Novel concepts to study conformation and association dynamics of biomolecules

Martin Held

Januar 2012

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

1. Gutachter Dr. Frank Noé
2. Gutachter Prof. Dr. Wolfgang Wenzel

Tag der Disputation: 2012-03-16

Ehrenwörtliche Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen verwendet habe.

Martin Held
Berlin, Januar 2012

Bibliographic Information

Parts of this thesis have been published or submitted for publication:

Chapter 2

- JH Prinz, H Wu, M Sarich, B Keller, M Fischbach, **M Held**, JD Chodera, Ch Schütte, F Noé (2011) Markov models of molecular kinetics: Generation and Validation. *J. Chem. Phys.* 134 (17) pp. 174105

  Link `http://link.aip.org/link/doi/10.1063/1.3565032`

- JH Prinz, **M Held**, JC Smith, F Noé (2011) Efficient Computation, Sensitivity and Error Analysis of Committor Probabilities for Complex Dynamical Processes. *Multiscale Model. Simul.* 9 pp. 545-567

  Link `http://dx.doi.org/10.1137/100789191`

Chapter 3

- **M Held**, P Imhof, B Keller, F Noé (2011) Modulation of a ligand's energy landscape and kinetics by the chemical environment. *submitted*

Chapter 4

- **M Held**, P Metzner, JH Prinz, F Noé (2011) Mechanisms of protein-ligand association and its modulation by protein mutations. *Biophys. J.* 100 (3) pp. 701-10

  Link `http://dx.doi.org/10.1016/j.bpj.2010.12.3699`

- **M Held**, F Noé (2011) Calculating Kinetics and Pathways fo Protein-Ligand Association, *European Journal of Cell Biology* 91 (4) pp. 357-364

  Link `http://dx.doi.org/10.1016/j.ejcb.2011.08.004`

Chapter 5

- K Faelber, Y Posor* , S Gao* , **M Held*** , Y Roske* , D Schulze, V Haucke, F Noé, O Daumke (2011) Crystal structure of nucleotide-free dynamin, *Nature*, 477 (7366) pp. 556-60 (* - equal contribution)

  Link `http://dx.doi.org/10.1038/nature10369`

Acknowledgements

I would like to thank all people who supported and inspired me during the past four years. I owe special gratitude to my supervisor Frank Noé, without his continuous guidance and support the completion of this work would have been near to impossible. My sincere thanks go further to Christof Schütte for creating the unique interdisciplinary *biocomputing habitat* which provides the breeding ground for many scientific explorations. Wolfgang Wenzel deserves special recognition for volunteering to review this thesis.

In addition I am very grateful to Katja Fälber and Oliver Daumke for the interesting and fruitful cooperation we had on the Dynamin protein, the loop modeling sessions at the crystallographer dungeon forming the most enjoyable memories. Further I thank all people in the International Max Planck Research School for Computational Biology and Scientific Computing for social support, especially Kirsten Kelleher and Hannes Luz (†) for their true dedication.

Many thanks are given to all current and past members of the biocomputing group for creating this friendly atmosphere. Special thanks go to Tim Conrad for initiating my engagement in the biocomputing group and his support as well as to my office mates Jan Wigger, Johannes Schöneberg and Bettina Keller for always having a sympathetic ear and interesting discussions and Jan-Hendrik Prinz for providing his home in the final writing stages.

Finally and foremost, I want to express my deepest gratitude to my parents, my brother and Sabine for their limitless and unconditioned backup during this laborious journey.

This work was supported by Deutsche Forschungsgemeinschaft Sonderforschungsbereich 449 and 740 and the International Max Planck Research School - Computational Biology and Scientific Computing.

Contents

1 Introduction

Life is complex and dynamic. A comprehensive understanding of a living system is desirable as it would allow for directed manipulation of the system, e.g., to cure diseases with a particular treatment, or to engineer artificial systems that exhibit a certain behavior. However, the usually large complexity found in living systems permits to study only individual parts rather than the whole system. Thus, it is inevitable to explore the mechanisms of interaction between these individual parts to gain knowledge about the entire system and to understand and eventually predict the system's behavior.

In order to gain an understanding of partial aspects of the system and its dynamical nature, experimental and computational models are employed. Experiments provide a direct observation of reality while computational models provide a representation of our current knowledge and understanding of biochemical and physical processes. Both strategies are complementary with respect to their strengths and limitations. An experiment probes aspects of the system under investigation that are accessible by the experimental technique. For example, a fluorescence resonance electron transfer (FRET) experiments can monitor particular distances in a molecule but not directly reveal the complete picture of its high-dimensional conformational dynamics. Likewise, when fluorescence labeling is used to monitor the movement of a particular protein, unlabeled proteins that contribute to the dynamics of the labelled one remain hidden. A model, on the other hand, integrates the knowledge gained from different experiments and thus summarizes the current understanding of reality at an appropriate level. By making assumptions about unknown model parameters and relationships, it can be used to predict the system's behavior and unmeasurable system quantities. Further, it allows to generate hypotheses which then can be evaluated by experiments. Hence, it generates a discovery cycle that is fueled by the interplay of experiment and computational modeling (see Figure 1.1). The combination of both strategies can be regarded as a prerequisite for understanding complex living systems.

However, with the current knowledge and methods available, it is prohibitive to model an entire organism in its full complexity. Hence, the characteristics of the system that enter the model should depend on the research question tackled. It is also the scientific question that determines the appropriate level of detail at which an experiment is performed or a computational model is built. For example, on the one hand, to answer the question if administration of a substance will result in loss of pain, a simple model that yields yes/no answers might suffice [1, 2]. On the other hand, this black-box model does not generate

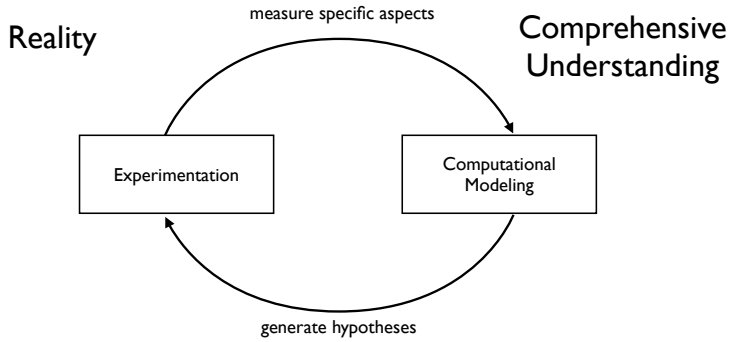


Figure 1.1: Discovery cycle to understand reality by integrating knowledge gained from experiments and computational modeling.

or utilize understanding of the underlying mechanism of action. To shed light on this question, in particular to reveal detailed molecular interactions of the substance with its target, modeling atomistic details would be necessary. That kind of model however, does not in turn necessarily provide any information about the desired effect of the substance in the context of a functional organism.

To allow for the study of particular aspects of the biological system under investigation, the modeling strategy has to be chosen carefully as a biological system can be represented on different scales. For example, on an atomistic scale, all atoms of a molecule are represented while on a cellular level the molecule itself is regarded as a single entity. Hence, the scientific question determines the scale of the computational model as well as the components of the system that are included in the model. Several scientific disciplines have emerged that differ with respect to the scale of the models employed. In particular for metabolic processes, the following disciplines prominently exist: pharmacokinetic/pharmacodynamic (PK/PD) modeling, systems biology and computational biophysics (see 1.2a). In that respect, PK/PD is the most high-level description where a whole organism is represented as a set of connected distributional spaces with low level of detail. These models aim at describing and predicting the concentration and effect of single or multiple substances on a whole-organism level [3, 1, 4]. A finer level of detail is considered in systems biology modeling [5, 6]. Here, dynamical properties of a biological system are studied on the cellular level by interaction networks, e.g., signaling pathways, gene-regulatory networks and metabolic networks. These networks represent the interplay of biological entities, e.g., proteins, cofactors, ligands, mostly by means of coupled differential equations that describe their time dependent concentrations [5]. The mathematical description allows for studying the dynamical response of the system after perturbation or direct manipulation. However, the molecular detail of interaction as well as their spatial aspects remain unconsidered. To describe these molecular details, computational biophysics approaches are employed. Computational biophysical models aim at

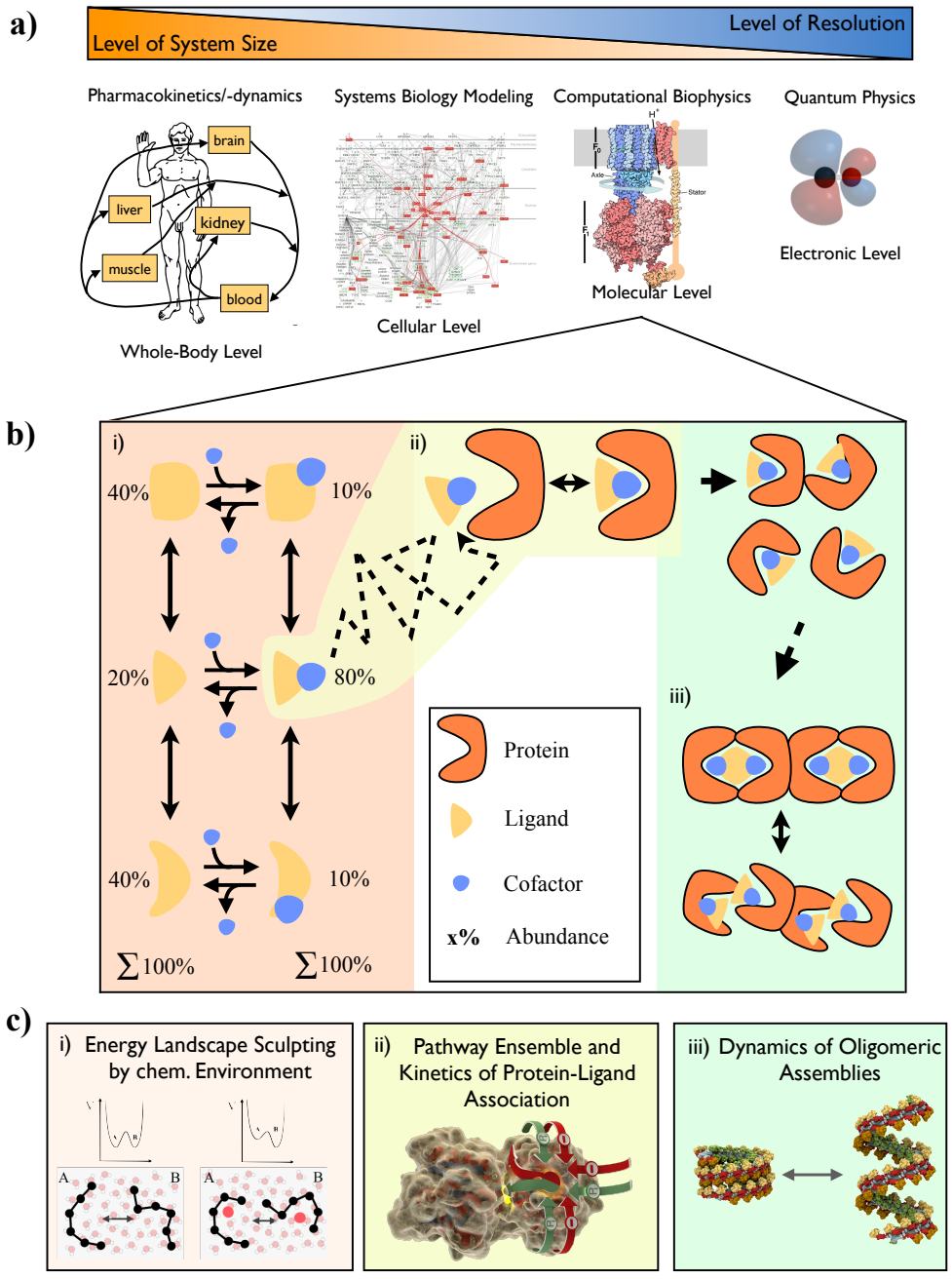


Figure 1.2: Modeling of biological systems. a) Various complexity levels with increasing resolution and decreasing spatial size: pharmacokinetics/pharmacodynamics, systems biology, computational biophysics and quantum physics. b) Dynamical aspects of protein-ligand association and protein complexes addressed in this thesis. Left column shows conformational dynamics of a small ligand and changes therein upon cofactor binding. Left and middle column show the protein-ligand association phase of the cofactor stabilized ligand. Right column shows the formation of protein complexes and their conformational dynamics. c) The three different biophysical processes modeled in this thesis: i) the effects of the chemical environment on thermodynamic and kinetic properties of a ligand molecule, ii) the mechanisms and pathways of protein-ligand association, iii) dynamical properties of oligomer assemblies. The relation to b) is indicated by numbering and coloring.

describing molecular details of the considered system based on fundamental physical laws [7, 8, 9]. Currently, the most accurate description of a molecular system is provided by a quantum-physical description which accounts for the distribution of individual electrons. For example, this approach is employed when enzymatic reactions are studied that involve atomic bound breaking and forming [10]. However, when one is interested in the speed of enzyme-substrate association rather than in the details of the reaction mechanism, the quantum-physical treatment might not be appropriate. On the one hand, it would be infeasible due to the huge system size. On the other hand, the time and length scales relevant for the association process are well beyond the scales described by the quantum-physical treatment.

Considering the above examples, it becomes apparent that a careful selection of the modeling strategy is inevitable to gain insights about particular aspects of a biological system. As Einstein already pointed out: "A model should be as simple as possible, but not simpler".

The focus of this thesis is on biophysical modeling. The last decade has seen a tremendous development in computational power, efficiency of molecular dynamics software as well as in techniques to analyze molecular dynamics trajectories. Taken together, these improvements enabled the study of complex biological processes in the cell, e.g., binding of small ligands to a protein [11], the folding of proteins [12, 13, 14, 15, 16] or conformational stabilization of peptides [17, 18].

In a living cell all these processes are interconnected and influence each other. To gain a comprehensive understanding of the greater biological system, their interplay has to be taken into account rather than studied in isolation. Unfortunately, it is yet impossible to simulate the system as a whole, due to the different time and length scales biological processes occur on. Hence, as already pointed out earlier, different levels of modeling are required to account for the individual processes. Once these processes have been adequately modeled and carefully analyzed, relevant process parameters can be extracted and used further to construct a model that captures more aspects of the system.

In this context, this thesis is concerned with the critical study of models of molecular association and dynamics of biomolecules. In particular, it focuses on biophysical processes that are interconnected with each other: conformational changes of ligands, the association pathways of ligands to proteins and the dynamics of protein assemblies (Figure 1.2b,c). The interface between these processes is investigated and critically discussed. Each of the chapters focuses on a particular model for the different processes: Chapter 3 investigates and illustrates the effects of the chemical environment on thermodynamic and kinetic properties of a ligand molecule, Chapter 4 studies the mechanisms and pathways of protein-ligand association and Chapter 5 reveals dynamical properties of oligomer assemblies. For each of the employed models, it is systematically investigated how changes in the system

setup affect the model behavior. Such investigations are essential to understand how "perturbations" arising from interactions with other individually modeled processes would affect the behavior of the whole system. They hence form the basis for a combination of the individual models to obtain a model that captures larger parts of the system.

The three main contributions of this thesis are shortly outlined below:

i) Modulation of a ligand's energy landscape and kinetics by the chemical environment (Chapter 3) Understanding how the chemical environment modulates the predominant conformations and kinetics of flexible molecules is a core interest of biochemistry and a prerequisite for the rational design of synthetic catalysts. This study combines molecular dynamics simulation and Markov state models (MSMs) to a systematic computational strategy for investigating the effect of the chemical environment of a molecule on its conformations and kinetics. MSMs allow quantities to be computed that are otherwise difficult to access, such as the metastable sets, their free energies, and the relaxation timescales related to the rare transitions between metastable states. Additionally, MSMs are useful to identify observables that may act as sensors for the conformational or binding state of the molecule, thus guiding the design of experiments. In the present study, the conformation dynamics of UDP-GlcNAc are studied in vacuum, water, water+Mg^{2+} and in the protein UDP-GlcNAc 2-epimerase. It is found that addition of Mg^{2+} significantly affects the stability, energetics and kinetics of UDP-GlcNAc. In particular, the slowest structural process - puckering of the GlcNAc sugar - depends on the overall conformation of UDP-GlcNAc and may thus act as a sensor of whether Mg^{2+} is bound or not. Interestingly, transferring the molecule from vacuum to water makes the protein-binding conformations UDP-GlcNAc first accessible, while adding Mg^{2+} further stabilizes them by specifically associating to binding-competent conformations. While Mg^{2+} is not co-crystallized in the UDP-GlcNAc 2-epimerase complex, the selectively stabilized Mg^{2+}:UDP-GlcNAc complex may be a template for the bound state, and Mg^{2+} may accompany the binding-competent ligand conformation to the binding pocket. This serves as a possible explanation of the enhanced epimerization rate in the presence of Mg^{2+}. This role of Mg^{2+} has previously not been described and opens the question whether "binding co-factors" may be a concept of general relevance for protein-ligand binding. This contribution provides a new conceptual approach that allows for a systematic analysis of effects the chemical environment has on kinetic and thermodynamic properties of small ligand molecules.

ii) Mechanisms of protein-ligand association and its modulation by protein mutations (Chapter 4) Protein-ligand interactions are essential for nearly all biological processes, and yet the biophysical mechanism that enables potential binding partners to associate before specific binding occurs remains poorly understood. Fundamental questions include which factors influence the formation of protein-ligand encounter complexes, and whether

designated association pathways exist. To address these questions, we develop a computational approach to systematically analyze the complete ensemble of association pathways. Here, we use this approach to study the binding of a phosphate ion to the Escherichia coli phosphate-binding protein. Various mutants of the protein are considered, and their effects on binding free-energy profiles, association rates, and association pathway distributions are quantified. The results reveal the existence of two anion attractors, i.e., regions that initially attract negatively charged particles and allow them to be efficiently screened for phosphate, which is subsequently specifically bound. Point mutations that affect the charge on these attractors modulate their attraction strength and speed up association to a factor of ten of the diffusion limit, and thus change the association pathways of the phosphate ligand. It is demonstrated that an anion that pre-binds to such an attractor neutralizes its attraction effect to the environment, making the simultaneous association of a second phosphate ion unlikely. This chapter suggests ways in which structural properties can be used to tune molecular association kinetics so as to optimize the efficiency of binding via an anion attractor "filtering device", and highlights the importance of kinetic properties.

iii) Revealing Dynamical Properties of Oligomer Assemblies: Dynamin - A case study (Chapter 5) Proteins can associate and form complexes or oligomeric assemblies. In addition to dynamics of individual proteins these complexes exhibit their "own" dynamics and can switch between different conformations/functions. A protein that builds such complexes is Dynamin, which forms helical assemblies to mediate the process of vesicle scission. For the exact scission mechanism, a number of models have been proposed [19, 20, 21]. The doubt about the exact process was partly due to the absence of a high resolution structure of the Dynamin protein. With the recent discovery of the structure [22] this shortcoming is resolved. In this chapter, it is described how the recently obtained structures were utilized in order to describe dynamical properties of helical assemblies of the Dynamin protein. It is investigated how the type of nucleotide bound to Dynamin G domains affects their interaction strength and how this influences the dynamics of the Dynamin oligomer helix. Furthermore, unresolved loop regions in the original crystal structure are completed by modeling and the resulting structure is fitted into a cryo-EM density of the Dynamin helix. From the modeled structure, the central stalk element could be identified as central mechanical building block of the helix. To assess the stability of the modeled loop regions and the dynamical mechanical properties of the building block, a number of molecular dynamics simulations are performed. The dynamical information about the local building block dynamics is then extrapolated to the entire helix dynamics. This enables the calculation of a stationary probability distribution of various helix conformations using a Markov state modeling approach. Based on the free energy landscape computed from the stationary distribution it is further possible to embed existing Dynamin mediated vesicle

scission mechanisms. Using the modeled helix structure as well as the calculated energy landscape, we are further enabled to propose an alternative scission mechanism.

2 Theory and Methods

"[...]
Markovian process lead us not in vain
Prove to our descendants what we did to them
Then make us go away"

Markovian Process - Bad Religon

2.1 Introduction

This chapter is concerned with the description of theory and methods that have been employed to carry out the research presented in this thesis. All insights gained are a result of careful modeling, simulation and analysis of physical processes. In this work, two types of models were used: microscopic and Markov state models (MSMs). Microscopic models aim at modeling a system by correctly representing its dynamical behavior at atomistic resolution and Markov state models describe the long-time statistical ensemble dynamics of a molecular system by means of a Markov chain on a discrete partition of the configuration space of the system. The first part of this chapter is concerned with microscopic models and their simulation and the second part describes Markov state models of dynamical processes.

2.2 Microscopic Models of Dynamical Processes

2.2.1 Models

Atomistic Molecular Modeling by using an Empirical Potential Function Molecular modeling is a process that aims to realistically describe and predict macroscopic properties of complex chemical systems based on empirical information. The most accurate model of a molecular system at the electronic scale is a quantum mechanical one. In this treatment nuclei as well as individual electrons are considered which in principle allows to predict a wide range of complex chemical behavior. However, by using this high level of detail the computational costs associated with the evaluation of a quantum mechanical model increases dramatically with the system size. In fact, with currently available computational power this circumstance renders the quantum mechanical treatment of biophysically

relevant systems, e.g., proteins prohibitive. Overcoming this limitation is possible by the use of a more simplistic molecular mechanics model. Here the energy of a molecular system is computed using an empirical force field which only takes positions of individual atom into consideration, without accounting for the motion of electrons. It is clear that such a model cannot describe properties of the system that depend on the electronic distribution such as for example formation or breaking of covalent bonds. But it may be very well suited to investigate problems where a fixed typical molecular topology can be assumed, e.g., protein-folding or protein association phenomena. Note that the "fixed molecular topology" assumption does not always hold, e.g., when groups of the system are likely to be protonated, or the partial charges of atoms are likely to change due to polarization effects. However, with the development of next generation force fields [23, 24] these effects are likely to be accounted for, too.

The molecular mechanics model is based on two approximations that allow the reduction from a quantum dynamical system. The first one is the *Born-Oppenheimer* approximation. Based on the fact that nuclei are much heavier than electrons, it states that the nuclei motions can described independently from the motions of the electrons as the electron degrees of freedom relax almost instantaneously. The second approximates the nuclei as point particles that are assumed to behave according to laws of classical Newtonian mechanics, i.e., do not exhibit relativistic effects. While it is now possible to define an empirical potential function to describe the atomistic behavior of a molecular system, one should be aware that high frequency oscillations of covalent molecule bonds often violate the assumptions. However, it has been observed from spectroscopic studies of protein dynamics that there exist a timescale separation between fast bond oscillations and slower degrees of freedom, e.g., angle and dihedral oscillations. Hence, it is possible to fix the length of a covalent bond to its mean length value by using constraints. The simulation than only accounts for the slower degrees of freedom. Practically, this approach has the advantage that a larger integration time step can be used, as the unconstrained degrees of freedom are slower, it thus allows for a more efficient simulation of the model.

In its basic form an empirical potential describes the energy of a system V in terms of bonded and non-bonded energy contributions:

$$V = V_{\text{bonded}} + V_{\text{nonbonded.}} \tag{2.1}$$

Bonded terms account for the part of the energy that originates from covalently bound atoms like bond (V_{bond}), angle (V_{angle}), dihedral (V_{dihedral}) and improper dihedral (V_{improper}) interactions:

$$V_{\text{bonded}} = V_{\text{bond}} + V_{\text{angle}} + V_{\text{dihedral}} + V_{\text{improper}}. \tag{2.2}$$

Non-bonded terms describe energy contributions that arise from non-covalently bound

interactions such as electrostatic (Coulomb) or Van der Waals (VdW) interactions:

$$V_{\text{nonbonded}} = V_{\text{Coulomb}} + V_{\text{VdW}} \tag{2.3}$$

See Figure2.1 for an overview of commonly employed force field terms and potential functions. In the following the individual contributions are described in more detail.

Bond Interactions In models that do not use constrained covalent bonds, they are usually modeled by a harmonic potential. Here the energy of a bond varies with the square of the displacement from the reference bond length:

$$V_{\text{bond}}(r_{ij}) = \frac{k}{2}(r_{ij} - r_0)^2, \tag{2.4}$$

where r_{ij} is the distance between atoms i and j, k is the bond's force constant and r_0 the reference length. Reference length in this context means the length of the bond in absence of other interactions. The equilibrium length of a bond can therefore differ from their reference length, as the equilibrium length can be affected by interactions with surrounding atoms that make a different bond length more energetically favorable. For a list of typical reference bond lengths and force constants see Table 2.1. As expected

Bond	r_0 **[Å]**	**k [kcal/mol/^{2}]**
Csp^3-Csp^3	1.523	317
Csp^2=Csp^2	1.337	690
Csp^2=O	1.208	777
C-N	1.345	719

Table 2.1: Reference lengths and force constants of typical bonds. Table partly adopted from [8] and [25].

from intuition the double bond of sp^2 hybridized carbons Csp^2=Csp^2 has a stronger force constant than the Csp^3-Csp^3 bound and is slightly shorter.

While the harmonic potential is the most commonly used potential to model covalent bonds, it is in principle incorrect as it does not allow for bond breaking. A potential which allows for the breaking of covalent bonds is the Morse potential:

$$V_{\text{bond}}(r_{ij}) = K\left(1 - \exp(-a(r_{ij} - r_0)\right)^2, \tag{2.5}$$

where K is the well depth, r_0 the reference bond length and a a parameter that controls the width of the potential well. See Figure 2.2 for its shape. However, the Morse potential requires more parameters and would be computationally more expensive to evaluate in a molecular dynamics simulation than the harmonic potential. Moreover, when bonds do not significantly deviate from their reference length the harmonic potential is a good approximation, as it can be also seen in Figure 2.2.

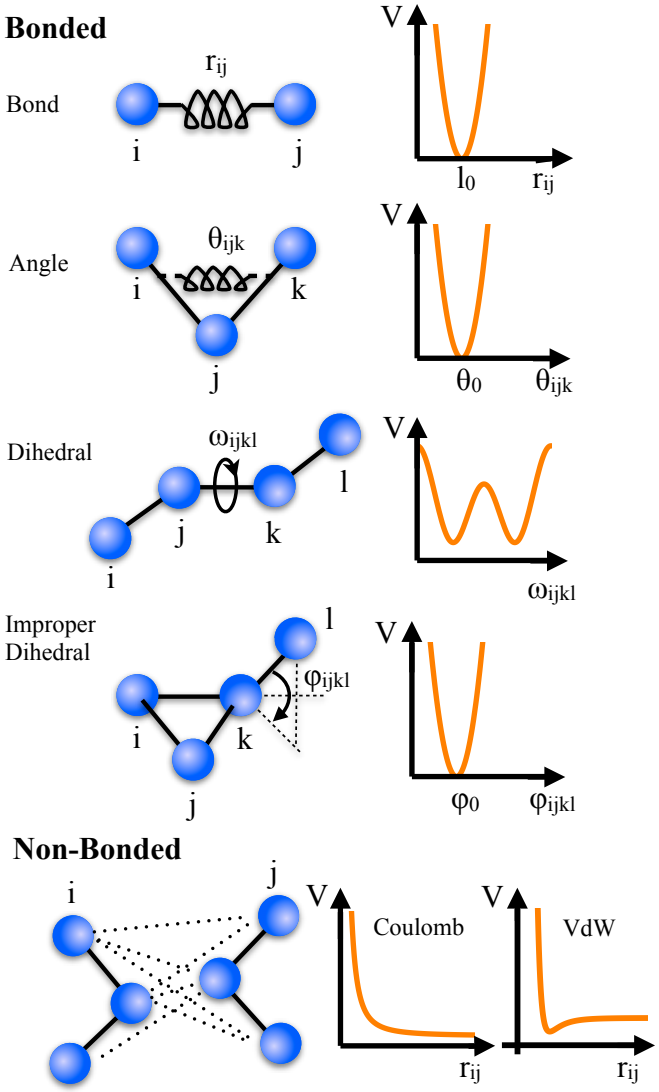


Figure 2.1: Bonded and non-bonded components of an empirical potential. On the left-hand side occurring bonded and non-bonded interactions are shown. The right-hand side depicts exemplary potential functions V of the respective degree of freedom..

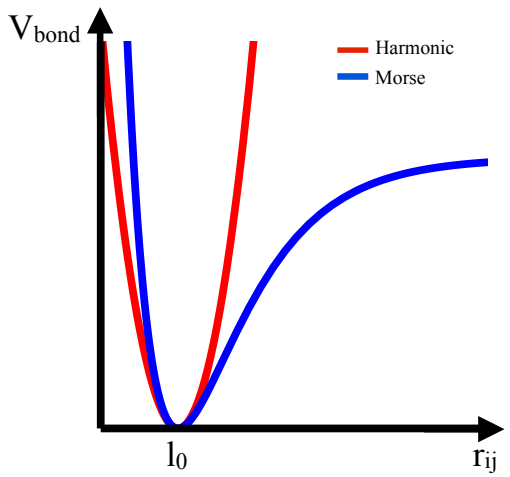


Figure 2.2: Exemplary potential graphs of a harmonic (red) and a Morse (blue) potential. r_{ij} is the bond length between atom i and j and l_0 its reference length. The minima of the Morse potential is well approximated by a harmonic potential.

Bending Angles The potential function that describes the energy of angles between three atoms is commonly modeled by a harmonic potential:

$$V_{\text{angle}}(\theta_{ijk}) = \frac{K}{2}(\theta_{ijk} - \theta_0)^2, \tag{2.6}$$

where θ_{ijk} is the angle spanned by atoms i, j and k , K is the angle's force constant and θ_0 is the reference angle. See Table 2.2 for typical values of θ_0 and K.

Angle	Θ_0 **[deg]**	**K [kcal/mol/deg]**
Csp^3-Csp^3-Csp^3	109.47	0.0099
Csp^3-Csp^3-H	109.47	0.0079
Csp^3-Csp^2-Csp^3	117.2	0.0099
Csp^3-Csp^2=O	122.5	0.0101

Table 2.2: Reference angles and force constants of typical angles. Table partly adopted from [8] and [25].

Dihedral Angles Bonds and angles are considered to be the "stiff" degrees of freedom of a molecular structure that are energetically expensive to change. Most of the structural and energetic variation that is found in molecules is hence due to dihedral degrees of freedom or due to the sum of bond and angle changes in large structural elements. The potential of a dihedral angle is commonly expressed as a cosine series expansion:

$$V_{dihedral}(\omega_{ijkl}) = \sum_{n=0}^{N} \frac{K_n}{2} \left(1 + \cos(n\omega_{ijkl} - \gamma)\right), \tag{2.7}$$

where ω_{ijkl} is the dihedral angle defined by atoms i, j, k and l, K_n can be interpreted as barrier height if only one term is present, otherwise it gives a qualitative interpretation of relative barriers present in the rotation of the dihedral angle. γ determines where the dihedral angle attains its minimum value. A dihedral angle is defined by 4 atoms A-B-C-D, this means parameterization of an ethane molecule (8 atoms) would require 9 dihedral terms to be defined if all atoms would be discriminable ($H_{1,\alpha} - C_\alpha - C_\beta - H_{1,\beta}$, $H_{1,\alpha} - C_\alpha - C_\beta - H_{2,\beta}$, $H_{1,\alpha} - C_\alpha - C_\beta - H_{3,\beta}$, $H_{2,\alpha} - C_\alpha - C_\beta - H_{1,\beta}$, ... , $H_{3,\alpha} - C_\alpha - C_\beta - H_{3,\beta}$). For that reason it is common to use general dihedrals. In force fields like AMBER [26] for example the energy profile of a dihedral angle depends only on the atom types of the central bond, i.e., H-C-C-H, C-C-C-C and H-C-C-C would use the same dihedral parameterization and for ethane only one dihedral parameter would be needed.

Improper Dihedrals For the modeling of certain molecules it can be necessary to keep an atom in a coplanar location with other atoms or in a particular angle. To achieve this improper dihedrals can be utilized. The idea is to define the dihedral not as sequence of bonded atoms A-B-C-D but rather to put the atoms in a sequence such that the thus defined improper dihedral measures the deviation from a plane. As potential functions either periodic functions:

$$V_{improper}(\phi_{ijkl}) = K(1 - cos(2\phi_{ijkl})), \tag{2.8}$$

or harmonic functions:

$$V_{improper}(\phi_{ijkl}) = \frac{K}{2}(\phi_{ijkl} - \phi_0)^2, \tag{2.9}$$

are employed. Here ϕ_{ijkl} is the improper angle define by atoms i, j, k and l, K is the force constant of the angle.

Non-Bonded Interactions Non-bonded interactions describe interactions between atoms that are not specifically bound. Hence they are acting through space and are normally described by an inverse distant relation. Usually non-bonded interactions are divided into two groups: electrostatic interactions and Van der Waals (VdW) interactions.

Electrostatic Interactions - The charge distribution of a molecule arises from elements that have a different electronegativity. The most common way to represent this charge distribution in molecular modeling is to use an arrangement of fractional point charges. This arrangement is designed to such that it covers the molecule and reproduces it electrostatic properties. Often the charges are placed at the positions of the atom nuclei of the system and are referred to as partial atomic charges. The electrostatic contribution to the energy of a system that consists of these partial charges can be computed by using Coulomb's law:

$$V_{Coloumb}(r_{ij}) = \frac{q_i q_j}{4\pi\epsilon_0\epsilon_r r_{ij}} \tag{2.10}$$

where q_i, q_j are partial charges of atom i and j, ϵ_0 is the vacuum permittivity, ϵ_r is the relative permittivity and r_{ij} is the charge distance.

Van der Waals Interactions - VdW interactions are much weaker and only short ranged than the electrostatic interactions. However, for an energetic modeling of a molecular system they have inevitably to be considered. The VdW potential consists of two parts an attractive part and a repulsive part. The attractive part arises from a dispersive force that is due to local short lived dipoles that are created by fluctuations in the electron clouds. The repulsive part becomes relevant when the distance between two nuclei becomes very close. Here, the Pauli exclusion principle prevents that the two nuclei collapse. To model the contribution of VdW interactions to the total potential of a modeled system the Lennard -Jones 12-6 potential is commonly used:

$$V_{VdW}(r_{ij}) = 4\epsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right), \tag{2.11}$$

where σ_{ij} is the collision diameter at which the potential is zero, ϵ_{ij} is the well depth of the potential and r_{ij} the distance of atoms i and j. The $\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}$ term represents the repulsive part and the $\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}$ term accounts for the attractive part.

Poisson-Boltzmann Electrostatics In the previous section, a molecular system was modeled at atomistic detail with various interactions between the atoms. This high level of modeling detail is appropriate to study systems with sizes in the order of 10^5 atoms, e.g., a protein in a box of water molecules. It becomes inefficient for larger systems, e.g., multiple interacting proteins due to the exponentially growing computational costs. This circumstance requires relaxing the level of detail when questions related to such systems are to be studied. A common way to reduce the complexity of the system is to treat the solvent as a continuum with a dielectric constant instead of explicitly representing each solvent molecule. The use of such a continuum representation for the solvent makes the calculation of electrostatic interactions in the system more challenging since Eq. 2.10 only accounts for homogenous systems with a single dielectric constant. Such a description is thus inadequate when the electrostatics of a solvated protein are to be computed, as protein and solvent have different dielectric constants.

An additional challenge is the potential presence of ions in the solvent. While explicitly represented in an atomistic solvent, a continuum representation is not straightforward at first sight. A theory that provides a continuum solution to both the changing dielectricity and the presence of ions problems is the Poisson-Boltzmann equation. It describes the electrostatic potential of molecules in an implicit ionic solution.

The classical Poisson equation relates the variation in the electrostatic potential $V(\mathbf{x})$ at a point $\mathbf{x}$ in a medium with uniform dielectric constant ϵ to the charge density ρ via:

$$\nabla^2 V(\mathbf{x}) = -\frac{\rho(\mathbf{x})}{\epsilon}, \tag{2.12}$$

where $\epsilon = \epsilon_r \epsilon_0$ is the permittivity with ϵ_0 being the vacuum permittivity and ϵ_r being the permittivity of the material. In case where the permittivity changes with position, e.g., inside/outside the protein, Eq. 2.12 changes to:

$$\nabla \cdot [\epsilon(\mathbf{x}) \nabla V(\mathbf{x})] = -\rho(\mathbf{x}). \tag{2.13}$$

Furthermore, if ions in the solution are to be considered it has to be accounted for the fact that the ions will distribute in response to the electrostatic potential present. This ion distribution can be modeled by a Boltzmann distribution and when incorporated into the Poisson equation the Poisson-Boltzmann equation is obtained:

$$\nabla \cdot [\epsilon(\mathbf{x}) \nabla V(\mathbf{x})] = -\rho(\mathbf{x}) - \sum_i c_i^\infty z_i \lambda(\mathbf{x}) \exp\left(\frac{-z_i V(\mathbf{x})}{k_B T}\right). \tag{2.14}$$

c_i^∞ is the bulk concentration of ion species i, z_i is the charge of ion i and $\lambda(\mathbf{x})$ is a delta function that indicates if a position $\mathbf{x}$ is accessible by an ion. This equation is impossible to solve analytically for nontrivial geometries as they are found in common biophysical applications. One has hence to rely on numerical methods such as finite elements, finite difference or boundary element methods to approximate it. Throughout this thesis we used the Adaptive Poisson Boltzmann Solver (APBS) [27] to calculate electrostatic potentials.

