Markov Modeling of Nonequilibrium Biomolecular Data

Georg Gabriel Diez

Supervisor

Prof. Dr. Gerhard Stock

Master's Thesis
Albert-Ludwigs-Universität Freiburg
Faculty of Mathematics and Physics
Institute of Physics
June 2020

Abstract

Allostery plays a fundamental role in regulatory biological processes. It describes the phenomenon that a functional change at one site of the protein is triggered by the binding of a ligand to another, distant site. It yet remains unclear what forces drive the underlying mechanisms, and in particular whether the mechanisms are of structural or dynamical nature. Here we study molecular dynamics simulations of the PDZ2 domain, in which the binding of a ligand causing the allosteric transition is mimicked by an azobenzene photoswitch.

The allosteric transition captured in a vast ensemble of equilibrium and non-equilibrium trajectories, covering more than 400 μs, is investigated by applying a state-of-the-art workflow. After employing a dimensionality reduction, the low dimensional space is partitioned into meaningful metastable clusters by density based clustering. Subsequent Markov state modeling allows to approximate the dynamics of the protein by memory-less jumps between those metastable conformations. We introduced *iterative dynamical coring* as a novel method for the correction of artefacts resulting from dimensionality reduction which significantly improves the validity of the Markov state model. The resulting Markov state model is self-consistent and predicts that the allosteric transition obeys an order-disorder-order pattern. Furthermore, it is suggested that allostery is neither driven exclusively dynamically nor exclusively conformationally but is rather governed by an interplay of both.

Acknowledgement

First of all, I want to thank Prof. Dr. Gerhard Stock not only for granting me the possibility to do my master thesis in his group but most of all for the excellent supervision, in which personal encouragement always comes first. Gerhard's door is always open for questions, ideas and discussions and his continuous and calm way of communicating, discussing and teaching physics has helped me learning a lot throughout this year without ever feeling lost.
At this point, I would also like to add and thank Benjamin Lickert and Daniel Nagel, who spend hours and hours with Gerhard and me racking their brains over problems related to the stubborn PDZ2 and figuring ways to crack it. Besides, many thanks to Daniel for an always excellent atmosphere in the office and for helping me with countless little and not so little problems. I highly appreciate his incredible patience even for the most annoying questions regarding Python and Markov state modelling.
Furthermore, I am grateful to all members of the AG Stock for their eagerness to help. I would like to mention three names in particular, namely Matthias Post for providing explanations and coordinate predictions, Adnan Gulzar for helping me a great deal with GROMACS and Anna Weber for helping me out with my early difficulties with the topic. Last but not least, I am grateful to Frieder Lindel, Daniel Nagel and Benjamin Lickert for proofreading this thesis.

Contents

Acronyms

ACF	autocorrelation function
CVs	collective variables
dPCA+	dihedral angle principal component analysis
MCMC	Markov Chain Monte Carlo
MD	molecular dynamics
PCA	principal component analysis
PDZ	PSD95/Disc large/ZO-1

I. Introduction

> *...if we were to name the most powerful assumption of all, which leads one on and on in an attempt to understand life, it is that all things are made of atoms, and that everything that living things do can be understood in terms of the jigglings and wigglings of atoms.*
>
> Richard P. Feynman

Derived from the Greek word πρωτειος "proteios" which translates as "standing in front", proteins are biology's workhorse molecules. It was Gerrit Mulder, a Dutch physician, who discovered proteins and foretold their crucial importance in the early 19th century. Even though there were several important steps and advances in protein research, such as e.g. the accidental discovery of haemoglobin by Friedrich Ludwig Hünefeld in 1840, the detailed structure of a protein was revealed for the first time in the year 1958 when J. Kendrew studied myoglobin by X-ray crystallography [1]. This groundbreaking experimental work was awarded with the Nobel Prize in Chemistry 1962.
Over the next years, more and more proteins with different structures were discovered which formed the dogma that the function of a protein is strictly governed by its structure which is encoded in the proteins amino acid sequence and ultimately shaped in the folding process:

Sequence → Structure → Function

It was almost 200 years after the first mention of a protein when Platt et al. showed that even the smallest modifications in the structure of a protein may lead to an entirely different function [2]. This was a first hint that there must be something crucial in the proteins nature which was overlooked before. Soon the protein paradigm was extended by dynamics, as this was identified to be the missing link:

Sequence → Structure → Dynamics → Function

Since then, experimental techniques have become more and more elaborated so that they now can not only provide an exceptionally informative picture of the structure of proteins (nuclear magnetic resonance spectroscopy, X-ray crystallography or circular dichroism spectroscopy) but also picturing dynamical information (ultrafast infrared spectroscopy or time depended X-ray via free electron lasers) [3, 4].

On the theoretical side, Molecular Dynamics (MD) simulations have emerged as another effective approach to tackle protein dynamics due to rapid advances in computer technology. MD simulations are a very powerful tool in providing the atomistic picture of dynamics with desirable small temporal and structural resolution by solving Newtons equations of motion for all atoms of a protein.

This advanced techniques on both experimental and theoretical side have led to an ever deeper understanding of protein dynamics, especially in the field of protein folding [3, 5, 6]. In order to execute their vital cell processes, proteins need to fold in their unique structure. Understanding this phenomenon is, besides academic interest for knowledge, of utmost interest for medicine since (mis)folding is prominently involved in neurodegenerative diseases such as Alzheimer's disease or Huntington's chorea [7]. Besides, misfolding of proteins also inhibits their function which e.g. prevents them from fulfilling their duty of repairing damaged snippets of DNA potentially leading to abnormal cell growth (cancer) [8]. At the same time, hopes are high that understanding protein dynamics can accelerate and catalyze state of the art drug research. At the time of writing this thesis, scientists are simulating potentially drug-treatable protein targets, so called spike proteins or *Demogorgons*, of the SARS-CoV-2 virus (responsible for the COVID-19 pandemic). They are using a network of private and scientific computers with a total computational power of currently more than 2.4 exaFLOPs ($2.4 \cdot 10^{18}$ floating point operations per second) and still counting, which made it the worlds first exaFLOPS computing system [9]. In terms of computation power, it easily dwarfs the worlds top 10 best supercomputers combined.

Allostery

Over several decades scientists, in experimental and theoretical work, have developed a well-established picture of protein folding and unfolding. However, an understanding on such a detailed level has not yet been achieved for allostery. Allostery (ἄλλος, στερεὸς ≙ other site) describes a universal phenomenon whereby a ligand binds to the allosteric site of a protein. Thereby, a conformational change at a distant active side of the protein through alteration of conformation and/or dynamics is triggered [10] (see Fig. 1.1).

On the very basic level, allostery is related to the protein itself, but its fundamental importance arises on the cellular level by affecting different phenomena such as signal transport, transcriptional regulation

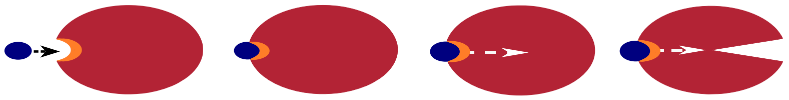

Figure 1.1.: Simple scheme of allostery: A ligand (blue) docks to the allosteric site (orange) of the protein (red) which triggers a conformational change at a distant side in the protein.

or metabolism [11]. It takes place in all dynamic proteins and was identified as being responsible for a number of diseases arising due to changes in the allosteric binding site or by creating sites for allosteric posttranslational modifications [12]. This makes it a promising candidate for paving the way for innovative drug research and design as well as a deep understanding of diseases related to allostery [12]. In contrast to protein folding, the relatively small structural and/or dynamical changes makes it very difficult to identify the underlying principles of allostery. Originally, the interpretation of allostery as a hierarchy of several structural changes prevailed. However, more recent studies point to the fact that an allosteric transition can be induced with very little changes in the structural conformation shifting the focus mainly onto the dynamics of the protein [13, 14]. Despite several decades of ongoing debate the nature of allostery remains an open question [15] which also shows the need for a deeper understanding of the underlying mechanisms on a molecular level.

PDZ2

Concerning the research on allostery, PDZ (PSD95/Disc large/ZO) domains in particular have proven to be suitable because they are relatively small compared to other allosteric proteins and are widespread in a large variety of proteins. Binding to the C-terminus of their targets, they are involved in a wide range of signal transduction pathways in the human body [10].

Combining state of the art experimental and theoretical techniques (NMR and IR spectroscopy, as well as MD simulations), Buchli et al. studied an allosteric transition in a photoswitchable PDZ2 domain. Here, an azobenzene photoswitch was linked covalently across the allosteric site, which mimics the binding of a ligand [16]. Applying a laser pulse with a certain wavelength, the conformation of the PDZ2 protein was switched from its Cis to Trans conformation and the allosteric transition was this way enforced.

S. Buchenberg et al. provided additional MD simulations of the same system, which showed good agreement with the experimental findings [17]. Those simulations were later extended to a total length of 408 μs by A. Gulzar. It is this data set, which we investigate within this thesis.

Markov State Models

Simulating such extensively long trajectories is computationally very expensive and modern computer systems—due to their processor structure—are far better suited for the computation of multiple short trajectories instead of a single large trajectory. Nevertheless, one is usually interested in the full global dynamics instead of local information in the single trajectories.

Markov State Models (MSMs) represent an effective remedy as they can combine several separate, only locally converged trajectories and thereby allow predictions of the global dynamics [18–25]. The transition matrix, representing one of the main pillars of the MSM, approximates the time evolution of the MD data as memory-less jumps between metastable states in the conformational space of the protein. This does the trick of extracting global dynamical and structural information out of an ensemble of locally converged trajectories and therefore allows us to understand global mechanisms of the protein by combining only local information. In other words, we can exploit the structure of such MSMs to overcome the bottleneck of modern computer systems and shed light on long time scale dynamics only by using many predictions on much shorter time scales.

In this thesis

The aim of this thesis is to predict the complete nonequilibrium transition mechanism of the Cis → Trans allosteric transition of PDZ2 by constructing a MSM on the data set mentioned above. We hope to gain insights about the timescales involved and to understand the most important pathways describing the transition in order to learn more about the mechanisms that govern this process. Last but not least we seek to put our findings into perspective in order to classify which of the two theories about allostery—conformational or dynamical—is better supported by our data.

After giving a short introduction into the underlying theory and methods in Ch. 2, we continue with the extraction of the proteins internal motion from the MD data by selecting suitable contact-distances in Ch. 3. For an effective description of the dynamics in terms of the protein's metastable conformations—so called microstates— it is inevitable to project the internal coordinates onto relevant, low dimensional reaction coordinates. During this process, it can happen that particularly frames on the boundaries between microstates are incorrectly assigned. As a remedy to this problem we present *iterative dynamical coring* in Ch. 4, an advancement of dynamical coring [26] and apply it to the data after we investigated and validated its effect. Also in Ch. 4, we will construct a MSM in order investigate the mechanisms involved in the Cis→Trans transition. In Ch. 5, we confirm and substantiate the results by using a machine learning decision tree to identify the most important contact-coordinates and then establishing another MSM on the basis of these findings. Improved input coordinates will allow us to analyze the nonequilibrium transition between Cis and Trans in detail. In the last chapter, Ch. 6, we conclude this thesis by discussing our findings, the problems we were faced with and point out which new developments in the future could facilitate Markov modeling.

2. Theory & Methods

A protein is a set of coordinates.

Andreas P. Heiner

In order to aid the readers understanding of this topic, we will first introduce the allosteric system under study and then move on to give a short introduction to the theory which is required for tackling the decipherment of the allosteric transition by studying molecular dynamics simulations.

Allostery is indispensable for processes occurring in living cells [27] and plays a fundamental role in all dynamical proteins [12]. Due to its complexity and its impact which is not only limited on the protein itself but affects the cellular level as well, an attempt is made to trace the origin of allostery back to a single allosteric domain.

The system investigated here, namely a modified photoswitchable PSD95/Disc large/ZO-1 (PDZ) domain [28, 29] (see Fig. 2.1), performs its general function by clustering different proteins. For

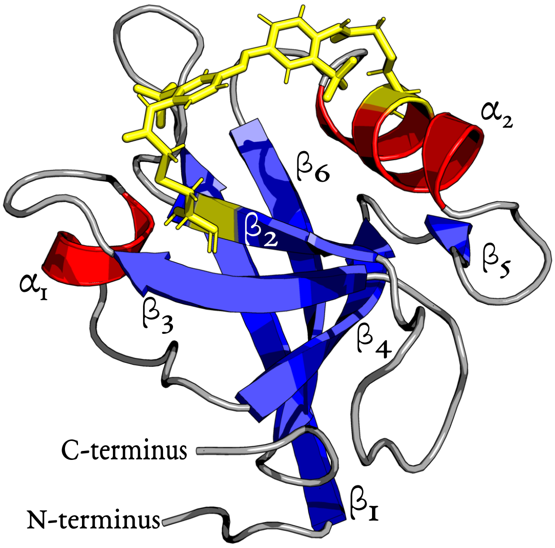

Figure 2.1.: Cartoon representation of the PDZ2-S domain in Cis-conformation. PDZ2-S has two α-helices marked in red and 6 β-sheets (marked in blue). The loops are in grey color and the azobenzene photoswitch connecting the residue 22 with residue 77 is marked in yellow.

