2007, pp. 41–59.

[202] J. Wang, Y. Deng, and B. Roux. "Absolute Binding Free Energy Calculations Using Molecular Dynamics Simulations with Restraining Potentials". *Biophysical Journal* 91.8 (Oct. 2006), pp. 2798–2814.

[203] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman. "The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method". *Journal of Computational Chemistry* 13.8 (Oct. 1992), pp. 1011–1021.

[204] H.-J. Woo and B. Roux. "Calculation of Absolute Protein-Ligand Binding Free Energy From Computer Simulations". *Proceedings of the National Academy of Sciences* 102.19 (May 2005), pp. 6825–6830.

[205] G. A. Huber and S. Kim. "Weighted-Ensemble Brownian Dynamics Simulations for Protein Association Reactions". *Biophysical Journal* 70.1 (Jan. 1996), pp. 97–110.

[206] B. W. Zhang, D. Jasnow, and D. M. Zuckerman. "Efficient and Verified Simulation of a Path Ensemble for Conformational Change in a United-Residue Model of Calmodulin." *Proceedings of the National academy of Sciences* 104.46 (Nov. 2007), pp. 18043–8.

[207] D. Bhatt, B. W. Zhang, and D. M. Zuckerman. "Steady-State Simulations using Weighted Ensemble Path Sampling." *Journal of Chemical Physics* 133.1 (July 2010), p. 014110.

[208] E. Suárez, S. Lettieri, M. C. Zwier, C. A. Stringer, S. R. Subramanian, L. T. Chong, and D. M. Zuckerman. "Simultaneous Computation of Dynamical and Equilibrium Information Using a Weighted Ensemble of Trajectories". *Journal of Chemical Theory and Computation* 10.7 (2014). PMID: 25246856, pp. 2658–2667.

[209] D. Bhatt and D. M. Zuckerman. “Heterogeneous Path Ensembles for Conformational Transitions in Semiatomistic Models of Adenylate Kinase”. *Journal of Chemical Theory and Computation* 6.11 (2010), pp. 3527–3539.

[210] J. L. Adelman, A. L. Dale, M. C. Zwier, D. Bhatt, L. T. Chong, D. M. Zuckerman, and M. Grabe. “Simulations of the Alternating Access Mechanism of the Sodium Symporter Mhp1.” *Biophysical Journal* 101.10 (Nov. 2011), pp. 2399–407.

[211] M. C. Zwier, J. W. Kaus, and L. T. Chong. “Efficient Explicit-Solvent Molecular Dynamics Simulations of Molecular Association Kinetics: Methane/Methane, Na+/Cl-, Methane/Benzene, and K+/18-Crown-6 Ether”. *Journal of Chemical Theory and Computation* 7.4 (2011), pp. 1189–1197.

[212] J. L. Adelman and M. Grabe. “Simulating Rare Events Using a Weighted Ensemble-Based String Method”. *The Journal of Chemical Physics* 138.4 (2013), p. 044105.

[213] M. C. Zwier, J. L. Adelman, J. W. Kaus, A. J. Pratt, K. F. Wong, N. B. Rego, E. Suárez, S. Lettieri, D. W. Wang, M. Grabe, D. M. Zuckerman, and L. T. Chong. “WESTPA: An Interoperable, Highly Scalable Software Package for Weighted Ensemble Simulation and Analysis”. *Journal of Chemical Theory and Computation* 11.2 (2015), pp. 800–809.

[214] A. Dickson and I. Charles L. Brooks. “WExplore: Hierarchical Exploration of High-Dimensional Spaces Using the Weighted Ensemble Algorithm”. *Journal of Physical Chemistry B* 118.13 (2014), pp. 3532–3542.

[215] A. Dickson, A. M. Mustoe, L. Salmon, and C. L. Brooks. “Efficient in silico exploration of RNA interhelical conformations using Euler angles and WExplore”. *Nucleic Acids Research* 42.19 (Oct. 2014), pp. 12126–12137.

[216] R. W. Pastor and M. Karplus. “Inertial Effects in Butane Stochastic Dynamics”. *Journal of Chemical Physics* 91.1 (1989), pp. 211–218.

[217] R. J. Loncharich, B. R. Brooks, and R. W. Pastor. “Langevin dynamics of peptides: The frictional dependence of isomerization rates ofN-acetylalanyl-N?-methylamide”. *Biopolymers* 32.5 (May 1992), pp. 523–535.

[218] Y. Zhang and R. W. Pastor. “A Comparision of Methods for Computing Transition Rates from Molecular Dynamics Simulation”. *Molecular Simulation* 13.1 (1994), pp. 25–38.

[219] J. C. Maxwell. “On the Dynamical Theory of Gases”. *Philosophical Transactions of the Royal Society of London* 157 (Jan. 1867), pp. 49–88.

[220] J. E. Gentle. *Numerical Linear Algebra for Applications in Statistics*. Springer New York, 1998.

[221] B. W. Zhang, D. Jasnow, and D. M. Zuckerman. “The “weighted Ensemble” Path Sampling Method Is Statistically Exact for a Broad Class of Stochastic Processes and Binning Procedures”. *The Journal of Chemical Physics* 132.5, 054107 (2010).

[222] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé. “Markov Models of Molecular Kinetics: Generation and Validation”. *The Journal of Chemical Physics* 134.17, 174105 (2011).

[223] S. Kirmizialtin and R. Elber. “Revisiting and Computing Reaction Coordinates with Directional Milestoning”. *The Journal of Physical Chemistry A* 115.23 (2011). PMID: 21500798, pp. 6137–6148.

[224] **M. P. Luitz** and M. Zacharias. “Protein–Ligand Docking Using Hamiltonian Replica Exchange Simulations with Soft Core Potentials”. *Journal of Chemical Information and Modeling* 54.6 (May 2014), pp. 1669–1675.

[225] H. Park, J. Lee, and S. Lee. “Critical Assessment of the Automated AutoDock as a New Docking Tool for Virtual Screening”. *Proteins: Structure, Function, and Bioinformatics* 65.3 (2006), pp. 549–554.

[226] T. J. A. Ewing, S. Makino, A. G. Skillman, and I. D. Kuntz. “DOCK 4.0: Search Strategies for Automated Molecular Docking of Flexible Molecule Databases”. *Journal of Computer-Aided Molecular Design* 15.5 (2001), pp. 411–428.