2.2.2 Microscopic Dynamics

The dynamics of a molecular system can be described in the realms of a Markov process. Consider the state space Ω which is thought to contain all dynamical variables necessary to describe a molecular system. In the following Ω is assumed to be continuous (but it might also be discrete). The dynamical process that describes the temporal evolution of the molecular system is denoted as $\mathbf{x}(t)$. This process can be either regarded as time continuous (for theoretical elaborations) or time discrete (for computational purposes). For the remainder of this thesis we assume that $\mathbf{x}(t)$ has the following properties:

1. $\mathbf{x}(t)$ is a Markov process in full state space Ω, i.e., the instantaneous change is calculated only based on $\mathbf{x}(t)$ and does not depend on the history of the process. As a consequence the transition probability density of the process is well defined:

$$p(\mathbf{x}, \mathbf{y}; \tau) d\mathbf{y} = \mathbb{P}[\mathbf{x}(t+\tau) \in \mathbf{y} + d\mathbf{y} \,|\, \mathbf{x}(t) = \mathbf{x}], \tag{2.15}$$

   i.e., the probability to find the system started at time t in point $\mathbf{x}$ in an infinitesimal

region $d\mathbf{y}$ around a point $\mathbf{y}$ at time $t+\tau$. Refer to Figure 2.3 for an illustration of the above probability function for a diffusion process in a one dimensional potential . Given $p(\mathbf{x}, \mathbf{y}; \tau)$ is a smooth function, the transition probability into a set $A \subseteq \Omega$ can also be defined and is given by:

$$\begin{aligned} p(\mathbf{x}, A; \tau) &= \mathbb{P}[\mathbf{x}(t+\tau) \in A \,|\, \mathbf{x}(t) = \mathbf{x}] \\ &= \int_{\mathrm{y} \in A} p(\mathbf{x}, \mathbf{y}; \tau) d\mathbf{y}. \end{aligned} \tag{2.16}$$

2. $\mathbf{x}(t)$ is an ergodic process in Ω, i.e., for $t \to \infty$ each point $\mathbf{x}$ will be visited infinitely often. Assuming an infinitely long trajectory the fraction of time the process spends in a point $\mathbf{x}$ is given by its stationary density $\mu(\mathbf{x})$. Under the states assumptions this stationary density is unique. For molecular dynamics simulations at constant volume and temperature (NVT ensemble) $\mu(\mathbf{x})$ is a function of temperature T and can be expressed by the Boltzmann distribution:

$$\mu(\mathbf{x}) = Z(\beta)^{-1} \exp(-\beta V(\mathbf{x})), \tag{2.17}$$

with $V(\mathbf{x})$ denoting the potential energy of the system and $\beta = (k_B T)^{-1}$ where k_B is the Boltzmann constant and T the temperature. $Z(\beta) = \int \exp(-\beta V(\mathbf{x})) d\mathbf{x}$ denotes the partition function. For an illustration of the stationary density of the aforementioned diffusion process refer to Figure 2.3.

3. $\mathbf{x}(t)$ is reversible, i.e., the transition probability obeys the condition of detailed balance:

$$\mu(\mathbf{x}) p(\mathbf{x}, \mathbf{y}; \tau) = \mu(\mathbf{y}) p(\mathbf{x}, \mathbf{y}; \tau). \tag{2.18}$$

This means the fraction of systems transitioning from $\mathbf{x}$ to $\mathbf{y}$ per time τ is equal to the fraction of systems transitioning from $\mathbf{y}$ to $\mathbf{x}$ when the system is in equilibrium. Note that this concept is more general than "time-reversibility" found in microscopic descriptions of dynamics.

Note that many different dynamic realizations, i.e., $p(\mathbf{x}, \mathbf{y}; \tau)$ functions, may exist which result in the same correct Boltzmann distribution. Examples to obtain realizations of a process $\mathbf{x}(t)$ with the properties described above are, e.g., molecular dynamics with the right thermostat or Brownian dynamics which are both described below.

Thermostated Molecular Dynamics In the previous section, we defined models to create an abstract description of reality at various levels of detail. How can these models be utilized to gain insights about the system under consideration? Let us consider the empirical potential defined by various energetic contributions (bonds, angles etc.). This potential defines an energy landscape for our system, from which each system configuration can get

an energy assigned. For very small and simple systems it is possible to enumerate a large fraction of all possible structures and then to perform an energy minimization to obtain a set of structures that are energetically favorable. Given this set of likely structures it is theoretically possible to calculate thermodynamic properties of the system. However, this approach is infeasible for systems of biophysical interest as their size normally prohibits the direct enumeration of all possible structures. Furthermore, the understanding of biophysical processes often requires to consider time-dependent properties of the system under investigation. Prominent examples include protein folding or protein interaction processes.

In order to obtain dynamical information about a system the previously introduced empirical potential can be employed. It can be used to simulate the system behavior in time by using a molecular dynamics simulation algorithm. The algorithm allows the dynamics of the system to be calculated on a microscopic level, which can be translated into macroscopic observables using statistical mechanics.

How are these MD trajectories are generated in practice? MD simulations are classical, which means that they are governed by Newtonian mechanics:

$$\mathbf{f}_i = m_i \mathbf{a}_i, \tag{2.19}$$

where $\mathbf{f}_i$ is the force on particle i, m_i its mass and $\mathbf{a_i}$ its acceleration. Equivalently this force can be expressed as:

$$\mathbf{f}_i = -\nabla_i V. \tag{2.20}$$

$\nabla_i V$ denotes the gradient of the potential energy of particle i. The potential energy can be computed using the empirical potential defined above (Eq. 2.1). Combining Eq. 2.19 and Eq. 2.20 leads to:

$$-\frac{\mathrm{d}V_i}{\mathrm{d}\mathbf{x}_i} = m_i \mathbf{a}_i = m_i \frac{\mathrm{d}\mathbf{v}_i}{\mathrm{d}t} = m_i \frac{\mathrm{d}^2 \mathbf{x}_i}{\mathrm{d}t^2}, \tag{2.21}$$

where $\mathbf{x}_i$ is the position of particle i. For very short times, the acceleration $\mathbf{a}_i$ is assumed to be constant and the velocity $\mathbf{v}_i$ of a particle i can be computed as:

$$\mathbf{v}_i = \mathbf{a}_i t + \mathbf{v}_{i0} = \frac{\mathrm{d}\mathbf{x}_i}{\mathrm{d}t}, \tag{2.22}$$

where $\mathbf{v}_{i0}$ is the initial velocity of particle i. By integration, we obtain:

$$\mathbf{x}_i = \mathbf{v}_i t + \mathbf{x}_{i0} = \frac{\mathbf{a}_i}{2} t^2 + \mathbf{v}_{i0} t + \mathbf{x}_{i0}, \tag{2.23}$$

where $\mathbf{x}_{i0}$ is the initial position of particle i. Equation 2.23 provides a rule to compute the movement of a particle based on its acceleration, velocity and position. The initial positions of the particles are commonly *a priori* given by experimentally determined structural data

or theoretical models. The acceleration that acts on a particle can be directly computed using the empirical potential and the relation of Eq. 2.21:

$$\mathbf{a}_i = -\frac{1}{m_i}\frac{\mathrm{d}V_i}{\mathrm{d}\mathbf{x}_i}. \tag{2.24}$$

The initial velocities of the particles are commonly drawn from a Maxwell-Boltzmann distribution:

$$p(v_{ix}) = \sqrt{\left(\frac{m_i}{2\pi k_B T}\right)} \exp\left(-\frac{1}{2}\frac{m_i v_{ix}^2}{k_B T}\right). \tag{2.25}$$

This distribution provides the probability for an atom with mass m_i to have a velocity v_{ix} in the x direction. k_B denotes the Boltzmann constant and T is the temperature at which the simulation is carried out.

The high dimensionality of the system and the diverse contributions to the empirical potential render it impossible to find an analytical solution to the equations of motion. Hence, it is necessary to use a numerical integration scheme to obtain a solution for the trajectory of a system in its empirical potential. Several of such schemes exist and they are mostly all based on a Taylor series expansion of the position, velocity and acceleration equations for the particle under consideration:

$$\mathbf{x}_i(t+\Delta t) = \mathbf{x}_i(t) + \frac{\mathrm{d}\mathbf{x}_i(t)}{\mathrm{d}t}\Delta t + \frac{1}{2}\frac{\mathrm{d}^2\mathbf{x}_\mathrm{i}(\mathrm{t})}{\mathrm{d}t^2}\Delta t^2 + \ldots \tag{2.26}$$

$$\mathbf{v}_i(t+\Delta t) = \mathbf{v}_i(t) + \frac{\mathrm{d}^2\mathbf{x}_\mathrm{i}(\mathrm{t})}{\mathrm{d}t^2}\Delta t + \frac{1}{2}\frac{\mathrm{d}^3\mathbf{x}_\mathrm{i}(\mathrm{t})}{\mathrm{d}t^3}\Delta t^2 + \ldots \tag{2.27}$$

$$\mathbf{a}_i(t+\Delta t) = \mathbf{a}_i(t) + \frac{\mathrm{d}^3\mathbf{x}_\mathbf{i}(\mathrm{t})}{\mathrm{d}t^3}\Delta t + \frac{1}{2}\frac{\mathrm{d}^4\mathbf{x}_\mathrm{i}(\mathrm{t})}{\mathrm{d}t^4}\Delta t^2 + \ldots \tag{2.28}$$

A popular integrator is the *leap-frog* algorithm [28]. By using the equality

$$\mathbf{v}_\mathbf{i}(t) + \frac{1}{2}\frac{\mathrm{d}^2\mathbf{x}_\mathrm{i}(\mathrm{t})}{\mathrm{d}t^2}\Delta t = \mathbf{v}_i(t+\frac{1}{2}\Delta t) \tag{2.29}$$

we obtain

$$\mathbf{x}_i(t+\Delta t) = \mathbf{x}_i(t) + \mathbf{v}_i(t+\frac{1}{2}\Delta t)\Delta t \tag{2.30}$$

and

$$\mathbf{v}_i(t+\frac{1}{2}\Delta t) = \mathbf{v}_i(t-\frac{1}{2}\Delta t) + \frac{\mathrm{d}^2\mathbf{x}_\mathrm{i}(\mathrm{t})}{\mathrm{d}t^2}\Delta t. \tag{2.31}$$

The algorithm is called "leap-frog" as positions and velocities are "leaping" over each other. It uses positions at time t and velocities at time $t - \frac{1}{2}\Delta t$ at the same time to calculate

updates of position and velocity. The leap-frog algorithm has the strength to be time reversible and symplectic, i.e., it conserves the energy of the simulated system. The leap-frog integrator is the default integrator used by GROMACS [29].

Using the techniques describes so far we are able to perform a classical MD simulation at constant energy, i.e., to generate trajectories that sample from the microcanonical ensemble [1]. However, to obtain ergodic trajectories of the system, and to assure a unique invariant distribution, temperature coupling is needed. For MD simulations many algorithms exist to control the temperature. One of the simplest is the velocity rescaling method [30]. It is based on the fact that the average kinetic energy $\langle V_{kin} \rangle$ of an unconstraint system is given by

$$\langle V_{kin} \rangle = \frac{3}{2} N k_B T = \frac{1}{2} \sum_i m_i |\mathbf{v}_i|^2. \tag{2.32}$$

Using this relation it is possible to derive a scaling factor that adjusts the particle velocities of a system to meet the desired temperature. The difference between the current temperature $T(t)$ and the desired temperature T can be expressed using a scaling factor λ as:

$$\Delta T = \frac{1}{2} \sum_i \frac{2}{3} \frac{m_i (\lambda |\mathbf{v}_i|)^2}{N k_B} - \frac{1}{2} \sum_i \frac{2}{3} \frac{m_i |v_i|^2}{N k_B} \tag{2.33}$$

$$\Delta T = (\lambda^2 - 1) T(t). \tag{2.34}$$

Employing $T = T(t) + \Delta T$ the scaling factor λ can be simply determined via

$$\lambda = \sqrt{\frac{T}{T(t)}}. \tag{2.35}$$

A variation of this scaling scheme is the Berendsen method [31]. Here, the scaling factor is given by

$$\lambda = \sqrt{1 + \frac{\Delta t}{\tau} \left(\frac{T}{T(t)} - 1 \right)}, \tag{2.36}$$

where Δt is the integration time step and τ is the coupling constant. The larger τ, the weaker is the temperature coupling of this thermostat. While these velocity scaling thermostats are very simple and easy to implement, they are not recommended to be used in production simulations as they do not strictly reproduce the correct canonical ensemble. This can cause error when quantities are to be computed that depend on variations in the temperature, e.g., the heat capacity of the system. There are alternative thermostats that yield the right canonical ensemble, e.g., Andersen [32] or Bussi-Donadio-Parrinello

[1] Also called NVE ensemble as it describes systems with constant number of particles (N), constant volume (V) and constant energy (E).

[33] thermostats which are both based on stochastic velocity rescaling. For a discussion of currently available thermostat techniques refer to [34]. In this thesis, Langevin dynamics based thermostat approaches were used. These do in principle generate the right canonical ensemble, but the generated dynamics may be only reversible on slow processes. When using Langevin dynamics Newton's equation of motion is modified by adding a noise and friction term:

$$m_i \frac{\mathrm{d}^2 \mathbf{x}_i}{\mathrm{dt}^2} = -m_i \gamma_i \frac{\mathrm{d}\mathbf{x}_i}{\mathrm{dt}} - \nabla_i V + \sqrt{2\gamma_i k_B T m_i} W_t, \tag{2.37}$$

where γ_i is the friction constant and W_t is a multivariate Wiener process, with $\langle W_t \rangle = 0$ and $\langle W_t W_{t+s} \rangle = s$.

When this dynamics is used as a thermostat for an explicit solvent simulation, γ_i should be chosen around 0.5 ps^{-1}. This choice results in an overall friction that is lower than the internal friction of water, because Langevin dynamics is here not used to replace the water molecules, but rather just to act as a thermostat.

Periodic Boundary Conditions Molecular dynamics simulations are carried out to compute macroscopic properties of a system that usually consists only of a small number of particles. It is therefore of great importance to correctly treat the boundaries of such a small system as they differ substantially from a "bulk" system. The disparity becomes apparent if one imagines a protein simulation in explicit water. The protein is usually put into a box of water molecules, where the box size is chosen such that 1 nm of water surrounds the protein at each face. When this is compared to an experimental condition it becomes clear that the fraction of water molecules that are at an interface is much higher in the simulation. The most popular method to overcome this discrepancy is to use periodic boundary conditions. In doing so the simulated system is copied and repeated infinitely often in space. The simplest geometry used is a cubic box. Starting from a center box the box is repeated into each direction in a space filling manner. Thus, the simulation system has no interface with empty space anymore. Note that for certain systems geometries different volumes might be more appropriate. Other commonly used alternatives are hexagonal prisms or truncated octahedron geometries. By imposing periodic boundary conditions in principle an infinite number of non-bonded interactions would have to be considered. A way to circumvent this is to consider only non-bonded interactions with the closest image particle (*minimum image convention*). This is appropriate for short ranged Van der Waals interactions but can lead to errors in longer ranged electrostatic interactions. For these interactions the commonly used method is Particle Mesh Ewald (PME) [35]. This method is based on Ewald summation [36] were the periodicity is used to replace the real space summation of interaction energies into a summation in Fourier space. The advantage of this procedure is that the infinite sums converge much quicker in that space.

Brownian Dynamics Simulations In molecular dynamics simulations all atoms of a molecular system are generally represented explicitly. This results in a large number of particles for biologically relevant systems, e.g., solvated proteins. Consequently, a huge number of interactions have to be evaluated at each MD simulations step, rending it computationally expensive. Together with the discrepancy between short simulation time step and long timescales on which interesting biological phenomena occur, studying very complex systems by explicit MD is currently beyond the scope using commonly available computational power. This limitation can be overcome by reducing the number of degrees of freedom in the system, i.e., reducing the number of considered particles. A valuable approach in that regard is to separate important from unimportant degrees of freedom and to focus on the important ones. Clearly, the division in important and unimportant degrees of freedom is inherently subjective and linked to the actual research questions to be investigated. However, a prevailing strategy in this context is to replace the solvent molecules in the system by an implicit solvent, i.e., replaying the contributions by individual solvent atoms by a continuum that is described by a noise and friction term. A widely used formulation to describe this dynamics is the Langevin equation as already introduced above (Eq. 2.37):

$$m_i \frac{\mathrm{d}^2 \mathbf{x}_i}{\mathrm{d}t^2} = -m_i \gamma_i \frac{\mathrm{d}\mathbf{x}_i}{\mathrm{d}t^2} - \nabla_i V + \sqrt{2\gamma_i k_B T m_i} W_t, \tag{2.38}$$

where γ_i is the friction coefficient of particle i and W_t is a multivariate Wiener process with $\langle W_t \rangle = 0$ and $\langle W_t W_{t+s} \rangle = s$. In Eq. 2.38 Newton's second law (Eq. 2.21) is extended by a noise term ($\sqrt{2\gamma_i k_B T m_i} W_t$) that accounts for random kicks the water molecules exert on the solute, and a friction ($-m_i \gamma_i \frac{\mathrm{d}\mathbf{x}_i}{\mathrm{d}t}$) term which models the solute's friction in the solvent. The relation of these two terms is given by the fluctuation dissipation theorem [37], which ensures that the energy that is added by the noise term is dissipated again by the friction term. If this would not be the case, the system would heat up and the equilibrium condition would be violated.

Many biophysical questions are related to protein dynamics in solvent. This dynamics can be well described by Brownian dynamics, which is also referred to as Smoluchowski dynamics or Langevin dynamics in the high friction limit. The reason why this dynamics can be regarded as inertia free is grounded in the fact that the velocity of a protein immersed in a dense solvent will relax very quickly after it got a kick by a solvent molecule. When the observation time step is bigger as this fast relaxation time it appears as if the protein performs a random walk or Brownian motion. The Brownian dynamics equations can be obtained from the Langevin equation by scaling the friction γ and the time t by introducing a parameter ϵ. The friction is scaled via $\frac{\gamma}{\epsilon}$:

$$m_i \frac{\mathrm{d}^2 \mathbf{x}_i}{\mathrm{d}t^2} = -m_i \frac{\gamma_i}{\epsilon} \frac{\mathrm{d}\mathbf{x}_i}{\mathrm{d}t} - \nabla_i V + \sqrt{2\frac{\gamma_i}{\epsilon} k_B T m_i} W_t \tag{2.39}$$

Scaling also the time according to $t = \epsilon t$ leads:

$$m_i \epsilon^2 \frac{\mathrm{d}^2 \mathbf{x}_i}{\mathrm{d}t^2} = -m_i \gamma_i \frac{\mathrm{d}\mathbf{x}_i}{\mathrm{d}t} - \nabla_i V + \sqrt{2\gamma_i k_B T m_i} W_t. \tag{2.40}$$

Further letting $\epsilon \to 0$ we obtain:

$$\frac{\mathrm{d}\mathbf{x}_i}{\mathrm{d}t} = -\frac{1}{m_i \gamma_i} \nabla V_i(\mathbf{x_i}) + \sqrt{2\frac{k_B T}{\gamma_i m_i}} W_t \tag{2.41}$$

By using the Einstein relation $D = \frac{k_B T}{\gamma m_i}$, we obtain the Brownian dynamics equation:

$$\frac{\mathrm{d}\mathbf{x}_i}{\mathrm{d}t} = -\frac{D_i}{k_B T} \nabla V_i(\mathbf{x_i}) + \sqrt{2D_i} W_t \tag{2.42}$$

D is the diffusion constant of the particle, i.e., protein. For a precise description D would have to be a tensor quantity and express correlations as the accelerated solvent molecules might affect the movement of other solute atoms. For work on the inclusion of these hydrodynamic interactions refer to [38] and [39]. Throughout this work we assume that the simulated particles are rigid and their separation is large enough such that hydronamic effects can be neglected. To obtained Brownian dynamics simulation trajectories for a particle the Euler–Maruyama method can be used:

$$\mathbf{x}_i(t + \Delta t) = \mathbf{x_i}(t) + -\frac{D_i}{k_B T} \nabla V_i(\mathbf{x_i}) \Delta t + \sqrt{2D_i \Delta t} \mathcal{N}(0, 1), \tag{2.43}$$

where Δt is the simulation time step and $\mathcal{N}(0, 1)$ is a normally distributed random number. The potential V is usually assumed to arise from electrostatic interactions between the particles and can be for example computed via Eq. 2.14.

In addition to significantly reducing the number of considered particles, this approximation allows to use a larger simulation time step due to the smoother nature of the used potential function. In summary, while providing less detail implicit solvent models allow to access time scales for very large systems that are out of reach for current all atom simulations.

2.2.3 Free Energy Change and Umbrella Sampling

Free energy is an important quantity to describe the behavior of macrostates in a molecular system. Changes in the macrostate of a molecular system, e.g., a protein changing from unfolded state to folded state can be characterized by accompanying changes in the free energy. However, calculating the free energy difference and height of the energetic barrier between states is often difficult. While the considered states might themselves have a low free energy, i.e., likely to be observed in a simulation, they are often separated by high energetic barriers. As a consequence the probability of observing transitions from one state to another is very small, which renders the estimation of the correct barrier

height from simulations challenging. A technique to overcome this sampling problem is umbrella sampling [40]. Here, by *a priori* assuming a reaction coordinate between states, a number of biasing potentials is introduced that guide the MD sampling along this reaction coordinate to improve the sampling in energetically unfavorable regions. In this thesis we used umbrella sampling to calculate free energy differences between the bound and unbound state of two interacting proteins (Chapter 5). In the following we briefly introduce the theoretical foundations of umbrella sampling.

The free energy profile (also referred to as potential of mean force) along a certain reaction coordinate ξ, e.g., a dihedral angle, or the center of mass distance between two proteins can be expressed as [41]:

$$F(\xi) = -\beta^{-1} \ln[P(\xi)] + const., \tag{2.44}$$

where β is the inverse temperature $1/k_BT$ and $P(\xi)$ the probability distribution function along coordinate ξ, which can be calculated from a Boltzmann weighted average:

$$P(\xi) = \frac{\int \exp[-\beta V(\mathbf{x})]\delta(\xi'(\mathbf{x}) - \xi)d\mathbf{x}}{\int \exp[-\beta V(\mathbf{x})]d\mathbf{x}}, \tag{2.45}$$

where $V(\mathbf{x})$ is the total energy of the system as function of its coordinates.

$P(\xi)$ can in principle be obtained from MD simulations. However, this might be impractical as reaction coordinates often contain high energy barriers which will be only poorly sampled by standard MD and thus lead to large errors in $F(\xi)$. To circumvent that problem the potential can be biased in order to force the simulation also to sample areas that are energetically unfavorable. This idea has been termed umbrella sampling and was originally developed by Torrie and Valleau [40].

Consider the following biased potential:

$$\tilde{V}(\xi) = V(\xi) + \omega(\xi). \tag{2.46}$$

Here, another function ω is added to V that biases the potential. In practical cases omega is often chosen to be a harmonic function. When for example $\omega(\xi) = K(\xi - \xi_0)^2$, with K being the force constant, is used as function, positioning ξ_0 at an energy barrier will cause the simulation to increase the sampling in this region, as deviations from ξ_0 are penalized.

The probability for ξ is altered by the biasing potential and it is hence necessary to correct for the biasing term in the free energy calculations. The probability distribution of ξ considering the biased potential is given by:

$$P^{\text{Umbrella}}(\xi) = \frac{\int \exp[-\beta V(\mathbf{x}) + \omega(\xi'(\mathbf{x}))]\delta(\xi'(\mathbf{x}) - \xi)d\mathbf{x}}{\int \exp[-\beta V(\mathbf{x}) + \omega(\xi'(\mathbf{x}))]d\mathbf{x}}. \tag{2.47}$$

As the biasing potential ω depends only on ξ, Eq. 2.47 can be reformulated to:

$$P^{\text{Umbrella}}(\xi) = \exp[-\beta\omega(\xi)] \times \frac{\int \exp[-\beta V(\mathbf{x})]\delta(\xi'(\mathbf{x}) - \xi)d\mathbf{x}}{\int \exp[-\beta V(\mathbf{x}) + \omega(\xi'(\mathbf{x}))]d\mathbf{x}}. \tag{2.48}$$

Extending this equation with Eq. 2.45, the unbiased probability distribution $P(\xi)$ is obtained:

$$\begin{aligned} P(\xi) &= P^{\text{Umbrella}}(\xi)\exp[\beta\omega(\xi)] \times \frac{\int \exp[-\beta V(\mathbf{x}) + \omega(\xi'(\mathbf{x}))]d\mathbf{x}}{\int \exp[-\beta V(\mathbf{x})]d\mathbf{x}} && (2.49)\\ &= P^{\text{Umbrella}}(\xi)\exp[\beta\omega(\xi)] \times \frac{\int \exp[-\beta V(\mathbf{x})]\exp[-\beta\omega(\xi'(\mathbf{x}))]d\mathbf{x}}{\int \exp[-\beta V(\mathbf{x})]d\mathbf{x}} && (2.50)\\ &= P^{\text{Umbrella}}(\xi)\exp[\beta\omega(\xi)]\left\langle\exp[-\beta\omega(\xi)]\right\rangle. && (2.51) \end{aligned}$$

The Helmholtz free energy of along an umbrella biased coordinate ξ is thus given by:

$$F(\xi) = -\beta^{-1}\ln[P^{\text{Umbrella}}(\xi)] - \omega(\xi) + const. \tag{2.52}$$

$P^{\text{Umbrella}}(\xi)$ can be readily obtained from biased molecular dynamics simulations and $\omega(\xi)$ can be analytically calculated. Often the use of one umbrella potential is not enough to ensure a sampling of the whole reaction coordinate. Hence it is common to use multiple umbrellas / biasing potentials along the reaction coordinate. By using this procedure one obtains different values of $A_i(\xi)$ for each of the umbrella windows i. Note that the additive constants also differ for each window. In order to obtain these constants and thus to reconstruct the free energy along the entire coordinate additional methods have to be employed. A popular method in this context is the weighted histogram analysis method (WHAM) [42], which was also used in this thesis.

2.3 Markov State Models of Dynamical Processes

The dynamics of a molecular system can be very complex. By employing the methods described above it is in principle possible to generate trajectory data that contains all the relevant information of the system dynamics. However, the large degree of freedom, i.e., number of positions and velocities, present in the system is usually too large to be directly interpreted by the human mind. Hence, a commonly undertaken approach is to project this high dimensional data onto a few observables to analyze and extract the relevant processes. For example, protein folding is often studied by measuring the RMSD change of the system with respect to a starting structure. Unfortunately, this approach can have a number of pitfalls. By projecting the system to a small number of observables energetic barriers can become hidden or kinetic features might be missed as kinetically distinct structures are lumped together. Nevertheless, some dimension reduction is inevitable to gain insights into the system dynamics.

The least subjective approach is to partition the entire configuration space of the system into a finite number of sets and to project the dynamics onto these sets. Even though such an approach may still be prone to hide important kinetic aspects, it can in principle reveal all relevant details by choosing a fine enough discretization. The models that are obtained from such a procedure are called Markov state models (MSMs) or simply Markov models. Given n discrete sets that represent the whole configuration space of the system, the underlying kinetics is modeled by a $n \times n$ transition matrix. An entry of this matrix represents the conditional probability of the process to make a transition from one set i to another set j within a time τ, given the process is in set i. MSMs have a number of neat properties that make them appealing to be used in the analysis of complex dynamical systems. First of all, MSMs replace the single trajectory view of dynamics with an ensemble view, i.e., they describe dynamics of the entire ensemble rather than only the dynamics of a few observations. The fact that they use only conditional probabilities allows to estimate individual transition probabilities from short trajectories that are specifically started in a certain set, rather then to use a very long ergodic trajectory. This circumstance allows in practice to generate a huge number of trajectories in parallel to obtain a description of the complete dynamics of the whole system. In fact, it is possible to recover timescales that are much longer than the individual short trajectories by using a MSM. An example where these properties are utilized is the Folding@Home [43] project where worldwide distributed computing is used to generate a vast amount of short trajectories to parameterize MSMs. Furthermore, recent advances [44, 45] allow to compute statistical uncertainties in the entries of the MSM transition matrix permitting the derivation of an adaptive MSM construction scheme. Here, more sampling is performed or discrete sets are introduced in areas that show a poor statistics. Finally, MSMs allow the straightforward computation of statistical quantities: equilibrium properties such as the stationary distribution or free energy differences between conformations and kinetic properties such as metastable conformations or the transition pathway distribution of the system. An example where the latter property was used to study folding pathways of proteins can be found in [46].

Based on [47] a brief outline of theoretical foundations of MSMs is given and it is described how MSMs can be obtained from MD simulations. Further, it will be outlined how to test the validity of a MSM with respect to the input and how to compute transition pathways.

After having introduced Markov processes based on the conception of an individual trajectory in the beginning of Section Microscopic Dynamics we proceed to an ensemble perception of the systems dynamics, i.e., instead of considering only one trajectory we now take into account the ensemble of all possible trajectories of a system.

Let us consider an ensemble of molecular systems that is at time t distributed with probability density $p_t(\mathbf{x})$ in state space Ω. Assuming further an underlying ergodic process with stationary distribution $\mu(\mathbf{x})$, $p_t(\mathbf{x})$ will change and eventually reach $\mu(\mathbf{x})$ for $t \to \infty$.

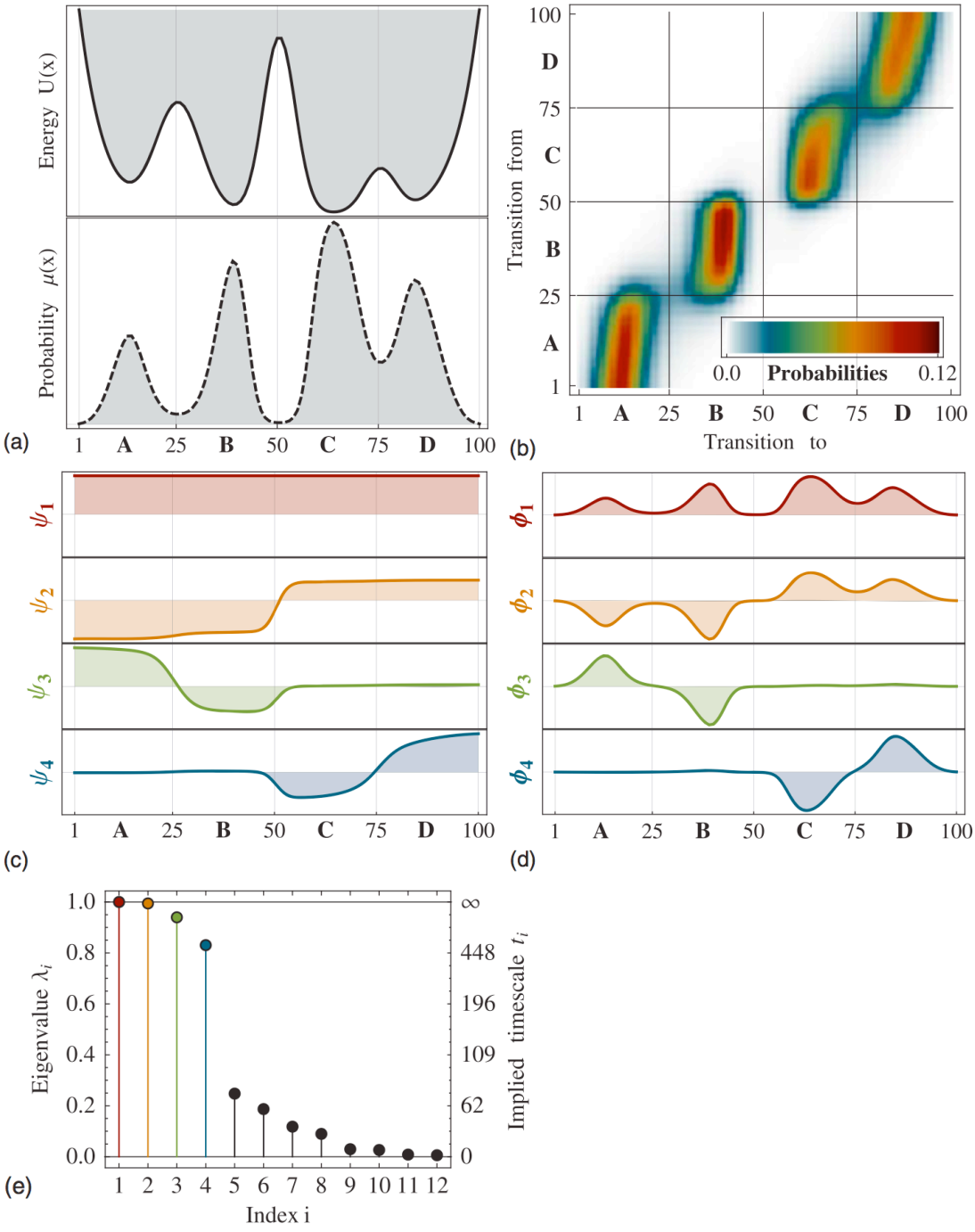


Figure 2.3: a) Potential energy function with four local minima and corresponding stationary probability density $\mu(x)$. b) Density plot of a transfer operator that models a diffusion dynamics in the depicted potential energy landscape on the domain $\Omega = [1 \dots 100]$. Red indicates high transition probability, white transition probability close to zero. The block-diagonal structure indicates a high probability to move inside the potential minima (metastable basins) and a very low probability to move between them. c) The four dominant eigenfunctions of the transfer operator, $\psi_1, \dots, \psi_4$. ψ_1 is associated with the stationary process, ψ_2 to a transition between $\{A, B\} \leftrightarrow \{C, D\}$, ψ_3 to a transition between $A \leftrightarrow B$ and ψ_4 to a transition between $C \leftrightarrow D$. d) The four dominant eigenfunctions of the transfer operator, $\phi_1, \dots, \phi_4$, weighted with the stationary probability density. e) Eigenvalues of the transfer operator, mind the clearly visible gap between the four metastable processes ($\lambda_i \approx 1$) and the fast processes. Figure and description adopted from [47].

The change of probability distribution $p_t(\mathbf{x})$ in time τ can be quantified using the transition probability density (Eq. 2.15). Precisely the change $p_t(\mathbf{x})$ to $p_{t+\tau}(\mathbf{x})$ is given by the action of the continuous propagator Q:

$$p_{t+\tau}(\mathbf{y}) = Q(\tau) \circ p_t(\mathbf{y}) = \int_{\mathbf{x}\in\Omega} p_t(\mathbf{x}) p(\mathbf{x}, \mathbf{y}; \tau) d\mathbf{x}. \tag{2.53}$$

The application of this propagator to a density p_t results in a modified $p_{t+\tau}$ that will be closer to $\mu(\mathbf{x})$ and reach it in finite time. It is computed by integrating over all transitions $\mathbf{x} \to \mathbf{y}$ that are possible within τ and weighting them by the probability density of $\mathbf{x}$. While this is a straightforward description an alternative description with more appealing mathematical properties is the transfer operator $\mathcal{T}(\tau)$:

$$u_{t+\tau}(\mathbf{y}) = \mathcal{T}(\tau) \circ u_t(\mathbf{y}) = \frac{1}{\mu(\mathbf{x})} \int_{\mathbf{x}\in\Omega} \mu(\mathbf{x}) u_t(\mathbf{x}) p(\mathbf{x}, \mathbf{y}; \tau) d\mathbf{x}, \tag{2.54}$$

where $p_t(\mathbf{x})$ and $u_t(\mathbf{x})$ are related via the stationary density $\mu(\mathbf{x})$ via:

$$p_t(\mathbf{x}) = \mu(\mathbf{x}) u_t(\mathbf{y}). \tag{2.55}$$

The above descriptions propagate probability densities in discrete time steps τ. It is also possible to define time continuous propagators, e.g., the infinitesimal generator that encodes the time continuous dynamics of a stochastic process and is related to the Fokker-Planck equation [48].

The defined operators have some general properties that will be of use later to introduce specific features of MSMs:

1. The operators fulfill the Chapman-Kolmogorov equation:

$$p_{t+k\tau}(\mathbf{x}) = [Q(\tau)]^k \circ p_t(\mathbf{x}) \tag{2.56}$$
$$u_{t+k\tau}(\mathbf{x}) = [\mathcal{T}(\tau)]^k \circ u_t(\mathbf{x}) \tag{2.57}$$

   $[\cdot]^k$ refers to the k-fold application of the operator to the density. Hence the Chapman-Kolmogorov equation states that the operators can be used to propagate the evolutions of the densities to arbitrary long times $k\tau$.

2. $Q(\tau)$ has eigenfunctions $\Phi_i(\mathbf{x})$ and eigenvalues λ_i:

$$Q(\tau) \circ \Phi_i(\mathbf{x}) = \lambda_i \Phi_i(\mathbf{x}). \tag{2.58}$$

   $\mathcal{T}(\tau)$ has the same eigenvalues but eigenfunctions $\Psi_i(\mathbf{x})$:

$$\mathcal{T}(\tau) \circ \Psi_i(\mathbf{x}) = \lambda_i \Psi_i(\mathbf{x}). \tag{2.59}$$

3. For reversible dynamics all eigenvalues λ_i are real-valued and lie in the interval $-1 < \lambda_i \leq 1$ [49] and the eigenfunctions are related via the stationary density $\mu(\mathbf{x})$:

$$\Phi_i(\mathbf{x}) = \mu(\mathbf{x})\Psi_i(\mathbf{x}). \tag{2.60}$$

4. Both operators possess a continuos spectrum of eigenvalues. In the following we will consider the m dominant eigenvalues and sort them in descending order. For both operators there is always an eigenvalue $\lambda = 1$, the order is hence:

$$\lambda_1 = 1 > \lambda_2 \geq \lambda_3 \cdots \geq \lambda_m. \tag{2.61}$$

   The eigenfunction that is associated with λ_1 is the stationary distribution:

$$Q(\tau) \circ \mu(\mathbf{x}) = \mu(\mathbf{x}) = \Phi_i(\mathbf{x}) \tag{2.62}$$

   The corresponding function for $\mathcal{T}(\tau)$ is:

$$\mathcal{T}(\tau) \circ \mathbb{I} = \mathbb{I} = \Phi_1(\mathbf{x}), \tag{2.63}$$

   as it can be seen from Eq. 2.60.