Table 2.1.: The residues of the PDZ2-S domain forming metastable structures.

Residue	7–13	21–24	36–41	46–50	58–62	65–66	74–81	85–91
Structure	β_1	β_2	β_3	α_1	β_4	β_5	α_2	β_6

this purpose, it binds to other proteins involved in multi-domain modular enzymes with their C- and N-terminus what mediates a protein network. Doing so, PDZ domains regulate multiple signal transduction pathways [30].

Within this thesis, the PDZ2 domain occurring in the human tyrosine phosphatase 1E (hPTP1E) is investigated. It is a small 96-residue protein and, besides other things, regulates cell growth and apoptosis in breast cancer cells [31]. PDZ2 folds into a metastable structure of two α-helices and six β-sheets (see Tab. 2.1 and Fig. 2.1), with the β_2-sheet and α_2-helix forming a binding groove for the ligand. Buchli et. al mimicked the binding of a ligand by covalently linking a photoswitchable azobenzene molecule to the residues 22 and 77 since the length of the photoswitch here closely matches the C_α-distances which are occupied in the free and bound state of the protein [16]. This modification of the PDZ2 domain is referred to as PDZ2-S, but we will simply call it PDZ2 for reasons of convenience and call the PDZ2 without azobenzene photoswitch "wild-type" PDZ2. By the stimulus of a femtosecond laser pulse the conformation of the photoswitch (and thus of the protein) can be changed from its Cis conformation (free state) to the Trans conformation (bound state) and vice versa.

Interestingly, the removal of a short α_3-helix close to the C-terminus reduced the affinity of binding a ligand by a factor of 21 [29]. Therefore PDZ domains are a very promising candidate to study the role of dynamics in allostery as this α_3-helix is located well outside of the ligand binding site. This indicates that fast side chain dynamics are the main driving force for allosteric transitions within PDZ domains.

As mentioned above, we aim to trace back the origin of allostery to a single allosteric protein which requires very information about the motion of every single atom—a task well suited for molecular dynamics simulations.

2.1. Molecular Dynamics Simulation

MD simulations are a widely used computer simulation technique which allows examining the atomic spatiotemporal details of complex systems such as e.g. proteins [32–34].

Since MD data contains the time evolution of every single atom of the molecule under study in phase space, MD simulations can be consulted to clarify question which can no longer be addressed in experiments due to their restrictions in temporal and/or spatial resolution.

Given the structure of a protein (often known from crystallization experiments), a simulation box with periodic boundary conditions is set up which contains the investigated protein/molecule and a solvent (e.g. water). Within this simulation box, Newtons equations of motion for all N atoms (combined in the position vector of all atoms $\boldsymbol{R} = (\boldsymbol{r}_1, ..., \boldsymbol{r}_i)$, $i = 1, ..., N$) are solved by numerical integration thus yielding the *trajectory* of the protein

$$\mathbf{F}_i = \boldsymbol{\nabla}_i\left[V^{\text{bonded}}(\boldsymbol{R}) + V^{\text{non-bonded}}(\boldsymbol{R})\right]. \tag{2.1}$$

The interactions between all atoms of the protein/molecule of interest are combined in the two force fields V^{bonded} and $V^{\text{non-bonded}}$, which describe local interactions and far-ranging interactions respectively [35].

In the bonded potential V^{bonded}, bond stretching ($\mathit{\Delta} r_{ij}^{b}$), bond bending ($\alpha$) as well as bond rotations for the proper (ϕ) and improper (ω) dihedrals are summarized.

$$\begin{aligned} V^{\text{bonded}} &= \underbrace{\sum K^{\text{b}}\left(\mathit{\Delta} r^{\text{b}} - \mathit{\Delta} r^{\text{eq}}\right)^2}_{\text{bond stretching}} \\ &+ \underbrace{\sum K^{\alpha}[\cos(\alpha) - \cos(\alpha^{\text{eq}})]^2}_{\text{bond bending}} \\ &+ \underbrace{\sum \{K^{\phi}[1 + \cos(m\phi - \phi^{\text{eq}})]^2 + K^{\omega}(\omega - \omega^{\text{eq}})^2\}}_{\text{dihedrals}} \end{aligned} \tag{2.2}$$

The non-bonded potential $V^{\text{non-bonded}}$ consists of the Coulomb interactions of two atoms, $\mathit{\Delta} r_{ij}$ apart, with the charges δ_i and δ_j and the Lennard-Jones potential:

$$\begin{aligned} V^{\text{non-bonded}} &= \underbrace{\sum_{ij} \frac{1}{4\pi\varepsilon_0} \frac{\delta_i \delta_j}{\mathit{\Delta} r_{ij}}}_{\text{Coulomb}} \\ &+ \underbrace{\sum_{ij} \varepsilon_{ij}\left[\left(\frac{\mathit{\Delta} r_{\text{m}}}{\mathit{\Delta} r_{ij}}\right)^{12} - 2\left(\frac{\mathit{\Delta} r_{\text{m}}}{\mathit{\Delta} r_{ij}}\right)^{6}\right]}_{\text{Lennard Jones}}. \end{aligned} \tag{2.3}$$

The Ks in V^{bonded} represent force constants which are typically either determined in experiment or calculated from semi-empirical quantum mechanics calculations. Same applies also for ε_{ij}, which denotes the depth of the potential well and $\mathit{\Delta} r_{\text{m}}$ which denotes the distance at which the Lennard-Jones potential reaches its minimum. Equations (2.2) - (2.3) show one possible example [35] for the force fields, but depending on the application, different expressions and force constants for the force fields have to be considered.

MD Simulation Details for PDZ2

For the MD simulations of the allosteric transition in PDZ2, GROMACS [36] was used. The Amber ff99SB*-ILDN force field [37–40] was applied in combination with the rigid TIP3P water model [41] and the temperature was set to room temperature, i.e. $T = 300\,\mathrm{K}$. In order to integrate Newtons equations of motion, a Leapfrog integrator was used. Frames were written out with a frequency of 50 frames per nanosecond.

S. Buchenberg has laid the foundation stone with 7 equilibrium trajectories of a length of 2.5 μs for both Cis and Trans [17]. We will see later, that these 2.5 μs trajectories did mostly not yet reach their equilibrium state and feature little overlap. Due to the latter issue, A. Gulzar extended the simulations of 6 out of 7 trajectories to a length of 10 μs each.

To provide a complete picture of the allosteric transition, S. Buchenberg set up 100 1.1 μs nonequilibrium trajectories for the transition between Cis and Trans. Here, nonequilibrium trajectories means that the simulations were initiated in a nonequilibrium conformation but then propagated under equilibrium conditions. As the great majority of them did not reach their Trans destination, 20 of them were elongated to a total of 10 μs each [42].

The total data of the PDZ2-S allosteric transition includes simulations of the total length of 20400112 frames which corresponds to a total simulation time of 408 μs.

2.2. Internal Coordinates

The data obtained from the MD simulations involves Cartesian coordinates of all atoms of the protein and its solvent. One needs to discriminate the internal motion of the proteins, which one is interested in from the global motion of the protein through the solvent in the simulation box. The latter is usually not of interest.

Sticking with the Cartesian coordinates, a mixing between internal and global motion is generally inevitable [25]. To circumvent this problem, the coordinates of the protein can be transformed to a set of internal coordinates. A number of options for the internal coordinates are available and it was shown that the choice drastically influences the outcome of the dimensionality reduction and therefore the state definition of the MSM [43].

In this thesis, contact distances are used. Typically, one refers to a contact if the distance between two residues falls below 4.5 Å as this indicates the distance where a hydrogen bond is usually formed. It was recently shown that native contacts, that is, contacts which are formed in the folded state of the protein, play a major role in protein folding [44]. Of course there are more types of contacts beside hydrogen bonds, which is why the distance cutoff range extents from around 4 Å to 8 Å in the literature [43].

Here, we consider a contact as formed if the distance $|\boldsymbol{r}|$ between the closest lying atoms of each residue i and j falls below 4.5Å, i.e.:

$$\mathrm{r}_{i,j} = \min_{k,l}(|\boldsymbol{r}_{i,k} - \boldsymbol{r}_{j,l}|) \leq 4.5\,\text{Å}. \tag{2.4}$$

The indices k and l run over all atoms (hydrogen atoms included) of the selected residue pair and two examples are shown in Fig. 2.2.

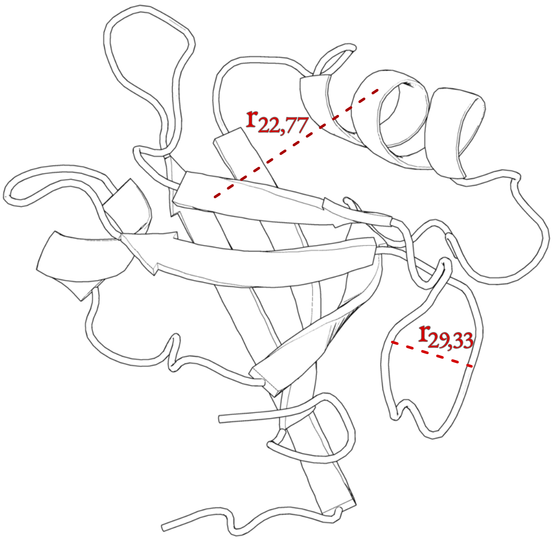

Figure 2.2.: Example of two distances between the residues 22 and 77 and the residues 29 and 33. The protein shown is PDZ2.

For a complete coverage of a protein, one would need n^2 distances, where n denotes the number of all residues present in the protein. This would make it numerically very expensive and thus not feasible for large systems like PDZ2. However, it is sufficient to include only relatively few distances as they are mostly highly correlated [43]. The distances can be either chosen intuitively by hand or being preselected by machine learning algorithms [45].

Apart from contact distances, other internal coordinates can also be considered, such as e.g.

- *C_α-distances.* In case one is not interested in sidechain dynamics, C_α-distances can be used as internal coordinates. C_α-distances describe the distance between the C_α-atoms in the backbone of the proteins and as well as contact distances they describe relative motion within the protein. Being an effective way to reconstruct the proteins backbone, sufficiently many C_α-distances are a helpful tool for tackling protein folding [43].

- *reciprocal contact- or C_α-distances.* By the usage of reciprocal distances, one can shift the focus from large scale motions towards smaller distances where the formation of contacts occurs [46].

- *dihedral angles.* In contrast to distance based coordinates, dihedral angles represent "local" coordinates of the protein instead of the "global" distances. Dihedral angles (ϕ_i, ψ_i) describe the conformation of the protein in terms of the rotation of its backbone [47, 48]. This works well for biomolecules such as proteins since bond length and angles are usually not subject to major changes.

Similar to C_α-distances, sidechain dynamics can not be investigated directly with dihedral angles. It has been found that although dihedral angles allow a high resolution of metastable states, they require more principal components to account for the same cumulative flux compared to distances [43].

2.3. Principal Component Analysis

After transforming the data obtained from the MD simulations to internal coordinates, the data is usually given in form of a $m \times n$ matrix, where m denotes the number of sampled timesteps and n is the amount of chosen input coordinates $\boldsymbol{r} = (r_1, ..., r_n)$.

For a meaningful statistical analysis of such vast data and to make it accessible for MSM, we use a clustering approach which uses a free energy ΔG estimate [49]

$$\Delta G(\boldsymbol{r}) = -k_\mathrm{B} T \ln[P(\boldsymbol{r})]. \tag{2.5}$$

k_B represents the Boltzmann constant, T denotes the temperature and $P(\boldsymbol{r})$ is the population density at the coordinate $\boldsymbol{r}$ [see Eq. (2.11)]. However, such an approach only works well for relatively well covered population densities $P(\boldsymbol{r})$. Obviously, convergence can not be expected for $\sim 10^7$ data points in $n \approx 10^2$–10^3 dimensions since the full space is only sparsely sampled (curse of dimensionality).

Fortunately, for proteins one can find that most of the dimensions used for the description of the trajectory are insignificant. Nonlinear couplings in the protein lead to cooperative effects which drastically reduce the *effective dimension* d_eff, i.e. the number of dimensions needed to provide a accurate picture of the MD trajectory in a reduced space [25].

Using different approaches for the dimensionality reduction of the data, ranging from deep neural networks [50, 51] to classical methods like principal component analysis (PCA) [52] and several modifications thereof [53–55], one typically ends up with around $\mathrm{d}_\mathrm{eff} \lesssim 10$ dimensions which cover the systems essential dynamics [25, 56, 57]. Eventually, all of them increase the reliability of the population density $P(\boldsymbol{r})$ significantly.

Within this work, we use PCA, a linear transformation which maximizes the variance and therefore yields a high-resolution structural picture of the protein.