[227] M. Totrov and R. Abagyan. “Flexible Ligand Docking to Multiple Receptor Conformations: A Practical Alternative”. *Current Opinion in Structural Biology* 18.2 (2008), pp. 178–184.

[228] M. Zacharias. “ATTRACT: Protein–protein Docking in CAPRI using a Reduced Protein Model”. *Proteins: Structure, Function, and Bioinformatics* 60.2 (2005), pp. 252–256.

[229] G. Kuzu, O. Keskin, A. Gursoy, and R. Nussinov. “Expanding the Conformational Selection Paradigm in Protein-Ligand Docking”. English. In: *Computational Drug Discovery and Design*. Ed. by R. Baron. Vol. 819. Methods in Molecular Biology. Springer New York, 2012, pp. 59–74. ISBN: 978-1-61779-464-3.

[230] A. H. Keeble, L. a. Joachimiak, M. J. Maté, N. Meenan, N. Kirkpatrick, D. Baker, and C. Kleanthous. “Experimental and Computational Analyses of the Energetic Basis for Dual Recognition of Immunity Proteins by Colicin Endonucleases.” *Journal of Molecular Biology* 379.4 (June 2008), pp. 745–759.

[231] S. Samsonov, J. Teyra, G. Anders, and M. T. Pisabarro. “Analysis of the impact of solvent on contacts prediction in proteins”. *BMC Struct. Biol.* 9.1 (2009), pp. 1–11.

[232] M. H. Ahmed, F. Spyrakis, P. Cozzini, P. K. Tripathi, A. Mozzarelli, J. N. Scarsdale, M. A. Safo, and G. E. Kellogg. “Bound Water at Protein-Protein Interfaces: Partners, Roles and Hydrophobic Bubbles as a Conserved Motif”. *PLOS ONE* 6.9 (Sept. 2011), e24712.

[233] R. O. Dror, A. C. Pan, D. H. Arlow, D. W. Borhani, P. Maragakis, Y. Shan, H. Xu, and D. E. Shaw. “Pathway and Mechanism of Drug Binding to G-Protein-Coupled Receptors”. *Proc. Natl. Acad. Sci.* 108.32 (2011), pp. 13118–13123.

[234] Y. Shan, E. T. Kim, M. P. Eastwood, R. O. Dror, M. A. Seeliger, and D. E. Shaw. “How Does a Drug Molecule Find Its Target Binding Site?” *Journal of the American Chemical Society* 133.24 (2011), pp. 9181–9183.

[235] J. Hritz and C. Oostenbrink. “Hamiltonian Replica Exchange Molecular Dynamics Using Soft-Core Interactions”. *Journal of Chemical Physics* 128.14, 144121 (2008), p. 144121.

[236] L. Wang, R. A. Friesner, and B. J. Berne. “Replica Exchange with Solute Scaling: A More Efficient Version of Replica Exchange With Solute Tempering (REST2)”. *Journal of Physical Chemistry B* 115.30 (2011), pp. 9431–9438.

[237] S. G. Itoh, H. Okumura, and Y. Okamoto. “Replica-Exchange Method in van der Waals Radius Space: Overcoming Steric Restrictions for Biomolecules”. *Journal of Chemical Physics* 132.13, 134105 (2010), p. 134105.

[238] S. Kannan and M. Zacharias. “Enhanced Sampling of Peptide and Protein Conformations Using Replica Exchange Simulations With a Peptide Backbone Biasing-Potential”. *Proteins: Structure, Function, and Bioinformatics* 66.3 (2007), pp. 697–706.

[239] J. Curuksu and M. Zacharias. “Enhanced Conformational Sampling of Nucleic Acids by a New Hamiltonian Replica Exchange Molecular Dynamics Approach”. *Journal of Chemical Physics* 130.10, 104110 (2009), p. 104110.

[240] M. Zacharias. “Combining Elastic Network Analysis and Molecular Dynamics Simulations by Hamiltonian Replica Exchange”. *Journal of Chemical Theory and Computation* 4.3 (2008), pp. 477–487.

[241] I. Buch, T. Giorgino, and G. De Fabritiis. “Complete Reconstruction of an Enzyme-Inhibitor Binding Process by Molecular Dynamics Simulations”. *Proc. Natl. Acad. Sci.* 108.25 (2011), pp. 10184–10189.

[242] M. Zacharias, T. P. Straatsma, and J. A. McCammon. “Separation–Shifted Scaling, a New Scaling Method for Lennard–Jones Interactions in Thermodynamic Integration.” *Journal of Chemical Physics* 100.12 (1994), pp. 9025–9031.

[243] T. C. Beutler, A. E. Mark, R. C. van Schaik, P. R. Gerber, and W. F. van Gunsteren. “Avoiding Singularities and Numerical Instabilities in Free Energy Calculations Based on Molecular Simulations.” *Chemical Physics Letters* 222 (1994), pp. 529–539.

[244] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen. “GROMACS: Fast, Flexible, and Free”. *Journal of Computational Chemistry* 26.16 (2005), pp. 1701–1718.

[245] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. “Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters”. *Proteins: Structure, Function, and Bioinformatics* 65.3 (2006), pp. 712–725.

[246] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. “Comparison of Simple Potential Functions for Simulating Liquid Water”. *Journal of Chemical Physics* 79.2 (1983), pp. 926–935.

[247] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case. “Development and Testing of a General Amber Force Field”. *Journal of Computational Chemistry* 25.9 (2004), pp. 1157–1174.

[248] J. Wang, W. Wang, P. A. Kollman, and D. A. Case. “Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations”. *Journal of Molecular Graphics and Modelling* 25.2 (2006), pp. 247–260.

[249] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije. “LINCS: A Linear Constraint Solver for Molecular Simulations”. *Journal of Computational Chemistry* 18.12 (Sept. 1997), pp. 1463–1472.

[250] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen. “A Smooth Particle Mesh Ewald Method.” *Journal of Chemical Physics* 103.19 (1995), pp. 8577–8593.

[251] W. F. van Gunsteren and H. J. C. Berendsen. “A Leap-Frog Algorithm for Stochastic Dynamics”. *Molecular Simulation* 1 (1988), pp. 173–185.