The eigenfunction/eigenvalues of the operators provide a wealth of information about the dynamical processes present in a system. To extract this information it is instructive to consider the following elaboration: The entire dynamics of a molecular system can be decomposed into a superposition of m individual slow dynamical processes and a number of remaining fast processes. With respect to $\mathcal{T}(\tau)$ this can be expressed as:

$$\begin{aligned} u_{t+\tau k} &= \mathcal{T}_{slow}(k\tau) \circ u_t(\mathbf{x}) + \mathcal{T}_{fast} \circ u_t(\mathbf{x}), & (2.64)\\ &= \sum_{i=1}^{m} \lambda_i^k \langle u_t, \phi_i \rangle \Psi_i(\mathbf{x}) + \mathcal{T}_{fast} \circ u_t(\mathbf{x}), & (2.65)\\ &= \sum_{i=1}^{m} \lambda_i^k \langle u_t, \Psi_i \rangle_\mu \Psi_i(\mathbf{x}) + \mathcal{T}_{fast} \circ u_t(\mathbf{x}). & (2.66) \end{aligned}$$

The weighted scalar product $\langle u_t, \Psi_i \rangle_\mu$ is defined as $\langle f, g \rangle_\mu = \int \mu(\mathbf{x}) f(\mathbf{x}) g(\mathbf{x}) d\mathbf{x}$. $\mathcal{T}_{slow}$ contains the dominant slowly decaying processes and $\mathcal{T}_{fast}$ contains infinitely many fast processes with $\lambda_i < \lambda_m$ that are usually not of interest. Applying this decomposition requires that the subspaces of $\mathcal{T}_{slow}$ and $\mathcal{T}_{fast}$ are orthogonal. It can be shown that this is indeed the case when detailed balance is fulfilled.

Using this decomposition it is possible to physically interpret the slow dynamics as a superposition of processes that can be associated to eigenfunctions Φ_i / Ψ_i with corre-

sponding eigenvalue λ_i. For an illustrations of these eigenfunctions and their corresponding processes refer to Figure 2.3 c), d).

The closer an eigenvalue is to 1, the slower decays the corresponding process; the closer an eigenvalue is to 0, the faster it decays. The λ_i can thus be interpreted as a physical timescale which indicates how quickly the process transports probability density towards the equilibrium distribution. The timescale of a process i can be computed via:

$$t_i = -\frac{\tau}{\ln \lambda_i}, \tag{2.67}$$

which is also called the ith implied timescale of the system. With this equation it becomes apparent that, if the λ_m eigenvalues show a large separation, the system has different dynamical processes that act simultaneously on different timescale. For example, consider a system that switches between two stable states. The corresponding eigenvalues are $\lambda_1 = 1$, $\lambda_2 \approx 1$, $\lambda_{3,\ldots} \ll \lambda_2$. For an additionally introduced intermediate state that is as stable as the two other states, the eigenvalues would change to $\lambda_2 \approx \lambda_3$. For an example with four metastable states refer to Figure 2.3.

At this point the theoretical foundations of a Markov process and some of its properties are defined. How can one make use of this theory in practice to analyze MD simulation data? As we have outlined in the beginning, the state space is represented by a number of discrete sets. In the following the discretization and discrete state space dynamics is described.

Discrete State Space Dynamics A common way to discretize the state space is to use a Voronoi tessellation [50]. Let us assume the state space is dissected into n sets $S = \{S_1, S_2, \ldots, S_n\}$ which completely partition the state space, i.e., $\Omega = \bigcup_{i=1\ldots n} S_i$ and are non-overlapping, i.e., $S_i \bigcap S_j = \emptyset \; \forall_{i \neq j}$. Then the former continuous stationary probability density $\mu(\mathbf{x})$ is now discrete and declared by π_i. It denotes the probability to find the system in set S_i and is related to the continuous stationary probability density via:

$$\pi_i = \int_{\mathbf{x} \in S_i} \mu(\mathbf{x}) d\mathbf{x}. \tag{2.68}$$

The discrete analogue of the continuous transfer operator $\mathcal{T}(\tau)$ is a row stochastic transition matrix $T(\tau) \in \mathbb{R}^{n \times n}$. The matrix element $T_{ij}(\tau)$ represents the conditional probability to find the system in state j at time $t + \tau$, given it was in state i at time t. Mathematically

this is expressed as

$$T_{ij}(\tau) = \mathbb{P}[\mathbf{x}(t+\tau) \in S_j \,|\, \mathbf{x}(t) \in S_i] \tag{2.69}$$

$$= \frac{\mathbb{P}[\mathbf{x}(t+\tau) \in S_j \cap \mathbf{x}(t) \in S_i]}{\mathbb{P}[\mathbf{x}(t) \in S_i]} \tag{2.70}$$

$$= \frac{\int_{\mathbf{x}\in S_i} \mu_i(\mathbf{x}) p(\mathbf{x}, S_j; \tau) d\mathbf{x}}{\int_{\mathbf{x}\in S_i} \mu_i(\mathbf{x}) d\mathbf{x}}, \tag{2.71}$$

where $\mu_i(\mathbf{x})$ denotes the local stationary density:

$$\mu_i(\mathbf{x}) = \begin{cases} \frac{\mu(\mathbf{x})}{\pi_i} & \mathbf{x} \in S_i \\ 0 & \mathbf{x} \notin S_i \end{cases}. \tag{2.72}$$

Note that the integrals run over local properties, i.e., points in set S_i. It is hence only necessary to obtain the local equilibrium distribution and transition probability densities out of S_i in order to estimate matrix entries $T_{ij}(\tau)$. Based on a number of trajectories of length τ that are started according to μ_i it is hence possible to construct a discrete transfer operator that accounts for the whole system dynamics, including timescales greater than τ.

The functions that are transported by the continuous operator $\mathcal{T}(\tau)$ can be related to vectors that are multiplied to the discrete transfer operator $T(\tau)$. Suppose that $p(t) \in \mathbb{R}^n$ is a column vector whose elements denote the probability to be within any set j at time t. The change in this vector after time τ can be expressed as:

$$p_j(t+\tau) = \sum_{i=1}^{n} p_i(t) T_{ij}(\tau), \tag{2.73}$$

or in matrix notation:

$$p^T(t+\tau) = p^T(t) T(\tau). \tag{2.74}$$

The discrete stationary density is thus given by:

$$\pi^T = \pi^T T(\tau). \tag{2.75}$$

The equations presented so far are exact and contain no error, provided $T(\tau)$ has been estimated based on probability densities obtained at lag time τ. However, when the transfer operator is used to propagate long time dynamics, i.e.:

$$p^T(t+k\tau) \approx p^T(t) T^k(t), \tag{2.76}$$

the result will be only approximative. This is a consequence of the state space discretization

which results in a loss of the Markov property. To estimate the discrepancy between propagated density and density observed in the input trajectory data the Chapman-Kolmogorov test can be used, as described in the paragraph "Chapman-Kolmogorov Test".

Relation between Transition Matrix and Rate Matrix In many chemical physics applications of Markov state models a rate matrix $\mathbf{K}$ is commonly used instead of transition matrix $\mathbf{T}$. The transition probabilities t_{ij} between states i and j are replaced by transition rates $k_{ij} \geq 0$ in the off-diagonals and the diagonal elements k_{ii} are given by the negative row sum $k_{ii} = \sum_j k_{ij}$. Using $\mathbf{K}$ the change in probability $p(t)_i$ to be in set i is described by a Master equation:

$$\begin{aligned} \frac{dp_i(t)}{dt} &= \sum_{i \neq j} p_j(t) k_{ji} - \sum_{i \neq j} p_i(t) k_{ij} && (2.77) \\ &= \sum_{j}^{n} p_j(t) k_{ji} && (2.78) \\ \frac{d\mathbf{p}(t)}{dt} &= \mathbf{p}^T(t)\mathbf{K} && (2.79) \end{aligned}$$

The solution of Equation 2.79 for a time step τ is given by

$$\mathbf{p}^T(t+\tau) = \mathbf{p}^T(t) \exp(\tau \mathbf{K}), \tag{2.80}$$

where $\exp(\cdot)$ is the matrix exponential. Using $\mathbf{p}^T(t+\tau) = \mathbf{T}(\tau)\mathbf{p}(t)$ this provides the relation between rate and transition matrices:

$$\mathbf{T}(\tau) = \exp(\tau \mathbf{K}). \tag{2.81}$$

Estimation

In the following, it is explained how the discrete analogue of the continuous transfer operator is obtained from molecular dynamics data in practice. Suppose an equilibrium trajectory that contains the temporal sequence of N system configurations which have been stored at fixed time intervals Δt:

$$X = \{\mathbf{x}_1 = \mathbf{x}(t=0), \mathbf{x}_2 = \mathbf{x}(t=\Delta t), \ldots, \mathbf{x}_N = \mathbf{x}(t=(N-1)\Delta t)\}. \tag{2.82}$$

Assume further a discrete state space which has been defined such that each of the system configurations can be assigned to one of the discrete states, i.e., $\mathbf{x}_k \in S_i \rightarrow S_k = i$. The trajectory information can hence be stored as a sequence of discrete states $S_1, \ldots, S_N$. Given such a discrete trajectory a transition count matrix $C^{obs}(\tau)$ can be obtained by calculating each entry $C^{obs}_{ij}(\tau)$ as:

$$C_{ij}^{obs}(\tau) = C_{ij}^{obs}(m\Delta t) = |\{k \in \{1, \ldots, N-m\} \,|S_k = i \wedge S_{k+m} = j\}|\,, \tag{2.83}$$

where τ is the lag time at which is counted. Note that here τ has to be chosen such that it is a whole number multiple of Δt. $C_{ij}^{obs}(\tau)$ denotes the number of transitions from i to j in lag time τ that are observed in the given trajectory. In the literature this counting scheme is referred to as "overlapping window counting". While this scheme has the advantage that all of the available trajectory information is used to estimate the count matrix, it has the disadvantage that these counts are statistically dependent. This has to be taken into account when the transition matrix and the distribution of transition matrices is estimated from the count matrix. The dependence of the overlapping window count method will indeed lead to asymptotically correct expectations of transition probabilities but can lead to errors in the distribution. This in particular relevant for example if variances of individual transition probabilities are to be estimated. To obtain a correct transition matrix distribution an independent counting scheme must be employed. Here, the trajectory information is only evaluated at every $m\Delta t$ step.

Regardless of the counting scheme used, the discrete transition operator $T(\tau)$ can be estimated based on the count matrix via:

$$T_{ij}(\tau) = \frac{c_{ij}}{\sum_j c_{ij}.} \tag{2.84}$$

Note that this estimator for T does not necessarily fulfill the detailed balance condition ($\pi_i T_{ij} = \pi_j T_{ji}$) even if the underlying dynamics is in equilibrium. However, it is possible to obtain estimates of T that surely fulfill the detailed balance condition by using iterative algorithms. See [16] or Algorithm 1 in [47] for details.

Chapman-Kolmogorov Test In the following, we introduce the Chapman-Kolmogorov test to estimate the long time propagation error of the MSM with respect to the input data. A possible way to estimate the long time propagation error is to measure the deviation between a MSM propagated probability density and a probability density directly estimated from the simulation data. In particular it is tested if the Chapman-Kolmogorov condition

$$T(\tau)^k \approx T(k\tau) \tag{2.85}$$

holds within statistical uncertainty. $T(\tau)^k$ is the k-times propagated transition matrix that is estimated from a trajectory at lag time τ. $T(k\tau)$ is the transition matrix directly estimated at lag time $k\tau$. Testing this relation for each of the entries might involve large uncertainties due to a potentially small number of observations. Additionally, it might be cumbersome to compare $n \times n$ elements. It has hence been suggested to test the condition for a predefined set of states A. The general idea is to test how much of an initially

defined probability distribution in A is left after propagating it using $T(\tau)^k$ and $T(k\tau)$, respectively. To carry out this test the following procedure has been proposed:

Let w_i^A denote the stationary probability distribution of state i that is restricted to a set A:

$$w_i^A = \begin{cases} \frac{\pi_i}{\sum_{j\in A}\pi_j} & i \in A \\ 0 & i \notin A \end{cases}, \tag{2.86}$$

where π is the overall stationary probability distribution. Using this conditional density a "relaxation" like experiment can be performed: Therefore w^A is considered as the initial probability vector and it is tested how much of this initial probability is left after $k\tau$ time using (i) the MD trajectory and (ii) the MSM model.

The MD based time-dependent probability density can be estimated via:

$$p_{MD}(A, A; k\tau) = \sum_{i\in A} w_i^A p_{MD}(i, A; k\tau), \tag{2.87}$$

where $p_{MD}(i, A; k\tau)$ is the trajectory based estimate of Eq. 2.15. Using the MSM parameterized at lag time τ the probability density $p_{MSM}(A, A; k\tau)$ is given by:

$$p_{MSM}(A, A; k\tau) = \sum_{i\in A} \left[(w^A)^T T^k(\tau)\right]. \tag{2.88}$$

Comparing these two estimates the Markov model can be tested for its correctness in long time propagation by examining how well the equality:

$$p_{MD}(A, A; k\tau) = p_{MSM}(A, A; k\tau) \tag{2.89}$$

holds. This equality is not to be expected to hold exactly as both estimates have an associated error. For the MD based estimate it is a statistical error as only a finite number of observations are made to estimate true transition probabilities. The standard error of $p_{MD}(A, A; k\tau)$ can be computed as:

$$\epsilon_{MD}(A, A; k\tau) = \sqrt{k\frac{p_{MD}(A, A, k\tau) - [p_{MD}(A, A, k\tau)]^2}{\sum_{i\in A}\sum_{j=1}^{n} C_{ij}^{obs}(k\tau)}}. \tag{2.90}$$

Uncertainties associated with $P_{MSM}(A, A, k\tau)$ can also be computed, for details refer to [47]. However, if the test reveals that the estimates of p_{MSM} are within the error of p_{MD} the MSM is consistent with the MD data, as the expectation value of p_{MD} is fix and not affected by uncertainty considerations.

2.4 Transition Path Theory

Having defined a Markov state model for the molecular system under investigation gives access to various properties, e.g., the relaxation timescales and associated dynamical processes as well as metastable sets and the long time behavior of the system. However, for many applications it is desirable to study specific transitions of the system. In protein folding for example one is interested in the transition from an unfolded peptide chain to the native protein conformation. Similarly in the study area of protein interactions the transition of a protein-protein / protein-ligand system from an unbound to a bound state is of interest.

A mathematical method that allows the extraction of specific transition information from a MSM is Transition Path Theory (TPT) [51, 52, 53]. TPT describes statistical quantities of *reactive pathways* between two disjoint subsets of states, in the following termed A and B. Reactive pathways, in the TPT sense, are defined as all the parts of an hypothetical, infinitely long, phase space trajectory that leave set A and reach set B without revisiting set A in between (see Figure 2.4). In the sense of the above examples, reactive pathways would describe the folding and association transitions of the system under consideration. In this scenario, set A would contain the unfolded / unbound states and set B the folded/bound state(s) of the system.

The key quantity of TPT is the committor probability q_i^+. It denotes the conditional probability that the system, given it is in state i, will next reach set B rather than set A. When for example applied in the context of protein-ligand association, q_i^+ denotes the probability that the ligand will reach the binding site of the protein from a given position rather than reaching a defined unbound region. By definition q_i^+ is 1 for all states in set B ($q_i^+ = 1\ \forall i \in B$) and 0 for all states in set A ($q_i^+ = 0\ \forall i \in A$). The committor values for the remaining intermediate states i that are not in A or B can be computed by solving the following system of equations:

$$-q_i^+ + \sum_{k \in S \setminus (A \bigcup B)} T_{ij} q_k^+ = -\sum_{k \in B} T_{ik}, \tag{2.91}$$

where T_{ij} denotes the transition probability from state i to j as given by the MSM model.

A related quantity is the backward committor probability q_i^-. It denotes the conditional probability that the system, when in state i, came from set A rather then from set B. For systems in equilibrium this probability is simply:

$$q_i^- = 1 - q_i^+. \tag{2.92}$$

The transition probabilities T_{ij} used in Eq. 2.91 to compute the committor probabilities contain contributions from all trajectory fragments, including fragments that leave A and

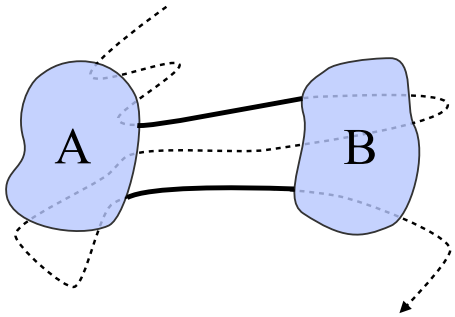


Figure 2.4: Transition Path Theory describes statistical properties of reactive/productive fragments (thick solid line) of an hypothetical infinitely long trajectory (dotted line). A fragment is thought to be reactive/productive if it directly leads from subset A to subset B.

return to A without reaching B, or fragments that go to from B to A. As we are interested in describing statistical properties of *reactive pathways* that go directly from A to B, the remaining transitions have to be neglected. Formally this quantity can be expressed as:

$$q_i^- T_{ij} q_j^+. \tag{2.93}$$

Further the reactive probability flux between two states i and j that contributes to the transition $A \to B$ is given as:

$$f_{ij} = \pi_i q_i^- T_{ij} q_j^+, \tag{2.94}$$

where π_i is the stationary probability to find the system in state i. Note that f_{ij} accounts also for recrossing events of the reactive trajectory, i.e., $A \to \cdots \to i \to j \to i \to j \to \cdots \to B$. To compute the net reactive probability flux from A to B, the reactive flux f_{ji} that is associated with recrossings is subtracted from the forward flux f_{ij}:

$$f_{ij}^+ = \max\{0, f_{ij} - f_{ji}\}. \tag{2.95}$$

f_{ij}^+ defines a network of fluxes leaving set A and entering set B. The total expected probability flux from A to B can be computed from this network and is given by:

$$F_{AB} = \sum_{i \in A} \sum_{j \notin A} f_{ij}^+ = \sum_{i \notin B} \sum_{j \in B} f_{ij}^+. \tag{2.96}$$

Note that the flux out of A equals the flux into B, i.e., the flux network f_{ij}^+ is flux-conserving. The $A \to B$ transition rate k_{AB} of the system can be directly computed from Eq. 2.96. However, it needs to be taken into account, that the system first has to move back to A to do another transition to B. This is achieved by dividing F_{AB} by a correction term [46] yielding:

$$k_{AB} = \frac{F_{AB}}{\tau \sum_{i \in S} \pi_i q_i^-}. \tag{2.97}$$

From the flux network f_{ij}^+ it is possible to determine both the system transition pathways and their relative probabilities as well as transition rates of the system. TPT is hence a versatile tool that allows the in detail analysis of information provided by MSMs. Note that TPT can also be used to analyze dynamics given by the rate matrix $\mathbf{K}$ (Eq. 2.81). The expressions are similar to the ones given above and can be found in Chapter 4.

3 Modulation of a Ligand's Energy Landscape and Kinetics by the Chemical Environment

3.1 Introduction

The chemical or biological roles of molecules depend on the conformations they can access and on the transition rates, i.e. the kinetics between these conformations. The conformations and kinetics of biomolecules are strongly affected by their chemical environment, such as the type of solvent or the presence of ions. For example, it has been shown that i) the stability of proteins is affected by small changes in the polarity of the solvent [54], ii) the folding kinetics of loop-forming peptides depend on the viscosity of the solvent [55], and iii) the Mg^{2+} concentration affects the folding dynamics [56, 57] as well as the conformation of the binding pocket [58] in ribozymes.

Understanding which effect the chemical environment has on the conformations and kinetics of biomolecules is one of the core interests of biochemistry and physical chemistry. Being able to model this relationship is crucial in order to comprehend how biochemical steps are driven in a cell. Here, we construct a computational strategy towards this end. In particular, we are interested in understanding how a substrate that specifically binds to an enzyme can be stabilized in its binding conformation. From a chemical point of view, such an understanding may be useful in order to mimic efficient biological catalysis by designing a synthetic scaffold.

The conformations and kinetics of a molecule may be understood to arise from its energy landscape. Such an energy landscape may be "sculpted" by the chemical environment, thus changing both the accessible conformations and the kinetics. These changes also affect experimental observables, such as intensities and relaxation timescales of fluorescence signals or scattering functions, and can thus be monitored with stationary and kinetic ensemble experiments [59, 60, 61, 62, 63, 64, 65]. Single-molecule experiments may even monitor changes of the transition rates between conformations, provided these conformations are distinguishable in the experimental observables employed [66, 67, 68, 57, 69, 56]. While such experimental indication is crucial, it is, however, indirect as it only traces changes in a particular experimental observable that may not be sensitive to all kinetic processes of the system. Hence, complementary computer simulation studies are needed, in which chemi-

cal environmental effects on conformations and kinetics can be captured in a microscopic model. Such a microscopic model may prove useful to both understand the environmental effects onto conformational stability and kinetics in detail, and also to guide experimental investigation by predicting interesting observables.

Consider the hypothetical scenario given in Figure 3.1 for an illustration of how the chemical environment may sculpt a free energy landscape. Here, a ligand molecule is depicted in four different chemical environments: vacuum, solvent, solvent plus an ion, as well as bound to a macromolecule. In vacuum, strong electrostatic intra-molecular interactions may stabilize compact ligand conformations [70, 71]. In the associated free energy landscape this is reflected by well defined minima and high energetic barriers between the states. Adding a polar solvent such as water effectively weakens electrostatic interactions by creating a reaction field and the availability of alternative hydrogen bonding partners. This effect increases the conformation space accessible by the ligand and lowering the free energy barriers between the states. The addition of some ion may stabilize specific conformations. In this way, the ion may be involved in conformational selection of binding competent ligand conformation prior to the specific binding of the ligand to a protein. The protein environment may further stabilize this conformation, here represented by a deeper energy well. Concepts such as binding by induced or selected fit [72, 73] are usually discussed in terms of the influence of ligands on protein conformations. However, biologically relevant ligands can potentially form specific interactions with ions present in the cell (charged ligands, aromatic moieties etc.). It is therefore quite likely that selection of ligand conformations by ions or small molecules does play a role in protein-ligand binding.

In order to arrive at a more thorough understanding of how the chemical environment sculpts the energy landscape of specific ligand molecules, thus affecting their conformations and kinetics, three components are needed: i) a microscopic model of the molecule that allows changes of the chemical environment to be implemented, ii) a way to map the interesting features of this model in a way that does not depend on subjective choices such as pre-defined reaction coordinates, and iii) a way to link this model to experimental observables. As a microscopic model (i), we chose atomistic molecular dynamics simulations, which provides temporally resolved trajectories of all the atoms in a molecular system. Classical molecular dynamics force fields have been demonstrated to be able to match experimentally-measurable quantities in many cases and are steadily improved [74, 75, 76, 65]. However, given molecular dynamics data, extracting the relevant features of the system's energy landscape, such as meta-stable structures and dynamical processes can be a non-trivial task [77]. A rather objective analysis that largely avoids bias from user-defined reaction coordinates are Markov state models (MSMs). In MSMs, the molecular state space is clustered into many (e.g. several 1000) "microstates" of similar configurations, and the transition probabilities among these microstates are estimated from the MD trajectories [49, 78, 79, 80, 15, 81, 82, 83, 84, 85]. See [85] for an overview of the state

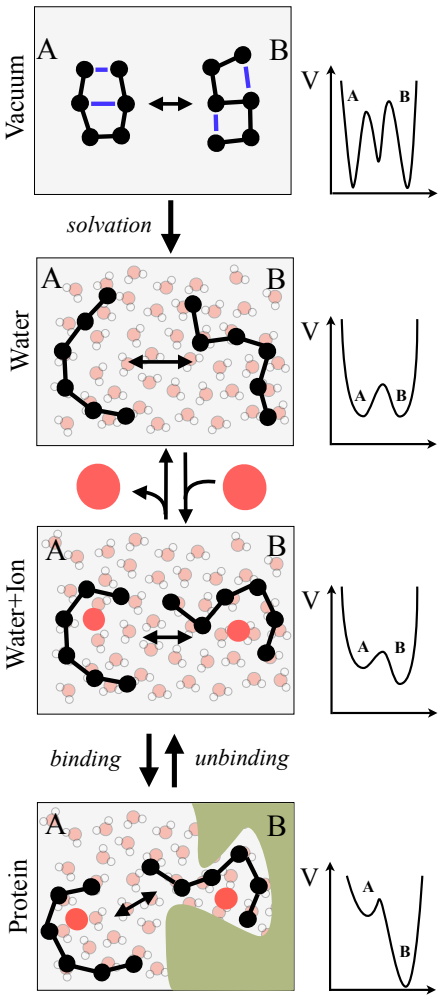


Figure 3.1: Scheme of different chemical environments and their influence on structural and energetic aspects of a hypothetical ligand molecule that can here attain two conformations A and B. The left hand side shows a pathway that describes the solvation, co-factor association and protein binding of the ligand molecule. Correspondingly the right hand side illustrates the hypothetical energy landscapes of the ligand molecule in the different environments. In vacuum the energy landscape is rough with a high barrier between the two conformations. When the molecule is solvated the landscape is smoothed and the barrier is reduced. The subsequent binding of a co-factor stabilizes conformation B with respect to A, resulting in a deeper energy well for conformation B. In the last step of, the binding to a protein further stabilizes conformation B by specific ligand-protein interactions.

of the art in MSM methodology and [86, 87] for software to construct MSMs. The MSM transition probability matrix can then be analyzed so as to yield the metastable states [49, 88, 81], their thermodynamic properties [89], the slowest relaxation processes [65], transition pathways [90, 91, 92], etc. In particular, kinetic experimental observables and their uncertainties may often be computed in terms of the eigenvalues and eigenvectors of the transition matrix, thus facilitating a comparison to experimental data and their interpretation [93, 94, 95, 96, 97].

In this chapter, a systematic MSM approach is presented to study how the energy landscape and kinetics of a small ligand molecule are affected by changes of its chemical environment. In particular we investigate Uridine diphosphate N-acetylglucosamine (UDP-GlcNAc) (Fig. 3.2a), a key player in the sialic acid synthesis pathway, whose products, sialic acids, are involved in a number of important biological processes, e.g., cellular adhesion or glycoprotein stabilization [98]. UDP-GlcNAc is simulated in different chemical environments: vacuum, water, water with an Mg^{2+} ion and in the binding pocket of an epimerase protein where UDP-GlcNAc is specifically bound. MSMs are constructed for each of the setups to investigate the effects that changes in the chemical environment have on the ligand molecule's thermodynamics and kinetics. Given the generated MSMs we are able to access quantities that are otherwise difficult to compute, such as metastable conformations, experimentally-measurable relaxation timescales as well as thermodynamic quantities such as free energies, internal energies and entropies of the identified metastable conformations in each of the considered chemical environments. We find a number of unexpected effects of the chemical environment onto the conformations and kinetics of the UDP-GlcNAc molecule and identify a potential conformational selection mechanism by the interaction of UDP-GlcNAc with the divalent ion Mg^{2+} that may be relevant for binding to the protein.

3.2 Methods

System Preparation and Molecular Dynamics Simulations

The parameters for UDP-GlcNAc were constructed by combining GlcNAc parameters of the Amber-Glycam [99, 100] database and the UDP parameters of GAFF [101]. The protein structure coordinates for the UDP-GlcNAc 2-epimerase (E.coli) with bound UDP-GlcNAc were obtained from the Protein Data Bank (accession number 1VGV). MD simulations of UDP-GlcNAc were carried out in vacuum (64 atoms), pure water (5902 atoms), pure water and one Mg^{2+} ion (5900 atoms) and with the solvated protein system ($\sim$47.000 atoms). For the ion simulations the Mg^{2+} ion was positioned in the vicinity of the two phosphates by replacing one of the previously added water molecule that were the closest to the two

phosphates. During the simulations the ion was not distance restrained. However, it was observed to reside close to the phosphate oxygens for the whole simulation time of 4 μs. The following simulation protocol was applied for all four setups: The simulation program Gromacs 4.5.3 [29] and the amber-99ff [102] force field were used. All covalent bonds to hydrogens were constrained using the LINCS algorithm [103], permitting an integration time-step of 2 fs. As integrator the stochastic integrator with a coupling constant of 1 ps^{-1} and a temperature of T=300K was used (NVT ensemble). To handle electrostatic interactions PME [35] with a real space cut-off of 1.0 nm was applied. In order to study the effect of the absence of an ion in the pure water setup, no counter-ions were used. The non-zero net charge in the pure water setup is corrected by a virtual background charge in the Gromacs PME implementation. The trajectory data was stored every 5 ps. For each water and water+Mg^{2+} 2 x 2 µs were obtained, for vacuum 4 x 2 µs and for the protein system 90 ns; resulting in 8×10^5 , 8×10^5 , 18×10^3 and 16×10^5 saved frames for each simulation.

Markov state models: Construction, Analysis and Validation

State space discretization To capture the dynamics of UDP-GlcNAc in the different simulation environments Markov models were constructed using EMMA 1.2 [87]. A single state space discretization was defined for all chemical environments by clustering the trajectory data using regular space clustering with minimal RMSD metric using all 39 heavy atoms of UDP-GlcNAc and d=0.15 nm as threshold (EMMA command `mm_discretize`). To obtain a single discretization, the clustering was performed on the unification of the trajectory data from all chemical environments. Ion and solvent molecules were not taken into account in the distance metric used for clustering.

Markov state model estimation The all-atom trajectories of each environment were projected onto the discretized state space and from the resulting discretized trajectories an MSM was constructed for each environment (vacuum, water, water+Mg^{2+}). The MSM lag time $\tau = 2$ ns was identified by calculating the "implied" relaxation timescales [78] (`mm_timescales`). Reversible transition matrices $\mathbf{T}(\tau)$ were subsequently calculated for all of the three simulation environments (`mm_estimate`).

Timescale estimation Using the relation $t_i = -\frac{\tau}{\ln \lambda_i}$ the relaxation timescales, t_i, of the slowest processes were computed based on the eigenvalues λ_i of the individual MSMs (`mm_transitionmatrixAnalysis`). These slowest processes were assigned to structural rearrangements by investigating the sign structure of the corresponding eigenvectors [85].

Metastable sets Based on the estimated system timescales four slow relaxation processes could be identified for each of the chemical environments. The associated five metastable sets of states were extracted using the improved Perron cluster cluster analysis (PCCA) [88] (`mm_pcca`).

Chapman-Kolmogorov tests To validate the estimated MSMs, it was tested whether the

Chapman-Kolmogorov condition $\mathbf{T}(n\tau) = \mathbf{T}^n(\tau)$ holds within statistical error. This was done using the procedure suggested in [85] where the system is assumed to start in each of the five metastable sets at time $t = 0$. The probability to stay in the starting metastable set is propagated to later times using the MSMs and compared to a distribution estimated directly from the trajectory data. Here this condition is tested for time ranges from 0 to 40 ns using the EMMA command `mm_chapman`. The test results are presented in SI Figure 7.2.

Energetics

For each metastable set determined for the different chemical environments, the associated free energies, and their decompositions into internal energies and entropies were calculated. All energies are estimated up to an arbitrary additive constant, which renders the direct comparison of energies between different simulation environments impossible. However, energies can be compared between different metastable sets of the same simulation environment. The free energy of a metastable set k was calculated using the relation $F_k = -k_B T \ln(\sum_{i \in S_k} \pi_i)$, where k_B is the Boltzmann constant, T the absolute Temperature and $\sum_{i \in S_k} \pi_i$ denotes the sum of stationary weights over all states i in metastable set k. To evaluate the statistical uncertainty in F_k, the probability distribution of stationary distributions, $\boldsymbol{\pi}$, was sampled using the reversible Monte-Carlo sampling described in [83]. For each sample, the corresponding estimate of F_k was calculated, and its statistical uncertainty is then given by the direct sampling estimate of the standard deviation of F_k from the Monte-Carlo sampling.

The internal energy U_k of a metastable set k was computed as mean total potential energy of all simulation trajectory frames that are assigned to a metastable set. The statistical uncertainty of these values is calculated as standard error of the mean potential energy, using the number of assigned trajectory frames as sample size. For this calculation, it was validated that the potential energies of subsequently stored trajectory frames are nearly statistically independent (correlation about 0.05 in subsequent frames) in all chemical environments.

The entropy S_k was computed using the relation $F_k = U_k - TS_k$ using F_k and U_k as given above. The statistical uncertainty $\mathrm{SE}(S_k)$ is simply computed from the standard errors in U_k and F_k via $\mathrm{SE}(S_k) = \sqrt{\mathrm{SE}(F_k)^2 + \mathrm{SE}(U_k)^2}$.

Diffusion Constant Calculations

Translational diffusion constants for UDP-GlcNAc in the water and water+Mg^{2+}environment were computed based on the center-of-mass mean-square displacement observed in the sim-

ulation trajectories by utilizing the Einstein relation:

$$\left\langle (x(t) - x(t_0))^2 \right\rangle = 6D(t - t_0), \tag{3.1}$$

where $\left\langle (x(t) - x(t_0))^2 \right\rangle$ is the mean square displacement within time $t - t_0$ and D the diffusion constant. The obtained mean square displacements and linear fits for the two environments are depicted in SI Figure 7.4.

3.3 Results and Discussion

Structures and metastable sets

The chemical environment modulates the size of the accessible conformation space. The conformation space was discretized using equidistant RMSD clustering producing "microstates" of approximately equal diameter. Thus the number of visited microstates gives a rough indication of the size of conformation spaces and of the conformational flexibility of UDP-GlcNAc in the different chemical environments (See Table 3.1). The smallest conformational freedom of UDP-GlcNac is found *in vacuo*, and would likely also be found in nonpolar solvents. *In vacuo*, unshielded intramolecular electrostatic interactions result in a strong conformational confinement which manifests in only 40 populated microstates. In pure water these electrostatic interactions are effectively reduced by the reaction field created in the polar solvent, or, structurally speaking, by the availability of water molecules as alternative hydrogen bonding partners. This effect increases the conformational freedom to 2281 populated microstates. In the water+Mg^{2+} environment this conformational flexibility is reduced to 1544 populated microstates, indicating that the ion restricts the flexibility of UDP-GlcNAc compared to pure water solvent.

System	Total Simulation Time	Number of Microstates	p($^4C_1$)	D_{trans}
Vacuum	8 µs (2 x 4 µs)	40	98 %	
Water	4 µs (2 x 2 µs)	2281	76.5 %	$4.54 \times 10^{-6}\, cm^2 s^{-1}$
Water+Mg^{2+}	4 µs (2 x 2 µs)	1544	96.2 %	$3.84 \times 10^{-6}\, cm^2 s^{-1}$

Table 3.1: Available simulation data, determined number of microstates, the stationary per cent fraction of the dominant GlcNAc pucker $^4C_1$ and the translational diffusion constant of UDP-GlcNAc for the water, water+Mg^{2+} and vacuum environments.

In all chemical environments five metastable sets exist on timescales of 4 ns or slower. A metastable set is characterized by fast kinetics within the state, i.e., conformations within that state interconvert quickly, while transitions between different metastable sets are rare, thus giving rise to slow kinetics. To facilitate a comparison of metastable sets in different chemical environment, the conformation spaces of UDP-GlcNAc in vacuum, water, and water+Mg^{2+} were decomposed into five metastable sets. This selection corresponds to

investigating the kinetics occurring on timescales of a few nanoseconds to microseconds (see Figure 3.5). The number of microstates pertaining to each metastable set quantifies the sizes of these sets and their relative probability (see Tables 3.2 and 3.3). Figure 3.3 illustrates the structures found in each metastable set. In vacuum, only two of the identified five metastable sets have high populations. This changes to three of five metastable sets when the ligand is immersed in water. Addition of the Mg^{2+} ion alters this scenario again resulting in mainly two populated metastable sets. Note how the volume (approximated as number of microstates) of each metastable changes with addition of Mg^{2+}: While there is almost no change in set 5, considering the overall difference in populated microstates, the volumes of states 3 and 4 are significantly reduced.