The covariance matrix describes the correlated motion of the system under study

$$\mathrm{Cov}(r_i, r_j) = \langle (r_i - \langle r_i \rangle)(r_j - \langle r_j \rangle) \rangle = \langle r_i r_j \rangle - \langle r_i \rangle \langle r_j \rangle, \tag{2.6}$$

where $\langle ... \rangle$ denotes the average over the sampled data. Considering distances, we are often interested in relative differences rather than absolute ones. In this case we use the correlation matrix instead of the covariance matrix:

$$\mathrm{Corr}(r_i, r_j) = \frac{\mathrm{Cov}(r_i, r_j)}{\sigma_{r_i}\sigma_{r_j}}, \tag{2.7}$$

where σ_{r_i} denotes the standard deviation of r_i. Diagonalization of (2.6) [or (2.7)] yields n eigenvectors $\boldsymbol{v}^{(i)}$ and their corresponding eigenvalues λ_i, which indicate the direction and variances of the principal motion (see Figure 2.3)

$$\mathbf{C} \cdot \boldsymbol{v}^{(i)} = \lambda_i \boldsymbol{v}^{(i)}. \tag{2.8}$$

By projecting the input data $\boldsymbol{r}$ onto the eigenvectors

$$x_i = \boldsymbol{v}^{(i)} \cdot \boldsymbol{r}, \tag{2.9}$$

we obtain the principal components x_i which describe the data along the directions of maximum variance.

After sorting the principal components by descending eigenvalues, the first principal component features the largest possible variance with each subsequent component featuring the highest possible variance that is orthogonal to its predecessors $\langle x_i x_j \rangle = \delta_{ij}\langle x_i^2 \rangle$.

A high often indicates dynamics of paricular interest, which is why truncating the principal components after the first d' dimensions yields in good approximation a low dimensional description of the systems

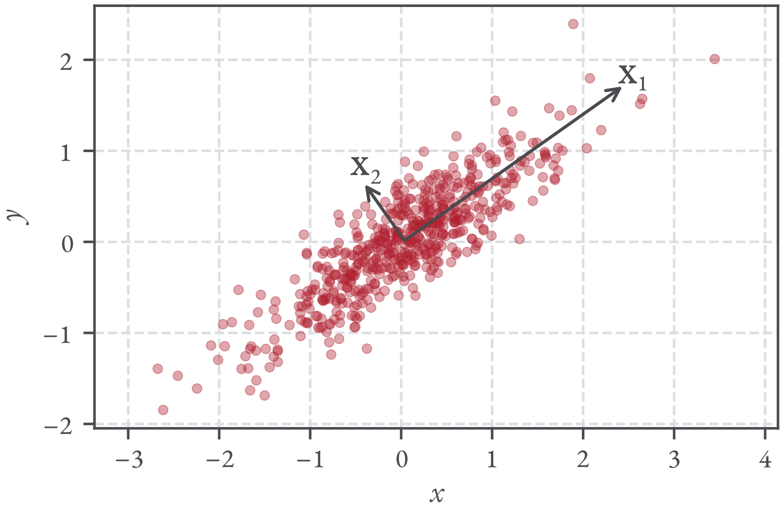

Figure 2.3.: Two dimensional example of the PCA. The arrows represent the two principal components x_1 and x_2 resulting from the diagonalization of the covariance matrix $\mathrm{Cov}(x, y)$. x_1 and x_2 indicate the directions of maximum variance in the data (under the condition that they are orthogonal to each other).

essential dynamics. Following this logic, all other dimensions can be neglected, as their motion drifts more and more towards Gaussian noise. In reference to the Brownian motion, these PCs are often called "bath". The set of chosen principal components $\boldsymbol{x} = (x_1, ..., x_{d'})$ is usually referred to as collective variables (CVs).

By summing up the corresponding eigenvalues (cumulative fluctuation) $\sum_i^{d'} \lambda_i$, we know how much percent of the total variance is covered. However, sometimes it is wiser to choose the d' components not only by decreasing eigenvalues but by some criteria of interest such as e.g. a non-harmonic free energy profile or well distinguishable dynamics in this PC.

2.4. Robust Density-Based Clustering

The application of the PCA with subsequent choice of the most interesting PCs yields a low dimensional representation of the initial $3N$ dimensional Cartesian MD data featuring the most important underlying processes in our data set. In this low dimensional coordinate space Γ, the system under study can be described by some stationary density f_0. Usually we are mostly interested in dynamics, which we can describe for an initial configuration $\boldsymbol{x}(t = 0) = \boldsymbol{x}_0$ and $\boldsymbol{x} = (x_1, ..., x_{d'})$ with the formal solution $\boldsymbol{x}(t) = \Phi^t \boldsymbol{x}_0$ of certain Hamiltonian equations of motion, where Φ^t denotes the flow [58].

Later in MSMs, we consider the conditional transition probability p between two metastable conformations of the protein, that is two subsets of the conformation space $S_1 \in \Gamma$ and $S_2 \in \Gamma$

$$p(S_1, S_2, t) = \frac{1}{\int_{S_1} f_0(\boldsymbol{x})\, \mathrm{d}^{d'}x} \int_{S1} \chi_{S_2}(\Phi^t \boldsymbol{x}) f_o(\boldsymbol{x})\, \mathrm{d}^{d'}x, \tag{2.10}$$

where χ_S denotes the characteristic function of the set $S \in \Gamma$, which means $\chi_S(\boldsymbol{x}) = 1$ for $\boldsymbol{x} \in S$ and $\chi_S(\boldsymbol{x}) = 0$ otherwise.

MSMs are based on the assumption that a separation of timescales between fast interstate fluctuations and slow interstate transitions exists and they are therefore very effective as they model the dynamics of the system by jumps between those states. Thus, we aim to identify some metastable subsets $S_i \in \Gamma$ which feature a high probability to stay within itself during the observed time τ, that is $p(S_i, S_i, \tau) \approx 1$.

Different approaches for clustering the data into a set of such (metastable) subsets are available. The most widely used cluster algorithm is k-means [59], which is relatively simple in a sense that it performs a Voronoi partitioning of the whole data set. A chosen number N_k of cluster centers is randomly distributed within the data set and frames (a frame is one data point in this d'-dimensional space, i.e. $\boldsymbol{x}(t) = (x_1(t), ..., x_{d'}(t))$ are subsequently assigned to the nearest cluster center by minimizing the sum of squared distances between them and the center. The position of the cluster center is rearranged iteratively until no further minimization of the sum of squared distances is achievable.

However, this method has several downsides because it is e.g. not self consistent as the input parameter

N_k is not a consequence of the underlying data, but rather a choice of the user. k-means is also not deterministic as the N_k cluster centers are initially randomly distributed in the coordinate space. Even more problematic is the fact, that the clusters resulting from k-means do not cut the states at their energy barriers since they are cut purely geometrical by Voronoi partitioning.
Density-based clustering methods are an elegant remedy to these problems as

- they cut the states at their energy barriers,
- apart from the desired final number of microstates, they do not need additional input parameters and are therefore quasi self-consistent and
- they are deterministic.

One of these techniques, namely robust density-based clustering, developed by Sittel and Stock [49], is used in this work:
Its basic idea is to construct a free energy landscape $\mathit{\Delta} G(\boldsymbol{x})$ [see Eq. (2.11)] [60] based on the local free energy estimates of each frame and subsequently identify metastable clusters as states which are well separated by their local energy barriers (see Fig. 2.4)

$$\mathit{\Delta} G(\boldsymbol{x}) = -k_\mathrm{B}T\ln\left(P_R(\boldsymbol{x})/P_R^\mathrm{max}\right), \tag{2.11}$$

where $P_R^\mathrm{max} = \max_t(P_R[\boldsymbol{x}(t)])$ is the maximum local neighborhood population. In a first step, the neighborhood population for every frame at point $\boldsymbol{x}'$ is determined by simply counting the number of frames which are located within a d'-dimensional hypersphere with the radius R around $\boldsymbol{x}'$

$$P_R(\boldsymbol{x}') = \sum_{m=1}^{N} \Theta\left[R - d(\boldsymbol{x}_m, \boldsymbol{x}')\right]. \tag{2.12}$$

Here, $\Theta[]$ is the Heaviside step function, which equals 1 if the frame lies within the hypersphere and 0 otherwise. $d(\boldsymbol{x}, \boldsymbol{x}') = \sqrt{\sum_{n=1}^{N}(x_n - x_n')^2}$ denotes the Euclidean norm in the N-dimensional space. For the hypersphere radius R, Nagel et al. showed that $R = d_\mathrm{lump}$ (see below for d_lump) empirically works very well for various model systems as d_lump represents a lower bound for R because it effectively limits the resolution of the clustering [26].
By repeating this procedure, a local free energy estimate can be assigned to every single frame through Eq. (2.11). Sorting the frames by their free energy from lowest to their highest value, the free energy landscape is constructed. First, a free energy cutoff is set to a very small value (e.g. $\mathit{\Delta} G_\mathrm{cutoff} = 0.1\ k_\mathrm{B}T$) in order to identify those frames which feature free energy estimates below (see Fig. 2.4, 1). Now the free cutoff is slowly increased (see Fig 2.4, 2) and frames are assigned to the same cluster if they feature a geometric distance below d_lump.

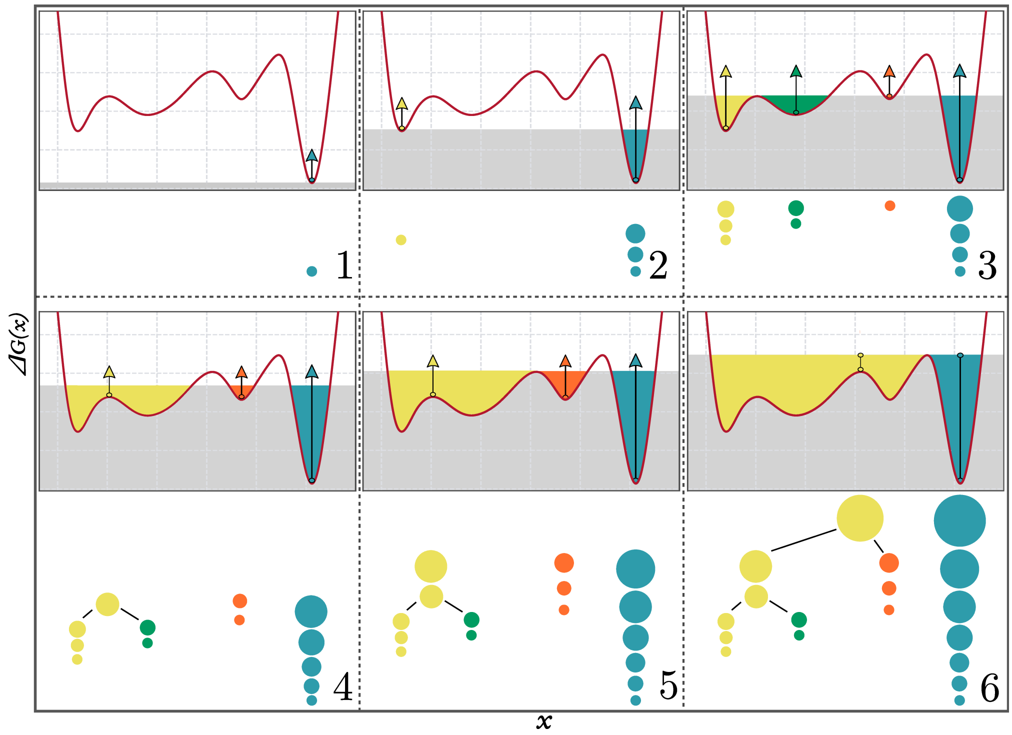

Figure 2.4.: One dimensional visualization of the density-based clustering method. The free energy profile is shown in red ●. The free energy cutoff (in grey ●) is slowly increased and geometrically isolated clusters are identified. For certain free energy values (see 4 and 6), the free energy barrier, which separates two clusters is lower than ΔG_{cutoff}—in this case the two clusters are merged. For a specific choice of ΔG_{cutoff}, the (still) isolated clusters are identified as states.
Underneath each free energy profile: Procedure of partition the coordinate space into microstates: Starting at the (bottom) end of every leaf, one follows it until it is merged at a node (free energy of barrier). If one of this branches aggregates enough frames ($P_{\text{branch}} \geq P_{\text{min}}$), a branch is regarded as a microstate.

This lumping distance is chosen as $d_{\text{lump}} = 2\sqrt{\langle x_{\text{NN}}^2 \rangle}$, where $\langle x_{\text{NN}} \rangle$ denotes the mean nearest-neighbor distance. For Gaussian distributed data, this choice guarantees a probability of 95% that a randomly picked frame has a neighbouring frame within the hypersphere spanned by the lumping radius as $d_{\text{lump}} \geq \langle x_{\text{NN}} \rangle + 2\sigma$, where σ denotes the standard deviation.
With an increasing free energy cutoff, those clusters absorb more frames and therefore grow closer together (see Fig 2.4, 3). If the distance between two frames of different clusters eventually falls below the lumping distance d_{lump}, the two clusters are merged together (see Fig 2.4, 4 and 6). Once this procedure is completed, the free energy is scanned again and clusters form microstates if they have higher populations than some desired value P_{min} prior to being merged with another cluster. Uniquely assigning every frame to such a microstate leads to a discretization of the trajectory by the resulting set of microstates.