[252] D. A. Holt, J. I. Luengo, D. S. Yamashita, H. J. Oh, A. L. Konialian, H. K. Yen, L. W. Rozamus, M. Brandt, and M. J. a. Bossard. “Design, Synthesis, and Kinetic Evaluation of High-Affinity FKBP Ligands and the X-Ray Crystal Structures of their Complexes with FKBP12”. *Journal of the American Chemical Society* 115.22 (1993), pp. 9925–9938.

[253] D. Fremont, M. Matsumura, E. Stura, P. Peterson, and I. Wilson. “Crystal Structures of Two Viral Peptides in Complex With Murine MHC Class I H-2Kb”. *Science* 257.5072 (1992), pp. 919–927.

[254] C. Dominguez, R. Boelens, and A. M. J. J. Bonvin. “HADDOCK: A Protein-Protein Docking Approach Based on Biochemical or Biophysical Information”. *Journal of the American Chemical Society* 125.7 (2003), pp. 1731–1737.

[255] A. D. J. van Dijk and A. M. J. J. Bonvin. “Solvated Docking: Introducing Water Into the Modelling of Biomolecular Complexes”. *Bioinformatics* 22.19 (2006), pp. 2340–2347.

[256] G. R. Smith, M. J. Sternberg, and P. A. Bates. “The Relationship Between the Flexibility of Proteins and Their Conformational States on Forming Protein–Protein Complexes With an Application to Protein–Protein Docking”. *Journal of Molecular Biology* 347.5 (2005), pp. 1077–1101.

[257] H. Fujitani, Y. Tanida, M. Ito, G. Jayachandran, C. D. Snow, M. R. Shirts, E. J. Sorin, and V. S. Pande. “Direct Calculation of the Binding Free Energies of FKBP Ligands”. *Journal of Chemical Physics* 123.8, 084108 (2005), p. 084108.

[258] M. Zacharias and S. Springer. “Conformational Flexibility of the MHC Class I $\alpha$1-$\alpha$2 Domain in Peptide Bound and Free States: A Molecular Dynamics Simulation Study”. *Biophysical Journal* 87.4 (2004), pp. 2203–2214.

[259] H. Neuweiler and M. Sauer. “Using Photoinduced Charge Transfer Reactions to Study Conformational Dynamics of Biopolymers at the Single-Molecule Level”. *Current pharmaceutical biotechnology* 5.3 (2004), pp. 285–298.

[260] S. Doose, H. Neuweiler, and M. Sauer. “Fluorescence Quenching by Photoinduced Electron Transfer: A Reporter for Conformational Dynamics of Macromolecules”. *ChemPhysChem* 10.9-10 (2009), pp. 1389–1398.

[261] E. A. Jares-Erijman and T. M. Jovin. “FRET Imaging”. *Nature biotechnology* 21.11 (2003), pp. 1387–1395.

[262] S. Weiss. “Fluorescence Spectroscopy of Single Biomolecules”. *Science* 283.5408 (1999), pp. 1676–1683.

[263] T. Ha. “Single-Molecule Fluorescence Resonance Energy Transfer”. *Methods* 25.1 (Sept. 2001), pp. 78–86.

[264] T. Förster. “Zwischenmolekulare Energiewanderung und Fluoreszenz”. *Annalen der Physik* 437.1-2 (1948), pp. 55–75.

[265] G. J. Kavarnos. *Fundamentals of Photoinduced Electron Transfer.* Vol. 98. Wiley-VCH, Oct. 1994.

[266] G. J. Kavarnos and N. J. Turro. “Photosensitization by Reversible Electron Transfer: Theories, Experimental Evidence, and Examples”. *Chemical Reviews* 86.2 (Apr. 1986), pp. 401–449.

[267] A. Weller. “Photoinduced Electron Transfer in Solution: Exciplex and Radical Ion Pair Formation Free Enthalpies and their Solvent Dependence”. *Zeitschrift für Physikalische Chemie* 133.1 (Jan. 1982), pp. 93–98.

[268] A. P. de Silva, H. Q. N. Gunaratne, T. Gunnlaugsson, A. J. M. Huxley, C. P. McCoy, J. T. Rademacher, and T. E. Rice. “Signaling Recognition Events with Fluorescent Sensors and Switches”. *Chemical Reviews* 97.5 (Aug. 1997), pp. 1515–1566.

[269] D. Zhong, S. K. Pal, C. Wan, and A. H. Zewail. “Femtosecond Dynamics of a Drug–Protein Complex: Daunomycin with Apo Riboflavin-Binding Protein”. *Proceedings of the National Academy of Sciences* 98.21 (2001), pp. 11873–11878.

[270] X. Li, R. Zhu, A. Yu, and X. S. Zhao. “Ultrafast Photoinduced Electron Transfer Between Tetramethylrhodamine and Guanosine in Aqueous Solution”. *The Journal of Physical Chemistry B* 115.19 (2011), pp. 6265–6271.

[271] X. Michalet, A. N. Kapanidis, T. Laurence, F. Pinaud, S. Doose, M. Pflughoefft, and S. Weiss. “The Power and Prospects of Fluorescence Microscopies and Spectroscopies”. *Annual review of biophysics and biomolecular structure* 32.1 (2003), pp. 161–182.

[272] P. Tinnefeld and M. Sauer. “Branching Out of Single-Molecule Fluorescence Spectroscopy: Challenges for Chemistry and Influence on Biology”. *Angewandte Chemie International Edition* 44.18 (2005), pp. 2642–2671.

[273] E. L. Elson and D. Magde. “Fluorescence Correlation Spectroscopy. I. Conceptual Basis and Theory”. *Biopolymers* 13.1 (1974), pp. 1–27.

[274] E. L. Elson. “Fluorescence Correlation Spectroscopy: Past, Present, Future”. *Biophysical Journal* 101.12 (Dec. 2011), pp. 2855–2870.

[275] H. Sahoo and P. Schwille. “FRET and FCS-Friends or Foes?” *ChemPhysChem* 12.3 (Feb. 2011), pp. 532–541.