Different GlcNAc sugar puckers are in different metastable sets. The GlcNAc sugar is observed in two different pucker conformations: the more stable $^4C_1$ chair and the less stable $^1C_4$ chair. These puckers are located in different metastable sets, indicating that puckering is a slow process. The pucker is illustrated in Fig 3.3 and manifests in the distribution of UDP-GlcNAc dihedral angles 10,11,12 shown in Figure 3.2 and SI Figure 7.1. In Vacuum, the $^4C_1$ chair is found in states 4 and 5, which are clearly predominant with 98 % of the population (Table 3.1). The water+Mg^{2+} environment also has a predominance of $^4C_1$ with 96 % (states 2, 3, 5). These results are consistent with a previous study combining molecular dynamics and NMR on pure GlcNAc which predicted a 99.6% probability of $^4C_1$ [104]. However, we find a very different fraction when UDP-GlcNAc is solvated in pure water, where the $^4C_1$ probability (states 3, 5) drops to 77 %. This observation is due to compact structures found in the water environment in which GlcNAc and UDP form interactions that stabilize the otherwise unstable $^1C_4$ chair, yielding 23% probability in states 1, 2, 4 (see Figure 3.3e). When Mg^{2+} is added, the ion is coordinated by the oxygen atoms of the diphosphate group and the associated stretching of the diphosphate backbone separates the GlcNAc and UDP, preventing this interaction (see Figure 3.3e). Thus, the populated water+Mg^{2+} structures are more extended compared to those populated in water.

Metastable sets of UDP-GlcNAc in water and water+Mg^{2+} show structural similarities. To facilitate comparison of metastable sets between chemical environments, the similarity of metastable sets of each chemical environment was evaluated in terms of the fraction of microstates common to both of the compared metastable sets (see SI Figure 7.3). The metastable sets in vacuum do not have significant overlap with metastable sets of the solvated systems. However, water and water+Mg^{2+} environments share similar UDP-GlcNAc structures, allowing metastable sets 3, 4 and 5 to be roughly associated between these environments (see Figure 3.3 and SI Figure 7.3). These metastable sets are the most probable ones in both environments (see Table 3.3). To get a more detailed impression of the structural similarity refer to Figure 3.2 where the dihedral histograms of the similar PCCA sets are compared. Besides the sugar pucker discussed above, the orientation of

the Uracil ring is also the same in sets 3, 4, and 5 of the water and the water+Mg^{2+} environments. The structural difference between these environments mainly arises from the dihedrals of the phosphate link. In the pure water environment various orientations are possible, while in the Mg^{2+} setup only straight orientations that coordinate the ion occur. This selective stabilization of UDP-GlcNAc conformations by Mg^{2+} is effectively a conformational selection mechanism that will be of importance when studying the binding competence of UDP-GlcNAc (see below).

	PCCA Set				
	1	2	3	4	5
Vacuum	3	7	3	11	16
Water	158	78	933	388	724
Water+Mg^{2+}	148	15	554	201	626

Table 3.2: Number of microstates for each identified metastable set in vacuum, water, and water+Mg^{2+}.

	PCCA Set				
	1	2	3	4	5
Vacuum	0.002	0.002	0.01	0.61	0.37
Water	0.032	0.005	0.197	0.195	0.569
Water+Mg^{2+}	0.015	0.005	0.192	0.023	0.765

Table 3.3: Stationary probability for each identified metastable set in vacuum, water, and water+Mg^{2+}.

The chemical environment changes the flexibility within the metastable structures. Representative structures of the five metastable sets with an indication of their flexibility are shown in Figure 3.3. The structures presented are randomly drawn representations of structures present in the respective metastable set, the gray cloud indicates the conformational flexibility found in the set. Note that the metastable sets found in vacuum are much more confined than the sets found for the solvated chemical environments. A more detailed impression of the structures present in each set can be obtained from the distribution of UDP-GlcNAc dihedral angles shown in Fig. 3.2 and SI Fig. 7.1. Based on the dihedral histograms differences in conformational flexibility can be explained. The vacuum histograms show very well defined structures by sharply peaked histograms. Solvation in water permits a high flexibility in the phosphate link (dihedrals 5-8). This flexibility is reduced by addition of an Mg^{2+} ion that selects conformations that are suitable to coordinate the doubly positive charge, thus effectively "focusing" the metastable sets onto more well-defined subsets.

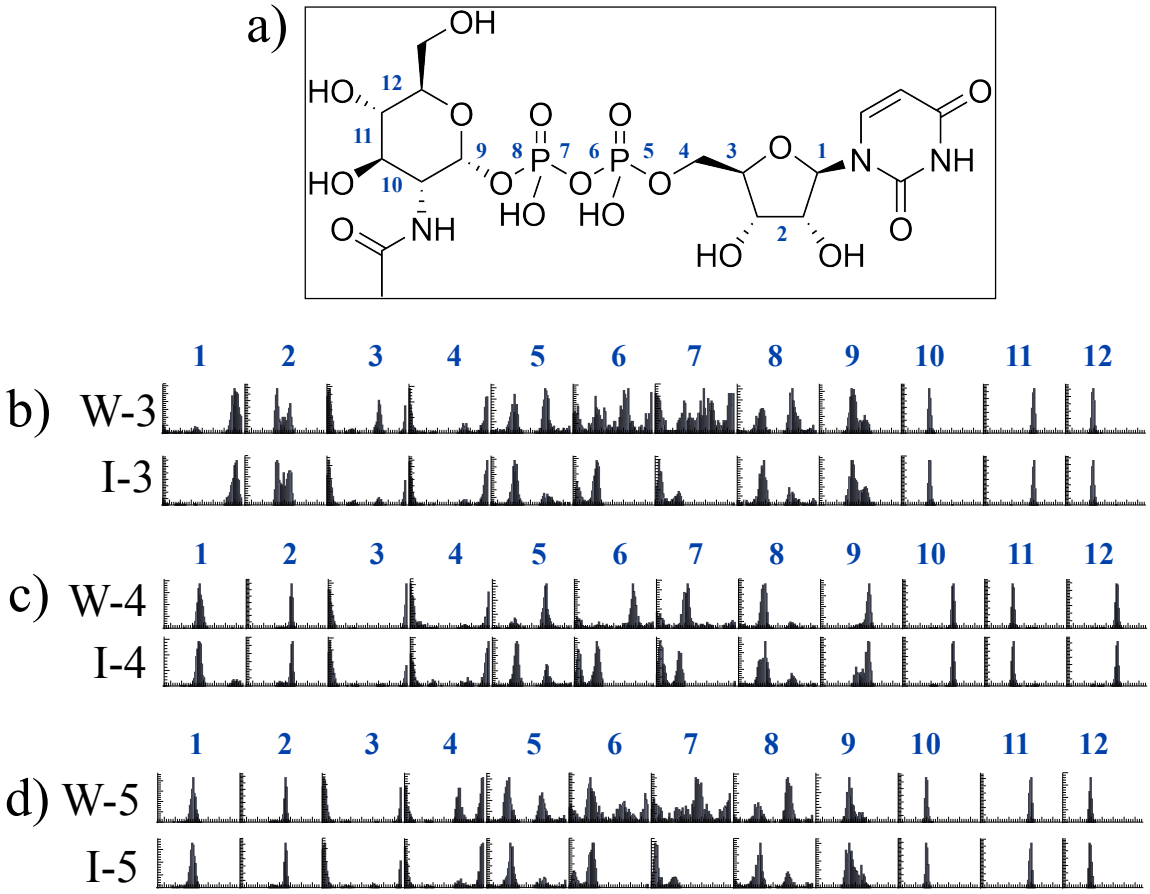


Figure 3.2: (a) Structure of Uridine diphosphate N-acetylglucosamine (UDP-GlcNAc) with dihedral annotation. (b-d) UDP-GlcNAc dihedral histograms of identified metastable sets that show a high microstate overlap. W-i indicates metastable set i of the water system, I-i the respective set for the water+Mg^{2+} system. The range of each dihedral histogram is from -180° to 180°, the bin size is 5°.

	1	2	3	4	5
Vacuum	0	0	0	0	0
Water	1 (0.005)	0	37 (0.056)	0	3 (0.0033)
Water+Mg^{2+}	1 (0.025)	0	32 (0.258)	0	2 (0.0037)

Table 3.4: Number of microstates in each metastable set where UDP-GlcNAc is in a binding conformation. The fraction of PCCA set probability that accounts for bound microstates is given in parentheses.

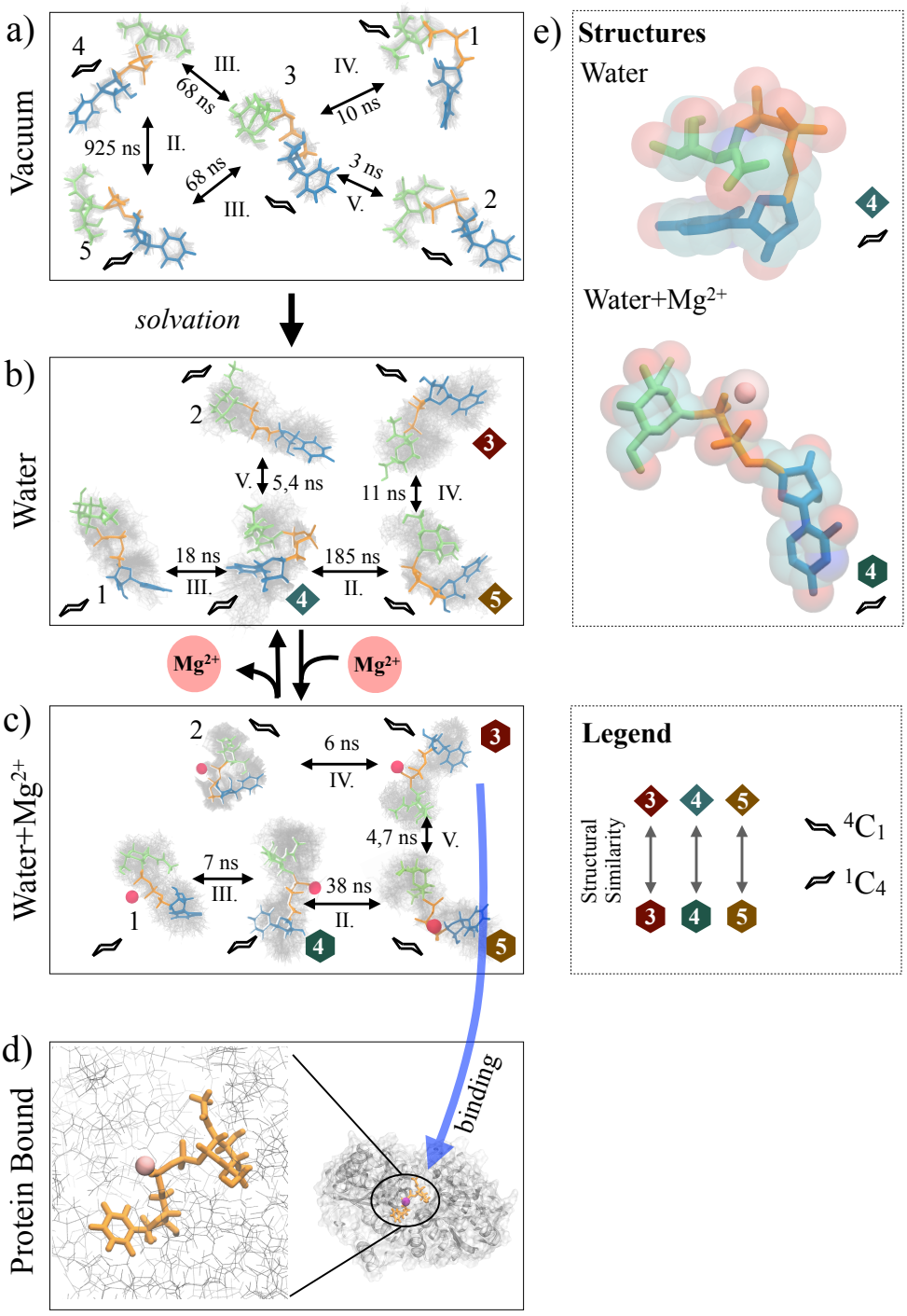


Figure 3.3: (a-d) Conformational dynamics of UDP-GlcNAc in different chemical environments. The structures presented are randomly drawn representations of structures present in the respective metastable set (1-5), the gray cloud indicates the conformational flexibility of that set. The pucker of the GlcNAc sugar is represented by a small icon. The arrows and timescales indicate the slowest relaxation processes of the system with the corresponding relaxation timescale calculated from the Markov model. Sets (3,4,5) in the water and water+Mg^{2+} environment are structurally similar and can thus be roughly associated with each other. Set (3) contains structures that are similar to protein-bound conformations. (e) representative structures from the predominant set of $^1C_4$ conformations (metastable set 4) in water and 4 water+Mg^{2+}. In pure water the $^1C_4$ ring conformations is likely to be stabilized by interactions with the uracil ring.

Thermodynamics

Having identified the metastable sets of structures in all chemical environments, we are now in a position to investigate their thermodynamic properties. To this end, free energies, internal energies and entropies of the metastable sets were calculated in their chemical environments (see Figure 3.4). We focus on states 3, 4 and 5 in the solvated environments as 1 and 2 are only rarely populated due to their high free energies.

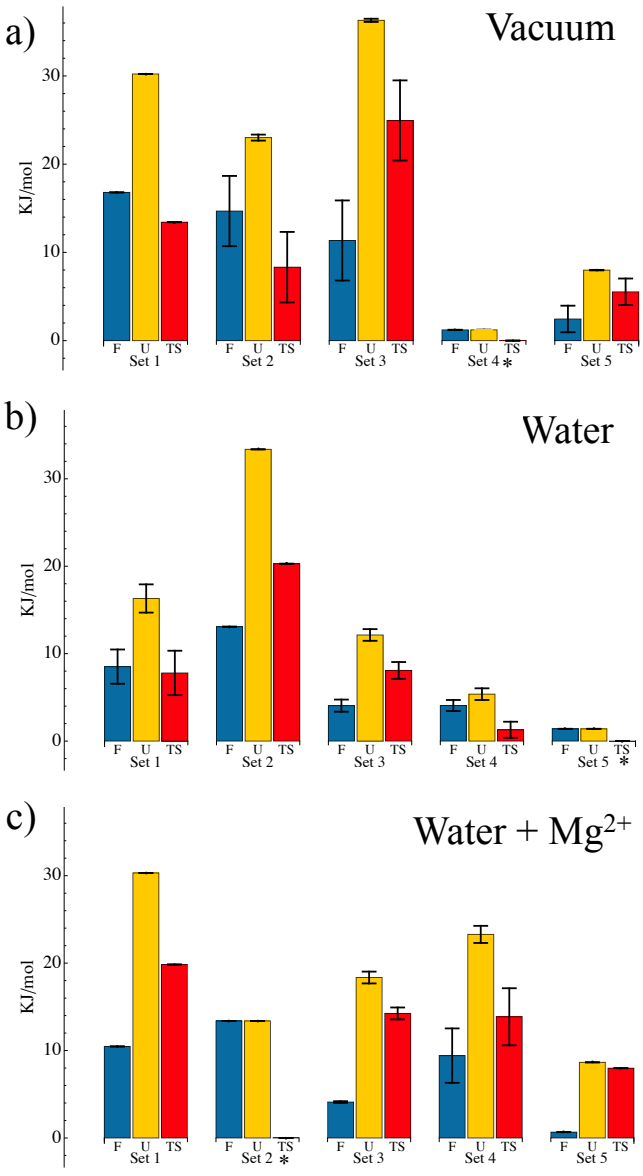


Figure 3.4: Energy composition of each identified metastable set of the different chemical environments (a) Vacuum, (b) Water and (c) Water+Mg^{2+}. F denotes the Helmholtz free energy, U the internal energy and TS entropy multiplied by absolute temperature. All values are scaled relative to the metastable set marked with *. For details on the computation and error analysis see methods section. Error bars are shown wherever the statistics was sufficient to compute them.

Low free energy states of UDP-GlcNAc in water have small entropies due to electrostriction. In pure water, sets 4 and 5 have the smallest free energies and also the smallest entropies. In contrast, set 3 also has a low free energy but a large internal energy that is compensated by an increased entropy. The reason of the small entropies in set 4 and 5 are likely due to electrostriction [105]: sets 4 and 5 have the charged phosphate oxygens exposed, while in set 3 they are shielded by interactions with other solute atoms. Such a solute charge exposure has been shown to induce ordering in the surrounding solvent with an accompanying entropy decrease of the solute-solvent system [89].

Mg^{2+} binding modulates the energy landscape and reduces electrostriction. The energy landscape of UDP-GlcNAc in water+Mg^{2+} is less uniform than in pure water. This manifests in a stronger separation of the free energies of the most likely sets 3, 4 and 5, with set 5 being the most stable (see Figure 3.4). Due to the high free energy of state 3 only a few realizations for this state were obtained which results in a relatively big error in the energy estimates. In contrast to the pure water environment the state with the smallest free energy in the Mg^{2+} scenario is not the one having the smallest entropy. This can be attributed to the attached Mg^{2+} ion. It compensates the negative charges of the oxygen atoms, thus presenting an effective net neutral charge to the environment. The solvent molecules thus become less ordered which increases the entropy of the solute-solvent system. Hence, the electrostriction effect is reduced.

Kinetics

MSMs allow the slowest relaxation timescales of the system and the associated structural rearrangements to be computed in terms of the eigenvalues and eigenvectors of the MSM transition matrix [106, 65]. This is of special interest from an experimental point of view as relaxation timescales can also be probed via kinetic experiments providing an experimental observable is found that is able to pick up the structural changes indicated by the corresponding eigenvector [107, 65].

Ring puckering and isomerization of the Uracil ring conformation are the slowest conformational changes. The structural rearrangements corresponding to the slowest relaxation timescales can be determined from the eigenvectors of the MSM transition matrices [85]. For the solvated environments (water and water+Mg^{2+}) we find that the slowest process (largest eigenvalue) describes a transition between sets with the more stable $^4C_1$ chair and sets with the less stable $^1C_4$ chair. Thus, the slowest process is found to be the GlcNAc sugar ring puckering. The second slowest process corresponds to isomerization of the Uracil ring. Refer to dihedrals 10, 11 and 12 in SI Figure 7.1 for an illustration of the ring puckering, and to dihedral 1 in the same figure for the Uracil ring turning. This explains the structural diversity found in the metastable sets of the solvated systems: Isomerizations in the phosphate link induce large conformational changes but occur on timescales that are faster (ns to sub-ns) than the kinetics of the metastable sets.

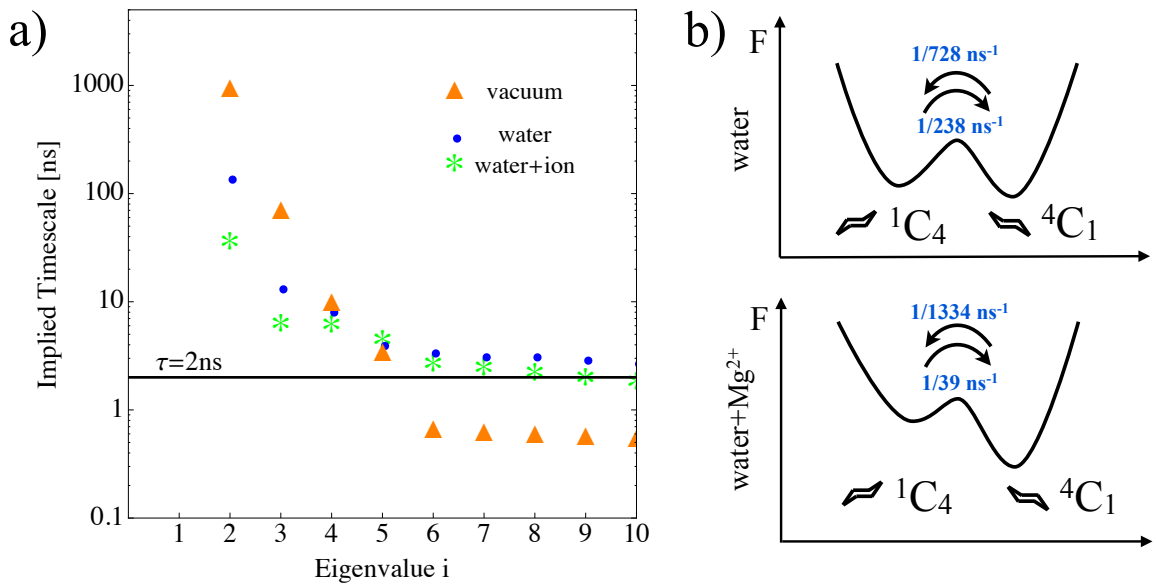


Figure 3.5: Kinetic properties of UDP-GlcNAc a) Implied timescales of Markov state models for vacuum, water, and water+Mg^{2+}. The X-Axis indicates the eigenvalue number, note that the first eigenvalue is 1 and hence has no associated timescale. The Y-Axis shows the timescale associated to the respective eigenvalue ($ITS_i = -\frac{\tau}{\ln \lambda_i}$). The thick black line indicates the lag time $\tau = 2$ ns at which the Markov state model was obtained. Timescales significantly below 2 ns are thus numerically unreliable. b) Schematic free energy landscapes of the GlcNAc puckering in the water and water+Mg^{2+} setups.

Chemical environment modulates kinetics. The slowest timescales for the three simulation setups (vacuum, water, water+Mg^{2+}) are shown in Figure 3.5. For the vacuum chemical environment the timescales are one order of magnitude slower than in the solvated environments. This indicates the presence of a rough energy landscape with high barriers that arise from unshielded electrostatics [70, 71]. Once the solute is immersed in water, the polar solvent shields these strong interactions. This smooths the energy landscape and results not only in a larger conformation space but also in faster dynamics. Comparing the water and water+Mg^{2+} environments, the pure water environment gives rise to slower timescales. This finding is interesting, as one might expect the dynamics of the Mg^{2+} system to be slower due to conformational stabilization of the phosphate link by the Mg^{2+} ion. However, given the energetics (Figure 3.4) and the structural changes corresponding to the slowest kinetic process, i.e., the process that switches between the puckers $^4C_1$ and $^1C_4$ (see Figure 3.3), this result can be understood in terms of a simple two-state rate theory argument. Let A denote the more stable $^4C_1$ chair and B denote the less stable $^1C_4$ chair. The timescale τ of a process that switches between the two states A and B is given by $\tau = \frac{1}{k_{AB}+k_{BA}}$ with puckering rates k_{AB} and k_{BA}. In the present system, we find this slowest timescale to be $\tau_{water} = 248\ ns$ for the pure water chemical environment and $\tau_{water+Mg^{2+}} = 38\ ns$ for the water+Mg^{2+} environment. From detailed balance we have the

relation $\frac{k_{BA}}{k_{AB}} = \exp(-\Delta F_{AB}/(k_B T))$ with the free energy differences $\Delta F_{AB} = 2.67$ kJ/mol for the water environment and $\Delta F_{AB} = 8.75$ kJ/mol for the water+Mg^{2+} environment. This allows effective two-state puckering rates to be computed for pure water ($k_{AB} = 1/248$ s^{-1} and $k_{BA} = 1/728$ s^{-1}) and for the water+Mg^{2+} environment ($k_{AB} = 1/39$ s^{-1} and $k_{BA} = 1/1334$ s^{-1}). Thus, the faster timescales in the water+Mg^{2+} environment are dominated by the increased transition rate from the less stable state $\{4\}$ into the more stable state $\{3, 5\}$, and thus are a result of the remodeling of the energy landscape that over-stabilizes state 5 (see Figure 3.5b). Note that all timescales smaller than 2 ns are unreliable in the present Markov models that were parametrized at a lag time of 2 ns and are hence not further investigated.

Slowest timescale is a sensor for Mg^{2+} binding As discussed above, the presence of Mg^{2+} could be experimentally detected by measuring the population of ${}^4C_1$ and ${}^1C_4$ which is significantly different with and without the Mg^{2+} due to the different interactions between GlcNAc and UDP in the two scenarios. However, the results above indicate that also the kinetics can be used as a sensor for the interaction of UDP-GlcNac with Mg^{2+}: In pure water, the stabilization of the otherwise unstable ${}^1C_4$ chair is predicted to increase the slowest relaxation timescale by a factor of 7. Kinetic experiments such as fluorescence correlation experiments or time-resolved IR experiments are able to measure the slowest relaxation timescales of the molecule, provided an appropriate observable is available [108, 109]. Thus, such kinetic experiments may be employed to measure the binding of UDP-GlcNac to Mg^{2+} *via* a change in the slowest relaxation timescale. It is likely such a sensor can also be constructed for other small molecules with ion binding sites.

Free UDP-GlcNAc diffuses faster than when Mg^{2+} is bound. Table 3.1 shows the translational diffusion constants for the solvated UDP-GlcNAc molecules. It has been found that in pure water UDP-GlcNAc shows a faster diffusion than when it has an Mg^{+2} ion attached. This can be explained by the different conformations that are stabilized depending on the presence of the ion. In the pure water chemical environment elongated and folded structures are present and can interconvert very quickly within a metastable set. When the Mg^{2+} ion is attached, compact structures that were present in the pure water chemical environment are destabilized. Thus a higher population of elongated structures is present in the water+Mg^{2+} chemical environment, leading effectively to a larger hydrodynamic radius and slower diffusion for UDP-GlcNAc.

Selection of protein-binding states

Mg^{2+} binding stabilizes binding-competent structures. Microstates containing structures that have an RMSD of less than 0.15 nm to the bound structure of UDP-GlcNAc in the UDP-GlcNAc-2-Epimerase protein were defined as binding-competent states. Binding-competent states comprise a ${}^4C_1$ pucker in GlcNAc, a specific Uracil ring isomer, and a stretched conformation of the diphosphate backbone (see Figure 3.3d). In the vacuum envi-

ronment the corresponding microstates have not been sampled at all, indicating negligible probability of binding-competent structures there. However, binding-competent structures are found in both pure water and water+Mg^{2+}. Interestingly, their total population is higher in the water+Mg^{2+} setup, indicating that addition of the Mg^{2+} cofactor brings the conformational ensemble of UDP-GlcNAc closer to the bound like ensemble. In the context of protein-ligand binding this can be interpreted as part of a conformational selection mechanism [72, 110]: The energy landscape of the ligand is changed by addition of Mg^{2+} such that the energy of the binding competent states is lowered, resulting in a higher population of these states (see "ion addition" in Figure 3.1).

Binding competent UDP-GlcNAc structures are found in a single metastable set. As discussed above, the Mg^{2+} "focuses" metastable sets, making them narrower. This "focusing" is especially interesting when considering the binding-competent structures. Table 3.4 shows in which metastable sets the binding competent microstates (described above) are located. Interestingly, nearly all binding-competent conformations can be assigned to metastable set 3 in both the water and the water+Mg^{2+} chemical environment. It is at first sight counterintuitive that the probability of metastable set 3 decreases when Mg^{2+} is added, while the probability of binding-competent conformation increases. This is explained by the fact that the probability fraction of binding competent microstates within the respective metastable water+Mg^{2+} set is significantly bigger than in pure water (see Table 3.3). Adding the Mg^{2+} thus focuses the metastable set onto the binding-competent conformations, such that the surrounding energy barriers that prevent rapid exit out of the metastable set are much closer around the binding-competent conformations. Thus, addition of Mg^{2+} may be understood as a kinetic conformational selection as the increased probability of binding-competent states is associated with an increase of the exit times out of these conformations.

Mg^{2+}-UDP-GlcNAc complex could bind to the protein There is no indication that the UDP-GlcNAc-2-Epimerase protein binds a divalent ion such as Mg^{2+} as a co-factor in the UDP-GlcNAc binding site [111]. Is it nevertheless biologically significant that Mg^{2+} stabilizes the binding-competent conformation of UDP-GlcNAc? In the water+Mg^{2+} simulations, Mg^{2+} is bound during the entire simulation time at a well defined coordination site of the diphosphate group (See Figure 3.3e). When fitting this Mg^{2+}-UDP-GlcNAc structure into the protein complex, it is found that the Mg^{2+} would not hinder the binding of UDP-GlcNAc. Rather, Mg^{2+} would be located outside of the protein, in the UDP-GlcNAc entrance channel. It is thus conceivable, that Mg^{2+} or other divalent ions interact with UDP-GlcNAc in the solute, saturating its charges and stabilizing the binding-competent conformation, and accompany the ligand until it binds specifically to a protein such as the UDP-GlcNAc-2-Epimerase. In [112] it was found that divalent cations are not necessary for UDP-GlcNAc epimerase activity, but do increase its rate under some conditions. It is conceivable that this increased rate is due to an increased binding rate resulting from such

"ion-assisted" binding.

3.4 Conclusions

In the present chapter we present a general methodology to analyze the influence the chemical environment has on structure, thermodynamics and kinetics of ligand molecules. These questions are investigated by using a combination of molecular dynamics simulations and Markov state models (MSMs). As an example, the ligand UDP-GlcNAc is analyzed in different chemical environments. The utilization of MSMs has permitted the systematic extraction of quantities that are otherwise difficult to access, such as the system's metastable sets, their thermodynamics, the relaxation timescales and their link to structural rearrangements.

As expected, the conformational flexibility increases and the relaxation timescales reduce when the ligand is solvated in water. The reverse is observed when the ligand binds to the protein, where a specific binding conformation is stabilized by the binding pocket. Interesting changes occur when a Mg^{2+} ion is added to the water solvent. On the one hand, these changes are not dramatic, as the most populated metastable sets can be roughly associated in both scenarios. However, the metastable sets become smaller as the addition of Mg^{2+} focuses the conformations onto structures that are competent to interact with the ion. There is also a marked change in the conformational energetics: In the water environment, the most populated states have relatively low entropies, likely due to an ordering in the surrounding solvent molecules caused by exposed phosphate charges (electrostriction [105, 89]). This effect is reduced when the Mg^{2+} ion is attached, as the ion interacts with the negative phosphate charges, thus effective shielding them from the solvent.

Interestingly, the GlcNAc ring pucker is strongly affected by the presence of Mg^{2+} because the interaction of Mg^{2+} with the ligand stretches the phosphate backbone, thus preventing an interaction between GlcNAc and UDP that stabilizes the otherwise unlikely $^1C_4$ chair. Thus, the GlcNAc pucker may act as a sensor for Mg^{2+} binding. In water solvent, the fraction of the $^1C_4$ pucker is predicted to be 23% which is large enough to be detected and quantified by NMR [104]. In water+Mg^{2+} the fraction is predicted to drop below 5%, which would be effectively invisible with current NMR techniques.

However, MSMs permit to explicitly calculate the system's kinetics, i.e. its slowest relaxation timescales and the corresponding structural rearrangements. These can be linked using MSMs through the duality of eigenvalues and eigenvectors of the transition matrix. In the present system, the change of the pucker populations also has a dramatic effect on the system's kinetics: In presence of the Mg^{2+} ion, the slowest relaxation time is predicted to be reduced by a factor of 7 compared to pure water solvent, mainly resulting from an increased rate of the $^1C_4 \rightarrow {}^4C_1$ transition. In principle, all relaxation timescales of the

system that can be theoretically calculated via MSMs are also experimentally measurable by kinetic experiments such as correlation experiments (e.g. fluorescence correlation, neutron / X-ray scattering) or perturbation-relaxation experiments (e.g. temperature jump, time-resolved IR) [107, 65]. The timescales that actually enter the experimental curve with significant amplitude however crucially depend on how well the experimental observable is able to trace changes along the corresponding eigenvectors. Therefore, experiments in which these observables can be controlled, e.g. by site-specific labeling, are of special interest as they can be designed to specifically track relaxations predicted by an MSM analysis [107, 65]. In the present case of a small ligand with timescales in the nanoseconds range, time-resolved IR spectroscopy may be a good candidate to complement simulation studies. The slowest timescales in Water and Water+Mg^{2+} are in the range of tens to hundreds of nanoseconds and can thus be probed in terms of the IR spectrum relaxations after a sufficiently rapid temperature-jump [113]. An alternative approach which allows much better time resolution are time-resolved IR experiments where the trigger consists of the photoisomerization or -dissociation of a construct that traps the system in one of its configurations [114]. With either trigger types, IR spectroscopy can be combined with site-specific isotope labeling, thus offering the ability to select specific eigenvectors and measure specific timescales separately [114].

Arguably the most interesting finding is how binding-competent ligand conformations are stabilized by the chemical environment. By adding water to the ligand, the energy landscape is changed such that binding competent conformations become accessible. By further adding Mg^{2+}, these binding competent conformations are selectively stabilized since the Mg^{2+} makes specific interactions with the diphosphate backbone of UDP-GlcNAc. At the same time, the non-binding conformations that lie in the same metastable set are destabilized. Thus, the Mg^{2+} ion narrows the energy well of the corresponding metastable set such that it "focuses" on the binding competent structures. This explains the surprising finding that the binding-competent structures become more probable while at the same time the metastable set that contains these structures becomes less probable. Stabilization of this sort is mainly a kinetic effect: With an Mg^{2+} ion, the metastable set containing the binding-competent structures is less often visited, but when visited, the system spend more time in binding-competent structures than without an Mg^{2+} ion.

Association of divalent cations such as Mg^{2+} to phosphate groups is a well-known and important interaction in biomolecules. Mg^{2+} binds to pairs of phosphate groups in DNA and RNA, and is important for the stabilization of the three-dimensional fold of RNA [115, 56, 58]. In ligands such as ATP and GTP, Mg^{2+} is often needed as a co-factor. Mg^{2+} co-factors have not only an electrostatic and structural role (compensation of the negatively-charged phosphate groups), but are often needed for the catalytic reaction, hence taking the role of a specific protein residue. The role of Mg^{2+} association to UDP-GlcNAc found in the present study is related, but different: There is no evidence that the

UDP-GlcNAc-epimerase binds Mg^{2+} as a co-factor. The UDP-GlcNAc binding pocket of the epimerase is itself positively charged, hence Mg^{2+} is probably not needed to coordinate or stabilize UDP-GlcNAc in the binding pocket. Nonetheless it has been observed that Mg^{2+} has a positive effect on the catalytic rate of UDP-GlcNAc-2-Epimerase [112]. It is thus conceivable that binding-conformation selection of UDP-GlcNAc by Mg^{2+} is relevant before or during binding, but not after binding to the protein: The Mg^{2+} ion stabilizes the binding-competent conformation of UDP-GlcNAc and accompany the ligand into its binding pocket. In this role, Mg^{2+} acts as a "binding co-factor".

It remains to be investigated whether such an "ion-assisted" binding is a generally relevant mechanism for other protein-ligand pairs. Even if this is not the case, being able to understand how such co-factors may be used to bias the conformation of a ligand towards a desired state, e.g. the transition state of a substrate, is an important step towards creating a molecular toolbox for design of synthetic catalysts.

4 Mechanisms of Protein-Ligand Association and Its Modulation by Protein Mutations

4.1 Introduction

The ability of proteins to bind ligands, including ions, substrates, co-factors and other proteins, is essential to all life processes. For instance, protein-ligand interaction mediates uptake and storage of cargo such as oxygen uptake in Hemoglobin, molecular recognition leading to information transfer such as in sensing of neurotransmitters or growth hormones, and buildup of biological structures such as in RNA-Ribosome interactions [116, 117]. While the majority of the biochemical and pharmaceutical work investigated protein-ligand interactions in terms of equilibrium binding affinities, it is becoming increasingly evident that the effectiveness of such interactions crucially depends on dynamical and kinetic properties [118]. The dynamical properties of binding are inherently linked to structural aspects such as size, concentration and spatial distribution of the binding partners as well as their detailed atomic structures and changes therein.

Structure-dynamics relationships for binding processes have been studied a lot at binding site contact distance, both on relevant energetics such as detailed electrostatic complementarity of the binding surfaces and hydrophobic burial, as well as on the structural binding mechanisms such as induced fit versus conformational selection [119, 72]. In contrast, fundamental properties of the spatiotemporal mechanism of how this first contact of the protein-ligand binding process is established are still elusive. For example, does binding occur via a single dominant pathway, via multiple separated pathways, or via a funnel-shaped ensemble of pathways? Is it directed to the binding site or do metastable states exist which trap the binding partners in nonfunctional states? Can diffusion-limited binding be sped up by rapid binding to the surface and subsequent surface search?

From a theoretical point of view, the protein-ligand association process can be thought of as a diffusion in a high dimensional energy landscape that usually has an energetically favorable minimum at the configuration of the protein-ligand complex. In situations in which the interaction process takes place in dilute media, this energy landscape is flat at large protein-ligand distances, resulting in a purely diffusive motion of the molecules. When the interaction partners approach each other, electrostatic forces become relevant

and for favorably interacting molecules, the energy landscape funnels down towards the complex formation configuration [120]. Such a binding funnel may also possess complex features as local minimal or parallel pathways. All mechanistic questions can be answered when the binding funnel and the dynamics governing the motion upon it are understood. Protein-ligand binding has thus many similarities with protein folding and principles or methods worked out in the protein folding field are also likely to be useful here.

In the past decades, the field of molecular simulations has been increasingly successful assigning structural and mechanistical information to experimental observations [65]. A widely used computational approach to simulate protein-ligand association dynamics are Brownian dynamics (BD) (refer also to Paragraph *Brownian Dynamics Simulations* in Chapter 2) and Langevin dynamics (LD) simulations [121] of the diffusional motion of internally rigid protein models in implicit solvent. The BD approach has been proven useful to predict bi-molecular association rates [122, 123, 124, 125] in situations where binding is diffusion limited, as well as to provide detailed insights into how protein-ligand encounter complexes are formed [126]. However, a systematic analysis of the obtained simulation data is often difficult. In this work we present a simulation and analysis approach that directly reveals the ensemble of pathways of a ligand to the binding pocket, thus allowing mechanistic questions to be answered. The approach allows to identify the presence of metastable states in the binding procedure and to study how binding mechanism and rates are altered by mutations in the protein.