Reaching the highest value of the free energy cutoff, the great majority of all frames are usually assigned to one cluster. Yet often a few percent of the frames remain geometrically isolated and can therefore be regarded as noise. In fact, all clusters are assigned as noise if their population does not exceed 0.1% of all data. In this case, we kinetically assign affected frames to the microstate visited before.

This procedure allows a partitioning of the coordinate space into a set of microstates and consequently a discretization of the trajectory which from now on will be referred to as microstate trajectory. In Markov state modeling, this microstate trajectory is considered a Markov chain and applying the MSM framework in Sec. 2.6 enables to extract valuable information and therefore allows to obtain deeper insights into the underlying dynamics of the protein.

2.5. Dynamical Coring

Projecting high dimensional data onto a low dimensional coordinate space (see Sec. 2.3) can lead to projection errors. We consult Fig. 2.5 for an illustrative, two-dimensional example where the depicted trajectory describes a transition from the metastable state 1 to the stable state 2. The state barrier (dashed line) is hereby only crossed once and the region around the barrier is poorly sampled due to its higher

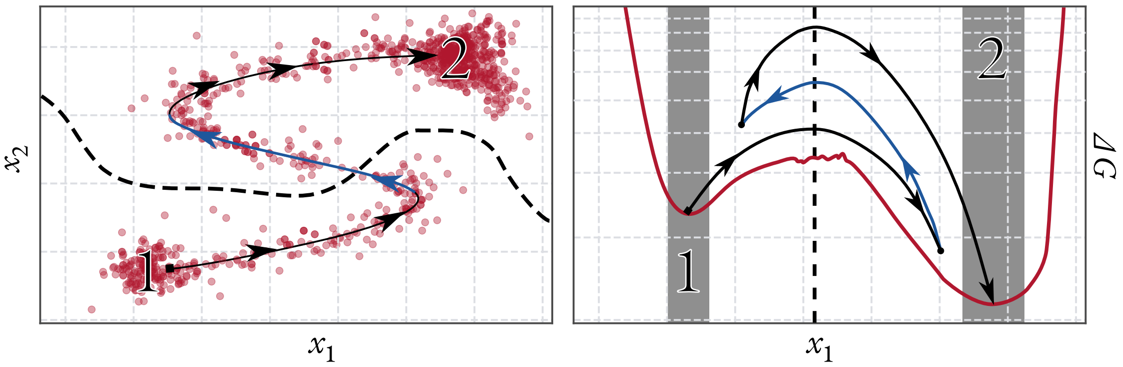

Figure 2.5.: Simple two dimensional model to illustrate one possible artifact of dimensionality reduction and coring as a possible remedy.
Left: Time evolution of a sampled trajectory between the metastable state 1 and stable state 2, separated by the energy barrier shown as a dashed line. During MD simulations, a protein typically spends a great majority of its time in (meta)stable conformations which is why the transition between both states is barely sampled.
Right: The free energy $\varDelta G(x_1)$ of the data projected on x_1. Due to the reduction of dimensionality, the part of the trajectory marked in blue is misinterpreted as a transition from state 2 to state 1, even though the energy barrier is only crossed once. Dynamical coring identifies a transition only on condition that the trajectory spends a specific time in a state.

free energy.

By projecting the trajectory on the one dimensional free energy profile along x_1 (left side of Fig. 2.5)

$$\varDelta G(x_1) = -k_\mathrm{B}T\ln\Big(\int P(x_1, x_2)\,\mathrm{d}x_2\Big), \tag{2.13}$$

it becomes evident that the transition between the two states is no longer well defined. The transition, which actually only occurs once and is marked in blue, is—due to the low dimensional projection—erroneously perceived as three transitions. In a high dimensional space, this becomes even worse which leads to a miss-classification of intrastate fluctuations as interstate transitions eventually resulting in artificially low metastability of the states and ruining the Markov property.

Fortunately, *coring* represents a simple remedy to these problems. Initially introduced by Buchete and Hummer [61], coring requires the trajectory to reach a certain zone around the center of a state in order to count the transition. Within the framework of this thesis, a similar ansatz called *dynamical coring*, invented by Jain and Stock is used [24] (for a detailed analysis compare Ref. [26]). Instead of defining the core of a state in a geometrical manner, dynamical coring requires the trajectory to spend at least a minimum time τ_cor in the new state before counting the transition. For any fluctuations which occur on faster timescales than τ_cor, the frames are assigned to the last visited state which fulfills the stable core criterion.

After coring, the shortest possible sequences of the trajectory being in one specific state is τ_cor. Apart from setting the minimal necessary time to be in a specific state, τ_cor simultaneously defines the maximal time resolution of the model. It is therefore advisable to set τ_cor as short as possible in order to access fast dynamical time scales as well. In the next section, we will see that Markov processes (see Sec. 2.6) can be used for describing the dynamics of proteins. A Markov process in a discrete state space is a homogeneous Poisson process which has the property that the probability of staying within a state $\mathrm{T}_{ii}(t)$ decays exponentially in time [62], representing a reasonable starting point for the choice of a appropriate coring time τ_cor [24]. Adopting this heuristic for large systems as e.g. PDZ2 consistently yields non-Markovian behaviour and consequently the coring time must be chosen larger. However, dynamical coring as described in Ref. [26] is not suitable for dealing with such large coring time τ_cor. Therefore, within this thesis the dynamical coring approach is supplemented by an iterative ansatz which highly improves the coverage of the original trajectory for long coring times (see Sec. 4.2).

2.6. Markov State Models

MSMs have been an important driving force for the modeling and analysis of MD simulation data, as they allow identify the most important processes in the investigated molecular system and to make statistically sound statements about them. They are extremely popular in modeling of protein dynamics

[18–24] as they facilitate a "divide and conquer" approach of extracting stationary quantities and long-time kinetics from an ensemble of short trajectories.

The key assumption of the Markov state modeling is that the dynamics of the system under study can be approximated by a Markov chain on the clustered, discrete set of microstates. In terms of the discrete microstates space Ω, we can write the Markov property as the independence of the prior history of the trajectory. Thus, the conditional probability $P(X_{T+1}|X_T)$ for the system to change its current conformation X_T given by the microstate $i \in \Omega$ to the conformation X_{T+1} represented by the microstate $j \in \Omega$ does only depend on the present microstate, i.e.:

$$\begin{aligned} P\big(X_{T+1} = j\,|X_1 = \tilde{i}, X_2 = i', \dots, X_T = i\big) \\ = P\big(X_{T+1} = j\,|X_T = i\big). \end{aligned} \tag{2.14}$$

Here, $\tilde{i}$ and i' denote arbitrary microstates. The power of the MSM lies in the fact that $P(|)$ is independent of the past which allows above mentioned usage of the ensemble of short trajectories.

This ensemble can now be used to construct the transition matrix $\mathbf{T}$ which represents apart from the state partitioning the essential ingredient of the MSM. In a first step, the transition count matrix $\mathbf{T}^c \in \mathbb{R}^{n\times n}$ which counts all transitions is set up

$$\mathbf{T}^c(\tau_{\text{lag}}) = \begin{pmatrix} \#(1,1,\tau_{\text{lag}}) & \dots & \#(1,n,\tau_{\text{lag}}) \\ \vdots & \ddots & \vdots \\ \#(n,1,\tau_{\text{lag}}) & \dots & \#(n,n,\tau_{\text{lag}}) \end{pmatrix}. \tag{2.15}$$

Here, $\#(i,j,\tau_{\text{lag}})$ is an operator which counts all transitions from state i to state j shifted by a lag time of τ_{lag}. This approach is referred to as *sliding window approach* as the transitions are counted shifted by a "sliding window" of length τ_{lag}, i.e. transitions are counted $X_0 \rightarrow X_{\tau_{\text{lag}}} \rightarrow X_{2\tau_{\text{lag}}}$ and then between $X_1 \rightarrow X_{\tau_{\text{lag}}+1} \rightarrow X_{2\tau_{\text{lag}}+1}$ and so forth.

We now seek the transition matrix $\mathbf{T}$ which maximizes the likelihood of describing the conditional probabilities of the transitions in the Markov chain, i.e. the microstate trajectory

$$\mathbf{T} = \underset{\hat{\mathbf{T}}}{\arg\max}\, p(\mathbf{T}^c|\hat{\mathbf{T}}), \tag{2.16}$$

where $p(|)$ is the likelihood. One can show [18] that the maximum likelihood estimator is simply the intuitively expected fraction of counts

$$\mathrm{T}_{ij}(\tau_{\text{lag}}) = \frac{\mathrm{T}^c_{ij}(\tau_{\text{lag}})}{\sum_{k=1}^{n} \mathrm{T}^c_{ik}(\tau_{\text{lag}})}. \tag{2.17}$$

Moreover, we assume following properties:

Ergodicity: The discrete microstate space Ω does not have two or more dynamically disconnected subsets. This implies that for an infinitely long sampling time $t \rightarrow \infty$, the system will visit every microstate i infinitely often. This allows us to define a stationary density $\mu(i) : \Omega \rightarrow \mathbb{R}_0^+$ in the whole state space Ω

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T \mathrm{d}t\langle i(t)\rangle_\mathrm{r} = \int_\Omega \mathrm{d}i\langle i\rangle_\mathrm{r}\,\mu(i), \tag{2.18}$$

with $\langle\rangle_\mathrm{r}$ denoting the running average.

The stationary density thus indicates the fraction of the time the system spent in a specific microstate during an infinitely long trajectory. It is normalized, therefore $\int_\Omega \mathrm{d}i\,\mu(i) = 1$. For finite sampled data, the stationary density μ depends on the lag time τ_lag: $\mu = \mu(\tau_\mathrm{lag})$

Detailed Balance: For equilibrium processes which take place in thermal equilibrium, reversibility must hold. This means that the number of transitions from state i to j, must correspond to the number of transitions from j to i:

$$\mu(i,\tau_\mathrm{lag})T_{ij}(\tau_\mathrm{lag}) = \mu(j,\tau_\mathrm{lag})T_{ji}(\tau_\mathrm{lag}), \tag{2.19}$$

$\mu(i,\tau_\mathrm{lag})$ is the stationary density in microstate i for a specific lag time τ_lag. This applies due to thermodynamical considerations: In case that Eq. (2.19) is not fulfilled, a set of microstates could form a loop in Ω which the system takes on primarily in one direction. Under certain conditions, such a system could fulfill the role of a perpetuum mobile as it would produce work. However, no external energy is added in thermal equilibrium which is why the production of work would violate the second law of thermodynamics.

For non-equilibrium processes this however is not the case as systems are externally driven out of their equilibrium conformation. Applying the principle of detailed balance in non-equilibrium data, certain backwards transitions could arise which only exist in forward direction.

Chapman-Kolmogorov-Equation: With the partitioning of the coordinate space into the microstate space and the construction of the transition matrix $\mathbf{T}$, the MSM is set up and ready to use.

The Chapman-Kolmogorov equation can be used in order to evaluate the quality of the MSM by comparing the predictions of the MSM to the data which was used to construct the MSM:

$$[\mathbf{T}(\tau_\mathrm{lag})]^k = \mathbf{T}(k\tau_\mathrm{lag}), \qquad k \in \mathbb{N}^+ \tag{2.20}$$

> The right hand side is directly computed from the MD simulations data while the left hand side is a simple matrix multiplication for the transition matrix set up at the lagtime τ_{lag}. A good agreement of both sides indicates Markovianity of the investigated system.
> This test is usually only performed for the the diagonal elements of $\mathbf{T}$ as comparing off-diagonal elements is very inconvenient and problems related to low sampling are likely to arise.

A very interesting kinetic property, which is often also accessible in experiment, are the relaxation timescales of the investigated system. In MSMs, this relaxation timescales correspond to implied timescales t_i, which can be calculated from the eigenvalues λ^{ts} of the transition matrix $\mathbf{T}(\tau_{\mathrm{lag}})$

$$t_i(\tau_{\mathrm{lag}}) = -\frac{\tau_{\mathrm{lag}}}{\ln \lambda_i^{\mathrm{ts}}(\tau_{\mathrm{lag}})}. \tag{2.21}$$

In a Markov system, the eigenvalues $\lambda_i^{\mathrm{ts}}(k\tau_{\mathrm{lag}})$ can be approximated by $\lambda_i^{\mathrm{ts}}(k\tau_{\mathrm{lag}}) = [\lambda_i^{\mathrm{ts}}\tau_{\mathrm{lag}}]^k$ of which follows that the timescales t_i should be relatively constant [63]

$$t_i(k\tau_{\mathrm{lag}}) = -\frac{k\tau_{\mathrm{lag}}}{\ln \lambda_i^{\mathrm{ts}}(k\tau_{\mathrm{lag}})} = t_i(\tau_{\mathrm{lag}}). \tag{2.22}$$

For complex systems, like e.g. PDZ2, this is however not the case for very short lag times. As implied timescales tend to become more constant for higher lag times, this test can determine a suitable lag time τ_{lag} for the construction of the MSM.