[276] S. Felekyan, S. Kalinin, H. Sanabria, A. Valeri, and C. A. M. Seidel. “Filtered FCS: Species Auto- and Cross-Correlation Functions Highlight Binding and Dynamics in Biomolecules”. *ChemPhysChem* 13.4 (Mar. 2012), pp. 1036–1053.

[277] M. Sauer and H. Neuweiler. “PET-FCS: Probing Rapid Structural Fluctuations of Proteins and Nucleic Acids by Single-Molecule Fluorescence Quenching”. *Fluorescence Spectroscopy and Microscopy: Methods and Protocols* (2014), pp. 597–615.

[278] S. Milles, S. Tyagi, N. Banterle, C. Koehler, V. VanDelinder, T. Plass, A. P. Neal, and E. A. Lemke. “Click Strategies for Single-Molecule Protein Fluorescence”. *Journal of the American Chemical Society* 134.11 (Mar. 2012), pp. 5187–5195.

[279] E. Haustein and P. Schwille. “Fluorescence Correlation Spectroscopy: Novel Variations of an Established Technique”. *Annual Review of Biophysics and Biomolecular Structure* 36 (2007), pp. 151–169.

[280] W. Moerner and D. P. Fromm. “Methods of Single-Molecule Fluorescence Spectroscopy and Microscopy”. *Review of Scientific Instruments* 74.8 (2003), pp. 3597–3619.

[281] I. Daidone, H. Neuweiler, S. Doose, M. Sauer, and J. C. Smith. “Hydrogen-Bond Driven Loop-Closure Kinetics in Unfolded Polypeptide Chains.” *PLoS Computational Biology* 6.1 (Jan. 2010), e1000645.

[282] G. F. Schröder, U. Alexiev, and H. Grubmüller. “Simulation of Fluorescence Anisotropy Experiments: Probing Protein Dynamics.” *Biophysical Journal* 89.6 (Dec. 2005), pp. 3757–3770.

[283] M. Hoefling, N. Lima, D. Haenni, C. A. M. Seidel, B. Schuler, and H. Grubmüller. “Structural Heterogeneity and Quantitative FRET Efficiency Distributions of Polyprolines through a Hybrid Atomistic Simulation and Monte Carlo Approach”. *PLOS ONE* 6.5 (May 2011), e19791.

[284] F. Finn, J. Dadok, and A. Bothner-By. “Proton Nuclear Magnetic Resonance Studies of the Association of Ribonuclease S-Peptide and Analogs with Ribonuclease S-Protein”. *Biochemistry* 11.3 (1972), pp. 455–461.

[285] M. Bastos, J. H. Pease, D. E. Wemmer, K. P. Murphy, and P. R. Connelly. “Thermodynamics of the Helix-Coil Transition: Binding of S15 and a Hybrid Sequence, Disulfide Stabilized Peptide to the S-Protein”. *Proteins: Structure, Function, and Bioinformatics* 42.4 (2001), pp. 523–530.

[286] R. Cole and J. P. Loria. “Evidence for Flexibility in the Function of Ribonuclease A”. *Biochemistry* 41.19 (2002), pp. 6072–6081.

[287] G. R. Marshall, J. A. Feng, and D. J. Kuster. “Back to the Future: Ribonuclease A”. *Peptide Science* 90.3 (2008), pp. 259–277.

[288] A. Bachmann, D. Wildemann, F. Praetorius, G. Fischer, and T. Kiefhaber. “Mapping Backbone and Side-Chain Interactions in the Transition State of a Coupled Protein Folding and Binding Reaction”. *Proceedings of the National Academy of Sciences* 108.10 (2011), pp. 3952–3957.

[289] F. Murtagh and P. Contreras. “Algorithms for Hierarchical Clustering: An Overview”. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2.1 (2012), pp. 86–97.

[290] Ü. Mets, J. Widengren, and R. Rigler. “Application of the Antibunching in Dye Fluorescence: Measuring the Excitation Rates in Solution”. *Chemical Physics* 218.1-2 (May 1997), pp. 191–198.

[291] D. Nettels, I. V. Gopich, A. Hoffmann, and B. Schuler. “Ultrafast Dynamics of Protein Collapse From Single-Molecule Photon Statistics.” *Proceedings of the National Academy of Sciences of the United States of America* 104.8 (Feb. 2007), pp. 2655–2660.

[292] S. Weiss. “Measuring Conformational Dynamics of Biomolecules by Single Molecule Fluorescence Spectroscopy”. *Nature Structural Biology* 7.9 (Sept. 2000), pp. 724–729.

[293] A. C. Vaiana, H. Neuweiler, A. Schulz, J. Wolfrum, M. Sauer, and J. C. Smith. “Fluorescence Quenching of Dyes by Tryptophan: Interactions at Atomic Detail From Combination of Experiment and Computer Simulation”. *Journal of the American Chemical Society* 125.47 (2003), pp. 14564–14572.

[294] F. Noé, S. Doose, I. Daidone, M. Löllmann, M. Sauer, J. D. Chodera, and J. C. Smith. "Dynamical Fingerprints for Probing Individual Relaxation Processes in Biomolecular Dynamics With Simulations and Kinetic Experiments". *Proceedings of the National Academy of Sciences* 108.12 (2011), pp. 4822–4827.

[295] S. Sindbert, S. Kalinin, H. Nguyen, A. Kienzler, L. Clima, W. Bannwarth, B. Appel, S. Müller, and C. A. M. Seidel. "Accurate Distance Determination of Nucleic Acids via Förster Resonance Energy Transfer: Implications of Dye Linker Length and Rigidity". *Journal of the American Chemical Society* 133.8 (Mar. 2011), pp. 2463–2480.

[296] M. J. Frisch et al. *Gaussian 09 Revision D.01*. Gaussian Inc. Wallingford CT 2009.

[297] A. D. Becke. "Density-Functional Exchange-Energy Approximation With Correct Asymptotic Behavior". *Physical Review A* 38 (6 Sept. 1988), pp. 3098–3100.

[298] C. Lee, W. Yang, and R. G. Parr. "Development of the Colle-Salvetti Correlation-Energy Formula Into a Functional of the Electron Density". *Physical Review B* 37 (2 Jan. 1988), pp. 785–789.