Two alternative approaches to simulate and analyze dynamics exist: Most commonly, one uses the *direct simulation* approach in which long trajectory realizations of the dynamical equations (e.g. BD) are generated, and then analyzed. This approach has the advantage that it allows complex geometries with many degrees of freedom to be simulated, such as large heterogeneous protein mixtures [127]. A disadvantage is that quantities computed from generated trajectories, e.g., association rates come with statistical uncertainty, or may be systematically biased when some rare events have not been sampled at all. Moreover, trajectory data are often tedious to analyze, involving the search for "interesting" observables which involves human subjectivity. Alternatively, one can describe the *ensemble dynamics* of the system, where the transition probabilities or rates between configurational substates of the system are characterized. This approach has been successfully used in modeling the conformational dynamics of proteins with Markov models [90, 83, 15, 16, 18] (refer also to Section 2.3 in Chapter 2), where the interstate transition probabilities are estimated from many short simulations that are initialized from different substates. In diffusion processes, such as BD and LD, the ensemble dynamics can be expressed directly via the Fokker-Planck equation. Based on this formulation, sampling of individual trajectories can be avoided and the sampling error can be made zero. However, the downside of this approach is that for solving the Fokker-Planck equation the configuration space must be discretized. When using a rectangular lattice, this is currently only

feasible for three-dimensional spaces. Nevertheless, with a three-dimensional space one can already address the biophysically interesting process of ion binding to proteins [128]. An extension to higher-dimensional problems such as protein-protein binding with internal dynamics is feasible by extending the approach to meshless discretization approaches [15, 129, 47].

In this chapter we show how the *ensemble dynamics* approach permits a straightforward and objective analysis of the protein-ligand association pathways by using the mathematical framework of Transition Path Theory (TPT) [53, 130] (refer also to Section 2.4 in Chapter 2) which provides a complete and quantitative description of association pathways that lead from a freely diffusing ligand towards a protein-ligand complex in a given molecular model. Here, we apply this approach to systematically study the binding of inorganic phosphate (P_i) to the Phosphate Binding Protein (PBP) of *Escherichia coli* [131, 132, 133]. This protein plays an important role in the phosphate supply of bacterial cells and is expressed in situations when the intracellular concentration of P_i is low. After it is transported to the bacterial periplasm, it scavenges for free P_i to subsequently pass it on to a membrane protein which transports the phosphate into the cytoplasm. While previous work on the binding of P_i to PBP was mainly concerned with investigating the binding kinetics by experimental means [134, 135] or direct simulation [136], to our knowledge this work for the first time provides a systematic description of the P_i binding pathway ensemble. Various mutations are studied and it is shown how they modulate the phosphate binding rates and pathways. It is also shown how PBP becomes saturated to a second binding attempt once a P_i has been bound.

The obtained findings highlight the importance of a positively charged patch of the PBP for the attraction of negatively charged ions. Our results suggest that this pre-binding site may be a general mechanism for efficiently organizing specific ion binding via a two-step mechanism that first selects by polarity and then by ion type.

4.2 Theory

Dynamical Model

Without loss of generality, the protein-ligand association process can be divided into two phases that are dominated by different forces [137] (see Figure 4.1). The association phase I is largely governed by electrostatic forces and thermal motion of solvent molecules which lead to a diffusive approach of the solutes studied, and does not depend on intramolecular flexibility. In the binding phase II the protein-ligand complex is formed, which involves more complex short range forces, intramolecular flexibility and the structural role of solvent molecules. This separation into two phases suggests two different computational models to describe them. The second phase requires a more detailed approach such as all-atom molec-

ular dynamics simulation with full structural resolution and flexibility. Here, we restrict ourselves on the association phase I where the motion of the ligand in the protein-ligand potential is described by rigid body Brownian (or Smoluchowski) dynamics in implicit solvent (refer also to Paragraph *Brownian Dynamics Simulations* in Chapter 2):

$$d\mathbf{x}(t) = -\frac{D}{k_B T}\,\boldsymbol{\nabla} V(\mathbf{x})\,dt + \sqrt{2D}\,d\mathbf{W}_t, \tag{4.1}$$

where $\mathbf{x}(t)$ is the position of the ligand at time $t \geq 0$, D is the joint translational diffusion constant of PBP and P_i, T the absolute temperature, k_B the Boltzmann constant, $V(\mathbf{x})$ the potential energy of the ligand at position $\mathbf{x}(t)$ and $\mathbf{W}_t$ is multivariate Wiener process, i.e. white noise with independent, normally distributed, increments. We assume isotropic diffusion for both the protein and the ligand, hence diffusion can be described by a scalar constant. The error introduced by neglecting hydrodynamic interactions between interaction partners is unlikely to affect the main findings of this chapter. However, in subsequent studies hydrodynamic interaction could be included following recent work of Geyer et. al [39]. The change in particle position $d\mathbf{x}(t)$ in a time interval dt is thus the result of the force from the potential, $-\nabla V(\mathbf{x})$, and a random displacement which implicitly models the collisions with solvent molecules. It is important to note that the solution $x(t)$ of the stochastic differential equation (Eq. 4.1) is a random sequence. Hence, for a given initial position $\mathbf{x}(0) = \mathbf{x}_o$, each realization of $\mathbf{x}(t)$ describes a possible ligand trajectory.

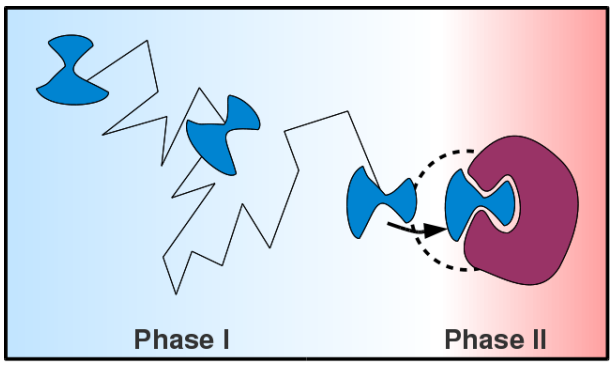


Figure 4.1: Phases of protein-ligand association. Phase I is largely characterized by diffusional association and ends upon encounter complex formation. Phase II involves actual binding of the ligand and might involve structural rearrangements in both interaction partners.

Interaction Potential

In order to compute the interaction potential between PBP and phosphate ion solely electrostatic forces are considered as they are the most important contributors during the association phase. An explicit modeling of Van der Waals forces can be omitted as the interaction partners can be thought of being immersed in dense media (water) and therefore

interact equally with all surrounding atoms.

Furthermore, the structure of the diffusing ligand P_i is approximated by a point charge of $-2e$ to represent the HPO_2^{-2} form of phosphate. This allows the energy of PBP-P_i configurations to be calculated by multiplying the electrostatic potential induced by the protein with the phosphate charge ($-2e$) at the respective positions. The protein potential $V(\mathbf{r})$ is computed using Poisson-Boltzmann theory [138] (refer also to Paragraph Poisson-Boltzmann Electrostatics in Chapter Theory and Methods), in which the solvent is modeled as a continuum with a specific dielectric constant. The Poisson-Boltzmann equation is given by:

$$\boldsymbol{\nabla}\cdot[\epsilon(\mathbf{x})\boldsymbol{\nabla}V(\mathbf{x})] = -\rho(\mathbf{x}) - \sum_i c_i^{\infty} z_i \lambda(\mathbf{x}) \exp\left(\frac{-z_i V(\mathbf{x})}{k_B T}\right) \tag{4.2}$$

where $\epsilon(\mathbf{x})$ is the dielectric constant at position $\mathbf{x}$, ρ indicates the charge density of the protein (given by assigning partial atom charges), c_i^{∞} denotes the concentration of ion species i at an infinite distance from the molecule, z_i is its partial charge, k_B the Boltzmann constant, T is the temperature and $\lambda(\mathbf{x})$ is a delta function that indicates the ion accessibility.

For the calculation of association rates to be correct, the volume considered around the protein has to be large enough such that the gradient of the potential approaches zero at its outer boundaries. At the same time it is crucial for a correct calculation that the potential close to the protein surface is well described. To comply with the large volume and high resolution requirements, we use the manual focusing mechanism (mg-manual) provided by APBS, solving the PB equation on differently sized grids ranging from $33 \times 33 \times 33$ with isotropic spacing of $d = 16$ Å to $193 \times 193 \times 193$ with isotropic spacing of $d = 0.35$ Å. The respective coarser solutions was used as an outer boundary condition for the finer one.

Transition Path Theory

While individual realizations of the stochastic dynamics (Eq. 4.1) are random, we are interested in the deterministic expectation values of this random process, such as transition rates, fluxes and pathway probabilities. In order to obtain these quantities, we apply Transition Path Theory (TPT) [139, 53, 52] (also described in Section Transition Path Theory of Chapter 2) to the rate matrix $\mathbf{K}$ (Markov jump process) that results from discretizing the Fokker-Planck equation that is associated with Eq. 4.1.

The TPT quantities already introduced in Section Transition Path Theory of Chapter 2 are slightly different in the rate matrix case. The reactive probability flux is given by

$$f_{ij} = \pi_i q_i^- k_{ij} q_j^+. \tag{4.3}$$

π_i denotes the Boltzmann weight of state i, i.e., the overall probability for the process

to be in the volume element represented by state i, q_i^- denots the backward committor of state i, k_{ij} the transition rate from i to j and q_j^+ the forward committor of state j. The net reactive flux is computed analogous as in the transiton matrix case:

$$f_{ij}^+ = \max\{0, f_{ij} - f_{ji}\}. \tag{4.4}$$

It is important to note that the flux is conserved, i.e., the amount of flux leaving A equals the amount entering B and for all intermediate states i the influx equals the outflux. Refer to Figure 4.2 for an illustration of TPT on a two-dimensional model of protein-ligand association. In the following the rate matrix based committor computation is explained.

Based on a discrete rate matrix (Eq. 2.81) the forward committor can be computed by solving the constrained linear problem

$$\begin{aligned} \mathbf{Kq} &= 0 \\ \text{s.t.} \quad & q_i = 0 \,\forall i \in A \\ & q_i = 1 \,\forall i \in B \end{aligned} \tag{4.5}$$

where A and B are the sets of discrete states corresponding to dissociated and associated states, respectively. This problem is solved by reordering the states in the order (S, A, B) where $S = (A \cup B)^C$, yielding the following structure in $\mathbf{K}$ and $\mathbf{q}$:

$$\mathbf{K} = \begin{pmatrix} \mathbf{K}_{SS} & \mathbf{K}_{SA} & \mathbf{K}_{SB} \\ \mathbf{K}_{AS} & \mathbf{K}_{AA} & \mathbf{K}_{AB} \\ \mathbf{K}_{BS} & \mathbf{K}_{BA} & \mathbf{K}_{BB} \end{pmatrix}, \quad \mathbf{q} = \begin{pmatrix} \mathbf{q}_S \\ \mathbf{q}_A = \mathbf{0} \\ \mathbf{q}_B = \mathbf{1} \end{pmatrix}, \tag{4.6}$$

Which allows Eq. (4.5) to be rewritten as:

$$\mathbf{K}_{SS}\mathbf{q}_S = \mathbf{K}_{SB}. \tag{4.7}$$

which can easily be solved by standard numerical methods to obtain the unknown $\mathbf{q}_S$. In the present application the number of unknowns is in the order of 10^6. In order to solve this task we use the implementation of the iterative BiCGStab algorithm provided by the Java Matrix Tookit [140]. A thorough discussion of efficient committor computations including error analysis can be found in [44].

Binding Rate Calculation

The expected number of $A \to B$ transitions per time unit is given by Eq. 2.96. This quantity includes the fact that the ligand needs to diffuse back to the A area until another

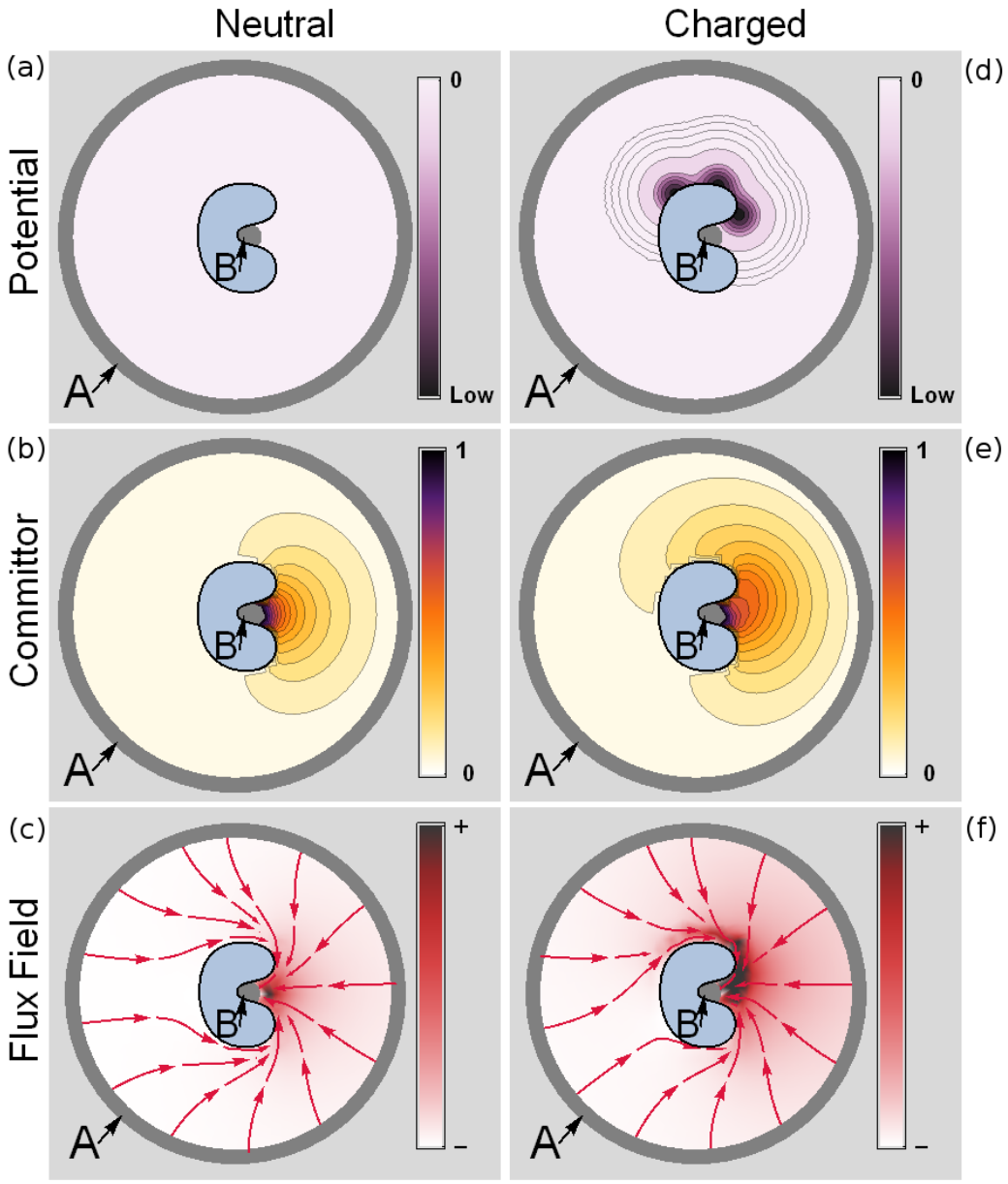


Figure 4.2: Illustration of TPT on a simple two-dimensional protein-ligand binding model. The dissociated state of the ligand A and the associated state B are shown. (a), (b), (c) show a situation in which no potential is present and the ligand can diffuse freely until it associates to the protein. (d), (e), (f) illustrate a situation where the protein has surface charges, generating energy minima that attract the ligand. (a) and (d) show the different potentials, (b), (e) show the forward committor q^+, revealing for areas on top of the charged protein a higher probability to reach the binding site than for the uncharged protein. (c) and (f) show the reactive flux density and integrated flux lines calculated from the flux field resulting from the fluxes f_{ij}^+. For the neutral protein it becomes apparent that the ligand diffuses freely and the binding pathways are only restricted by spatial constraints. In the charged scenario the ligand is strongly attracted by the top side of the protein, creating a high reactive flux density in that area that distorts the pathway ensemble accordingly.

transition to B is considered. Hence, in order to calculate the $A \to B$ transition rate, we need to take the probability into account, that the ligand is moving from A to B, i.e., it has been in A last:

$$\pi_A = \sum_{i \in S} \pi_i q_i^-, \tag{4.8}$$

where S is the set of all states. Therefore, the transition rate is given by [90]:

$$k_{AB} = \frac{F_{AB}}{\pi_A}. \tag{4.9}$$

k_{AB} is the rate at which a ligand molecule binds starting from set A. In order to compute the bimolecular association rate of PBP and P_i, the rate at which ligand molecules arrive at the A sphere has to be taken into account. Based on the assumption that protein and ligand diffuse freely upon a distance r, i.e., in our scenario the ligand enters the A sphere, according to Erban et al. [141] the diffusion limited association constant k_{On} can be obtained by:

$$k_{\mathrm{On}} = 4\pi D \left(r - \sqrt{\frac{D}{k_{AB}}} \tanh \left(r \sqrt{\frac{k_{AB}}{D}} \right) \right), \tag{4.10}$$

where D is the diffusion constant, and r denotes the radius of the A sphere. Note that k_{On} is a concentration dependent rate (e.g. in $nm^3 s^{-1}$), while k_{AB} is the rate of a single molecule event (in s^{-1}).

4.3 Methods

Molecular model and simulation setup

The coordinates of the open form mutant T141D of the Phosphate Binding Protein from *Escherichia coli* (PDB [142] code 1OIB, Chain A) served as a template to create several *in silico* mutants of the protein. The mutagenesis tool of PyMOL (vers. 0.99rc6) was used to create mutants D56N, D137T, K43M, K43Q, R134Q, R135Q, R134Q/K167Q/K175Q (3 mut.), R134Q/K167Q/K175Q/D21N/D51N/D61N (6 mut.), T141D, chosen in agreement with previous work on PBP [136]. The wild-type (wt) was modeled by replacing Asp141 with Thr141.

Energy minimization of the structures in a TIP3P water box was carried out by running 2000 steps of the steepest gradient algorithm using the Gromacs (version 4.5) program [29] employing the CHARMM [143] force-field. The protonation states of ionizable amino acids were determined by using the PROPKA [144] tool setting the pH to 7. The atomic partial charges were assigned using the PDB2PQR suite [145] using the CHARMM force-field as reference. The electrostatic potential of the resulting structures were calculated

using the Adaptive Poisson-Boltzmann Solver (APBS) [27], using dielectric constants of $\epsilon_P = 4.0$ for the protein interior and $\epsilon_S = 78.0$ for the solvent. As joint diffusion constant $D = 8 \times 10^{-6} cm^2 s^{-1}$ [146] was used.

Space Discretization

In order to calculate the TPT quantities for the protein-ligand binding process, a finite volume space discretization is required which extends over a large volume while at the same time having a high resolution close to the protein surface. Therefore, we developed a simple adaptive discretization scheme based on the numerical gradient of the electrostatic potential. The procedure starts from a coarse cubic $33 \times 33 \times 33$ grid with an edge length of 528 Å and refines interior grid points based on a local error criterion. By using central finite differences the potential derivatives in each Euclidean direction are computed for each point, at the one hand using the present discretization as well as using a finer discretization where additional grids points have been added halfway between each pair of initial grid points. Whenever at a given refinement point the two derivatives differ by more than a specified threshold (here 0.01 $kT/\mathring{A}$), the refinement is accepted and another grid plane is added intersecting with this refinement point and perpendicular to the connection between the two coarse grid points. This procedure is iterated until no more planes are added. Grid points that would lie inside the protein, defined by having a minimal distance to protein atoms of less than 3.2 Å, are not taken into account, and are dismissed from the final grid. The resulting grids had an average size of $173 \times 151 \times 177$ points (a total of 4.623.771 elements) with box lengths ranging from 16 Å for distant boxes to 0.5 Å in the vicinity of the protein.

Rate Matrix Computation

When considering Brownian dynamics (Eq. 4.1) the transition rates between volume elements of the regular grid defined in 4.3 can be computed using a discretization scheme introduced in [147]. The resulting matrix $\mathbf{K}$ (refer also to Paragraph Relation between Transition Matrix and Rate Matrix in Chapter 2) is a discrete model of the entire ensemble dynamics of the protein-ligand association process, and all subsequent analysis can be conducted based on this matrix. A matrix element k_{ij} specifies the number of transitions per time unit to a volume element j conditional on starting at element i, and is computed as follows:

$$k_{ij} = \begin{cases} \frac{D}{h_{i,j} d_{i \to j}} \exp\left(-\frac{1}{2k_B T}(V_j - V_i)\right) & j \in \{N_i\} \\ -\sum_j k_{ij}, & j = i \\ 0 & \text{otherwise,} \end{cases} \tag{4.11}$$

where N_i denotes the set of all volume elements that share a face with element i, D is

the joint diffusion constant, V_i designates the potential at grid point i, $h_{i,j}$ denotes the distance between grid points i and j and $d_{i\to j}$ stands for the length of the ith volume cell into the direction of j.

A and B Definition

After obtaining the space discretization of the volume around the protein, the A and B sets are assigned. For the set of free diffusing configurations of the phosphate ion (A set of states) all volume elements whose center is further than 250 Å away from the geometric center of the protein are chosen. Note that the choice of A is irrelevant as long as far enough away from the protein such that the electrostatic forces are zero in A. Defining A further away from this minimal distance will increase r but decrease k_{AB} resulting in the same concentration dependent binding rate as in Eq. 4.10. The set B of bound/pre-complex configurations is chosen to include all volume elements that are within a 3 Å radius of the geometric center of Thr10, Ser38 and Ser139, shown as yellow region in Figure 4.3. The choice of B will affect the pathways and association rates as it defines the bound state.

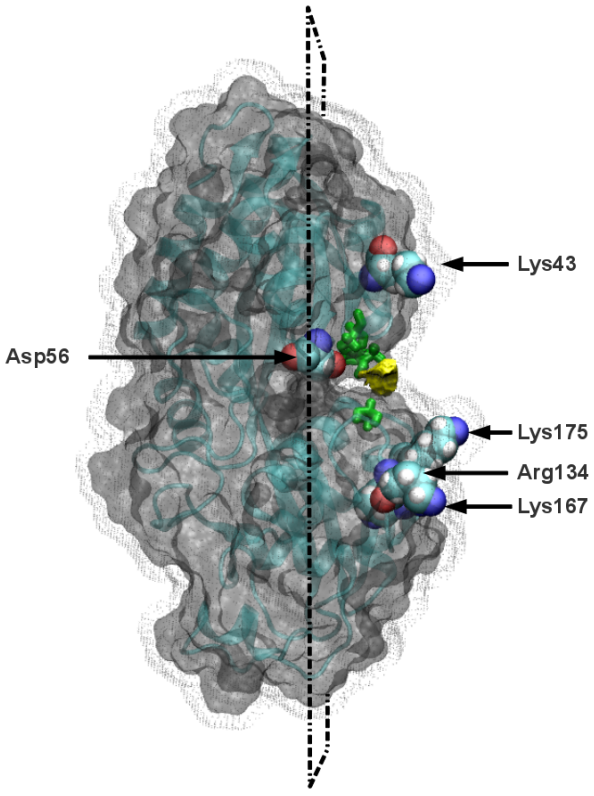


Figure 4.3: Transparent Connolly surface of the Phosphate Binding Protein (*Escherichia coli*) showing secondary structure elements. The yellow region depicts the B set. A subset of mutated amino acids are shown in Van der Waals representation. The dotted surface represents points acccessible by the phosphate ion. The indicated plane denotes the projection area used to visualize first hitting densities (Figs. 4.5 and 4.6).

Free Energy Profile of Ligand Association

As the forward committor is the probability to associate rather than dissociate, it measures the progress of the reaction and thus represents a "kinetic reaction coordinate" [148] with 0 representing the dissociated (A) and 1 representing the associated configurations (B). The free energy (refer to Subsection 2.2.3) along this coordinate is given by

$$F(q) = -k_B T \ln(\rho(q)) + const. \tag{4.12}$$

$\rho(q)$ denotes the stationary density of the set of states having a committor value q and is calculated in our discrete model by

$$\rho(q) = \sum_{i,\, q_i \in [q-\frac{\Delta}{2}, q+\frac{\Delta}{2}]} \exp\left(\frac{-V(x_i)}{k_B T}\right). \tag{4.13}$$

using a sliding window with width $\Delta = 0.005$ over the range of $q \in [\frac{\Delta}{2}, 1 - \frac{\Delta}{2}]$.

Binding flux field and visualization

For a visualization of phosphate association pathways a vector field of reactive fluxes was calculated. For this purpose, a total flux vector was assigned to each grid point i by vectorial summation of all outgoing fluxes f_{ij}^{+}. To visualize the resulting vector field, as in Figs. 4.5 and 4.6, the Mayavi2 program [149] was employed. Starting from a fixed number of points spherically distributed with distance 80 Å from the geometric center of the protein, the program follows the streamlines along the flux vectors, thus tracing out possible binding pathways. The streamlines are colored according to the local flux strength, i.e., norm of the total flux vectors. The lighter the coloring, the stronger the encountered flux.

In order to better visualize how the association pathways behave near the protein, we have calculated where they hit the protein surface for the first time. For this, a surface was defined in a distance of 10 Å around the phosphate accessible surface. At each surface element, the flux through the surface, quantified by the reactive TPT flux f_{ij}^{+} (Eq. 4.4), is calculated. For the sake of visualization the orthogonal projection of surface elements onto a two-dimensional plane which divides the surface into two halves is calculated. The plane is depicted in Figure 4.3. In the projection only surface elements on the half of the binding site are taken into account.

4.4 Results and Discussion

The results of the modeling and analysis of inorganic phosphate association to the phosphate binding protein and various *in silico* mutants are presented below. Selected mutants

are summarized in Figure 4.5 while the results for remaining structures are shown in the supplementary information (SI).

Free Energy Profiles and Association Rates

The left column of Figure 4.5 shows the free energy profile of phosphate associations along the committor coordinate. For most of the investigated mutants the free energy decreases with increasing committor value, indicating that binding of phosphate is energetically favorable. Inspecting the free energy profiles of different mutants shows the existence of several minima along the committor coordinate. Such minima indicate that the phosphate ion is more likely to be found at certain positions in space with corresponding committor values and these configurations may be metastable. Interestingly, the two committor iso-surfaces shown in Figure 4.4 are especially relevant for the phosphate binding process: for each mutant at least one of these two iso-surfaces describes configurations associated with a minimum in its free energy profile. Whenever a minimum could be assigned to one of the iso-surfaces it is marked with a red or blue dot in the free energy profile. Phosphate configurations represented by the outer iso-surface (red) are subsequently termed intermediate 1 and configurations described by the inner iso-surface (blue) are termed intermediate 2.

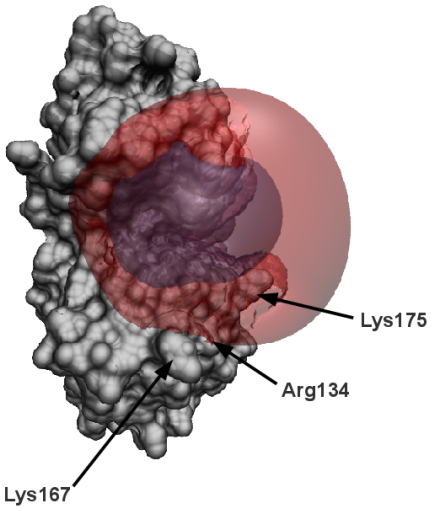


Figure 4.4: Connolly surface representation of PBP with transparent intermediate 1 (red) and intermediate 2 (blue) committor iso-surfaces.

In the wild-type protein, both intermediate 1 and intermediate 2 free energy minima indicate two metastable configurations of the phosphate before it reaches the binding site. The A197W (see SI) mutant exhibits a very similar profile and an almost equal association rate of 26.4 $M^{-1}s^{-1}$ compared to 27.9 $M^{-1}s^{-1}$ for the wild-type, indicating

that this mutation has little effect on the phosphate ion binding capability. For mutants R134Q/K167Q/K175Q (3 mut.) and R134Q/K167Q/K175Q/D21N/D51N/D61N (6 mut.) the intermediate 1 configuration is destabilized and thus only the configurations corresponding to intermediate 2 are found to be metastable. Both mutants have the same three positively charged amino acids replaced by neutral substitutes, but in the 6 mut. mutant the associated loss of charge is compensated by additionally replacing 3 negatively charged amino acids by neutral substitutes. The destabilization of intermediate 1 indicates that residues Arg134, Lys167 and Lys175 are necessary for holding the phosphate ion at the protein surface. Interestingly, losing this kinetic trap along the binding coordinate does not increase the association rate of phosphate: in contrast it is decreased by a factor of 3 for the 6mut. mutant (9.3 $M^{-1}s^{-1}$) and by a factor of ~ 10 for the negatively charged 3mut. mutant (2.5 $M^{-1}s^{-1}$). Due to its relevance for attracting phosphates and thereby enhancing the binding efficiency we henceforth abbreviate the positive charge patch around residues Arg134, Lys167 and Lys175 *"anion attractor"*.

To further assess the relevance of positive surface charges, single-point mutations R134Q and R135Q were considered. R134Q neutralizes one residue of the *anion attractor*, whereas R135Q neutralizes a residue which is found between the *anion attractor* and the phosphate binding site, therefore interfering with the phosphate transport route. While both mutants show a reduced association rate, this reduction is 5-fold in R135Q while it is only 2-fold for R134Q. The corresponding free energy profile of R135Q also reveals this effect by showing smaller binding (committor) probabilities for intermediate 1 and 2 configurations than for the R134Q mutant.

The mutants discussed so far have mainly affected residues in the vicinity of the *anion attractor*. For a more comprehensive assessment of phosphate association also mutations D56N, D137T, K43Q and K43M were considered. Mutations D56N and D137T both neutralize a negative charge and increase the association rate by a factor of about 3 compared to the wild-type. Due to thus stronger attraction of the negatively charged phosphate ion, the minimum associated to intermediate 2 configurations vanishes, while intermediate 1 trapping is still present although with an increased probability to reach the binding site from these configurations. The intermediate 2 minimum also disappears for the negatively-charged K43M/K43Q mutants. However, in contrast to the positively charged D56N/D137T mutants, the association rate is reduced by a factor of almost 3 and the binding probability associated with intermediate 1 configurations is strongly reduced, which can be seen from the left shifted minimum in the free energy profiles.

Finally, the T141D mutant is discussed. The free energy profile of this mutant is remarkably different from other mutants that also introduce a negative net charge of -1e. In fact, also a free energy minimum can be assigned to intermediate 1 configurations, but the associated binding probability is very small. Furthermore, the free energy difference between unbound and bound phosphate is positive, rendering phosphate binding unfavorable.

This can also be observed, at the corresponding association rate which is also drastically reduced, and a factor of 5 smaller than the smallest association rate found for almost all other mutants with a negative net charge of -1e. An explanation of this result might be the location of the mutation, which introduces a negative charge very close to the phosphate binding site, repelling the phosphate here. Unlike other mutations that introduce negative charges, in this case the phosphate ion cannot avoid the repulsive region via alternative pathways in order to reach the binding site. Consequently, this mutation has the largest effect on the association efficiency of the phosphate ion.

Table 4.1: Net charge and computed bimolecular association rates of considered mutant structures at ionic strength of $0mM$.

Mutant	Net Charge [e]	k_{on} [$10^8 M^{-1}s^{-1}$]
wt (modeled)	0	27.9
A197W	0	26.4
D56N	+1	73.9
D137T	+1	77.8
T141D	-1	1.6
R135Q	-1	5.9
K43M	-1	11.3
K43Q	-1	11.4
R134Q	-1	12.4
3mut	-3	2.5
6mut	0	9.3
PBP:P_{i}	-2	3.0

Stream Lines and First Hitting Density

The free energy profiles and rates described above provide information about macroscopic or effective properties of the phosphate association process, but they do provide little information about the fine details of phosphate association. More do the specific dynamical properties that can be accessed with the Transition Path Theory approach: the shape of the binding pathways and the distribution of where they hit the protein surface. Figure 4.5 and the following show representative pathways of the association pathway ensemble. These plotted pathways are streamlines that follow the reactive flux field of binding. The number of reactive trajectories that pass a volume element per unit of time is expressed by streamline coloring. The brighter the coloring, the more reactive trajectories pass through the surrounding volume elements. This manifests as almost white coloring in the vicinity of the binding site, where the increasing bundling of reactive trajectories leads to an increased flux density. In order to obtain additional information where the phosphate association pathways "attack" the protein, we measured how many reactive trajectories per unit of time hit surface elements in a distance of 10 Å around the protein. This hitting density

is visualized by a planar projection in the second column of Figure 4.5 along with the positions of the mutations.

The neutrally charged structures wt and A197W share a similar pattern in the first hitting density and distribution of pathways. The phosphate trajectories attack the protein on both sides of the phosphate binding side, with a preference for the side at which the *anion attractor* is located. The corresponding stream line illustrations show that some phosphates form first contact with the *anion attractor* and then crawl over the surface to the binding site. This picture is not qualitatively different for the positive D56N and D137T mutants. Here, also both sides of the protein are approached by the phosphate and the surface crawling still occurs. However, due to the increased net charge of the protein the number of reactive trajectories is strongly increased. A change in both hitting density and approach pathway distribution can be observed for in K43M/K43Q. In this case the number of pathways that attack the protein at the side of the mutation is reduced and the stream lines show that the phosphate is no longer attracted to the surface at the respective position. An even stronger distortion is observed when the positive patch is neutralized as in the *6mut.* and *3mut.* mutants. The number of pathways that hit the extended protein surface above the positive patch is significantly reduced in both cases. Furthermore, the flux lines show that the pathways are not attracted to the positive patch but rather straightly approach the phosphate binding site from the bulk. Due to the negative net charge of the *3mut.* mutant the number of phosphates that reach the binding site is per unit of time is reduced as visible from darker flux lines. While the T141D mutation was found to strongly reduce the association rate it neither exhibits a change in first hitting density nor is the topology of association pathways affected. The surface attraction of the phosphate ion is still present, however, the number of phosphates reaching the binding site is strongly reduced, i.e. this mutation only affects the last step of association. Although mutations R134Q and R135Q do not show a pronounced effect on the first hitting density, they do show a difference in the flux line picture. In comparison to R135Q, the surface attraction at the positive patch is less pronounced in the R134Q mutant.

In the results shown so far, we have investigated the binding dynamics of a single phosphate ion in the dilute limit, i.e. in absence of other solutes. In a biological scenario, the situation is much more complex as the cytosol is densely filled with various species of different sizes, shapes and charges. Although such a heterogeneous complexity is of limited interest to the biophysicist, it is very interesting to work out some of the principles that contribute to the phosphate binding dynamics, and more generally to potentially all ion-binding dynamics, in the cell. For example how does phosphate binding occur in a phosphate rich environment, i.e. where phosphates compete for binding? To model this, we investigate inorganic phosphate association in a model where a phosphate ion is already trapped at the positively charged surface patch. Therefore, a HPO_4^{-2} ion was placed in the vicinity of Arg134, Lys167 and Lys175 and the association dynamics were computed

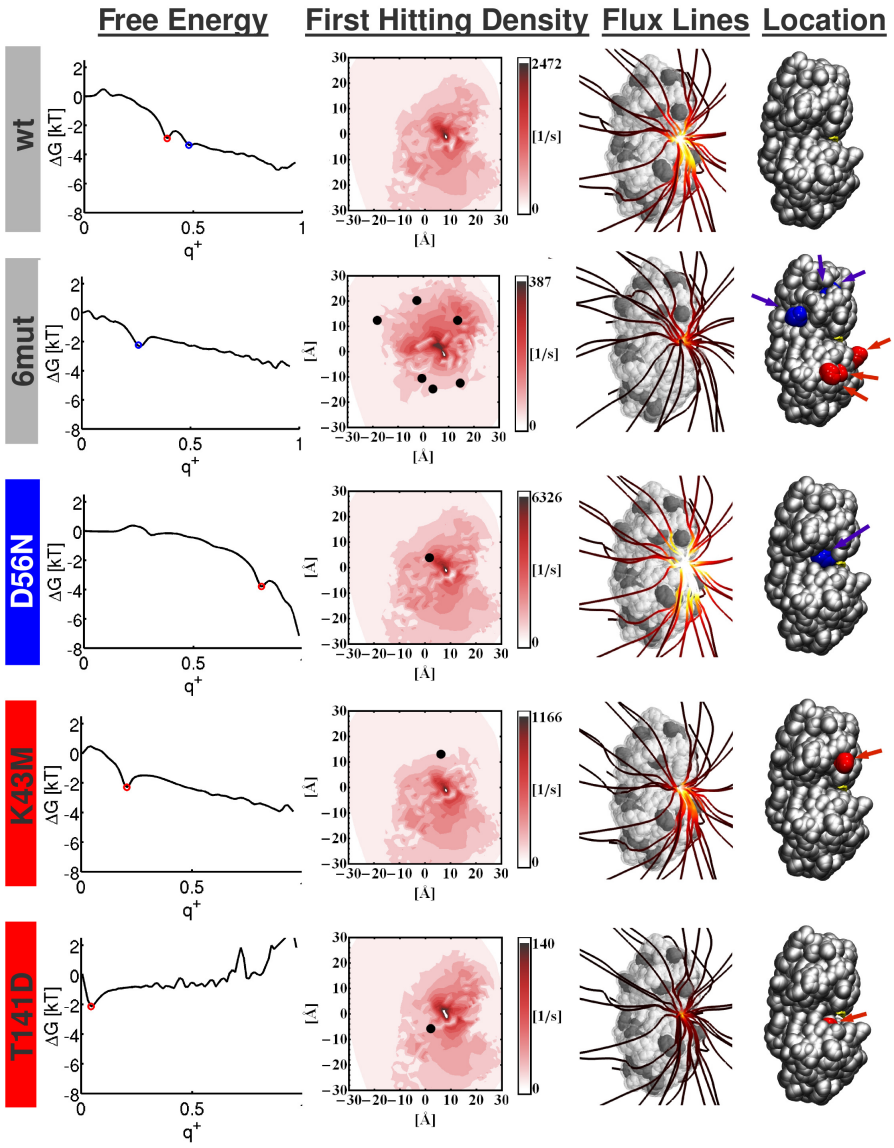


Figure 4.5: Free energy profiles, first hitting densities, and pathways for selected mutants of P_i associating to the Phosphate Binding Protein. Text background coloring: Gray - Protein has neutral net charge, Blue - positive net charge, Red - negative net charge. *First column* - Free energy profile of the ligand when it travels from the dissociated ($q^+ = 0$) to the associated state ($q^+ = 1$). A red or blue dot is shown whenever a minimum could be assigned to one of the two iso-surfaces shown in Figure 4.4. *Second column* - Surface density of reactive trajectories that hit the extended protein surface per unit of time. In each of the plots the black points represent projected C_α positions of mutated amino acids. Note that the color axis is scaled separately for each mutant for clarity. *Third column* - Streamlines of the reactive flux of ligand association (see Sec. 4.3) for the different mutants, which represent the ensemble of association pathways. A lighter streamline coloring corresponds to a higher local reactive flux. Here, the same color scheme is used for all pictures. *Fourth column* - Here the positions of mutated residues are shown. Red/blue corresponds to negative/positive charged mutations relative to the wild-type structure.