2.7. Typical Workflow

The theoretical background explained in this chapter allows us to move on to investigate the allosteric transition in PDZ2. In the following, all steps are briefly presented according to the order in which we will employ them.

1. *Selection of essential internal coordinates.* Starting with $3N$ Cartesian coordinates, where N denotes the number of atoms of the protein, we need to decouple the proteins dynamics from the global motion. Therefore we transform the coordinates to internal coordinates, which can be e.g. $C_\alpha-$distances, inter residual distances or backbone dihedral angles. By selecting a certain set of coordinates, one can shift the focus on areas of the protein which are of prior interest. In this thesis, interresidual contact distances were chosen.

2. *Dimensionality reduction.*

   a) *Projecting the data.* Since the later used clustering method is based on a free energy estimate, the density of the sampled points needs to be drastically increased in order to obtain well defined microstates. Therefore we apply a dimensionality reduction technique, which can be PCA, TICA or a machine learning based method. Here, PCA is used.

   b) *Selection of important PCs.* The resulting PCs from the PCA are investigated for properties such as cumulative flux or their free energy projections. A low dimensional representation of the simulated data is selected.

3. *Clustering.* The low-dimensional representation of the input data is clustered which yields a set of metastable microstates which can be assigned to certain conformations of the protein.
   Lumping. (Optional) In an additional optional step, these microstates can be lumped into a smaller set of macrostates by kinetically matching them together.

4. *Coring.* The reduction of dimensions in step 2 introduces artefacts that cause intrastate fluctuations to be misinterpreted as interstate transitions. Coring helps to correct this artefacts.

5. *Constructing the MSM.* After a suitable choice of the microstates and the coring time, the transition matrix and therefore the MSM can be constructed.

6. *Interpretation and validation of the MSM.* A row of predictions provided by the MSM such as timescales, and among other things, the most important pathways shed light on the underlying dynamics and mechanisms of the system under study.
   The predictions of the MSM must be compared with the MD data in order to verify their validity.

On the Choice of Presented Results

We will try to provide a well understandable and structured format also for a reader which may not be familiar with every aspect of Markov State Modeling. Therefore, one suitable model is used in the following two chapters, Ch. 3 and Ch. 4, for discussing each of the steps presented in the typical workflow (see Sec. 2.7) in detail. An attempt is also made to illustrate the multitude of possible decisions in order to show that there is no such thing as the ideal way. As the data used for this model mostly stayed the same during all the time spent on this model, we will refer to it as *data generation 1*. In Chapter 5 we consequently apply all the knowledge gained in the course of working with data generation 1 directly on a reduced data set what we then call *data generation 2* on which we will construct an additional, second MSM. This allows us to evaluate the reliability and the replicability of the clustering and the MSM predictions. Besides, we will see to what extend a smaller—but more targeted—set of input coordinates improves the quality of the MSM's predictions.
The findings of both models are then conclusively discussed in the last chapter, Ch. 6.

3. Data Generation 1: Preparative Steps Towards a Nonequilibrium-MSM of PDZ2

It always takes longer than you expect, even when you take into account Hofstadter's law.

Hofstadter's law

In this chapter, the preparing steps to construct a MSM for the full data set of PDZ2 (equilibrium and nonequilibrium trajectories) are explained. This includes in particular the choice of suitable internal coordinates in form of some distance metric (see Sec. 3.2) and the subsequent selection of adequate PCs depending upon various properties such as cumulative flux, free energy projections or autocorrelation functions (see Sec. 3.3).

In a subsequent step, the chosen PCs are clustered to get a set of microstates (see Sec. 3.4). The procedure performed in this chapter essentially comprises the steps 1–3 presented in the workflow in Sec. 2.7.

3.1. Previous Works

Several studies of PDZ2 have already been carried out and will shortly be discussed here. Besides Sebastian Buchenberg, who carried out the simulations of the equilibrium and nonequilibrium trajectories [17, 42], several persons were involved in PDZ2 Markov state modeling.

So far, all attempts of Markov modelling of PDZ2 were based on dihedral angles as input coordinates followed by a dihedral angle principal component analysis (dPCA+), a method which was developed to analyze periodic input coordinates with minimal projection error [53]. First, dihedral backbone angles were used as they are an intuitive choice for tackling the interesting dynamics of PDZ2, which mostly happen in its loop regions.

As the signal-to-noise ratio was one of the first problems which were encountered, F. Sittel proposed to take only those dihedrals into account which feature a clear multi-state behaviour [64]. This led to a reduction from the overall available 192 dihedral angles down to 104 angles. Neglecting almost half of

the overall available dihedral angles was possible since most of the excluded angles are located in stiff β-sheets which hardly change at all.
By performing dPCA+ on fourteen 2.5 μs long equilibrium trajectories, S. Ohnemus achieved a partial separation of the Cis and Trans regions when projecting the free energy onto the first two PCs x_1 and x_2 [65]. Yielding a good description for the Cis → Trans transition, the nonequilibrium data was subsequently projected on the first six PCs of the equilibrium dPCA+. A MSM was constructed which aimed to describe the Cis → Trans transition. After clustering, Cis and Trans microstates were not clearly separated which, among other things, considerably limited the validity of the MSM. Many problems can be traced back to the fact that the trajectories were not well converged. Therefore, the equilibrium trajectories were elongated to 10 μs each now matching the length of the long nonequilibrium trajectories.
This extended data set was analyzed by A. Weber in Ref. [66]. One of the seven Cis trajectories was dominated by two "trap microstates". Both microstates featured deformations in the α_2-helix leading to left-handed α-helical conformations. As PDZ2 naturally occupies right-handed helical conformations and the transition between left- and right-handed helical conformations is known to be very slow, it is reasonable to assume that this is the reason why the system is not able to leave those trap microstates. The azobenzene photoswitch, linked to a neighboring residue, might have caused this unnatural trap conformations [66]. Hence, this trajectory was removed from the data set and—to preserve equality for Cis and Trans—one Trans trajectory was removed as well. This is the actual state of the PDZ2 data, which now consists of 60 μs Cis, 60 μs Trans and 288 μs nonequilibrium trajectories.
It turned out that dihedral angles primarily work well in small proteins but are not well suited to tackle the conformational changes in a relatively large protein such as PDZ2. Thus, no explicit Cis and Trans regions could be determined which hindered any MSM to describe the Cis→Trans transition. Supported by the success for other large proteins such as T4 Lysozyme [43, 67], distances seem to be more promising internal coordinates for describing the dynamics in PDZ2.

3.2. Choice of Internal Coordinates

In the very first step, the intrinsic motion of the protein must be decoupled from the movement of the protein in the bath. Since we work with distance-based internal coordinates, two different options for them are discussed in Sec. 3.2.1 before a suitable set of coordinates is chosen in Sec. 3.2.2.

3.2.1. C_α-Distances vs. Contact-Distances

When working with distance-based internal coordinates (see Sec. 2.2), several different options to specify the start and end point of the distances are available. Depending on the problem at hand, one may e.g. focus on the backbone structure if dynamics on a large scale are involved (e.g. protein folding). In this case, C_α-distances would be appropriate since the difference between the relatively freely moving side chains and the backbone is negligible compared to the overall motion of the protein. In addition, the backbone (and therefore C_α-coordinates) tends to better describe the overall structure of the protein because unlike the side chains, it is not so susceptible to rapid fluctuations.

For PDZ2 however, fast dynamics within the side chains attached to the backbone seem to be the main driving force for the allosteric transition [29]. Contact-based distances are the shortest distance between atoms from two different residues [see Eq. (2.4)] and can therefore take dynamics of the side chains into account. As side chains often undergo rapid changes in their orientation, the shortest atoms between two residue can change several times in the course of the MD trajectory. The increased susceptibility in the side chains makes them the more promising candidate to describe the allosteric transition in PDZ2 compared to C_α-distances.

In order to verify this assumption, contact-distances and C_α-distances between two residues, evenly distributed throughout the protein, were calculated and then compared. In Fig. 3.1, the probability distribution of the distances P_d in Cis and Trans conformation of the protein is plotted for the shortest lying atoms of two residues (*left*) and for the C_α-atoms (*right*). Shown are two exemplary distances between the residues 17 and 22, covering the distance between the $\beta_1\beta_2$-loop and the stable β_2-sheet and the distance between residue 27 and 33 which represents dynamics in the very flexible $\beta_2\beta_3$-loop. The probability distribution of the distance between the residues 17 and 22 in particular reveals that C_α-distances may be too static for the subtle changes in the proteins conformation—the distribution almost looks Gaussian for the Cis and Trans conformation at mean values of ~1.28 nm and ~1.35 nm respectively. In contrast to contact-distances for which two well defined maxima are visible which indicates a higher resolution of dynamics in this picture. The side chains adopt different conformations in contrast to the relatively stiff backbone. Note that in general we find that the contact-distances are smaller compared to C_α-distances , as both side chains are oriented towards each other most probably resulting in the formation of hydrogen bonds (especially in the Cis conformation).

In the $\beta_2\beta_3$-loop, the contact-distance also provides a more differentiated picture emphasizing the advantage in resolution of the very flexible side chains over the stiff backbone. Although the probability distribution for the C_α-distances shows minor differences between Cis and Trans conformations as well, they are not that distinct as for contact-distances. To conclude, since the contact-distances seem to resolve the differences between Cis and Trans in more details, they are used in the following. One might note here, that the term “contact-distances” is misleading for long distances, since in that case no contacts are formed. But even for long distances, we expect contact-distances to perform equally well or slightly better, with relatively small differences compared to $C_\alpha-$distances.

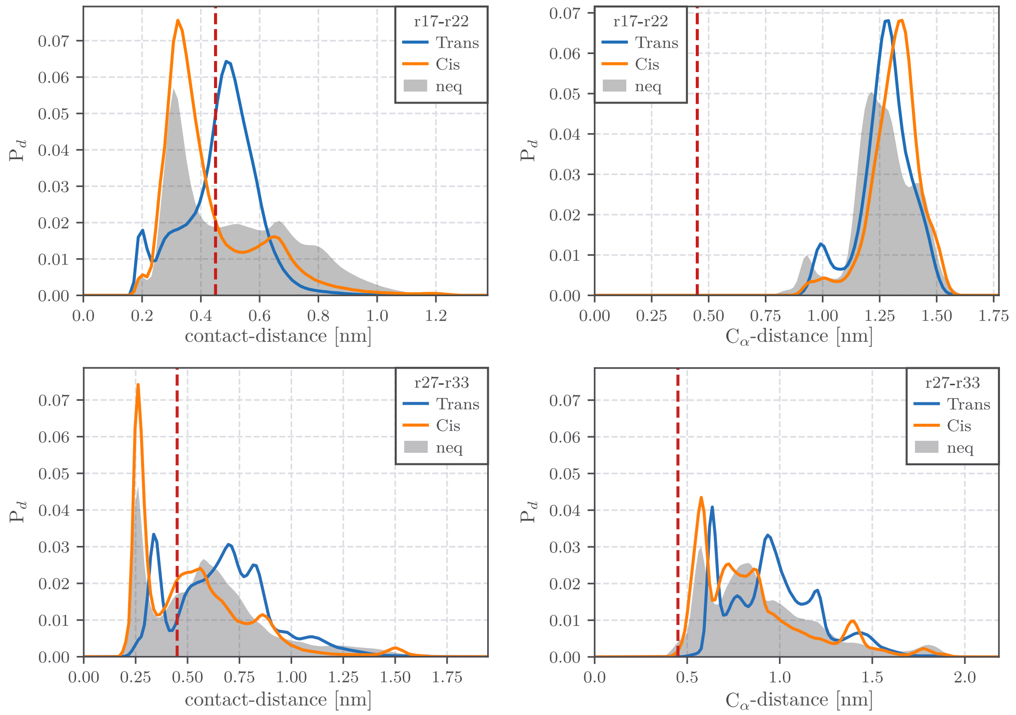

Figure 3.1.: Probability distribution for Cis (● orange), Trans (● blue) and nonequilibrium (● grey) using contact-distances (*left*) and the C_α-distances (*right*). *Top:* distances between residues 17 and 22. *Bottom:* distances between residues 27 and 33. ● The red dashed line indicates the threshold of 4.5 Å which indicates the maximum distance for the formation of hydrogen bonds.

3.2.2. Selection of Inter-Residual Contact-Distances

In the last subsection, it was illustrated that contact-distances are well suited as internal coordinates since they emphasize the differences between the Cis and Trans conformation of PDZ2. For the full description in terms of these internal coordinates one would need to transform the Cartesian coordinates of all 96 residues (more precise, the atoms of all these residues) into $(96-1)^2/2 \approx 4500$ inter-residual distances (factor $^1/2$ as $d_{ij} = d_{ji}$). Because the calculation of distances from the MD data is accompanied by a high computational effort, it is not feasible to calculate all those 4500 distances.