[299] P. Hariharan and J. Pople. "The Influence of Polarization Functions on Molecular Orbital Hydrogenation Energies". *Theoretica Chimica Acta* 28.3 (1973), pp. 213–222.

[300] C. I. Bayly, P. Cieplak, W. D. Cornell, and P. A. Kollman. "A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model". *Journal of Physical Chemistry* 97.40 (1993), pp. 10269–10280.

[301] G. Bussi, D. Donadio, and M. Parrinello. "Canonical Sampling Through Velocity Rescaling". *The Journal of chemical physics* 126.1 (2007), p. 014101.

[302] A. H. Crevenna, N. Naredi-Rainer, D. C. Lamb, R. Wedlich-Söldner, and J. Dzubiella. "Effects of Hofmeister Ions on the a-Helical Structure of Proteins". *Biophysical Journal* 102.4 (Feb. 2012), pp. 907–915.

[303] R. H. Brown and R. Twiss. "A Test of a New Type of Stellar Interferometer on Sirius". *Nature* 178.4541 (1956), pp. 1046–1048.

[304] C. Kurtsiefer, P. Zarda, S. Mayer, and H. Weinfurter. "The Breakdown Flash of Silicon Avalanche Photodiodes-Back Door for Eavesdropper Attacks?" *Journal of Modern Optics* 48.13 (Nov. 2001), pp. 2039–2047.

[305] S. Felekyan, R. Kühnemuth, V. Kudryavtsev, C. Sandhagen, W. Becker, and C. A. M. Seidel. "Full Correlation From Picoseconds to Seconds by Time-Resolved and Time-Correlated Single Photon Detection". *Review of Scientific Instruments* 76.8 (2005), p. 083104.

[306] H. Neuweiler, C. M. Johnson, and A. R. Fersht. "Direct Observation of Ultrafast Folding and Denatured State Dynamics in Single Protein Molecules." *Proceedings of the National Academy of Sciences* 106.44 (Nov. 2009), pp. 18569–18574.

[307] J. R. Lakowicz. *Principles of Fluorescence Spectroscopy*. Springer Science & Business Media, Dec. 2007.

[308] N. J. Greenfield. "Using Circular Dichroism Spectra to Estimate Protein Secondary Structure." *Nature Protocols* 1.6 (2006), pp. 2876–2890.

[309] **M. P. Luitz**, R. Bomblies, E. Ramcke, A. Itzen, and M. Zacharias. “Adenylylation of Tyr77 stabilizes Rab1b GTPase in an active state: A molecular dynamics simulation analysis”. *Scientific reports* 6 (Jan. 2016), p. 19896.

[310] J. Cherfils and M. Zeghouf. “Regulation of Small GTPases by GEFs, GAPs, and GDIs”. *Physiol. Rev.* 93.1 (2013), pp. 269–309.

[311] K. Aktories. “Bacterial protein toxins that modify host regulatory GTPases”. *Nat. Rev. Micro. Biol.* 9.7 (2011), pp. 487–498.

[312] M. P. Müller, H. Peters, J. Blümer, W. Blankenfeldt, R. S. Goody, and A. Itzen. “The Legionella Effector Protein DrrA AMPylates the Membrane Traffic Regulator Rab1b”. *Science* 329.5994 (2010), pp. 946–949.

[313] C. A. Worby, S. Mattoo, R. P. Kruger, L. B. Corbeil, A. Koller, J. C. Mendez, B. Zekarias, C. Lazar, and J. E. Dixon. “The Fic Domain: Regulation of Cell Signaling by Adenylylation”. *Molecular Cell* 34.1 (2009), pp. 93–103.

[314] M. L. Yarbrough, Y. Li, L. N. Kinch, N. V. Grishin, H. L. Ball, and K. Orth. “AMPylation of Rho GTPases by Vibrio VopS Disrupts Effector Binding and Downstream Signaling”. *Science* 323.5911 (2009), pp. 269–272.

[315] A. K. Haas, S.-i. Yoshimura, D. J. Stephens, C. Preisinger, E. Fuchs, and F. A. Barr. “Analysis of GTPase-activating proteins: Rab1 and Rab43 are key Rabs required to maintain a functional Golgi Complex in Human Cells”. *Journal of Cell Science* 120.17 (2007), pp. 2997–3010.

[316] A. Ingmundson, A. Delprato, D. G. Lambright, and C. R. Roy. “Legionella pneumophila proteins that regulate Rab1 Membrane Cycling”. *Nature* 450.7168 (2007), pp. 365–369.

[317] M. P. Müller, A. V. Shkumatov, L. K. Oesterlin, S. Schoebel, P. R. Goody, R. S. Goody, and A. Itzen. “Characterization of Enzymes from Legionella pneumophila Involved in Reversible Adenylylation of Rab1 Protein”. *J. Biol. Chem.* 287.42 (2012), pp. 35036–35046.

[318] M. Milburn, L. Tong, A. deVos, A. Brunger, Z. Yamaizumi, S. Nishimura, and S. Kim. “Molecular switch for signal transduction: structural differences between active and inactive forms of protooncogenic Ras Proteins”. *Science* 247.4945 (1990), pp. 939–945.

[319] C. Smit, J. Blümer, M. F. Eerland, M. F. Albers, M. P. Müller, R. S. Goody, A. Itzen, and C. Hedberg. “Efficient Synthesis and Applications of Peptides containing Adenylylated Tyrosine Residues”. *Angewandte Chemie International Edition* 50.39 (2011), pp. 9200–9204.

[320] C. Kötting, A. Kallenbach, Y. Suveyzdis, A. Wittinghofer, and K. Gerwert. “The GAP arginine finger movement into the catalytic site of Ras Increases the Activation Entropy”. *Proceedings of the National Academy of Science* 105.17 (2008), pp. 6260–6265.

[321] B. J. Grant, A. A. Gorfe, and J. A. McCammon. “Ras Conformational Switching: Simulating Nucleotide-Dependent Conformational Transitions with Accelerated Molecular Dynamics”. *PLoS Computational Biology* 5.3 (Mar. 2009), e1000325.

[322] R. A. Phillips, J. L. Hunter, J. F. Eccleston, and M. R. Webb. “The Mechanism of Ras GTPase Activation by Neurofibromin”. *Biochemistry* 42.13 (2003), pp. 3956–3965.