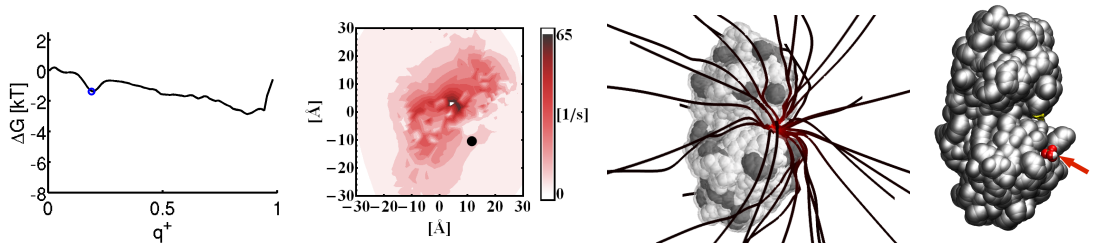


Figure 4.6: Phosphate Binding to PBP:P_i - free energy profile, first hitting density, association pathways, marked phosphate location

based on the resulting electrostatic potential. The computed free energy profile, the first hitting density and the binding pathways are depicted in Figure 4.6. The free energy profile shows that the trapping property of the positively charged patch is lost when it is already loaded with a negatively charged ion, the minima corresponding to the intermediate 1 iso-committor surface was not present anymore. Moreover, the overall binding free energy is nearly zero. The hitting density plot shows that the pathways avoid to attack the protein at the bound phosphate location and are redirected further down. The streamlines additionally reveal that the second phosphate does not "crawl" over the *anion attractor*, it rather reaches the binding site from space.

Rapid scanning of ligands

In vivo, the cell is densely filled with molecules of all sorts, including proteins, ligands, water, ions, RNA, etc. Most complexes, when formed, however, are very specific, e.g. a particular protein will be able to bind one particular ligand or a small class of ligands, but certainly not every ligand that has roughly the right size, shape and overall charge. This is because tight binding is specific in many sites, i.e. formation of hydrogen bonds, electrostatic complementarity or match of hydrophobic patches need to be favorable enough to overcome the associated entropic cost.

Let us remind the reader to the comparison to protein folding, where the entropically favorable unfolded chain is stabilized in a compact state by native interactions. In the early days of protein folding, Levinthal raised the so-called *Levinthal paradox*, which referred to the kinetics of protein folding: Levinthal assumed that each amino acid could assume at least two conformations (*via* its Backbone rotameric flexibility), all being equally probable except for the stable native state. If it then takes a certain waiting time (e.g. pico- or nanoseconds) to try one such combination, then, given the typical number of amino acids in the protein, so many trials would be necessary that the protein could not fold in the age of the universe. The resolution of this paradox is that different non-native states are *not* equally likely, but tend to become increasingly likely as the native state is approached - which has often been modeled by a protein folding funnel [150, 151, 152].

A similar problem seems to appear in protein-ligand binding. When each binding attempt of a wrong ligand to a protein takes a certain time (*e.g.* nano- or microseconds), can then a correct ligand be found within a reasonable timescale at all unless the concentration is very high? It is very likely that cells have developed efficient sorting and searching mechanisms such that this process is not governed by random trial alone. One aspect is certainly that proteins with related binding partners are often located in proximal positions in the cell. However, there might be additional mechanisms that guide binding partners to attract candidates and then rather quickly reject mismatches.

Consider the Phosphate Binding Protein discussed above in a mixture of anions, its ligand P_i at concentration c_L and other non-ligand anions with the same polarity at concentration c_0. The protein will attract all anions to its pre-binding site, the anion attractor, which is a quick way of screening anions, which are at least known to be in the same charge class in which the ligand also is in. Being at the pre-binding site, the non-ligand (0) cannot bind, and must dissociate some time later, while the ligand (L) can bind tightly. The reaction diagram below describes a kinetic model for this scenario:

$$\begin{array}{lccccc} & P+L & \overset{k_{\mathrm{AB}}}{\rightarrow} & PL & \overset{k_{\mathrm{bind}}}{\rightarrow} & PL^* \\ c_L k_{\mathrm{diff}} & \uparrow & \swarrow k_{\mathrm{off}} & & & \\ & \text{Bulk} & & & & \\ c_0 k_{\mathrm{diff}} & \downarrow & \nwarrow k_{\mathrm{off}} & & & \\ & P+0 & \overset{k_{\mathrm{AB}}}{\rightarrow} & P0 & & \end{array}$$

Figure 4.7: Reaction diagram showing a possible kinetic scenario of anion attraction to a pre-binding site. Starting from the bulk two possible reaction channels can be taken: In case of a ligand (L) the upper path is chosen, in case of a non-ligand (0) the lower one. Both species are attracted to the pre-binding site PL/P0 with their respective concentration dependent rate k_{on}, but only the ligand can move further to the bound state PL. If the concentration of non-ligand is increased with respect to the ligand concentration the lower channel will dominate the reaction system and the pre-binding site will be blocked by the non-ligand more often, which results in a lower yield of PL*.

Based on this reaction diagram, we can calculate the mean time needed to bind a ligand molecule from the bulk, depending on the relative concentration $c_L : c_0$ and the pre-binding affinity $K_a = k_{on}/k_{off}$ with k_{on} being given by Eq. 4.10. The results shown in Fig. 4.8 suggest that a high pre-binding affinity is optimal in the case where no non-ligand is around and thus each binding attempt results in success. In this case, increasing the pre-binding affinity reduces the expected time needed for binding as the associated state PL is then more likely to move on to the bound state PL^* rather than to dissociate. However, the picture drastically changes as soon as non-ligands are in bulk. While it is still true that the time needed for a trial with the true ligand is reduced, this is not the case for non-ligands. Whenever a non-ligand is associated, a high pre-binding affinity will force this non-ligand to stay at the pre-binding site for a long time before it can dissociate and free the site

for the next trial. The optimal settings in this case, i.e. the situation with minimal time needed to find and bind the ligand, is given for relatively small pre-binding affinities. This illustrates how decreasing the pre-binding affinity can be favorable to speed up binding in the sense that the waiting time to the next successful binding event is kept small since blocking times from „wrong“ ligands are avoided.

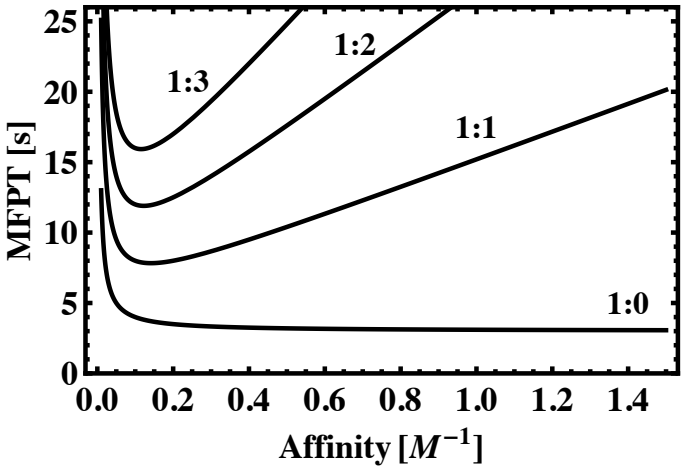


Figure 4.8: Mean binding time of a ligand in a mixture of “right” and “wrong” ligands at concentrations c_L and c_0 depending on the relative concentration $c_L : c_0$ and the pre-binding affinity $K_a = k_{on}/k_{off}$. Rates were set to $(c_0 + c_L)k_{\text{diff}} = k_{AB} = k_{bind} = k_u = 1s^{-1}$ and $k_{off} = k_{on}/K_a$.

4.5 Conclusion

In this study we have presented a computational approach to systematically investigate protein-ligand association kinetics. While existing computational approaches have permitted the calculation of binding energies and rates using a variety of molecular and dynamical models, the presented method goes beyond these by providing an extensive analysis of the entire ensemble of association pathways by which a ligand approaches its target protein, and their relative probabilities. While in the present study a simple electrostatic interaction model was used in combination with rigid body BD, our analysis approach can be readily applied to any molecular dynamical model that allows to calculate or estimate transition probabilities or rates between the substates of the protein-ligand configuration space.

The usefulness of the approach was demonstrated by studying the binding of inorganic phosphate to the Phosphate Binding Protein from *Escherichia coli* and several in silico mutants of it. The analysis reveals that protein mutations that affect surface charges may have subtle to drastic effects on the association kinetics and association pathways. Some mutations affect only association rates without significantly altering the associating pathways, i.e., they scale the fluxes. Other mutations change the association pathways of P_i, and the associated change of the rate, may be of very different magnitude depending

on the exact location of the mutation.

Overall, all systems studied here exhibit binding via a broad ensemble of parallel pathways, indicating a funnel-like energy landscape that narrows down towards the bound state, very similar to the situation in protein folding [153].

Consequently, only very few single-point mutations are able to effectively disable P_i binding. The only single point mutation observed to do this was next to the binding site and thus affected nearly all binding pathways at the bottleneck where they converge. Most other constructed single-point mutations only disabled a subset of pathways, allowing other parts of the pathway ensemble to take over, resulting in only a mild reduction of the association rate. Multiple mutations at critical positions, however, were much more effective and could efficiently disable binding.

The analysis of the mutagenetic behavior revealed the importance of two anion attractors on the surface of PBP which unspecifically attract all negatively charged molecules. This unspecific attraction brings anions closer to the phosphate binding site thus trapping them in a region of limited size. As a result, the PBP wild-type exhibits "superdiffusive association", i.e., associations with a rate that is about three fold compared to the free-diffusion association rate to the binding site that is estimated to be 9.2 $M^{-1}s^{-1}$. With favorable mutations, the association rate may be sped up to about ten times the free diffusion rate.

To experimentally verify our findings, the relevance of different pathways on the protein surface might be assessed by labeling of specific surface residues and P_i and investigation of their contact dynamics using, e.g., NMR.

The grid based discretization used here to define configurational substates is limited to few dimensions and is thus limited to study systems of a size like ligand approaching a rigid protein. However, in future work the approach will be extended to gridless data-based discretization of configuration spaces as they are frequently used in Markov model analysis of protein internal dynamics [90]. With this extension, a flux analysis of association pathways will be possible for complex protein-ligand and protein-protein binding with full dynamical treatment such as all-atom MD in explicit solvent.

One of the questions raised in this chapter is: How a protein can efficiently find its specific binding partner(s) in a vast and dense mixture of different possible binding partners present in the cell. Here we have suggested a simple sorting mechanism that promotes rapid identification of the right ligand by first quickly selecting all potential ligands that fall within the right category (here carrying the right charge), and then attempting to bind. It was found that in the presence of “wrong” ligands, binding is promoted most with a pre-binding site that has a low rather than a high affinity, in order to avoid creating kinetic traps with wrong ligands. This model is yet to be supported by experimental evidence and further sorting mechanisms, e.g. by size, shape, hydrophobicity etc. might exist.

5 Revealing Dynamical Properties of Oligomer Assemblies: Dynamin - A case study

5.1 Introduction

The previous two chapters of this thesis were concerned with developing concepts to study the dynamics of small molecules and the process of protein-ligand association as well as to determine factors that influence these. In the present chapter this conceptual framework is completed by introducing an approach that allows for studying the dynamics of large homo-oligomeric protein assemblies using the Dynamin oligomer helix as an example.

The transportation of molecules from one cellular compartment to another is an essential process for every living cell. In many cases the carried molecules are coated by a lipid bilayer vesicle that originates from the source compartment. To be transported these vesicles have first to be scissored from the source compartment and to be fused later at the target site. In both processes a number of players are involved. It is well known that vesicle fusion is mediated by SNARE proteins [154, 155] and for the vesicle scission process Dynamin and Dynamin related proteins have been identified as key players. Dynamins have first been discovered in temperature sensitive Drosophila melanogaster mutants [156].

Dynamin is a multi domain 100 kDa mechanochemical GTPase that catalyses the scission of clathrin coated vesicles from the plasma membrane in a GTP hydrolysis-dependent manner[157]. Several isoforms of Dynamin exist: Dynamin-1 is specifically expressed in the brain [158, 159, 160], Dynamin-2 is ubiquitously expressed [161] and Dynamin-3 was found at post-synaptic zones [162]. It is known that Dynamin is recruited to the membrane via its proline-rich domain (PRD), where it oligomerizes into helical assemblies around the neck of nascent vesicles. Once the helical assembly is formed it catalyzes the scission reaction in a GTP-hydrolysis dependent manner [163, 164]. The detailed dynamical mechanisms remain still elusive. However, a number of different possible mechanisms have been proposed for the scission reaction including constriction [19], extension [20] as well as twisting [21] of the vesicle neck.

The recently resolved structures of the Dynamin protein [22] (Figure 5.1) and the dimerized Dynamin G domains [165], have already contributed to a better structural understanding of the scission mechanism. In the present chapter we make use of this structural

knowledge and use molecular modeling and dynamics to extend the understanding of the scission process in terms of its dynamical component.

From previous studies it is known that the dimerization of Dynamin G domains plays an important role in the formation of higher order Dynamin assemblies [165], and, furthermore, that GTP hydrolysis is an essential step to perform the scission mechanism [20, 166]. For a comprehensive understanding of Dynamin mediated vesicle scission it is hence inevitable to i) gain a better understanding of the G domain interplay and their interactions and ii) to investigate dynamical properties of the Dynamin protein and consequences for higher order Dynamin assemblies. To pursue these questions we first estimated the free-energy change involved in the Dynamin G domain dissociation, depending on the bound nucleotide (i.e. GDP or GTP) by means of umbrella sampling (refer also to Section 2.2.3 in Chapter 2) along a potential G domain dissociation pathway. From these simulations it is found that approximately twice as much energy is needed for the dissociation of GTP bound Dynamin G domains as for GDP bound domains. This indicates that the interactions between GTP bound domains are very strong compared to GDP bound ones and that GTP hydrolyzation might be essential for a relative movement of the domains and thus for the scission process. To study the consequences of the domain interactions in the context of the whole Dynamin oligomer helix additional studies were performed. Based on the constricted conformation of the Dynamin stalk tetramer a Dynamin oligomer helix was modeled and its dynamics simulated. The constricted Dynamin stalk conformation was manually constructed by i) fitting Dynamin stalk elements (Figure 5.3) into a cryo-EM density of a constricted Dynamin helix [167] and ii) modeling of unresolved loop regions with the aid of evolutionary conserved interactions. To evaluate the internal flexibility of the tetramer as well as the structural stability of modeled regions a number of molecular dynamics simulations were performed. The generated trajectory data served further to investigate the dynamics of an oligomeric Dynamin helix structure. By considering the tetramer as mechanical building block and extrapolating its dynamics to the whole helix it was possible to obtain a free energy landscape in helix coordinates. With the aid of this landscape it was finally possible to relate previously proposed models of Dynamin dynamics to structural properties of the modeled helix and allows us to draw conclusions about the various proposed scission mechanisms.

5.2 Methods

Umbrella sampling of Dynamin G domain stability Using the crystal structure of the Dynamin GTPase [165] homodimer (pdb code 2X2E), the free energy change of G domain dissociation was estimated by means of molecular dynamics umbrella sampling simulations. To assess nucleotide induced stability differences, two dissociation scenarios were simulated: GTPase:GDP - GTPase:GDP and GTPase:GTP - GTPase:GTP. The GDP bound GTPase

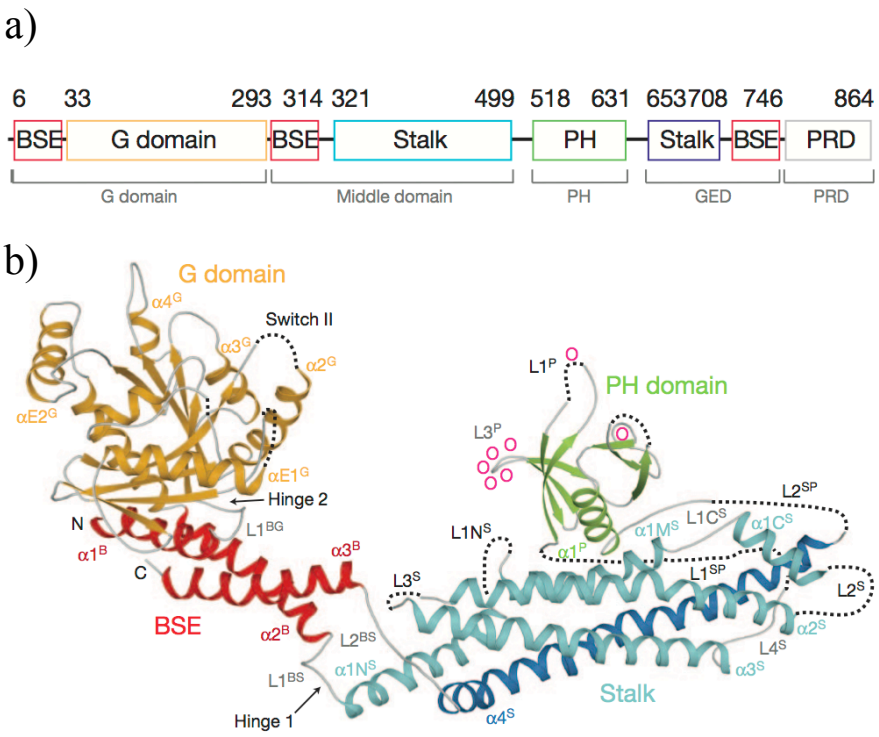


Figure 5.1: Structure of nucleotide-free human Dynamin-1. a) Domain architecture of Dynamin-1. b) Distinct structure elements of Dynamin-1 in cartoon representation. Dotted lines indicate regions not resolved in the crystal. Domains are colored and labelled. Residues binding to lipids are depicted as **O**.

was prepared by deleting the two ALF residues at the γ phosphate position from the crystal structure. The GTP bound GTPase was analogously prepared by replacing the ALF residues with a PO_3^- γ-phosphate. The water molecules present in the crystal structure were retained and the MSE residue was replaced by standard MET. As simulation suite Gromacs 4.5.3 [29] with the AMBER99SB [168] force field was employed. GDP and GTP force field parameters were taken from [169] and converted to be Gromacs compatible by using Antechamber [101, 170] and Acpype. The prepared systems were immersed in a $7.23 \times 8.54 \times 20.0$ Å^3 water box containing 37.000 water molecules. Further an ion concentration of 0.1M NaCl was assumed, including counter ions to neutralize the system (81 Na^+, 73 Cl^-). The simulations were carried out at constant temperature (T=300K) and pressure conditions (p=1atm) using the Parinello-Rahman barostat[171]. If not stated differently, the protocol described in paragraph *Loop modeling and molecular dynamics simulations* was used for simulation and equilibration.

As reaction coordinate in umbrella sampling sense, it was assumed that the homodimer dissociates along the z-coordinate of the given structure file. To construct umbrella window starting configurations along this coordinate the dimer was first pulled apart using the Gromacs pull code. For this purpose one domain (chain A) was positional restrained and

the other (chain D) dragged away by employing constant center of mass (COM) velocity pulling with a pulling speed of $0.01\,nm\ ps^{-1}$ for 600 ps. The pulling resulted in a final COM distance of about 8 nm for both structures. From the obtained pulling trajectories snapshots were taken as umbrella sampling windows. The window spacing ranged from 0.1 nm (COM separation 4.42 - 6.9 nm) to 0.2 nm (COM separation 6.9-8.1 nm). This led to 32 umbrella windows, in each of which a MD sampling of 10 ns was performed. The reconstruction of the free energy along this coordinate was carried out using the WHAM procedure of Gromacs [172]. The amount of sampling performed in each umbrella window is depicted in SI Figure 7.11.

Loop modeling and molecular dynamics simulations For modeling of the unresolved loop regions $L1N^S$ and $L2^S$, two stalk dimers of the Dynamin stalk in the constricted state served as scaffold. Employing Modeller (9v8) [173], the scaffold was fixed in position, whereas $L1N^S$ and $L2^S$ could freely sample the empirical potential function. To reduce the conformational search space, additional harmonic distance restraints were added between conserved residues Arg399-Asp406 and Glu355-Arg361. Based on the modeled stalk tetramer, five independent all-atom molecular dynamics simulations (NVT ensemble), each of 90 ns length, were conducted at T=300 K in a periodic boundary setting by using Gromacs 4.5.3 [29]. The model was immersed in a rectangular 20 nm x 10 nm x 9 nm box, containing approximately 56,400 water molecules, 21 sodium and 17 chloride ions to neutralize the system, resulting in a total number of 185,857 atoms. As force fields, Amber99 (protein and ions) [102] and TIP3P (water) [174] were applied. To treat long range interactions, the Particle-mesh Ewald method [175, 176] was used. A cut-off of 1 nm was used for the real parts of electrostatic and van der Waals interactions. All hydrogen bonds were constrained by using the LINCS [103] algorithm, allowing for an integration time-step of 2 fs. For the thermostatted integration, Langevin dynamics were used as implemented by the Gromacs SD integrator (tau_t=1). For the calculation of bending and twisting angles in Figure 5.2c, each of the four stalk monomers was represented by two geometric centers, defined as the mean position of Cα atoms of residues 366-377, 420-430, 468-481 and 671-683 for position A and residues 360-365, 428-445, 457-472 and 686-701 for position B (Figure 5.2a). The stalk bending angle α was defined as the mean angle between parallel stalks, and the twisting angle β by the minimal angle between the planes spanned by each dimer (positions A, B, B' in Figure 5.2a).

Dynamin helix construction For each simulation time step, the corresponding stalk tetramer structure describes a linear transformation of the first dimer onto the second dimer, consisting of a translation vector and a rotation matrix (Figure 5.2b). This linear transformation was used to reconstruct the structure of an ideal Dynamin helix by applying it to the Dynamin dimer model in the constricted state. The diameter, d, and the

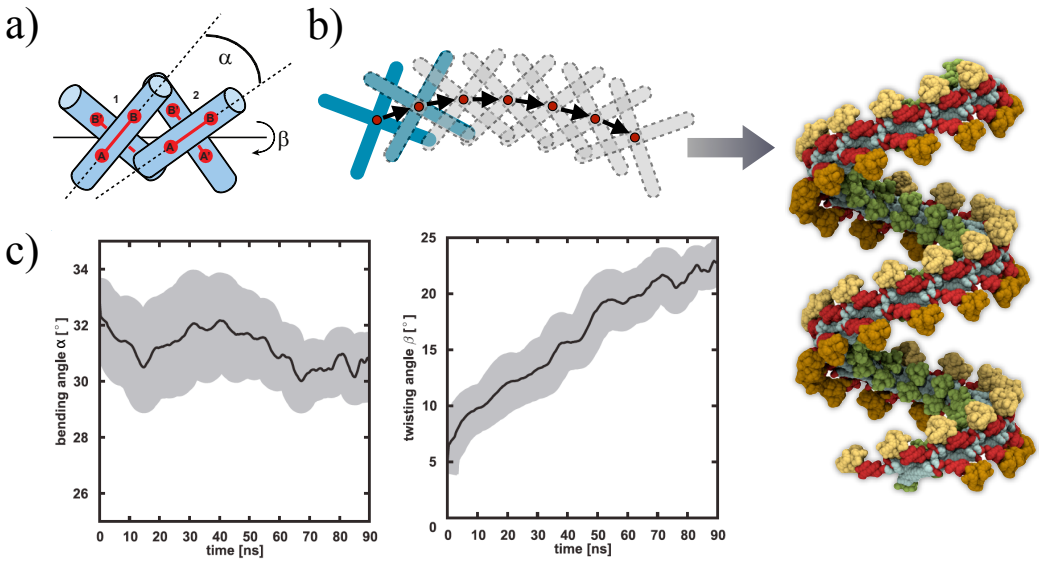


Figure 5.2: Stalk tetramer based modeling of the Dynamin helix. a) Dynamin stalk tetramer as mechanical building block with bending angle, α, and twisting angle β. b) Superposition scheme of the Dynamin stalk tetramer used to construct the helical Dynamin oligomer (right). c) Bending and twisting angle evolution over simulation time course. The black lines indicate the mean value over five independent simulations, the grey area denotes the associated standard error.

rise per turn, r, of these helices were measured by using the geometric centers of the stalk coordinates and obtaining trajectories in (d, r).

Free energy computations Based on these trajectories, the free energy surface of stalk helix conformations was calculated in the following way: The two-dimensional space (d, r) was discretized into boxes of size 25 x 25 Å. Based on the simulation trajectories, the transition probability between all pairs of boxes was computed, which allowed the calculation of an equilibrium probability of finding a single tetramer in a given box, $p_1(d, r)$ [47]. When more than two dimers are assembled, non-cooperative behavior of neighboring dimers has to be considered, e.g. the dimers of the helix can almost independently switch between different conformations. The resulting equilibrium distribution of two independent tetrameric units would be given by the convolution of two single-tetramer distributions, $p_2(d, r)$. It was found that for only about 3 such convolutions, the resulting probability distribution converges to $p_3(d, r) \approx p(d, r)$. Thus, assuming that the helix has at least three independently switching subunits, the free energy landscape is unique, and is given by $F(d, r) = -k_B T \ln(p(d, r))$, where k_B is the Boltzmann constant and T the temperature.

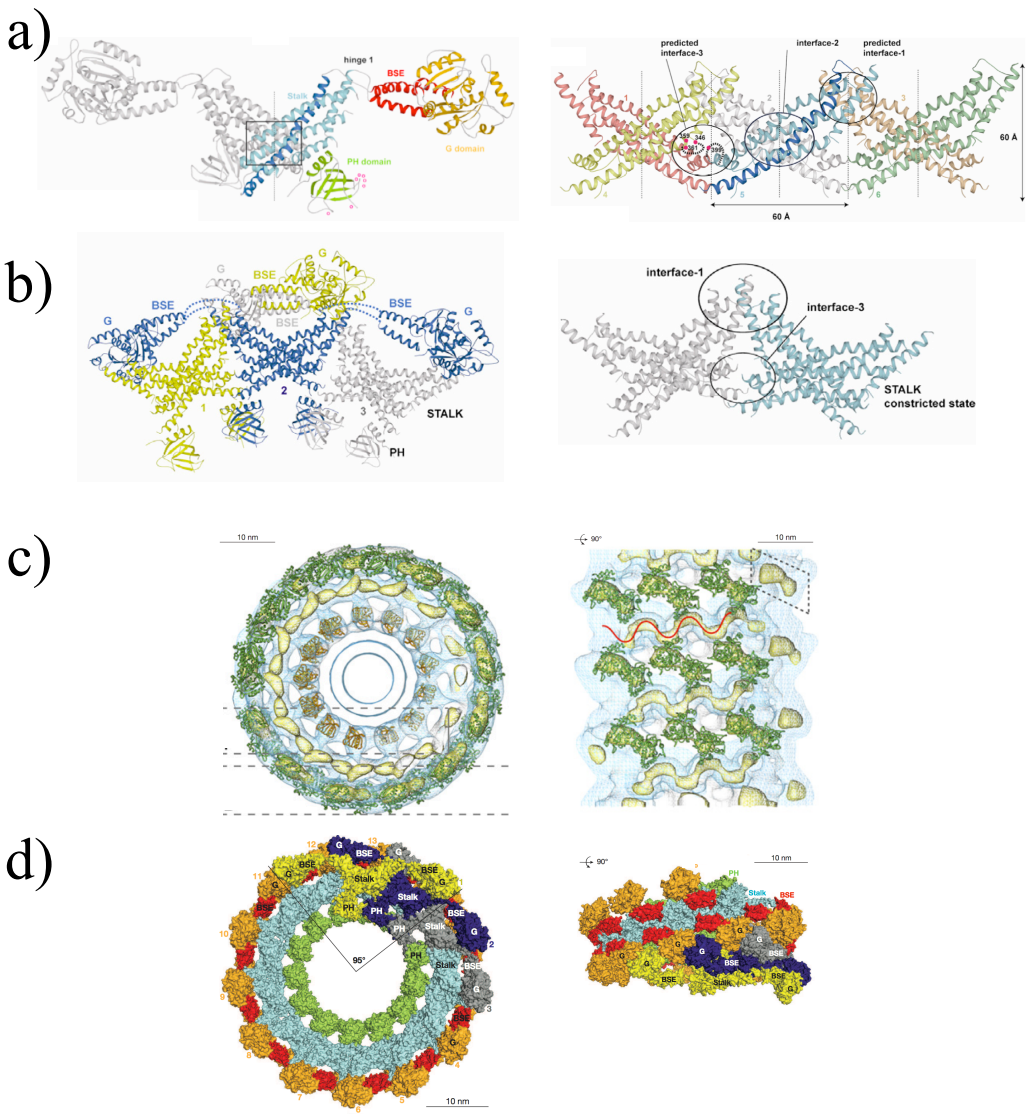


Figure 5.3: Dynamin dimer and assemblies a) left - Dynamin dimer as obtained from the crystal, right - criss-cross like arranged Dynamin stalk as present in the crystal b) left - Details of Dynamin dimer (same color) assembly and interaction in the constricted helix state, right - constricted Dynamin stalk tetramer bent to fit constricted Dynamin cryo-EM density c) Cryo-EM maps of constricted state Dynamin including fitted rat GTPase (green) and human PH (orange) structures. Left - looking down the helical axis, right - turned 90 degrees. d) Cryo-EM based model of the oligomerized Dynamin-1 helix in two perspectives, showing PH domain (green), stalk (light blue), BSE (red) and G domain (orange). Subfigure c) was adopted from [167].

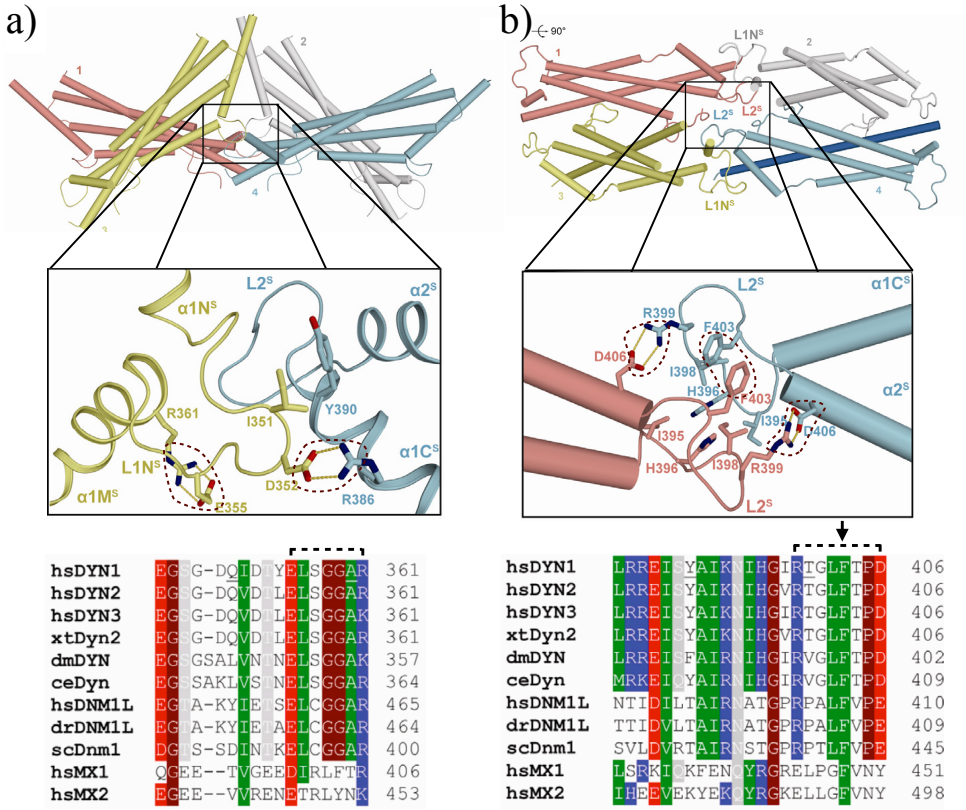


Figure 5.4: Two perspectives of the constricted Dynamin-1 stalk tetramer and detailed illustration of interface-3. a) bottom, showing the position of the modeled loop $L1N^S$ b) bottom, showing the position of the modeled loop $L2^S$. The modeling procedure was guided using distance constraints between residue pairs. The residue pairs were chosen based on sequence conservation revealed by a multiple sequence alignment (bottom). In the zoom-in the relevant pairs are encircled by dotted lines. On sequence level the interacting residues are marked by a dotted line and an arrow respectively.

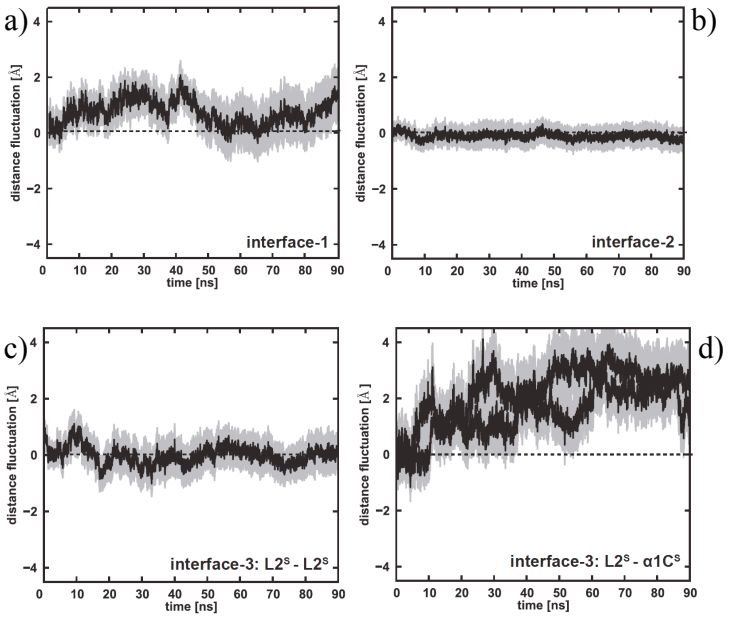


Figure 5.5: Distance fluctuations of predicted Dynamin interfaces-1, -2 and -3 based on 90 ns molecular dynamics simulations of the modeled stalk tetramer in the constricted state. The standard error is indicated in gray (n=5). a-b), Modeled interface-1 and interface-2 are stable, indicated by small fluctuations around the baseline. c) While the central $L2^S$ - $L2^S$ interaction of opposing stalks remains stable, d) the flanking $L1N^S$ - $\alpha1C^S$ interaction opens (increasing slope).

5.3 Results and Discussion

Stability of the G domain interaction and potential of mean force From previous experiments it was known that the interactions of Dynamin G domains and GTP hydrolyzation play an important role in the process of Dynamin mediated vesicle scission. In order to gain a better understanding of the G domain interaction, i.e., strength of the interaction depending on the nucleotide bound, the two domains were pulled apart (Figure 5.7a) by means of steered molecular dynamics simulations. This created a potential dissociation pathway that served as a coordinate for further umbrella sampling simulations. Using umbrella sampling along this coordinate the free energy change from bound to unbound was calculated for GDP and GTP bound domains. The resulting free energy profile is depicted in Figure 5.7b. Here it becomes apparent that the interaction between GTP bound Dynamin G domains is approximately twice as strong as if GDP is bound to the domain. This indicates that a hydrolysis of GTP might be necessary for changing the relative orientation of the two domains. This fact is especially relevant for understanding the dynamics of Dynamin in higher order helix assemblies. It might be for example possible that certain helix conformations and conformational changes are only possible if a defined number of G domains is in a certain state (GDP or GTP bound). To understand the consequences of the different strength in G domain interactions for the dynamics of the whole Dynamin

oligomer helix, the dynamics of this helix was investigated as described subsequently.