Additionally, problems arise when calculating the correlation matrix for the PCA as the amount of data would exceed the capacity of the computer systems available.
Besides, and more importantly, a reduced number of distances reduces the "noise" in the data which improves the quality of the PCA. Hence, we aim to lower this number drastically in order to perform a PCA which is affected by as little noise as possible while not exceeding computational capacities. A first preselection for the residue pairs used for the calculation of the distances is made by applying the following four criteria:

1. The residues could potentially form *hydrogen bonds* by featuring a lone pair on a electron-rich donor atom in the first residue (nitrogen (N), oxygen (O), fluorine (F)) while the second residue functions as acceptor.

2. The residues could potentially form *salt bonds/ionic bonds* by featuring both, positive and negatively charged ions, respectively.

3. The residues are located at the beginning or the end of a stable secondary structure element and could therefore cover the relative movement between secondary structure elements.

4. The distance between two residues falls below 4.5 Å in the course of the MD simulation. This criterion certainly has a large overlap with especially the hydrogen bond criterion, but it turned out that important pairs of residues which are not covered by one of the first three criteria can still be found this way.

For all criteria, pairs of residues were excluded if they are less than four residues apart from each other as those often form stable helical structure elements which are relatively stiff. Applying all four criteria to the data, little more than 1600 pairs of residues were identified which fulfill at least one of them. While already cutting the number of distances down to about one third, the number of remaining distances must be further decreased in order to reduce noise in the data and make the PCA computationally feasible. Therefore, the probability distribution in the Cis and Trans conformation of those 1600 contact-distances were evaluated (see Fig. A.1 on page 86). As we seek collective variables (in the form of principal components) which describe the allosteric transition between the Cis and Trans conformation of the protein, those contact-distances need to be retained which show a clear separation between their Cis and Trans probability distribution. This is the case for e.g. the distance between the residues 29 and 94 in Fig. A.1, while the distance on the between the residues 2 and 87 does not show a distinct separation and is thus discarded. By following this procedure, the ~1600 contact-distances could be further reduced down to 429 distances shown in a matrix representation in Fig. 3.2. The red dots indicate the contact-distances which are retained while the black dots represent pairs of residues which are less than four residues apart and therefore excluded. The yellow dots indicate the two α-helices.

As several clusters of residue pairs can be found in this matrix representation, high correlations between the contact-distances in these clusters are probable. In Ch. 5, a workflow is presented which aims at reducing correlations between the distances to a minimum.

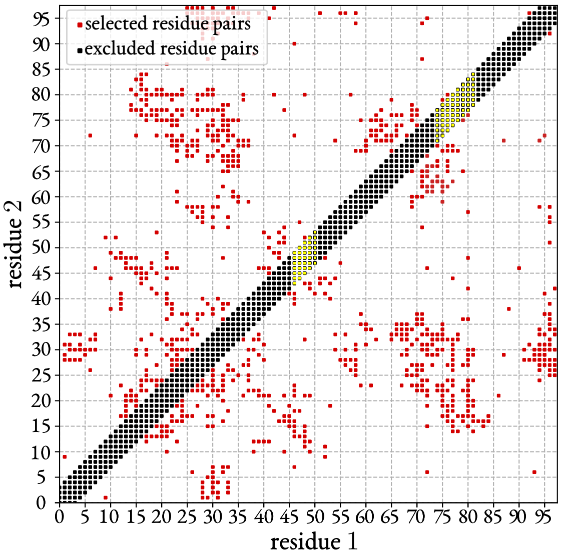

Figure 3.2.: (● red) Selected residue pairs used in the PCA due to their distinct difference between their Cis and Trans distribution. (○ yellow) The yellow fields indicate the two α-helices. (● black) Contacts between residues which are less than four residues apart are discarded as they often occur within helical structure elements or do not feature a distinct difference between their Cis and Trans distribution. The matrix is symmetric as $d_{ij} = d_{ji}$.

...in a nutshell.

Compared to C_α-distances, contact-distances seem to be better suited as internal coordinates for PDZ2 due to the importance of dynamics in the side chains. The potential ~4500 contact-distances were reduced to 429 contact-distances by excluding those distances which are not featuring a distinct difference between their Cis and Trans probability distribution.

3.3. PCA and Selection of PCs

In order to reduce the dimensionality of the data, which was preselected in the last section, a standard PCA is performed in two steps. We aim at identifying reaction coordinates which describe the conformational transition between Cis and Trans conformation. Since the input data was chosen in such a way that the most prominent difference in the data is a large variance between Cis and Trans conformation, we expect x_1 in particular to describe the allosteric transition. In order to obtain PCs which well separate Cis from Trans, the PCA was initially ran only on the equilibrium data and in a next step the nonequilibrium data was projected onto the eigenvectors of the equilibrium PCA. This procedure has been proven advantageous in previous works [65, 66]. However, we found that the equilibrium trajectories were not yet fully equilibrated and that the first 3 μs are the most affected ones. While we still retain these first 3 μs of all equilibrium trajectories in data generation 1, we completely discard the affected data from data generation 2 on.

The range of contact-distances extends from about ~0.2 nm to about ~3.5 nm. Consequently, the contact-distances for the residues far apart tend to undergo changes which are greater in absolute value than those of residues nearby. Because those larger changes are not necessarily more important than smaller ones, which e.g. also describe the formation of hydrogen bonds, we calculate the correlation matrix instead of the covariance matrix [see Eq. (2.7)].

Fig. 3.3 shows the cumulative fluctuations of the equilibrium data. The cumulative fluctuations provide information on how much of the systems total variance is covered within the first n PCs by cumulatively summing up the corresponding eigenvalues. One can see that the first seven PCs cover already more than 60% of the systems overall motion. It takes another 13 PCs (x_1–x_{20}) to cover 80% of the variance

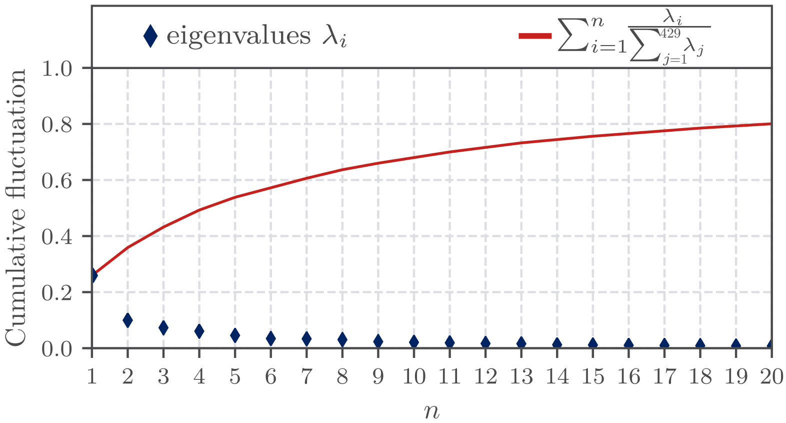

Figure 3.3.: The eigenvalues (here sorted from high to low and normalized) indicate how much of the systems variance much is covered by the corresponding PC (marked in ● blue). The cumulative fluctuation is the sum of the eigenvalues up to a specific number of PCs n (marked in ● red).

present in the data and 192 PCs in total to cover around 99%. Hence, the last 237 PCs only account for 1% of all variance in the data set (see Appendix, Tab. A.1 on page 85). Now, that 429 PCs are available to choose, we need to select a low number which can be clustered to obtain microstates. The selection will be based on four criteria:

1. *Eigenvalues of the PCs and the resulting cumulative flux (see Fig. 3.3 and Tab. A.1)*
2. *Temporal evolution of the data along the PCs (see Fig. 3.4)*
3. *Slow decay of the autocorrelation functions of the PCs (see Fig. 3.5)*
4. *Free energy projections on the PCs and their ability to separate Cis and Trans (see Fig. 3.6)*

The application of these four criteria to the examined PCs aims for the best possible description of the proteins dynamics. Since most of them generally go hand-in-hand, it is sufficient to only analyze to the first few PCs as those already cover the great majority of the systems dynamics (see Fig. 3.3).

Temporal Evolution of MD Trajectories

Our goal is to identify and isolate the important, slow dynamics of PDZ2 by selecting and retaining those PCs, which describe the system's essential motion and discarding those which show mostly uncorrelated motion. One approach to do this is the investigation of the temporal evolution of the systems trajectory along the different PCs.

We hope to identify PCs which can distinguish multiple metastable states along the trajectory because this is a strong indicator for non-random dynamics of the system as they should be clearly distinguishable from the mostly uncorrelated motion of the bath. For illustration, Fig. 3.4 shows the projected data along the first 8 PCs x_1–x_8 and additionally along x_{78} to offer some comparison of how non-essential dynamics might look like. x_{78} was chosen as a example because it only contributes 0.01% to the cumulative flux (see Tab. A.1).

Each grid-box in Fig. 3.4 represents one single trajectory with a length of 10 μs and the type of the trajectory is represented by orange for Cis (first 6 trajectories), blue for Trans (next 6 trajectories) and grey for nonequilibrium (last 20 trajectories). On the right hand side one can see the free energy projection along the corresponding PC x_i, i.e.

$$\varDelta G(x_i) = -k_{\mathrm{B}}T \frac{\ln P(x_i)}{P_{\max}}, \tag{3.1}$$

where $P_{\max} = \max_t P[x_i(t)]$ corresponds to the highest probability density. The free energy projection thus measures how often a certain value was sampled this way indicating metastability. Thus, PCs with several local minima are good candidates for resolving the slow dynamics of the protein while PCs with a single minimum usually do not provide any valuable information.

A good example is x_1: During its temporal evolution, most of the single trajectories occupy several plateaus corresponding to metastable conformations which arise again and again. This indicates that important dynamics are taking place here. In addition, Cis and Trans trajectories on average occupy different values while the nonequilibrium trajectories occupy values in between which suggests that the transition region is nicely covered. Consequently, the free energy projection onto x_1 shows multiple local minima located in regions far away from $x_1 = 0$. The next 4 PCs, x_2–x_5, all resolve several local minima in the free energy and feature plateaus as well.

In contrast, distinct plateaus are far less occupied along x_6 in the course of the evolution, both in the equilibrium and in the nonequilibrium trajectories. This is also reflected in the free energy plot which indicates that the frames are Gaussian distributed around $x_6 = 0$. However, x_7 seems to resolve more structure, especially in the Cis trajectories which is also notable in the slightly more structured free energy projection (*left*, at values around $x_7 \approx 8$ a local free energy minimum can be observed). This hints that x_7 may be better suited to capture important motion of the protein compared to x_6. Looking at our example for non essential PCs, x_{78}, one can see that hardly no structure is resolved and that only very small values are covered The free energy projection along x_{78} on the right confirms this impression as only a single minimum is apparent.

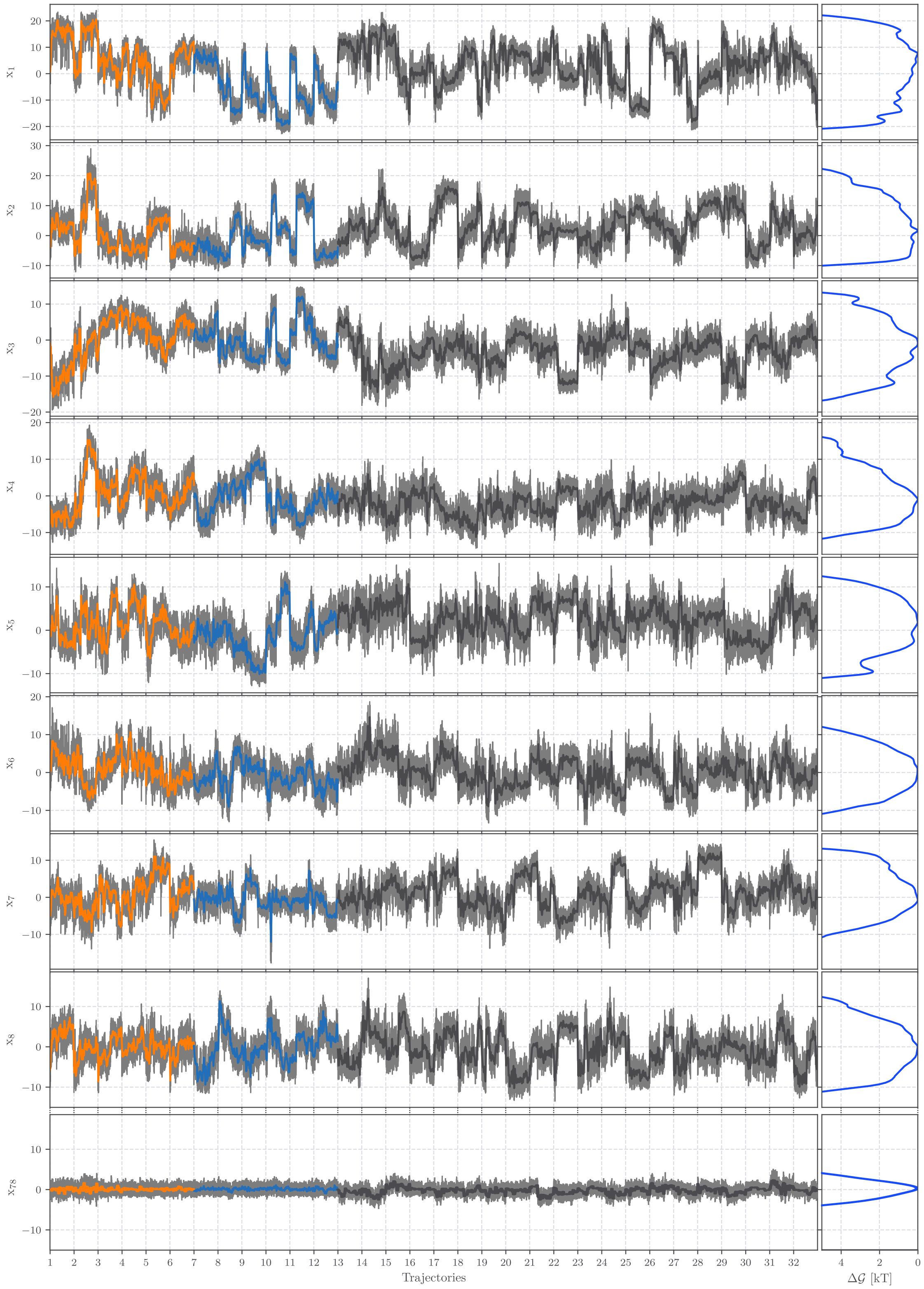

Figure 3.4.: *Left:* Temporal evolution along the first 8 PCs. Orange ● denotes the symmetric time average over 10^5 Cis frames, blue ● Trans frames and grey ● are nonequilibrium frames. As a reference, x_{78} is shown as well which contributes 0.01% to the cumulative flux. *Right:* On the right side, the free energy is projected onto the corresponding PC.