[323] G. Privé, M. Milburn, L. Tong, A. de Vos, Z. Yamaizumi, S. Nishimura, and S. Kim. “X-ray Crystal Structures of Transforming p21 Ras Mutants Suggest a Transition-State Stabilization Mechanism for GTP Hydrolysis”. *Proceedings of the National academy of Sciences* 89.8 (Apr. 1992), pp. 3649–3653.

[324] A. Diao, D. Rahman, D. J. Pappin, J. Lucocq, and M. Lowe. “The coiled-coil membrane protein golgin-84 is a novel rab effector required for Golgi Ribbon Formation”. *Journal of Cell Biology* 160.2 (2003), pp. 201–212.

[325] J. Fischer, T. Weide, and A. Barnekow. “The MICAL Proteins and Rab1: A Possible Link to the Cytoskeleton?” *Biochemical and Biophysical Research Communications* 328.2 (2005), pp. 415–423.

[326] N. Hyvola, A. Diao, E. McKenzie, A. Skippen, S. Cockcroft, and M. Lowe. “Membrane targeting and activation of the Lowe syndrome protein OCRL1 by rab GTPases”. *EMBO Journal* 25.16 (2006), pp. 3750–3761.

[327] B. D. Moyer, B. B. Allan, and W. E. Balch. “Rab1 Interaction with a GM130 Effector Complex Regulates COPII Vesicle cis-Golgi Tethering”. *Traffic* 2.4 (2001), pp. 268–276.

[328] M. Rosing, E. Ossendorf, A. Rak, and A. Barnekow. “Giantin interacts with both the small GTPase Rab6 and Rab1”. *Experimental Cell Research* 313.11 (2007), pp. 2318–2325.

[329] T. Weide, J. Teuber, M. Bayer, and A. Barnekow. “MICAL-1 Isoforms, Novel Rab1 Interacting Proteins”. *Biochemical and Biophysical Research Communications* 306.1 (2003), pp. 79–86.

[330] T. Weide, M. Bayer, M. Köster, J.-P. Siebrasse, R. Peters, and A. Barnekow. “The Golgi matrix protein GM130: a specific interacting partner of the small GTPase rab1b”. *EMBO Reports* 2.4 (2001), pp. 336–341.

[331] R. Valsdottir, H. Hashimoto, K. Ashman, T. Koda, B. Storrie, and T. Nilsson. “Identification of rabaptin-5, rabex-5, and GM130 as putative effectors of rab33b, a regulator of retrograde traffic between the Golgi apparatus and ER”. *FEBS Letters* 508.2 (2001), pp. 201–209.

[332] S. Schoebel, L. K. Oesterlin, W. Blankenfeldt, R. S. Goody, and A. Itzen. “RabGDI displacement by DrrA from Legionella is a Consequence of its Guanine Nucleotide Exchange Activity”. *Molecular Cell* 36.6 (2009), pp. 1060–1072.

[333] Y.-W. Wu, L. K. Oesterlin, K.-T. Tan, H. Waldmann, K. Alexandrov, and R. S. Goody. “Membrane targeting mechanism of Rab GTPases Elucidated by Semisynthetic Protein Probes”. *Nature Chemical Biology* 6.7 (2010), pp. 534–540.

[334] S. Piana, A. G. Donchev, P. Robustelli, and D. E. Shaw. “Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States”. *Journal of Physical Chemistry B* 119.16 (2015), pp. 5113–5123.

[335] R. B. Best and J. Mittal. “Free-energy landscape of the GB1 Hairpin in All-Atom Explicit Solvent Simulations with Different Force Fields: Similarities and Differences”. *Proteins* 79.4 (2011), pp. 1318–1328.

[336] M. Geyer, C. Wilde, J. Selzer, K. Aktories, and H. R. Kalbitzer. “Glucosylation of Ras by Clostridium sordellii Lethal Toxin: Consequences for Effector Loop Conformations observed by NMR Spectroscopy”. *Biochemistry* 42.41 (2003), pp. 11951–11959.

[337] I. R. Vetter, F. Hofmann, S. Wohlgemuth, C. Herrmann, and I. Just. “Structural consequences of mono-glucosylation of Ha-Ras by Clostridium Sordellii Lethal Toxin”. *Journal of Molecular Biology* 301.5 (2000), pp. 1091–1095.

[338] S. Mukherjee, X. Liu, K. Arasaki, J. McDonough, J. E. Galán, and C. R. Roy. “Modulation of Rab GTPase Function by a Protein Phosphocholine Transferase”. *Nature* 477.7362 (2011), pp. 103–106.

[339] K. L. Meagher, L. T. Redman, and H. A. Carlson. “Development of polyphosphate parameters for use with the AMBER Force Field”. *Journal of Computational Chemistry* 24.9 (2003), pp. 1016–1025.

[340] A. Sousa da Silva and W. Vranken. “ACPYPE - AnteChamber PYthon Parser interfacE”. *BMC Research Notes* 5.1 (2012), p. 367.

[341] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman. “Multidimensional Free-Energy Calculations Using the Weighted Histogram Analysis Method”. *Journal of Computational Chemistry* 16.11 (1995), pp. 1339–1350.

[342] A. Grossfield. *WHAM: An Implementation of the Weighted Histogram Analysis Method*. Version Revision: 7153.

[343] N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon. “Electrostatics of Nanosystems: Application to Microtubules and the Ribosome”. *Proc. Natl. Acad. Sci. U.S.A.* 98.18 (2001), pp. 10037–10041.

[344] L. Li, C. Li, Z. Zhang, and E. Alexov. “On the Dielectric “Constant” of Proteins: Smooth Dielectric Function for Macromolecular Modeling and Its Implementation in DelPhi”. *Journal of Chemical Theory and Computation* 9.4 (2013), pp. 2126–2136.

[345] P. R. Goody, K. Heller, L. K. Oesterlin, M. P. Müller, A. Itzen, and R. S. Goody. “Reversible phosphocholination of Rab proteins by Legionella Pneumophila Effector Proteins”. *EMBO Journal* 31.7 (2012), pp. 1774–1784.

[346] R. T. Raines. “Ribonuclease A”. *Chemical Reviews* 98.3 (1998). PMID: 11848924, pp. 1045–1066.