Loop Modeling and Interface Stability In order to obtain the constricted Dynamin stalk conformation it was necessary to bend the linear stalk crystal formation (Figure 5.3a-b) to fit the Dynamin helix cryo-EM density (Figure 5.3c). This procedure resulted in a relative shift of the stalk dimers such that residue contacts especially in interface-1 and -3 can be lost, formed or changed. While interaction in interface-1 were resolved in the crystal structure, the unstructured region of interface-3 was not. For an assessment of the stability and plausibility of interface-3 it was hence necessary to obtain a structural model of this region first. By utilizing the information about evolutionary conserved interactions in this interface a sound model could be obtained using an empirical energy function based conformational search procedure. The result is depicted in Figure 5.4, the loop regions of both stalk dimers obey symmetry principles while forming conserved interactions.

The applied modeling procedure is solely based on minimizing an energy functional, thereby not taking temperature or interactions with water molecules into account. Hence, molecular dynamics trajectories were generated to evaluate the interface stability under solvated and finite temperature conditions. Using five independent trajectories of the 90 ns length each, the dynamic stability of the three interfaces was measured as distance fluctuations between different dimers. Interface-1 and interface-2 (Figure 5.5) show both a high stability. This is expected as they are formed by well structured interactions. In modeled interface-3 the fluctuations differ for the two regions considered. The inner region formed by contacts of the inner loops $L2^S$-$L2^S$ (Figure 5.4b) does not show large fluctuations and is stable on average (Figure 5.5c). In contrast does the flanking interaction between loop $L1N^S$ and helix $\alpha 1C^S$ exhibit large fluctuations (Figure 5.5d) including a breaking of residue contacts. While in the modeled stalk tetramer this behavior is expected as the loop can alternatively interact with surrounding water molecules, in the formed Dynamin helix this flexibility will be reduced by neighboring domains.

Dynamin oligomer helix dynamics and relation to existing mechanisms The molecular dynamics simulations of the stalk tetramer were used to model the dynamics of the Dynamin oligomer helix and its accessible conformational space (see Methods - Dynamin helix construction). Given this model and the dynamical information we computed a free energy landscape in the two helix parameters: diameter, d, and radius r. The landscape is depicted in Figure 5.6, five representative helix structures are shown (Ia, Ib, II, III, IV) with their respective landscape position. Previously suggested models (poppase [20], twistase [21] and constrictase [19, 21]) for Dynamin helix dynamics and their potential trajectories are marked in the landscape by arrows. In the GDP-bound / nucleotide-free form, G domains of neighboring turns do not interact (state Ia, Ib). Following GTP binding, G domains interact and stabilize the constricted state of the Dynamin helix with a

helix rise per turn of approximately 100 Å (state II and IV). After dimerization-induced GTP hydrolysis, G domains dissociate, thereby releasing the G domain constraints. For a rigid Dynamin template of fixed diameter, the Dynamin helix follows a vertical trajectory (blue arrow) to the energy minimum (e.g. state II→Ia) apparent by an extension of the helix, as suggested in the poppase model (template size was chosen here to be similar as in Ref. [20]).

When the template is deformable, the diameter of the Dynamin helix can vary, leading to a constricted GTP-bound (state IV) and a non-constricted GDP-bound state (state II), as proposed in the constrictase model. During relaxation/constriction, the stalks of neighboring turns slide against each other via an intermediate state III until the new equilibrium position (constricted state) is reached. This sliding can be observed in live assays of Dynamin function and has been termed twistase [21]. This transition is facilitated by transiently breaking G domain interactions by the GTPase reaction, followed by partial helix constriction and re-establishment of G domain interactions after GTP rebinding. Note that for a full-length Dynamin oligomer assembled on a lipid template, the energy landscape is likely be further modified by BSE-stalk, PH domain-stalk and PH domain lipid interactions.

Based on these considerations, we propose the following model for Dynamin-catalyzed membrane fission. The Dynamin oligomer assembles in the GTP-bound form around the vesicle neck. After formation of a complete helical turn, G domain contacts are established, enforcing a helix with a rise per turn of 100 Å and a diameter that depends on the initial template size. In the continuous presence of GTP, the helix constricts via a constrictase/twistase mechanism towards the constricted state (IV). GTPase activity and rebinding of GTP might induce a local opening and twisting of the Dynamin helix. The resulting shear forces in the underlying membrane template might cause it to break. Whether scission occurs by a full-relaxation from the constricted state (poppase) or during transient openings is still an open question.

5.4 Conclusions

The last chapter of this thesis was concerned with the study of macromolecular interactions and dynamics of homo-oligomeric assemblies. Based on existing structural knowledge of the Dynamin protein and specific G domain interactions, molecular modeling, molecular dynamics and Markov state modeling were used to gain a better understanding of Dynamin's role in the yet poorly understood process of clathrin mediated endocytosis of vesicles. It was known that the interactions of Dynamin G domains and the hydrolysis of GTP play an important role in this process [20, 166]. In order to gain a better understanding of this macromolecular interaction computational pulling experiments were performed. Here the two interacting G domains of the Dynamin G domain dimer were pulled apart

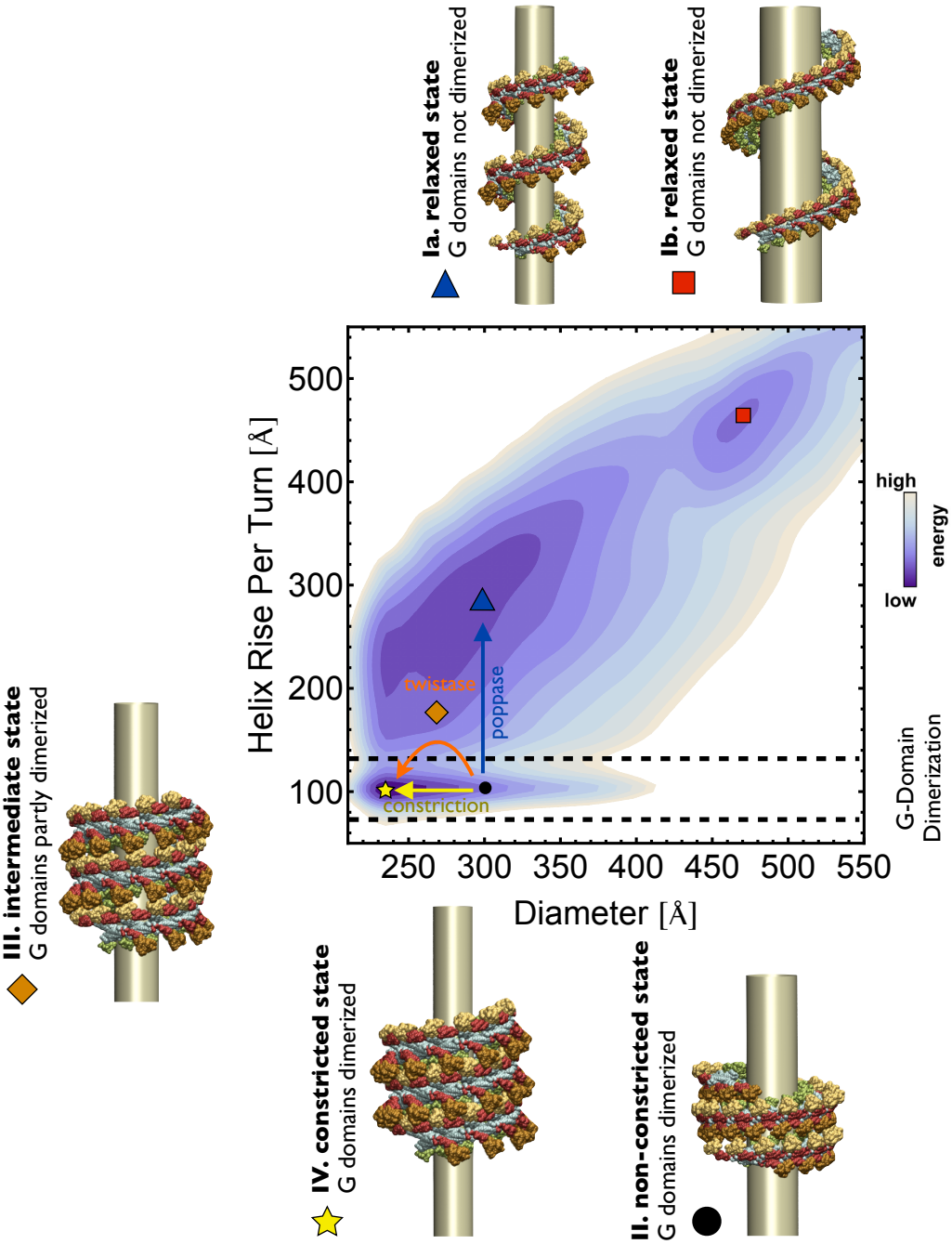


Figure 5.6: Free energy landscape of the Dynamin stalk tetramer in coordinates of a modeled Dynamin helix. The calculations are based on molecular dynamics simulations and tetramer superposition. Five representative Dynamin helix conformations and their respective energy landscape position are shown. Dotted lines mark the restraint imposed by dimerized G domains, it can be released when GTP is hydrolyzed by Dynamin. Previously suggested models of Dynamin dynamics and their respective trajectories in this energy landscape are depicted by named arrows.

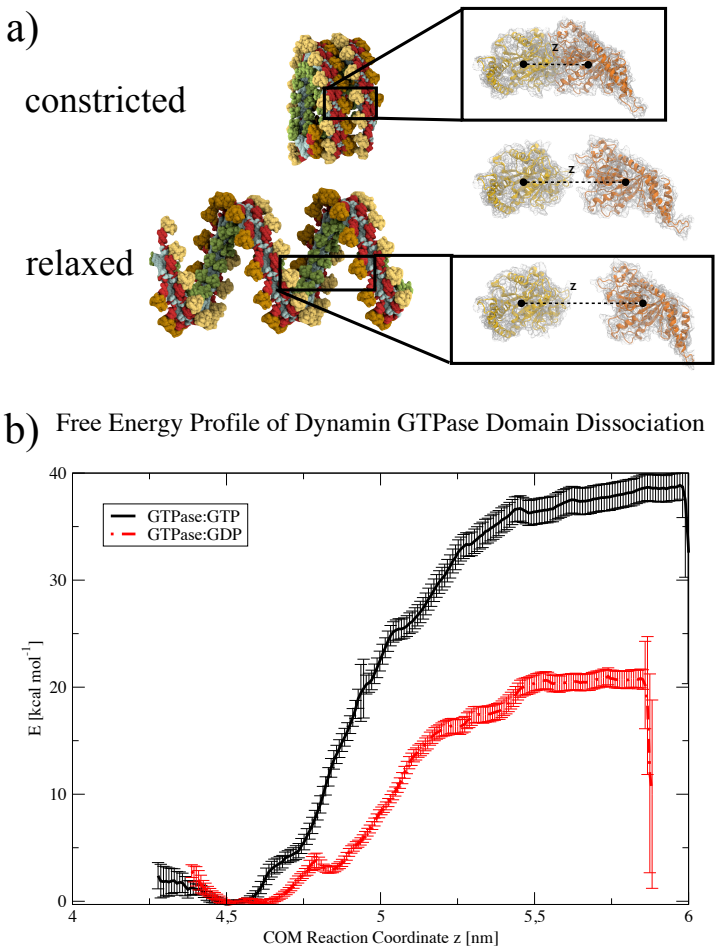


Figure 5.7: Stability of the Dynamin G domain dimers in complex with different nucleotides. a) Constricted and relaxed state of the Dynamin helix. The zoom-in shows the dissociating G domain dimer with the dissociation pathway used to calculate free energy profiles. b) Free energy profiles of Dynamin G domains in complex with GDP and GTP along a particular dissociation pathway. The profiles and error bars were computed via `g_wham` [172].

and the resulting dissociation pathway served as reaction coordinate for umbrella sampling simulations to calculate the free energy profile of Dynamin G domain dissociation. This procedure enabled an assessment of qualitative differences in the dimer interaction strength depending on the nucleotide bound, i.e., GDP or GTP. The analysis revealed that GTP bound G domains interact more strongly than the GDP bound ones. This indicates that the hydrolysis of GTP renders the dissociation of two G domains more likely, a fact important for understanding the complex interplay of Dynamins when found in an oligomeric helix. Note that the free energy calculations presented here were based only on a single dissociation pathway. Hence, entropic contributions, e.g., from different pathways and changes in accessible conformational space of the proteins are not adequately considered. The reported free energy values are thus not to be interpreted as absolute values. However, under the assumption that pathway ensemble and the change in accessible conformational space are similar for the GDP and GTP bound domains, the qualitative difference found for GDP and GTP bound domains is captured.

To gain further insights about the dynamics of the Dynamin protein when forming an oligomeric helix, this helix was modeled and simulated using the Dynamin stalk domain in a tetrameric arrangement. By extrapolation of the stalk tetramer dynamics to the entire Dynamin helix and application of Markov state model theory it was possible to derive a free energy landscape in Dynamin helix parameters. This energy landscape allowed the integration and relation of existing models of Dynamin mediated vesicle scission. By further incorporation of knowledge obtained from the G domain interactions study it was possible to link the mechanical properties of the Dynamin stalk element to the helix dynamics and to draw novel conclusions about the biophysical mechanism of Dynamin mediated endocytosis.

6 Conclusions

Modern biophysical research is driven by the fruitful interplay of experimentation and computational modeling (Figure 1.1). For both strategies, it is challenging to study biological systems as a whole. Due to the enormous complexity which arises from both their large heterogeneity as well as the large span of relevant time and length scales. Hence, experiments and modeling approaches usually target subsystems and seek to embed the gained insights into the context of the entire system. Such an embedding is usually challenging as subsystems can mutually interfere with each other. A common approach to face this challenge is to define a metamodel that incorporates interactions between subsystems and accounts only for the most relevant features of the subsystems. Describing the subsystems by only their relevant features reduces the time and length scale problem. However, it is not clear a priori which features of a subsystem are relevant to adequately model its behavior in the context of a larger system. For this reason, it is hence necessary to study individual subsystems at a high level of detail with the aim to identify properties that are most relevant in the interaction with other subsystems. In this context, this thesis employs a novel combination of recently developed methods in the field of conformational dynamics in order to identify relevant determinants in the biophysical processes of cellular protein interaction dynamics (Figure 1.2b). The modeling and simulation results were individually analyzed and interpreted but may also be embedded, e.g., in a higher level systems biology model of a metabolic network or serve as input to parameterize a cellular dynamics simulation.

In particular three problems of general interest were studied: i) How does the chemical environment affect thermodynamic and kinetic properties of a small ligand molecule? ii) How is the process of protein-ligand association characterized in terms of kinetics and binding pathway ensemble? iii) How can atomistic modeling be employed in order to predict the flexibility and dynamics of large macromolecular assemblies?

First, a systematic theoretical approach was developed to characterize the impact of the chemical environment on structural, energetic and kinetic properties of small ligand molecules. Based on molecular dynamics simulations and Markov state modeling, the conformational dynamics of UDP-GlcNAc, the key substrate in the sialic acid synthesis pathway, was investigated in four different chemical environments (vacuum, water, water+Mg^{2+}, protein). The systematic analysis revealed a number of unexpected phenomena. It was for example found that transition rates between metastable conformations

can be sped up in the presence of an Mg^{2+} ion. Furthermore, it was discovered that binding competent conformations of UDP-GlcNAc are stabilized in the presence of an Mg^{2+} ion. This suggests that ions associated to ligands might in general serve as "binding-cofactors" as they support the selection of the right ligand binding conformation.

This thesis further contributes an approach which enables the theoretical study of the protein-ligand association process. With the novel methodology, it is possible to elucidate questions like: Do specific association pathways exist when a ligand binds to a protein? What factors influence these and the association kinetics? To answer these questions Markov state models (MSMs) were used in conjunction with transition path theory (TPT). By applying both methods to the problem of diffusional two-body association, the usual application domain of MSMs and TPT was expanded to be applicable to an additional class of problems. We showed, as an example, how the derived method enables the description of the binding path ensemble of a phosphate ion to the phosphate binding protein in E.coli. By investigating the influence protein mutations have on the kinetics and binding path ensemble, a novel hypothesis for a ligand sorting mechanism was derived that may be relevant for efficient search of binding partners in the cell. It was found that an unspecific binding site might increase the concentration of potential candidate ligands (anions) close to the phosphate binding protein, thus, rendering the selection of the "right" ligand at the actual binding site more efficient. Using a simple model, we could further show the existence of an optimal affinity for the unspecific binding. Increasing the unspecific affinity by targeted protein mutations did not necessarily lead to more efficient binding of the "right" ligand.

In the last part of the thesis, existing modeling approaches were employed to combine data of protein structure experiments and to extend the present knowledge by a dynamical interpretation. In particular, two problems were addressed: i) use of molecular modeling to derive a structure prediction for a large protein complex based on experimental data, ii) use of atomistic molecular dynamics simulations of a small complex building block to make statements about the dynamics of the whole complex.

Based on the discovery of the crystal structure of Dynamin-1 and the existing coarse cryo-EM density of the Dynamin helix, it was possible to derive a detailed structural model of the Dynamin helix. To study the dynamics of this helix and to link it to existing biophysical models of Dynamin mediated endocyctosis, the helix dynamics was approximated by using molecular dynamics. To accomplish this approximation, we identified the tetrameric dynamin stalk as key building block of the Dynamin helix from the structural experimental data. Simulating the tetrameric building block by means of molecular dynamics allowed for the extrapolation of local building block dynamics to an entire Dynamin helix. The projected helix dynamics was further analyzed by Markov state modeling, which enabled the computation of a free energy landscape on the Dynamin helix conformational space. Using this landscape together with simulation results obtained for the stability of Dynamin

G domain interactions, a novel scission mechanism was proposed.

7 Appendix

7.1 Supplement - Chapter 3

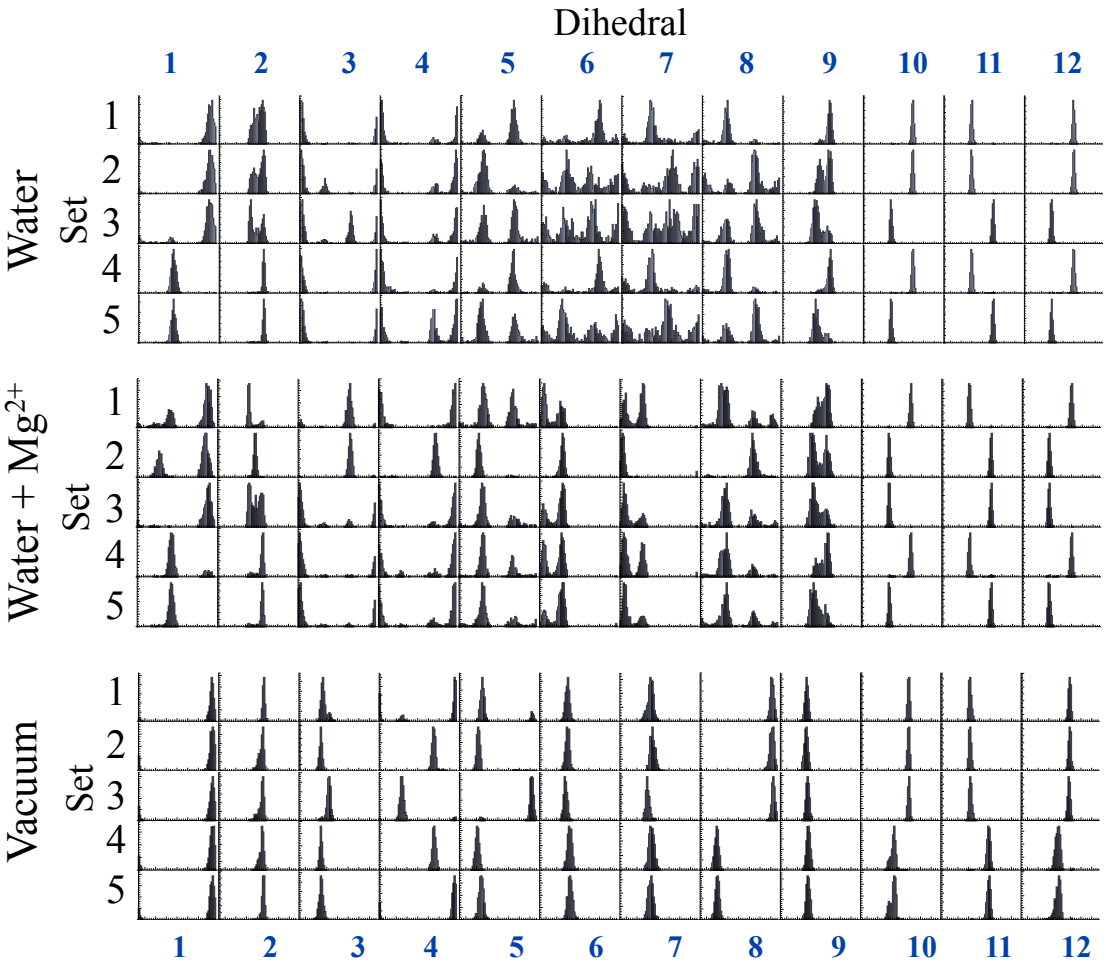
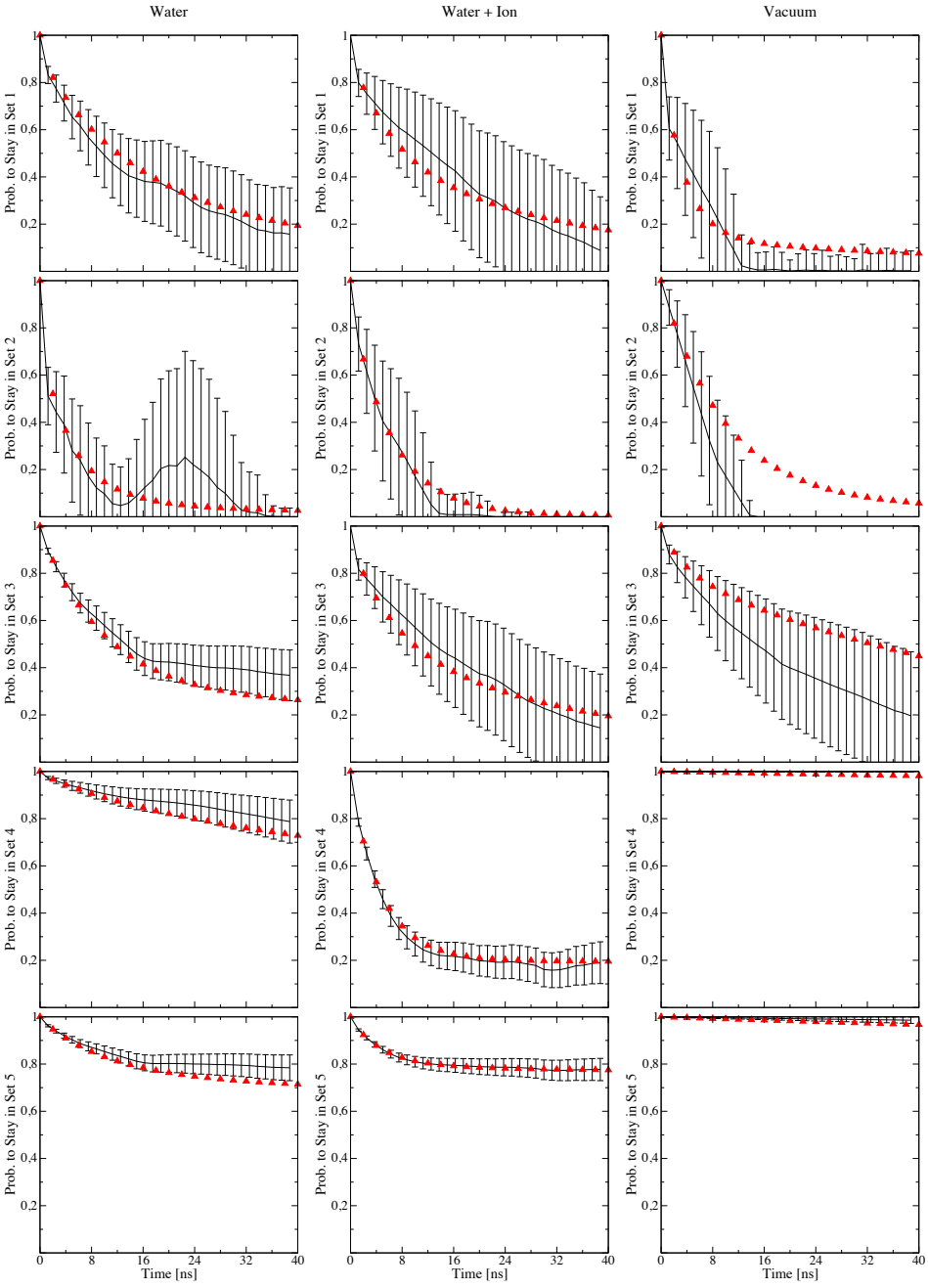


Figure 7.1: UDP-GlcNAc dihedral histograms of identified metastable sets of water, water+Mg^{2+} and vacuum systems. The range of each dihedral histogram is from -180° to 180°, the bin size is 5°.



(a) Water **(b)** Water+Mg^{2+} **(c)** Vacuum

Figure 7.2: Chapman-Kolmogorov tests for identified metastable sets for (a) water, (b) water+Mg^{2+} and (c) vacuum simulation systems. Given the configuration starts in one of the sets the plots show how many of the systems conformation is expected to stay in the respective set after a given time. The line with error bars shows the expected probability directly calculated from simulation data, the red triangles depict the probability predicted by the Markov model.

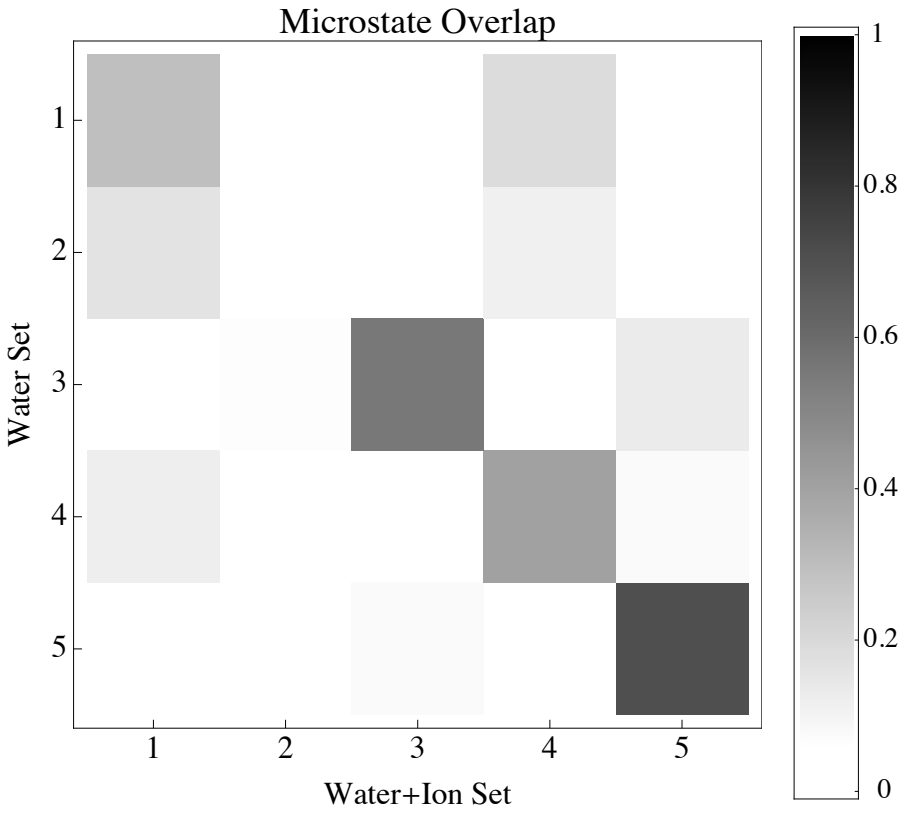


Figure 7.3: Microstate overlap between identified metastable sets of water and water+Mg^{2+} systems. The overlap O between two sets i, j is computed as $O(i,j) = |S_i \cap S_j|(\max(|S_i|, |S_j|)^{-1}$, where S_i denotes the set of all microstates in PCCA set i.

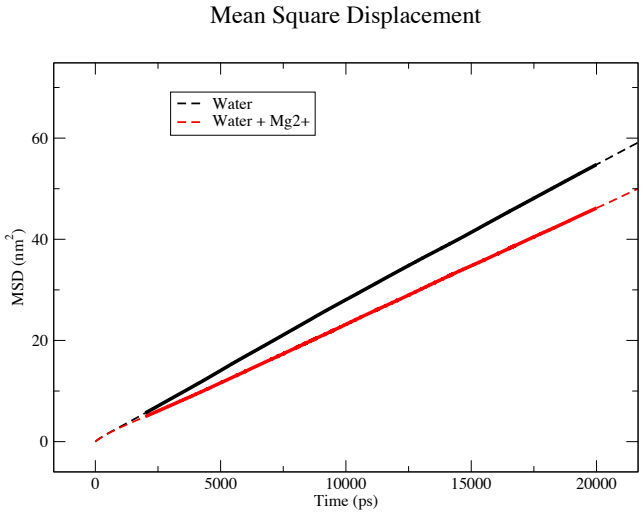


Figure 7.4: Mean Square Displacement of UDP-GlcNAc in water and in water+Mg^{2+} calculated from simulation trajectories. The diffusion constants were obtained by fitting linear functions ($\text{MSD}_{Water}(t) = 0.0027227t + 0.57764$, $\text{MSD}_{Water+Mg^{2+}}(t) = 0.0023041t + 0.14976$) to the mean square displacements calculated between 2.5 ns and 20 ns (thick parts of the line). Thus leading diffusion constants $D_{\text{Water}} = 4.53783 \times 10^{-6}\, cm^2 s^{-1}$ and $D_{\text{Water+Mg}^{2+}} = 3.84017 \times 10^{-6}\, cm^2 s^{-1}$.

7.2 Supplement - Chapter 4

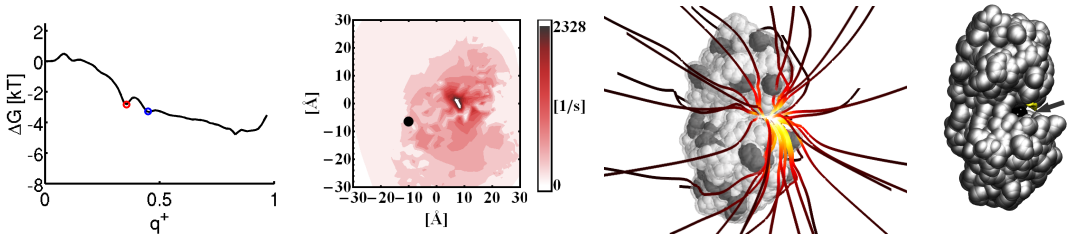


Figure 7.5: Mutant A197W - Committor Free Energy Profile, First Hitting Density, Association Pathways, Mutation Site

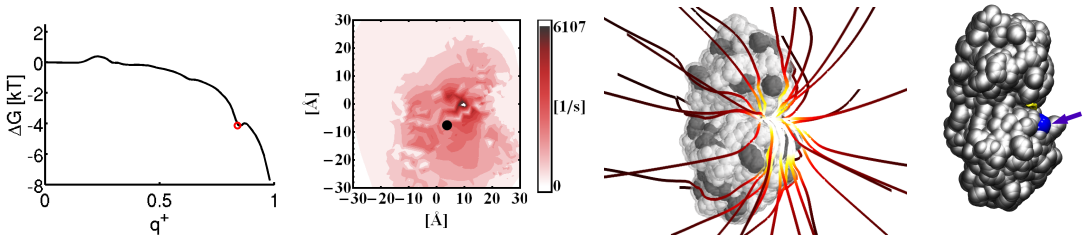


Figure 7.6: Mutant D137T - Committor Free Energy Profile, First Hitting Density, Association Pathways, Mutation Site

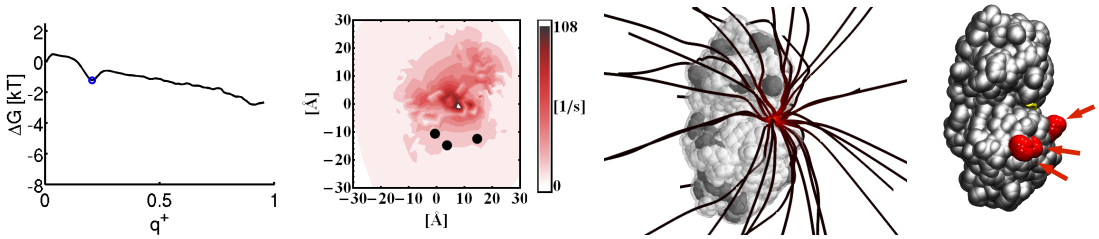


Figure 7.7: Mutant 3 mut. - Committor Free Energy Profile, First Hitting Density, Association Pathways, Mutation Site

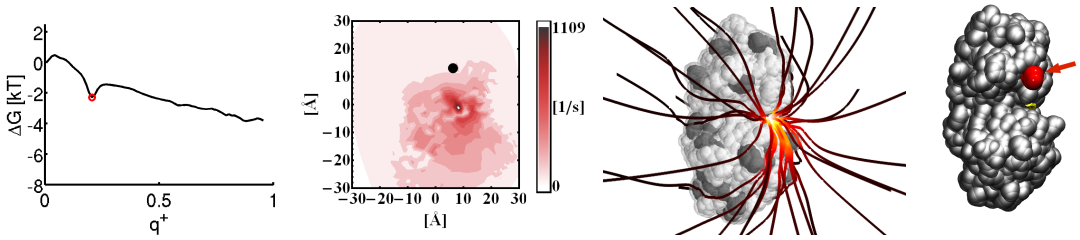


Figure 7.8: Mutant K43Q. - Committor Free Energy Profile, First Hitting Density, Association Pathways, Mutation Site

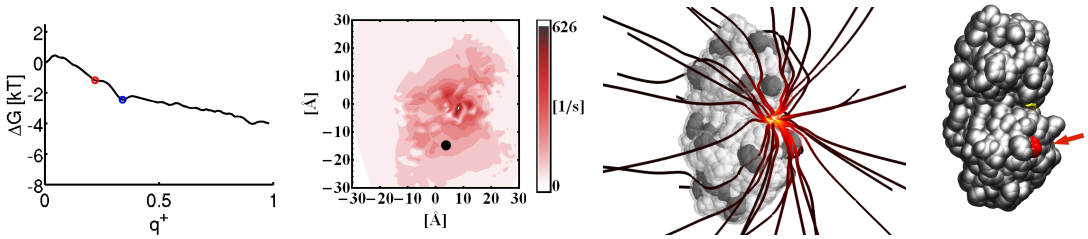


Figure 7.9: Mutant R134Q. - Committor Free Energy Profile, First Hitting Density, Association Pathways, Mutation Site

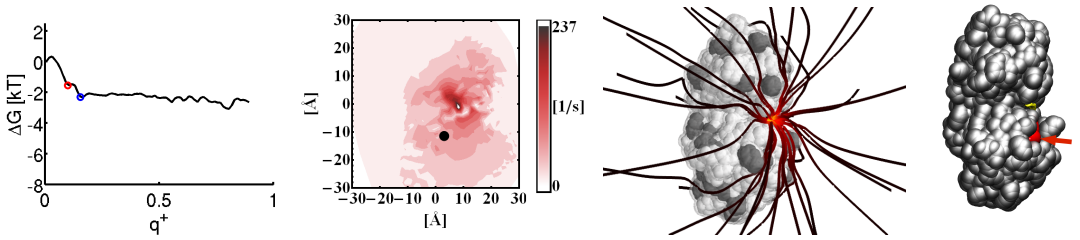


Figure 7.10: Mutant R135Q. - Committor Free Energy Profile, First Hitting Density, Association Pathways, Mutation Site

7.3 Supplement - Chapter 5

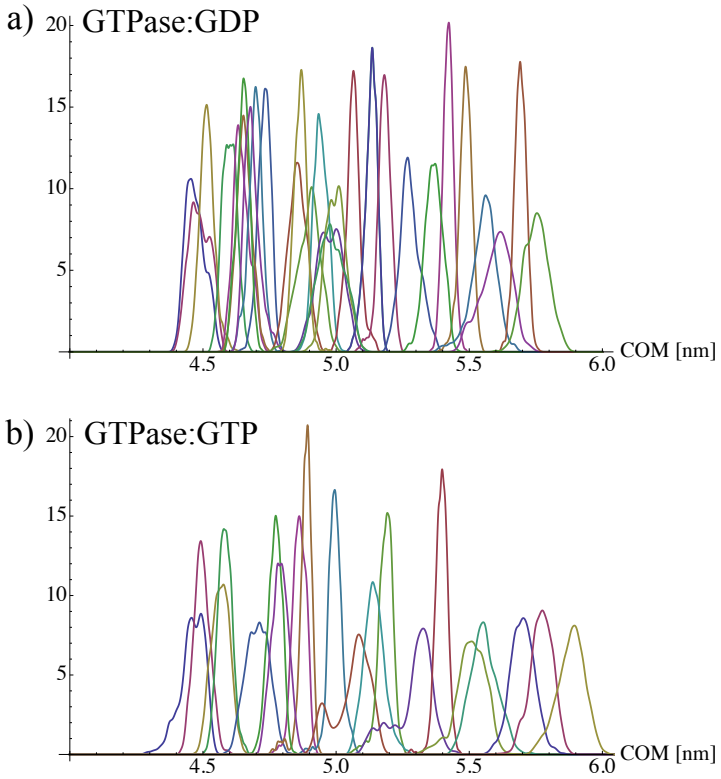


Figure 7.11: Umbrella sampling histogramss along the center of mass (COM) domain separation coordinate. a) Histograms of the Dynamin GTPase:GDP complex umbrella sampling simulations b) Histograms of the Dynamin GTPase:GTP complex umbrella sampling simulations.