Autocorrelation functions

The dynamics of proteins are often complex and evolve over several timescales. A plain analysis of the time evolution of the trajectory along the PCs, as it was performed in the last subsection, is often helpful for judging the ability of the PCs to separate essential motion from noise, but it is only partially suitable for making predictions about the time scales involved. The latter can be examined by calculating the autocorrelation function (ACF) for different principal components i, which, in our case, describe the correlation of the system with prior conformations in form of contact-distances along directions of maximum variance $x^{(i)}(t)$ as a function of the lag time τ

$$\mathrm{ACF}_i(\tau) = \frac{\left\langle \left(x^{(i)}(t) - \langle x\rangle\right)\left(x^{(i)}(t+\tau) - \langle x\rangle\right)\right\rangle}{\langle x^2\rangle - \langle x\rangle^2} = \frac{\mathrm{Cov}\left[x^{(i)}(t), x^{(i)}(t+\tau)\right]}{\mathrm{Cov}[x^{(i)}(t), x^{(i)}(t)]}, \tag{3.2}$$

where $\langle x\rangle$ denotes the mean of x and $\langle x^2\rangle - \langle x\rangle^2$ the variance. This can be rewritten for a discrete trajectory with N_T data points $x_k^{(i)}$ as

$$\mathrm{ACF}^{(i)}(\tau) = \frac{\frac{1}{N_T-\tau}\sum_k^{N_T-\tau}\left(x_k^{(i)} - \mu_T\right)\left(x_{k+\tau}^{(i)} - \mu_T\right)}{\frac{1}{N_T}\sum_k^{N_T}\left(x_k^{(i)} - \mu_T\right)^2}, \tag{3.3}$$

where $\mu_T = \frac{1}{N_T N_{\text{traj}}}\sum_i^{N_{\text{traj}}}\sum_k^{N_T} x_k^{(i)}$ denotes the mean of one equilibrium conformation (Cis or Trans) for $N_{\text{traj}} = 6$ trajectories each [68]. Calculating the ACF this way leads to meaningful results only for stationary equilibrium-processes, because they feature time-independent mean and variance. Instead of calculating the mean and variance for each trajectory separately, we calculate them for once Cis and once for Trans separately. The underlying reason for that is that all trajectories of one kind (Cis or Trans) are part of one conformation of the system and we do not anticipate that a single trajectory represents the entire dynamics of the system, but the complete ensemble instead [69].

As the sample size $N_T - \tau$ decreases linearly which results in a higher statistical error, values $\mathrm{ACF}(\tau > {}^{N_T}\!/_2 \approx 5\,\mu\mathrm{s})$ should be analyzed with caution. If the motion of the system along x_i stays correlated for some lag time τ, i.e. non-random, the ACF deviates from zero. Therefore we conduce the ACF-analysis in order to test whether a timescale separation between the slow dynamics of the system and uncorrelated fast motion of the remaining components exists. We expect a much slower decay for the first few PCs while higher PCs should decay faster as they are expected to represent bath dynamics.

Fig. 3.5 shows the ACF of the first 8 PCs for the Cis and Trans trajectories. Again, x_{78} is showed to offer a comparison of what a fast decaying PC looks like. We see that x_6 falls off comparatively fast for both Cis and Trans trajectories. Especially in the Cis ACFs, one can see that x_7 decays relatively slow which matches the observations made in the last section, where we notice a higher correlated motion of the trajectory along x_7 compared to x_6. For the Trans trajectories both, x_6 and x_7, decay in a similar manner

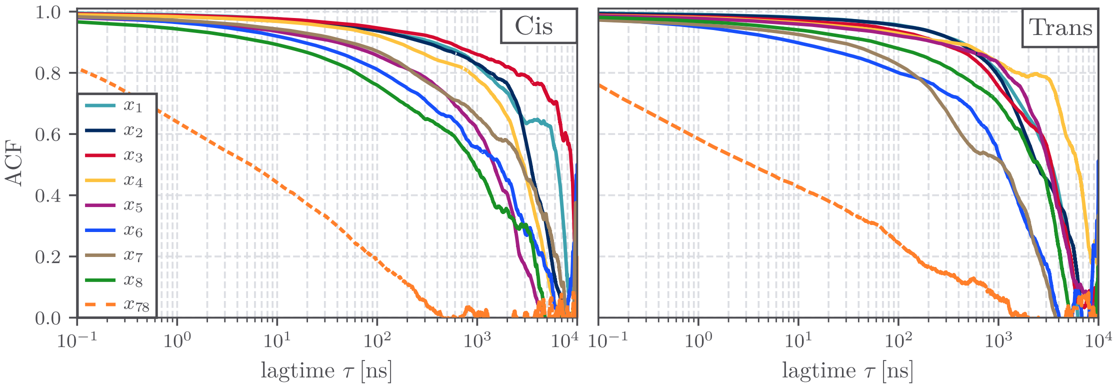

Figure 3.5.: Autocorrelation function of the equilibrium conformations (*left:* Cis, *right:* Trans) for the PCs x_1–x_8 and x_{78}.

which is not surprising as the time evolution of the trajectories along both PCs look similar (compare Fig. 3.4). As x_7 seems to be more important for resolving structural changes in the Cis conformation, this indicates that x_7 might be better suited to describe the systems dynamics as x_6. Not surprising, x_{78} decays orders of magnitude faster than the rest.

It is important to emphasize that neither the examination of the trajectory evolution along the PCs, nor the ACF analysis alone is sufficient when conduced single-handed since the former does not allow precise predictions about the time scales involved and the latter does not tell us anything on the frequency of events which both should be taken into consideration when selecting the PCs. This is further supported by the fact that such an investigation of ACFs, as it was performed here, is only meaningful for equilibrium trajectories since we presume relatively constant values for mean and variance which is however not fulfilled for nonequilibrium trajectories.

Free Energy Landscapes

After discussing the contribution of the individual PCs to the overall variance, the time evolution and the ACF of the first few PCs in the last three sections, the last missing piece is to discuss their ability to split the free energy landscape into Cis and Trans regions.

Further ahead in Sec. 3.3, the PCA setup was explained: The principal components were computed using only the equilibrium data in order to guarantee the best possible separation between Cis and Trans conformation. Here we want to illustrate to what extent such a separation has been achieved. To this end, the free energy landscape of the whole data set, as well as the Cis and Trans datasets separately, have been projected onto two PCs at a time. By adding a second dimension to the free energy projections, one can get insights into the interplay of both used PCs and therefore examine the connectivity between local minima which can not be resolved in only one dimension.

Fig. 3.6 shows the two dimensional free energy projections onto the first 7 PCs. We also investigated x_8

in detail but will refrain from showing it in the following for the sake of clarity. Regarding x_7 and x_8, no big differences are apparent, but since x_7 contributes more to the overall variance (see Tab. A.1), we favored x_7 over x_8.

One can see that x_1 nicely separates Cis and Trans data and that there are large regions present which are either populated by Cis or by Trans. It is worth noting that there are two equilibrium trajectories which fall out of the pattern, namely the fifth Cis trajectory and the first Trans trajectory. Both of them are located in the region typically occupied by the remaining trajectories of the other type (see Fig. A.3). Besides the possibility of being a projection artifact, this might indicate that the initial conditions for performing the equilibrium simulations were not yet fully equilibrated resulting in an inadequate propagation of the remaining trajectory. Interestingly, x_3 does also separate the Trans region from the Cis region to some extend.

Approaching higher PCs, the overlap of Cis and Trans raises since the free energy projections of both Cis and Trans increasingly approaches a Gaussian distribution. Keep in mind, however, that it is hard to judge the overall connectivity/separation only based on two dimensional projections. Concerning higher PCs, despite the overlap in two dimensions, Cis and Trans region might still be separated in the full dimensional space. Regarding the importance of x_6 or x_7, no big difference between both are notable in this two-dimensional representation.

...final choice of PCs

Over the last four sections, we discussed different properties of the PCs in order to analyze them with respect to their suitability to separate the system's essential motion from the bath dynamics. Regarding the contribution of the individual PCs towards the overall variance of the system and the time traces with the corresponding one-dimensional free energy projection, we could see a big difference between the first five PCs and the rest. As already more than half of the system's dynamics is covered with these five PCs we retained them for the subsequent clustering process. We aimed to cover around ~60% of the dynamics of the system which is why we included one additional PC. In most of the criteria we imposed, x_7 appears to be a more promising candidate compared to x_6, resulting in the final set of PCs consisting of x_1–x_5 and x_7. We have decided to not take any more PCs along, because the resulting quality of the microstates resolution would suffer with each additional PC. This would result in an even larger clustering radius which increases with each additional PC due to a larger mean nearest-neighbor distance.