[347] F. M. Richards. “On the Enzymic Activity of Subtilisin-Modified Ribonuclease”. *Proceedings of the National Academy of Sciences of the United States of America* 44.2 (1958), pp. 162–166.

[348] F. M. Richards and P. J. Vithayathil. “The Preparation of Subtilisin-Modified Ribonuclease and the Separation of the Peptide and Protein Components”. *Journal of Biological Chemistry* 234 (6 1959), pp. 1459–1465.

[349] H. Wyckoff, K. D. Hardman, N. Allewell, T. Inagami, L. Johnson, and F. M. Richards. “The Structure of Ribonuclease-S at 3.5A Resolution”. *Journal of Biological Chemistry* 242.17 (1967), pp. 3984–3988.

[350] H. Wyckoff, D. Tsernoglou, A. Hanson, J. Knox, B. Lee, and F. M. Richards. “The Three-Dimensional Structure of Ribonuclease-S Interpretation of an Electron Density Map at a Nominal Resolution of 2A”. *Journal of Biological Chemistry* 245.2 (1970), pp. 305–328.

[351] G. S. Ratnaparkhi, S. K. Awasthi, P. Rani, P. Balaram, and R. Varadarajan. "Structural and Thermodynamic Consequences of Introducing $\alpha$-aminoisobutyric Acid in the S peptide of Ribonuclease S". *Protein engineering* 13.10 (2000), pp. 697–702.

[352] J. M. Goldberg and R. L. Baldwin. "A Specific Transition State for S-peptide Combining With Folded S-protein and Then Refolding". *Proceedings of the National Academy of Sciences of the United States of America* 96.5 (1999), pp. 2019–2024.

[353] F. M. Finn and K. Hofmann. "Studies on Polypeptides. XXXIII. Enzymic Properties of Partially Synthetic Ribonucleases1-4". *Journal of the American Chemical Society* 87.3 (1965), pp. 645–651.

[354] A. Komoriya and I. Chaiken. "Sequence Modeling Using Semisynthetic Ribonuclease S." *Journal of Biological Chemistry* 257.5 (1982), pp. 2599–2604.

[355] H. C. Taylor, A. Komoriya, and I. M. Chaiken. "Crystallographic Structure of an Active, Sequence-Engineered Ribonuclease". *Proceedings of the National Academy of Sciences* 82.19 (1985), pp. 6423–6426.

[356] C. M. Cuchillo, M. V. Nogues, and R. T. Raines. "Bovine Pancreatic Ribonuclease: Fifty Years of the First Enzymatic Reaction Mechanism". *Biochemistry* 50.37 (2011), pp. 7835–7841.

[357] F. M. Finn and K. Hofmann. "S-peptide-S-protein System. Model for Hormone-Receptor Interaction". *Accounts of Chemical Research* 6.5 (1973), pp. 169–176.

[358] A. A. Schreier and R. L. Baldwin. "Mechanism of Dissociation of S-peptide From Ribonuclease S". *Biochemistry* 16.19 (1977), pp. 4203–4209.

[359] J. M. Goldberg and R. L. Baldwin. "Kinetic Mechanism of a Partial Folding Reaction. 2. Nature of the Transition State". *Biochemistry* 37.8 (1998), pp. 2556–2563.

[360] A. T. Alexandrescu, K. Rathgeb-Szabo, W. Jahnke, T. Schulthess, R. A. Kammerer, and K. Rumpel. "15N Backbone Dynamics of the S-peptide from Ribonuclease A in its free and S-protein Bound Forms: Toward a Site-Specific Analysis of Entropy Changes upon Folding". *Protein Science* 7.2 (1998), pp. 389–402.

[361] P. S. Kim and R. L. Baldwin. "A helix stop signal in the isolated S-peptide of ribonuclease A". *Nature* 5949 (1984), pp. 329–334.

[362] M. Rico, J. Santoro, F. Bermejo, J. Herranz, J. Nieto, E. Gallego, and M. Jimenez. "Thermodynamic Parameters for the Helix–Coil Thermal Transition of Ribonuclease-S-peptide and Derivatives From 1h-Nmr Data". *Biopolymers* 25.6 (1986), pp. 1031–1053.

[363] A. Bierzynski, P. S. Kim, and R. L. Baldwin. "A Salt Bridge Stabilizes the Helix Formed by Isolated C-peptide of RNase A". *Proceedings of the National Academy of Sciences* 79.8 (1982), pp. 2470–2474.

[364] K. R. Shoemaker, P. S. Kim, D. N. Brems, S. Marqusee, E. J. York, I. M. Chaiken, J. M. Stewart, and R. L. Baldwin. "Nature of the Charged-Group Effect on the Stability of the C-peptide Helix". *Proceedings of the National Academy of Sciences* 82.8 (1985), pp. 2349–2353.

[365] M. Schlosshauer and D. Baker. “A General Expression for Bimolecular Association Rates with Orientational Constraints”. *The Journal of Physical Chemistry B* 106.46 (2002), pp. 12079–12083.

[366] M. Schlosshauer and D. Baker. “Realistic Protein–protein Association Rates From a Simple Diffusional Model Neglecting Long-Range Interactions, Free Energy Barriers, and Landscape Ruggedness”. *Protein Science* 13.6 (2004), pp. 1660–1669.

[367] S. H. Northrup, S. A. Allison, and J. A. McCammon. “Brownian Dynamics Simulation of Diffusion Influenced Bimolecular Reactions”. *The Journal of Chemical Physics* 80.4 (1984), pp. 1517–1524.

[368] D. L. Ermak and J. McCammon. “Brownian Dynamics With Hydrodynamic Interactions”. *The Journal of chemical physics* 69.4 (1978), pp. 1352–1360.

[369] F. M. Richards, H. Wyckoff, W. Carlson, N. Allewell, B. Lee, and Y. Mitsui. “Protein Structure, Ribonuclease-S and Nucleotide Interactions”. In: *Cold Spring Harbor Symposia on Quantitative Biology*. Vol. 36. Cold Spring Harbor Laboratory Press. 1972, pp. 35–43.

[370] M. P. Luitz and M. Zacharias. “Role of tyrosine hot-spot residues at the interface of colicin E9 and Immunity Protein 9: A Comparative Free Energy Simulation Study”. *Proteins: Structure, Function, and Bioinformatics* 81.3 (2013), pp. 461–468.