7.4 Curriculum Vitae

Der Lebenslauf ist in der Online-Version aus Gründen des Datenschutzes nicht enthalten.
– CV removed for reasons of privacy protection.

Der Lebenslauf ist in der Online-Version aus Gründen des Datenschutzes nicht enthalten.
– CV removed for reasons of privacy protection.

7.5 Zusammenfassung

Das interdisziplinäre Gebiet der theoretischen Biophysik wird zunehmend erfolgreich angewendet, um grundlegende Forschungsfragen der molekularen Biologie zu entschlüsseln. Effiziente Moleküldynamiksoftware und verbesserte Analysetechniken ermöglichen die Untersuchung struktureller, energetischer und dynamischer Eigenschaften von makromolekularen Systemen auf atomarer Ebene. Hierbei werden zunächst meist einzelne Teilsysteme untersucht, ohne die Wechselwirkung der Teilsysteme oder die Einbettung in ein übergeordnetes, gröberes System zu betrachten. Begründet ist dieser Sachverhalt darin, dass unterschiedliche dynamische Prozesse die Betrachtung unterschiedlicher Längen- und Zeitskalen notwendig machen. Die individuelle Modellierung eines Teilsystems auf der entsprechen Längen- und Zeitskala ermöglicht es dann, die für die Wechselwirkung mit anderen Teilsystemen relevanten Eigenschaften zu extrahieren und in ein übergreifendes Modell zu integrieren.

In dieser Arbeit werden neuartige Konzepte zur Analyse und Modellierung von dynamischen Teilprozessen der Protei/Ligand- beziehungsweise Protein/Protein-Interaktion vorgestellt. Die Anwendung dieser Konzepte erlaubt die Identifikation neuer biophysikalischer Phänomene sowie Systemeigenschaften, die für die Einbettung in gröbere Modelle relevant sind.

Es werden drei Teilprozesse analysiert, die für die intrazelluläre Interaktionsdynamik von Proteinen von Bedeutung sind: Erstens wird systematisch am Beispiel eines Nukleotidzuckers untersucht, inwiefern die chemische Umgebung eines Liganden dessen strukturelle, kinetische und energetische Eigenschaften beeinflusst. Hierbei wird gezeigt, dass die Bindung eines Ions an den Liganden die proteinbindende Konformation stabilisiert und somit als Bindungskofaktor agieren kann. Zweitens wird in dieser Arbeit untersucht, ob ausgeprägte Bindungspfade existieren und welche Faktoren diese Bindungspfade sowie -kinetik beeinflussen. Dies wird am Beispiel des Phosphat-bindenden Proteins untersucht. Hier wird gezeigt, dass ein von der Bindungstasche entfernter, für Anionen attraktiver Bereich an der Proteinoberfläche existiert, der wahrscheinlich eine wichtige Filterfunktion ausübt, welche die Effizienz der Phosphatbindung steigert. Der dritte Teil der Arbeit untersucht am Beispiel des Proteins Dynamin, inwiefern der Typ des gebundenen Liganden, hier GTP oder GDP, die Stärke spezifischer Protein/Protein Interaktionen beeinflusst und wie diese sich schlussendlich auf die Konformationsdynamik größerer Proteinkomplexe auswirken. Aus den gewonnen Erkenntnissen wird ein Modell für die Konformationsynamik der Dynaminhelix abgeleitet. Dieses ermöglicht es, existierende Theorien zur Funktionsweise von Dynamin zu integrieren und eine alternative Theorie vorzuschlagen. Es wird somit demonstriert, wie Erkenntnisse aus detaillierten Simulationen von Teilsystemen benutzt werden können, um Aussagen über einen übergeordneten Prozess zu machen.

Bibliography

[1] TN Tozer and M Rowland. *Introduction to Pharmacokinetics and Pharmacodynamics*. Lippincott Williams & Wilkins, 2006.

[2] J Gabrielsson and D Weiner. *Pharmacokinetic and Pharmacodynamic Data Analysis: Concepts and Applications*. Swedish Pharmaceutical Press, 4th edition, 2000.

[3] EI Ette and PJ Williams. *Pharmacometrics: The Science of Quantitative Pharmacology*. John Wiley & Sons, 2007.

[4] S Pilari and W Huisinga. Lumping of physiologically-based pharmacokinetic models and a mechanistic derivation of classical compartmental models. *J. Pharmacokinet. Pharmacodyn.*, 37:1–41, 2010.

[5] E Klipp, R Herwig, A Kowald, C Wierling, and H Lehrach. *Systems Biology in Practice: Concepts, Implementation and Application*. Vch Verlagsgesellschaft Mbh, 2005.

[6] EC Butcher, EL Berg, and EJ Kunkel. Systems biology in drug discovery. *Nature Biotechnology*, 22(10):1253–1259, 2004.

[7] D Frenkel and B Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Academic Press, San Diego, 2002.

[8] AR Leach. *Molecular Modelling: Principles and Applications*. Pearson Education Ltd., Harlow, England, 2001.

[9] T Schlick. *Molecular Modeling and Simulation: An Interdisciplinary Guide*. Springer, Berlin/Heidelberg, 2002.

[10] J Manners. *Quantum physics: An introduction*. Institute of Physics Publishing, 2000.

[11] I Buch, T Giorgino, and G De Fabritiis. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA*, 108(25):10184–10189, 2011.

[12] DE Shaw, P Maragakis, K Lindorff-Larsen, S Piana, RO Dror, MP Eastwood, JA Bank, JM Jumper, JK Salmon, Y Shan, and W Wriggers. Atomic-level characterization of the structural dynamics of proteins. *Science*, 330(6002):341–346, 2010.

[13] I Kondov, A Verma, and W Wenzel. Folding path and funnel scenarios for two small disulfide-bridged proteins. *Biochemistry*, 48(34):8195–205, 2009.

[14] A Verma and W Wenzel. Protein structure prediction by all-atom free-energy refinement. *BMC Struct. Biol.*, 7(1):12, 2007.

[15] JD Chodera, KA Dill, N Singhal, VS Pande, WC Swope, and JW Pitera. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J. Chem. Phys.*, 126(15):155101, 2007.

[16] GR Bowman, KA Beauchamp, George Boxer, and VS Pande. Progress and challenges in the automated construction of Markov state models for full protein systems. *J. Chem. Phys.*, 131(12):124101, 2009.

[17] N Buchete and G Hummer. Structure and dynamics of parallel beta-sheets, hydrophobic core, and loops in Alzheimer's A beta fibrils. *Biophys. J.*, 92(9):3032–9, 2007.

[18] N Buchete and G Hummer. Coarse master equations for peptide folding dynamics. *J. Phys. Chem. B*, 112:6057–6069, 2008.

[19] JE Hinshaw and SL Schmid. Dynamin self-assembles into rings suggesting a mechanism for coated vesicle budding. *Nature*, 374(6518):190–2, 1995.

[20] MH Stowell, B Marks, P Wigge, and HT McMahon. Nucleotide-dependent conformational changes in dynamin: evidence for a mechanochemical molecular spring. *Nat. Cell. Biol.*, 1(1):27–32, 1999.

[21] A Roux, K Uyhazi, A Frost, and P De Camilli. GTP-dependent twisting of dynamin implicates constriction and tension in membrane fission. *Nature*, 441(7092):528–31, 2006.

[22] K Faelber, Y Posor, S Gao, M Held, Y Roske, D Schulze, V Haucke, F Noé, and O Daumke. Crystal structure of nucleotide-free dynamin. *Nature*, 477(7366):556–60, 2011.

[23] M Lill and V Helms. Molecular dynamics simulation of proton transport with quantum mechanically derived proton hopping rates (Q-HOP MD). *J. Chem. Phys.*, 115(17):7993–8005, 2001.

[24] S Patel and C L Brooks. CHARMM fluctuating charge force field for proteins: I parameterization and application to bulk organic liquid simulations. *J. Comput. Chem.*, 25(1):1–16, 2003.

[25] N Allinger. MM2 - A hydrocarbon force field utilizing V1 and V2 torsional terms. *J. Am. Chem. Soc.*, 99(25):8127–8134, 1977.

[26] DA Pearlman, DA Case, JW Caldwell, WR Ross, TE Cheatham, S Debolt, D Ferguson, G Seibel, and P Kollman. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structures and energies of molecules. *Comp. Phys. Commun.*, 91(1-3):1–41, 1995.

[27] NA Baker, D Sept, S Joseph, MJ Holst, and JA McCammon. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. USA*, 98(18):10037–10041, 2001.

[28] R Hockney. The potential calculation and some applications. *Methods Comp. Phys.*, (9):136–211, 1970.

[29] D Van Der Spoel, E Lindahl, B Hess, G Groenhof, AE Mark, and HJC Berendsen. GROMACS: fast, flexible, and free. *J. Comput. Chem.*, 26(16):1701–18, 2005.

[30] L Woodcock. Isothermal molecular dynamics calculations for liquid salts. *Chem. Phys. Lett.*, 10(10):257–261, 1971.

[31] HJC Berendsen, JPM Postma, WF van Gunsteren, A Dinola, and JR Haak. Molecular-dynamics with coupling to an external bath. *J. Chem. Phys.*, 81(8):3684–3690, 1984.

[32] HC Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.*, 72(4):2384–2393, 1980.

[33] G Bussi, D Donadio, and M Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126(1):014101, 2007.

[34] P Hünenberger. Thermostat algorithms for molecular dynamics simulations. *Adv. Polym. Sci.*, 173(130):105–149, 2005.

[35] T Darden, D York, and L Pedersen. Particle mesh Ewald: An N-log(N) method for Ewald sums in large systems. *J. Chem. Phys.*, 98(12):10089–10092, 1993.

[36] P Ewald. Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Annalen der Physik*, 369(3):253–287, 1921.

[37] H Nyquist. Thermal agitation of electric charge in conductors. *Phys. Rev.*, 32:110–113, 1928.

[38] D Ermak and JA McCammon. Brownian dynamics with hydrodynamic interactions. *J. Chem. Phys.*, 69(4):1352–1360, 1978.

[39] T Geyer and U Winter. An O(N2) approximation for hydrodynamic interactions in Brownian dynamics simulations. *J. Chem. Phys.*, 130(11):114905, 2009.

[40] GM Torrie and JP Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comp. Phys.*, 23(2):187–199, 1977.

[41] JG Kirkwood. Statistical mechanics of fluid mixtures. *J. Chem. Phys.*, 3:300–313, 1935.

[42] S Kumar, D Bouzida, R Swendsen, P Kollman, and J Rosenberg. The weighted histogram analysis method for free-energy calculations on biomolecues. I. The method. *J. Comput. Chem.*, 13(8):1011–1021, 1992.

[43] SM Larson, C Snow, and VS Pande. Folding@Home and Genome@Home: Using distributed computing to tackle previously intractable problems in computational biology. *Comp. Genom.*, 2003.

[44] JH Prinz, M Held, JC Smith, and F Noé. Efficient computation of committor probabilities and transition state ensembles. *submitted to SIAM Multiscale Model. Simul.*, 2010.

[45] M Sarich, F Noé, and C Schütte. On the approximation error of Markov state models. *SIAM Multiscale Model. Simul.*, 8(4):1154–1177, 2010.

[46] F Noé, C Schütte, E Vanden-Eijnden, L Reich, and T Weikl. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc. Natl. Acad. Sci. USA*, 106(45):19011, 2009.

[47] J Prinz, H Wu, M Sarich, B Keller, M Senne, M Held, JD Chodera, C Schütte, and F Noé. Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.*, 134(17):174105, 2011.

[48] H Risken. *The Fokker-Planck Equation: Methods of Solutions and Applications.* Springer, 2nd ed. 1989. 3rd printing edition, 1996.

[49] C Schütte, A Fischer, W Huisinga, and P Deuflhard. A direct approach to conformational dynamics based on hybrid Monte Carlo. *J. Comput. Phys.*, 151(1):146–168, 1999.

[50] F Aurenhammer. Voronoi diagrams: A survey of a fundamental geometric data structure. *ACM Comp. Surv.*, 23(3):345–405, 1991.

[51] E Vanden-Eijnden. Transition path theory. pages 453–493. 2006.

[52] P Metzner, Schütte C, and E Vanden-Eijnden. Transition path theory for Markov jump processes. *SIAM Multiscale Model. Simul.*, 7(3):1192–1219, 2009.

[53] P Metzner, C Schütte, and E Vanden-Eijnden. Illustration of transition path theory on a collection of simple examples. *J. Chem. Phys.*, 125(8), 2006.

[54] CN Pace, S Treviño, E Prabhakaran, and JM Scholtz. Protein structure, stability and solubility in water and other solvents. *Philos. T. Roy. Soc. B*, 359(1448):1225, 2004.

[55] H Neuweiler, M Löllmann, S Doose, and M Sauer. Dynamics of unfolded polypeptide chains in crowded environment studied by fluorescence correlation spectroscopy. *J. Mol. Biol.*, 365(3):856–869, 2007.

[56] AY Kobitski, A Nierth, M Helm, A Jäschke, and GU Nienhaus. Mg2+-dependent folding of a Diels-Alderase ribozyme probed by single-molecule FRET analysis. *Nucleic Acids Res.*, 35(6):2047–59, 2007.

[57] X Qu, GJ Smith, KT Lee, TR Sosnick, T Pan, and NF Scherer. Single-molecule nonequilibrium periodic Mg2+-concentration jump experiments reveal details of the early folding pathways of a large RNA. *Proc. Natl. Acad. Sci. USA*, 105(18):6602–6607, 2008.

[58] T Berezniak, M Zahran, P Imhof, A Jaeschke, and JC Smith. Magnesium-dependent active-site conformational selection in the Diels-Alderase ribozyme. *J. Am. Chem. Soc.*, 132(36):12587–12596, 2010.

[59] W Doster, S Cusack, and W Petry. Dynamical transition of myoglobin revealed by inelastic neutron scattering. *Nature*, 337(6209):754–756, 1989.

[60] LJ Lapidus, WA Eaton, and J Hofrichter. Measuring the rate of intramolecular contact formation in polypeptides. *Proc. Natl. Acad. Sci. USA*, 97(13):7220–7225, 2000.

[61] J Woenckhaus. Pressure-Jump Small-Angle X-Ray Scattering Detected Kinetics of Staphylococcal Nuclease Folding. *Biophys. J.*, 80(3):1518–1523, 2001.

[62] M Jäger, H Nguyen, JC Crane, JW Kelly, and M Gruebele. The folding mechanism of a beta-sheet: the WW domain. *J. Mol. Biol.*, 311(2):373–393, 2001.

[63] HD Kim, U Nienhaus, T Ha, JW Orr, JR Williamson, and S Chu. Mg2+-dependent conformational change of RNA studied by fluorescence correlation and FRET on immobilized single molecules. *Procl. Natl. Acad. Sci. USA*, 99(7):4284–4289, 2002.

[64] H Neuweiler, S Doose, and M Sauer. A microscopic view of miniprotein folding: Enhanced folding efficiency through formation of an intermediate. *Proc. Natl. Acad. Sci. USA*, 102(46):16650–16655, 2005.

[65] F Noé, S Doose, I Daidone, M Löllmann, JD Chodera, M Sauer, and JC Smith. Dynamical fingerprints: Understanding biomolecular processes in microscopic detail by combination of spectroscopy, simulation and theory. *Proc. Natl. Acad. Sci. USA*, 108:4822–4827, 2011.

[66] M Rief, M Gautel, F Oesterhelt, JM Fernandez, and HE Gaub. Reversible unfolding of individual titin immunoglobulin domains by AFM. *Science*, 276(5315):1109–1112, 1997.

[67] B Schuler, EA Lipman, and WA Eaton. Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature*, 419(6908):743–747, 2002.

[68] WJ Greenleaf, MT Woodside, and SM Block. High-resolution, single-molecule measurements of biomolecular motion. *Annu. Rev. Biophys. Biomol. Struct.*, 36(1):171–190, 2007.

[69] JD Chodera, P Elms, F Noé, B Keller, CM Kaiser, A Ewall-Wice, S Marqusee, C Bustamante, and N Singhal-Hinrichs. Bayesian hidden Markov model analysis of single-molecule force spectroscopy: Characterizing kinetics under measurement uncertainty. *http://arxiv.org/abs/1108.1430*, 2011.

[70] D W Miller and K A Dill. Ligand binding to proteins: The binding landscape model. *Protein Sci.*, 6(10):2166–79, 1997.

[71] S Kumar, B Ma, CJ Tsai, N Sinha, and R Nussinov. Folding and binding cascades: Dynamic landscapes and population shifts. *Protein Sci.*, 9(1):10–9, 2000.

[72] OF Lange, N Lakomek, C Fares, GF Schroder, KFA Walter, S Becker, J Meiler, H Grubmuller, C Griesinger, and BL De Groot. Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science*, 320(5882):1471–1475, 2008.

[73] TR Weikl and C von Deuster. Selected-fit versus induced-fit protein binding: Kinetic differences and mutational analysis. *Proteins*, 75(1):104–10, 2009.

[74] CD Snow, EJ Sorin, YM Rhee, and VS Pande. How well can simulation predict protein folding kinetics and thermodynamics? *Annu. Rev. Biophys. Biomol. Struct.*, 34(1):43–69, 2005.

[75] DD Schaeffer, A Fersht, and V Daggett. Combining experiment and simulation in protein folding: Closing the gap for small model systems. *Curr. Opin. Struct. Biol.*, 18(1):4–9, 2008.

[76] W van Gunsteren, J Dolenc, and A Mark. Molecular simulation as an aid to experimentalists. *Curr. Opin. Struct. Biol.*, 18(2):149–153, 2008.

[77] SV Krivov and M Karplus. Hidden complexity of free energy surfaces for peptide (protein) folding. *Proc. Nat. Acad. Sci. USA*, 101(41):14766–14770, 2004.

[78] WC Swope, JW Pitera, F Suits, M Pitman, and M Eleftheriou. Describing protein folding kinetics by molecular dynamics simulations: 2. Example applications to alanine dipeptide and beta-hairpin peptide. *J. Phys. Chem. B.*, 108(21):6582–6594, 2004.

[79] F Rao and A Caflisch. The protein folding network. *J. Mol. Bio.*, 342(1):299–306, 2004.

[80] N Singhal and VS Pande. Error analysis and efficient sampling in Markovian state models for molecular dynamics. *J. Chem. Phys.*, 123:204909, 2005.

[81] F Noé, I Horenko, C Schütte, and JC Smith. Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states. *J. Chem. Phys.*, 126(15):155102, 2007.

[82] F Noé and S Fischer. Transition networks for modeling the kinetics of conformational transitions in macromolecules. *Curr. Opin. Struc. Biol.*, 18(2):154–162, 2008.

[83] F Noé. Probability distributions of molecular observables computed from Markov models. *J. Chem. Phys.*, 128(24):244103, 2008.

[84] VA Voelz, GR Bowman, KA Beauchamp, and VS Pande. Molecular simulation of ab initio protein folding for a millisecond folder NTL9. *J. Am. Chem. Soc.*, 132(5):1526–1528, 2010.

[85] J Prinz, H Wu, M Sarich, B Keller, M Senne, M Held, JD Chodera, C Schütte, and F Noé. Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.*, 134(17):174105, 2011.

[86] KA Beauchamp, GR Bowman, TJ Lane, L Maibaum, IS Haque, and VS Pande. MSMBuilder2: Modeling conformational dynamics at the picosecond to millisecond scale. *J. Chem. Theor. Comp.*, 7(10):3412–3419, 2011.

[87] M Senne, C Schütte, and F Noé. EMMA - a software package for Markov model building and analysis. *submitted to J. Chem. Theo. Comp.*, 2011.

[88] P Deuflhard and M Weber. Robust Perron cluster analysis in conformation dynamics. *ZIB Report*, 03-09, 2003.

[89] F Noe, I Daidone, J C Smith, A di Nola, and A Amadei. Solvent electrostriction-driven peptide folding revealed by quasi-Gaussian entropy theory and molecular dynamics simulation. *J. Phys. Chem. B*, 112(35):11155–11163, 2008.

[90] F Noé, C Schütte, E Vanden-Eijnden, L Reich, and TR Weikl. Constructing the full ensemble of folding pathways from short off-equilibrium simulations. *Proc. Natl. Acad. Sci. USA*, 106(45):19011–19016, 2009.

[91] M Held, P Metzner, J Prinz, and F Noé. Mechanisms of protein-ligand association and its modulation by protein mutations. *Biophys. J.*, 100(3):701–710, 2011.

[92] GR Bowman, VA Voelz, and VS Pande. Atomistic folding simulations of the five-helix bundle protein Lambda 6-85. *J. Am. Chem. Soc.*, 133(4):664–667, 2011.

[93] JD Chodera and F Noé. Probability distributions of molecular observables computed from Markov models. II: Uncertainties in observables and their time-evolution. *J. Chem. Phys.*, 133(24):105102, 2010.

[94] F Noé, S Doose, I Daidone, M Löllmann, M Sauer, JD Chodera, and JC Smith. Dynamical fingerprints for probing individual relaxation processes in biomolecular dynamics with simulations and kinetic experiments. *Proc. Natl. Acad. Sci. USA*, 108(12):4822–7, 2011.

[95] D Sezer, JH Freed, and B Roux. Simulating electron spin resonance spectra of nitroxide spin labels from molecular dynamics and stochastic trajectories. *J. Chem. Phys.*, 128(16):165106, 2008.

[96] W Zhuang, RZ Cui, DA Silva, and Xi Huang. Simulating the T-jump-triggered unfolding dynamics of trpzip2 peptide and its time-resolved IR and two-dimensional IR signals using the Markov state model approach. *J. Phys. Chem. B*, 115(18):5415–5424, 2011.

[97] VA Voelz, M Jäger, L Zhu, S Yao, O Bakajin, S Weiss, LJ Lapidus, and VS Pande. Markov state models of millisecond folder ACBP reveals new views of the folding reaction. *Biophys. J.l*, 100(3):515a, 2011.

[98] R Schauer, S Kelm, G Reuter, P Roggentin, and L Shaw. *Biochemistry and role of sialic acids*. New York: Plenum Press, 1995.

[99] KN Kirschner and RJ Woods. Solvent interactions determine carbohydrate conformation. *Proc. Natl. Acad. Sci. USA*, 98(19):10541–5, 2001.

[100] M Basma, S Sundara, D Çalgan, T Vernali, and RJ Woods. Solvated ensemble averaging in the calculation of partial atomic charges. *J. Comput. Chem.*, 22(11):1125, 2001.

[101] J Wang, R Wolf, J Caldwell, P Kollman, and D Case. Development and testing of a general amber force field. *J. Comput. Chem.*, 25(9):1157–1174, 2004.

[102] J Wang, P Cieplak, and P Kollman. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.*, 21(12):1049–1074, 2000.

[103] B Hess, H Bekker, and H Berendsen. LINCS: a linear constraint solver for molecular simulations. *J. Comput. Chem.*, 18(12):1463–1472, 1997.

[104] BM Sattelle and A Almond. Is N-acetyl-D-glucosamine a rigid 4C1 chair? *Glycobiology*, 21(12):1651–62, 2011.

[105] KA Dill. Dominant forces in protein folding. *Biochemistry*, 29(31):7133–55, 1990.

[106] JH Prinz, B Keller, and F Noé. Probing molecular kinetics with Markov models: Metastable states, transition pathways and spectroscopic observables. *Phys. Chem. Chem. Phys.*, 2011 (in revision).

[107] B Keller, JH Prinz, and F Noé. Markov models and dynamical fingerprints: Unraveling the complexity of molecular kinetics. *Chem. Phys. (in press)*, 2011.

[108] S Weiss. Fluorescence spectroscopy of single biomolecules. *Science*, 283(5408):1676–1683, 1999.

[109] K Gerwert, G Souvignier, and B Hess. Simultaneous monitoring of light-induced changes in protein side-group protonation, chromophore isomerization, and backbone motion of bacteriorhodopsin by time-resolved fourier-transform infrared spectroscopy. *P. Natl. Acad. Sci. USA*, 87(24):9774–9778, 1990.

[110] DD Boehr, R Nussinov, and PE Wright. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.*, 5(11):789–796, 2009.

[111] RE Campbell, SC Mosimann, ME Tanner, and NC Strynadka. The structure of UDP-N-acetylglucosamine 2-epimerase reveals homology to phosphoglycosyl transferases. *Biochemistry*, 39(49):14993–5001, 2000.

[112] S Hinderlich, A Sonnenschein, and W Reutter. Metal ion requirement of bifunctional UDP-N-acetylglucosamine 2-epimerase/N-acetylmannosamine kinase from rat liver. *BioMetals*, 11(3):253–258, 1998.

[113] R Ballew, J Sabelko, C Reiner, and M Gruebele. A single-sweep, nanosecond time resolution laser temperature-jump apparatus. *Rev. Sci. Instrum.*, 67(10):3694–3699, 1996.

[114] JA Ihalainen, B Paoli, S Muff, EHG Backus, J Bredenbeck, GA Woolley, A Caflisch, and P Hamm. Alpha-helix folding in the presence of structural constraints. *Proc. Natl. Acad. Sci. USA*, 105(28):9588–93, 2008.

[115] VK Misra and DE Draper. The linkage between magnesium binding and RNA folding. *J. Mol. Biol.*, 317(4):507–21, 2002.

[116] A Ostermann, R Waschipky, FG Parak, and UG Nienhaus. Ligand binding and conformational motions in myoglobin. *Nature*, 404(6774):205–208, 2000.

[117] H Frauenfelder, SG Sligar, and PG Wolynes. The energy landscapes and motions of proteins. *Science*, 254(5038):1598–1603, 1991.

[118] PJ Tummino and RA Copeland. Residence time of receptor-ligand complexes and its effect on biological function. *Biochemistry*, 47(20):5481–5492, 2008.

[119] P Csermely, R Palotai, and R Nussinov. Induced fit, conformational selection and independent dynamic segments: An extended view of binding events. *Trends in Biochem. Sci.*, 35(10):539–546, 2010.

[120] N O'Toole and IA Vakser. Large-scale characteristics of the energy landscape in protein-protein interactions. *Proteins*, 71(1):144–152, 2008.

[121] J Schluttig, D Alamanova, V Helms, and US Schwarz. Dynamics of protein-protein encounter: A Langevin equation approach with reaction patches. *J. Chem. Phys.*, 129(15):155106, 2008.

[122] S Northrup, S Allison, and JA McCammon. Brownian dynamics simulation of diffusion-influenced bimolecular reactions. *J. Chem. Phys.*, 80(4):1517, 1984.

[123] SH Northrup and HP Erickson. Kinetics of protein-protein association explained by Brownian dynamics computer simulation. *Proc. Natl. Acad. Sci. USA*, 89(8):3338–42, 1992.

[124] RR Gabdoulline and RC Wade. Simulation of the diffusional association of barnase and barstar. *Biophys. J.*, 72(5):1917–29, 1997.

[125] RR Gabdoulline and RC Wade. Protein-protein association: Investigation of factors influencing association rates by Brownian dynamics simulations. *J. Mol. Biol.*, 306(5):1139–55, 2001.

[126] A Spaar, C Dammer, RR Gabdoulline, RC Wade, and V Helms. Diffusional encounter of barnase and barstar. *Biophys. J.*, 90(6):1913–1924, 2006.

[127] S McGuffee and A Elcock. Diffusion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm. *PLoS Comput. Biol.*, 6(3):e1000694, 2010.

[128] Y Song, Y Zhang, T Shen, CL Bajaj, JA McCammon, and NA Baker. Finite element solution of the steady-state Smoluchowski equation for rate constant calculations. *Biophys. J.*, 86(4):2017–2029, 2004.

[129] S Kube and M Weber. A coarse graining method for the identification of transition rates between molecular conformations. *J. Chem. Phys.*, 126(2):024103, 2007.

[130] C Schütte, F Noé, E Meerbach, P Metzner, and C Hartmann. Conformation dynamics. In R Jeltsch and G Wanner, editors, *Proceedings of the International Congress on Industrial and Applied Mathematics (ICIAM)*, pages 297–336. EMS publishing house, 2009.

[131] N Yao, PS Ledvina, A Choudhary, and FA Quiocho. Modulation of a salt link does not affect binding of phosphate to its specific active transport receptor. *Biochemistry*, 35(7):2079–2085, 1996.

[132] Z Wang, A Choudhary, PS Ledvina, and FA Quiocho. Fine tuning the specificity of the periplasmic phosphate transport receptor. Site-directed mutagenesis, ligand binding, and crystallographic studies. *J. Biol. Chem.*, 269(40):25091–4, 1994.

[133] H Luecke and FA Quiocho. High specificity of a phosphate transport protein determined by hydrogen bonds. *Nature*, 347(6291):402–406, 1990.

[134] M Brune, JL Hunter, SA Howell, SR Martin, TL Hazlett, JET Corrie, and MR Webb. Mechanism of inorganic phosphate interaction with phosphate binding protein from Escherichia coli. *Biochemistry*, 37(29):10370–10380, 1998.

[135] PS Ledvina, AL Tsai, Z Wang, E Koehl, and FA Quiocho. Dominant role of local dipolar interactions in phosphate binding to a receptor cleft with an electronegative charge surface: Equilibrium, kinetic, and crystallographic studies. *Protein Sci.*, 7(12):2550–2559, 1998.

[136] H Huang and JM Briggs. The association between a negatively charged ligand and the electronegative binding pocket of its receptor. *Biopolymers*, 63(4):247–60, 2002.

[137] RR Gabdoulline and RC Wade. On the protein-protein diffusional encounter complex. *J. Mol. Recognit.*, 12(4):226–34, 1999.

[138] F Fogolari, A Brigo, and H Molinari. The Poisson-Boltzmann equation for biomolecular electrostatics: a tool for structural biology. *J. Mol. Recognit.*, 15(6):377–392, 2002.

[139] W E and E vanden Eijnden. Towards a theory of transition paths. *J. Stat. Phys.*, 123(3):503–523, 2006.

[140] Matrix toolkits java - http://code.google.com/p/matrix-toolkits-java/.

[141] R Erban and SJ Chapman. Stochastic modelling of reaction-diffusion processes: Algorithms for bimolecular reactions. *Phys. Biol.*, 6(4):046001, 2009.

[142] HM Berman, J Westbrook, Z Feng, G Gilliland, TN Bhat, H Weissig, IN Shindyalov, and PE Bourne. The protein data bank. *Nucleic Acids Res.*, 28(1):235–242, 2000.

[143] AD MacKerell, D Bashford, Bellott, RL Dunbrack, JD Evanseck, MJ Field, S Fischer, J Gao, H Guo, S Ha, D Joseph-McCarthy, L Kuchnir, K Kuczera, FTK Lau, C Mattos, S Michnick, T Ngo, DT Nguyen, B Prodhom, WE Reiher, B Roux, M Schlenkrich, JC Smith, R Stote, J Straub, M Watanabe, J Wiorkiewicz-Kuczera, D Yin, and M Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins†. *J. Phys. Chem. B*, 102(18):3586–3616, 1998.

[144] H Li, AD Robertson, and JH Jensen. Very fast empirical prediction and rationalization of protein pKa values. *Proteins*, 61(4):704–721, 2005.

[145] TJ Dolinsky, JE Nielsen, JA McCammon, and NA Baker. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.*, 32(Web Server issue):W665–7, 2004.

[146] HS Kielman and JC Leyte. Selfdiffusion of phosphate and polyphosphate anions in aqueous solution. In *Proceedings of Congress AMPERE*, volume 2, pages 515–516, 1975.

[147] J Latorre, P Metzner, C Hartmann, and C Schütte. A structure-preserving numerical discretization of reversible diffusions. *submitted to Comm. Math. Sci.*, 2010.

[148] R Du, VS Pande, Alexander Y, T Tanaka, and ES Shakhnovich. On the transition coordinate for protein folding. *J. Chem. Phys.*, 108(1):334–350, 1998.

[149] P Ramachandran and G Varoquaux. Mayavi: Making 3D data visualization reusable. In *Proceedings of the 7th Python in Science Conference*, pages 51–56, 2008.

[150] D Thirumalai and SA Woodson. Kinetics of folding of proteins and RNA. *Acc. Chem. Res.*, 29(9):433–439, 1996.

[151] KA Dill. Polymer principles and protein folding. *Protein Sci.*, 8(6):1166–1180, 1999.

[152] JN Onuchic and PG Wolynes. Theory of protein folding. *Curr. Opin. Struc. Biol.*, 14(1):70–75, 2004.

[153] KA Dill and HS Chan. From Levinthal to pathways to funnels. *Nat. Struct. Biol.*, 4(1):10–19, 1997.

[154] TC Südhof. The synaptic vesicle cycle: A cascade of protein-protein interactions. *Nature*, 375(6533):645–53, 1995.

[155] PI Hanson, JE Heuser, and R Jahn. Neurotransmitter release - four years of SNARE complexes. *Curr. Opin. Neurobiol.*, 7(3):310–5, 1997.

[156] TA Grigliatti, L Hall, R Rosenbluth, and DT Suzuki. Temperature-sensitive mutations in Drosophila melanogaster. XIV. A selection of immobile adults. *Mol. Gen. Genet.*, 120(2):107–14, 1973.

[157] GJK Praefcke and HT McMahon. The dynamin superfamily: Universal membrane tubulation and fission molecules? *Nat. Rev. Mol. Cell. Biol.*, 5(2):133–47, 2004.

[158] AM van der Bliek and EM Meyerowitz. Dynamin-like protein encoded by the Drosophila shibire gene associated with vesicular traffic. *Nature*, 351(6325):411–4, 1991.

[159] SM Ferguson, G Brasnjo, M Hayashi, M Wölfel, C Collesi, S Giovedi, A Raimondi, L Gong, P Ariel, S Paradise, E O'Toole, R Flavell, O Cremona, G Miesenböck, TA Ryan, and P De Camilli. A selective activity-dependent requirement for dynamin 1 in synaptic vesicle endocytosis. *Science*, 316(5824):570–4, 2007.

[160] PJ Robinson, JM Sontag, JP Liu, EM Fykse, C Slaughter, H McMahon, and TC Südhof. Dynamin gtpase regulated by protein kinase c phosphorylation in nerve terminals. *Nature*, 365(6442):163–6, 1993.

[161] C Diatloff-Zito, AJ Gordon, E Duchaud, and G Merlin. Isolation of an ubiquitously expressed cDNA encoding human dynamin II, a member of the large GTP-binding protein family. *Gene*, 163(2):301–6, 1995.

[162] J Lu, TD Helton, TA Blanpied, B Rácz, TM Newpher, RJ Weinberg, and MD Ehlers. Postsynaptic positioning of endocytic zones and AMPA receptor cycling by physical coupling of dynamin-3 to homer. *Neuron*, 55(6):874–89, 2007.

[163] B Marks, MH Stowell, Y Vallis, IG Mills, A Gibson, CR Hopkins, and HT McMahon. GTPase activity of dynamin and resulting conformation change are essential for endocytosis. *Nature*, 410(6825):231–5, 2001.

[164] A Roux, G Koster, M Lenz, B Sorre, J Manneville, P Nassoy, and P Bassereau. Membrane curvature controls dynamin polymerization. *Proc. Natl. Acad. Sci. USA*, 107(9):4141–6, 2010.

[165] JS Chappie, S Acharya, M Leonard, SL Schmid, and F Dyda. G domain dimerization controls dynamin's assembly-stimulated GTPase activity. *Nature*, 465(7297):435–40, 2010.

[166] BD Song, M Leonard, and SL Schmid. Dynamin GTPase domain mutants that differentially affect GTP binding, GTP hydrolysis, and clathrin-mediated endocytosis. *J. Biol. Chem.*, 279(39):40431–6, 2004.

[167] JA Mears, P Ray, and JE Hinshaw. A corkscrew model for dynamin constriction. *Structure*, 15(10):1190–202, 2007.

[168] 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*, 65(3):712–25, 2006.

[169] K Meagher, L Redman, and H Carlson. Development of polyphosphate parameters for use with the Amber force field. *J. Comput. Chem.*, 24(9):1016–1025, 2003.

[170] J Wang, W Wang, PA Kollman, and DA Case. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.*, 25(2):247–260, 2006.

[171] M Parrinello and A Rahman. Polymorphic transitions in single-crystals - A new molecular-dynamics method. *J. Appl. Phys.*, 52(12):7182–7190, 1981.

[172] JS Hub, BL de Groot, and D Van Der Spoel. gwham-A free weighted histogram analysis implementation including robust error and autocorrelation estimates. *J. Chem. Theory Comput.*, 6(12):3713–3720, 2010.

[173] A Fiser, RK Do, and A Sali. Modeling of loops in protein structures. *Protein Sci.*, 9(9):1753–73, 2000.

[174] W Jorgensen, J Chandrasekhar, JD Madura, R Impey, and M Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2):926, 1983.

[175] T Darden, L Perera, L Li, and L Pedersen. New tricks for modelers from the crystallography toolkit: The particle mesh Ewald algorithm and its use in nucleic acid simulations. *Structure*, 7(3):55–60, 1999.

[176] U Essmann, L Perera, M Berkowitz, T Darden, H Lee, and L Pedersen. A smooth particle mesh Ewald method. *J. Chem. Phys.*, 103(19):8577–8593, 1995.