[371] A. Tjernberg, N. Markova, W. J. Griffiths, and D. Hallén. “DMSO-related Effects in Protein Characterization”. *Journal of biomolecular screening* 11.2 (2006), pp. 131–137.

[372] G. S. Ratnaparkhi and R. Varadarajan. “Thermodynamic and Structural Studies of Cavity Formation in Proteins Suggest That Loss of Packing Interactions Rather Than the Hydrophobic Effect Dominates the Observed Energetics”. *Biochemistry* 39.40 (2000), pp. 12365–12374.

[373] G. Kieseritzky and E.-W. Knapp. “Optimizing pKA Computation in Proteins with pH Adapted Conformations”. *Proteins: Structure, Function, and Bioinformatics* 71.3 (2008), pp. 1335–1348.

[374] J. L. Markley. “Observation of Histidine Residues in Proteins by Nuclear Magnetic Resonance Spectroscopy”. *Accounts of Chemical Research* 8.2 (1975), pp. 70–80.

[375] B. Borah, C. W. Chen, W. Egan, M. Miller, A. Wlodawer, and J. S. Cohen. “Nuclear Magnetic Resonance and Neutron Diffraction Studies of the Complex of Ribonuclease A with Uridine Vanadate, a Transition-State Analog”. *Biochemistry* 24.8 (1985), pp. 2058–2067.

[376] D. E. Walters and A. Allerhand. “Tautomeric States of the Histidine Residues of Bovine Pancreatic Ribonuclease A. Application of Carbon 13 Nuclear Magnetic Resonance Spectroscopy.” *Journal of Biological Chemistry* 255.13 (1980), pp. 6200–6204.

[377] J. E. Brown and W. A. Klee. “Helix-Coil Transition of the Isolated Amino Terminus of Ribonuclease”. *Biochemistry* 10.3 (1971), pp. 470–476.

[378] M. Dadlez, A. Bierzyński, A. Godzik, M. Sobocińska, and G. Kupryszewski. “Conformational Role of His-12 in C-peptide of Ribonuclease A”. *Biophysical chemistry* 31.1 (1988), pp. 175–181.

[379] K. R. Shoemaker, R. Fairman, D. A. Schultz, A. D. Robertson, E. J. York, J. M. Stewart, and R. L. Baldwin. “Side-Chain Interactions in the C-peptide Helix: Phe 8 – His 12+”. *Biopolymers* 29.1 (1990), pp. 1–11.

[380] A. M. Labhardt. “Secondary Structure in Ribonuclease: I. Equilibrium Folding Transitions Seen by Amide Circular Dichroism”. *Journal of molecular biology* 157.2 (1982), pp. 331–355.

[381] D. Loftus, G. O. Gbenle, P. S. Kim, and R. L. Baldwin. “Effects of Denaturants on Amide Proton Exchange Rates: A Test for Structure in Protein Fragments and Folding Intermediates”. *Biochemistry* 25.6 (1986), pp. 1428–1436.

[382] B. Rabenstein and E.-W. Knapp. “Calculated pH-Dependent Population and Protonation of Carbon-Monoxy-Myoglobin Conformers”. *Biophysical Journal* 80.3 (2001), pp. 1141–1150.

[383] D. A. Pearlman and B. G. Rao. “Free Energy Calculations: Methods and Applications”. In: *Encyclopedia of Computational Chemistry*. John Wiley & Sons, Ltd, 2002. ISBN: 9780470845011.

[384] C. H. Bennett. “Efficient Estimation of Free Energy Difference from Monte Carlo Data”. *Journal of Computational Physics* 22 (1976), pp. 245–268.

[385] C. W. Hopkins, S. Le Grand, R. C. Walker, and A. E. Roitberg. “Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning”. *Journal of Chemical Theory and Computation* 11.4 (2015). PMID: 26574392, pp. 1864–1874.

[386] M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. Oxford university press, 1989.

[387] G. Schreiber, G. Haran, and H. Zhou. “Fundamental Aspects of Protein−Protein Association Kinetics”. *Chemical Reviews* 109.3 (2009). PMID: 19196002, pp. 839–860.

[388] H.-X. Zhou. “Brownian Dynamics Study of the Influences of Electrostatic Interaction and Diffusion on Protein-Protein Association Kinetics.” *Biophysical journal* 64.6 (1993), p. 1711.

[389] K. Šolc and W. H. Stockmayer. “Kinetics of Diffusion-Controlled Reaction Between Chemically Asymmetric Molecules. II. Approximate Steady-State Solution”. *International Journal of Chemical Kinetics* 5.5 (1973), pp. 733–752.

[390] N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon. “Electrostatics of Nanosystems: Application to Microtubules and the Ribosome”. *Proceedings of the National Academy of Sciences* 98.18 (2001), pp. 10037–10041.

[391] H. Wyckoff, K. D. Hardman, N. Allewell, T. Inagami, L. Johnson, and F. M. Richards. “The structure of ribonuclease-S at 3.5 A resolution”. *Journal of Biological Chemistry* 242.17 (1967), pp. 3984–3988.

[392] H. R. Kunsch. “The Jackknife and the Bootstrap for General Stationary Observations”. *Annals of Statistics* 17.3 (Sept. 1989), pp. 1217–1241.

[393] W. Kabsch. “A Solution for the Best Rotation to Relate Two Sets of Vectors”. *Acta Crystallographica, Section A: Foundations of Crystallography* 32.5 (Sept. 1976), pp. 922–923.

[394] W. Kabsch. “A Discussion of the Solution for the Best Rotation to Relate Two Sets of Vectors”. *Acta Crystallographica, Section A: Foundations of Crystallography* 34.5 (Sept. 1978), pp. 827–828.

[395] S. Liem, D. Brown, and J. Clarke. “Molecular Dynamics Simulations on Distributed Memory Machines”. *Computer Physics Communications* 67.2 (1991), pp. 261–267.

[396] K. J. Bowers, R. O. Dror, and D. E. Shaw. “The Midpoint Method for Parallelization of Particle Simulations”. *The Journal of Chemical Physics* 124.18, 184109 (2006).