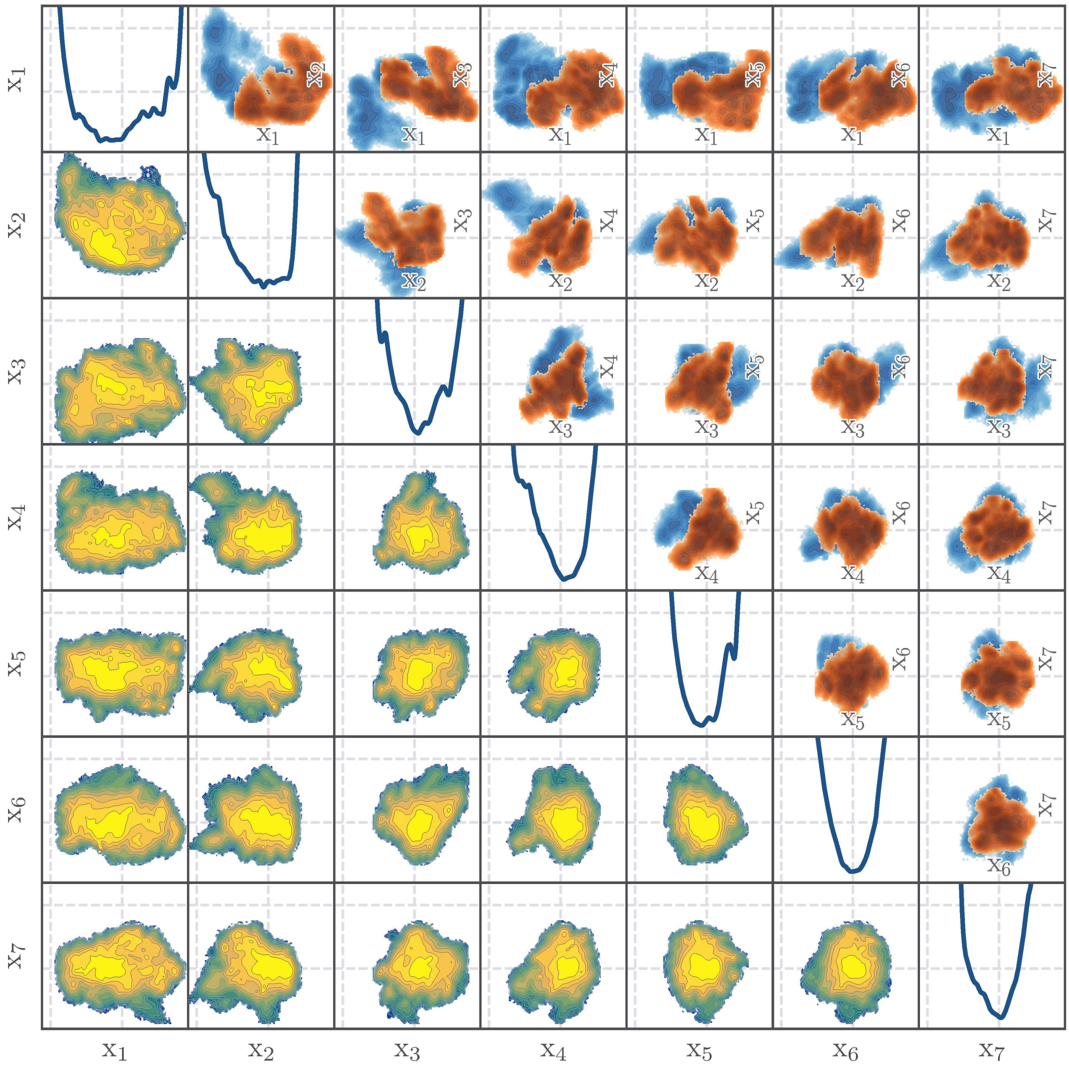

Figure 3.6.: *Lower diagonal elements:* Free energy landscapes for combinations of the PCs x_1–x_7. On the diagonal the one dimensional free energy profile in the respective PC x_i is shown. The bright yellow color indicates the free energy minima, while greenish colors announce increasing free energy values up to high free energy values which are displayed in blue.
Diagonal elements: One-dimensional free energy projections along the corresponding PC. *Upper diagonal elements:* (Non-)separation between Cis (● red) and Trans (● blue). For the sake of clarity, the upper diagonal elements are arranged in the same manner as the bottom diagonal elements (see small labels). Differences between upper and lower diagonal elements is due to the nonequilibrium data, which is included only in the lower diagonal elements. Dark colors indicate free energy minima while bright colors mean high values of the free energy.
Since we only discuss the free energy qualitatively, we refrain from showing the colorbar.

3.4. Clustering

In the last section, a reasonable set of collective coordinates was identified (x_1, x_2, x_3, x_4, x_5 and x_7). Now we cluster this data by the density-based clustering method described in Sec. 2.4 in order to discretize the high-dimensional MD-simulations trajectory into a discrete microstate trajectory, enabling one to construct a MSM.

Following the heuristic of Nagel *et al.* [26], the clustering radius was determined to be $R = d_{\text{lump}} = 0.94$ which ensures a probability of at least 0.95% of finding a neighbor for every data point within that radius. Subsequently, the minimal population P_{min} is the only parameter left to choose. P_{min} denotes the number of frames, typically given as percentage of all MD frames, which a cluster must at least contain to form a microstate. On the one hand, it must be large enough to prevent the formation of numerous microstates within one single local minimum due to fluctuations in the free energy. On the other hand, if chosen too high, the overall resolution suffers because several isolated free energy minima may not contain the required number of frames and are consequently merged. However, the latter can be tested and used deliberately if one desires a more coarse grained model since geometrically close, but different conformations are in that case lumped into a larger, more ambiguous state. Nevertheless, the amount of available data plays a decisive role and might prohibit a too detailed model due to low sampling of inter-state transitions. If one finds that the number of transitions between microstates becomes very scarce with an increasing number of microstates, it is reasonable to increase P_{min} in order to reduce the total number of microstates.

Geometrically close microstate are not always kinetically well connected which is why in this case the above mentioned approach is not necessarily useful. In this case, a better option for achieving a more coarse-grained model is clustering the data to a geometrically very detailed model first by choosing a low P_{min} and subsequently merge those microstates dynamically to macrostates [24, 49]. This procedure takes into consideration that geometrically close microstates can still be separated by high free energy barriers but are at the same time dynamically well connected to distant microstates.

Fig. 3.7 shows the number of resulting microstates as a function of P_{min}. The initially fast declining curve begins to slow down at a value of $P_{\text{min}} \approx 7 \cdot 10^{-5}\,\%$ indicating that fluctuations within one local minimum of the free energy are now clearly diminished. A second reduction of the rate of decline starts at about $P_{\text{min}} = 0.1\%$ and a narrow plateau can be seen (indicated by the grey-shaded area). We decided to set $P_{\text{min}} = 25000$ frames which corresponds to roughly 0.123% of the total population. This results in 58 microstates in total, a number which still allows to describe the allosteric transition in a sufficient level of detail while still providing sufficient inter-state transitions. If desired, the number of microstates is also large enough in order to allow dynamically lumping aiming for a more coarse-grained model.

As mentioned above, the lumping radius was chosen in such a way that ~95% of the points have at least one neighbour within the clustering radius, thus there must still be a rest of frames which are

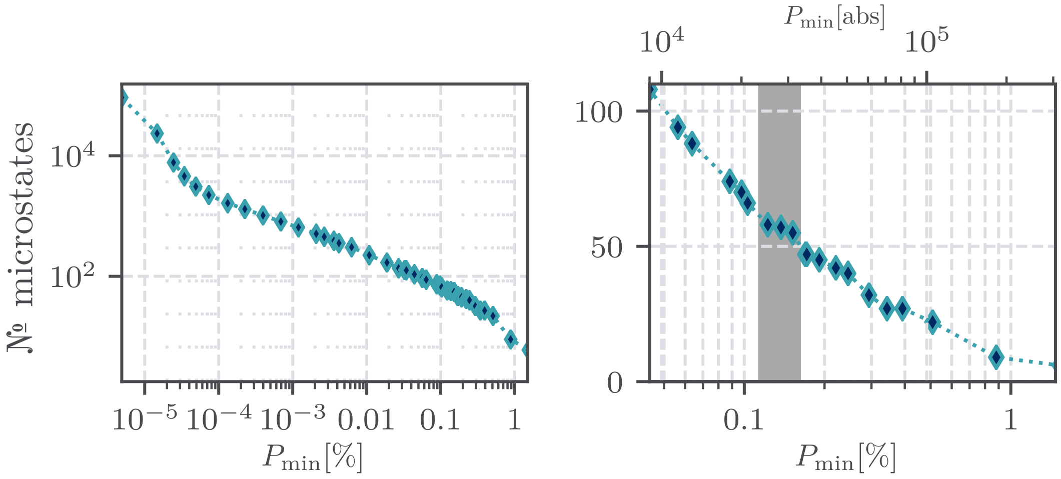

Figure 3.7.: *Left:* Number of microstates as a function of their minimal population P_{min}. *Right:* Excerpt of the area of interest from the left. ● The grey-shaded area indicates the plateau from which the final clustering was chosen.

completely isolated without a neighbouring frame within the clustering radius R resulting in a defective density estimation. Furthermore, completely isolated clusters incorporating frames in the range of $\sim 10^2$ are assigned by default to the geometrically closest microstate. In a subsequent step, these affected frames are first identified as "noise" and then dynamically reassigned to the prior microstate [26]. This affected 2.68% of all frames.

One of the 58 microstates was identified as a trap state as it was entered but not left until the end of the trajectory in which it appeared. With a resulting metastability of $T_{ii}(\tau_{\text{lag}}) = 1$, trap states are devastating in MSMs and this trap state was consequently assigned to the preceding microstate. Tab. 4.2 shows the resulting microstates and Fig. 4.8 illustrates the arrangement of the microstates within the free energy landscape along x_1 and x_2. However, both the table and the figure display microstates which have been subjected to a coring process, which is described in the following chapter.

The density-based clustering process yielded 57microstates in total after removing one microstate which was identified as a trap state.

4. Data Generation 1: MSM on PDZ2 Nonequilibrium Data

Torture numbers, and they'll confess to anything.

Gregg Easterbrook

In this chapter, we will construct a MSM on the clustered data, which was discussed in the last chapter. For this purpose, in a fist step a coring procedure is carried out to correct artifacts resulting from dimensionality reduction (see Sec. 4.1). On the basis of the cored microstate trajectory, the transition matrix is estimated which already represents a MSM. For a subsequent interpretation of the MSM, however, we must first classify the microstates in order to provide a deeper understanding of the mechanism involved (see Sec. 4.3). This includes, e.g., examining global properties such as e.g. to which conformation the system develops towards (Cis, Trans or nonequilibrium) or the investigation of the most important pathways (see Sec. 4.4).
The steps undertaken in this chapter consist of the steps 4–6 which are described in the workflow (see Sec. 2.7). A number of problems, especially in the coring process, have emerged in earlier works on PDZ2 [65, 66]. Section 4.1 therefore considers the coring and its impact on the microstate trajectory in detail and presents *iterative dynamical coring* as a remedy for artifacts resulting from too aggressive coring (see Sec. 4.2).

4.1. Dynamical Coring

After obtaining the noise-corrected microstate trajectory via density-based clustering in the last chapter, we apply dynamical coring in order to correct artifacts which have emerged from dimensionality reduction (see Sec. 2.4). Often, these artifacts manifest themselves in the fact that individual transitions are falsely perceived multiple times in form of spurious inter-state transitions (see Fig. 2.5). This will result in artificially increased transition probabilities which drastically reduce the metastability and, hence, the life time of the single microstates. By correcting them, we aim to eliminate these spurious transitions and only retain the actual transition which occurs according to the MD data. Without

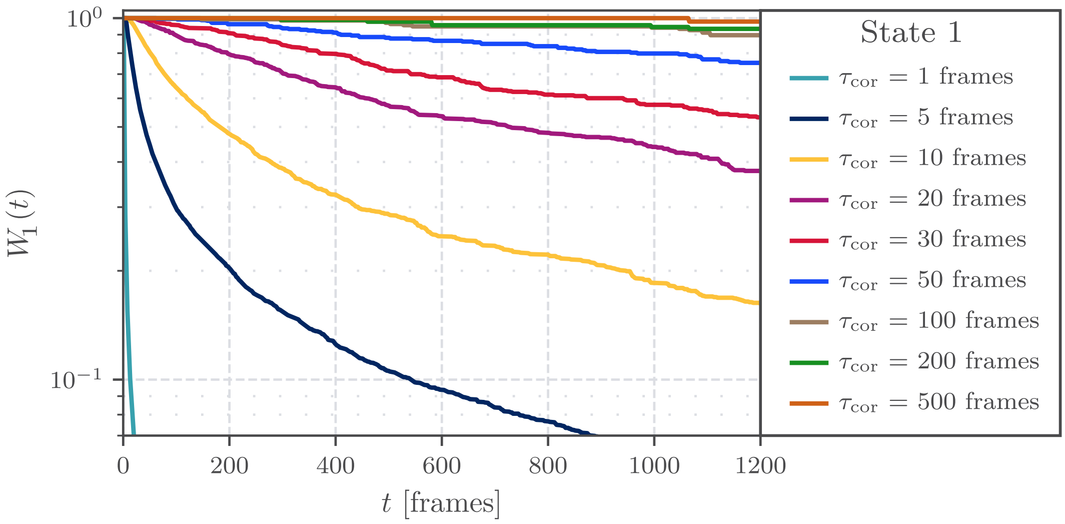

Figure 4.1.: Population probability $W_1(t)$ as a function of time t for different coring times τ_{cor}. In order to remove the strong initial decay, a coring time of $\tau_{\mathrm{cor}} \approx 30$ frames $= 0.6$ ns seems to be appropriate. Back-transitions are not included in this plot.

spurious inter-state transitions, and with a constant self-transition probability $P_{ii}(t)$, we expect a mono-exponential decay of the population probability over time (if backwards transitions into the microstate are forbidden).

We start by following the heuristic of Jain et. al [24] and investigate the probability $W_i(t)$ to stay in microstate i for at least the time t. This should give us a first idea about an appropriate coring time τ_{cor}. Figure 4.1 shows $W_1(t)$, which is a good example for the discussion of all microstates since $W_i(t)$ is very similar for all microstates. The strong initial drop, which is extremely distinctive for no or very little coring, is already removed for the most part by a coring time of $\tau_{\mathrm{cor}} = 20$ frames $= 0.4$ ns and practically completely diminished for $\tau_{\mathrm{cor}} = 30$ frames $= 0.6$ ns. Choosing $\tau_{\mathrm{cor}} = 0.6$ ns as the final coring time, we can examine the Markov property of the system by performing the Chapman-Kolmogorov test [see Eq. (2.20)].

The Chapman-Kolmogorov test for microstate 1 and a coring time of $\tau_{\mathrm{cor}} = 0.6$ ns is shown on the left side of Fig. 4.2 for different lag times τ_{lag}. The lag time τ_{lag} hereby specifies the width of the sliding window approach, i.e., transitions between $X_0 \rightarrow X_{\tau_{\mathrm{lag}}}$ are counted, then between $X_1 \rightarrow X_{\tau_{\mathrm{lag}+1}}$ and so on. All those transitions counted this way are then combined into one single transition count matrix which yields the transition matrix through normalization. By counting the transitions which happen in the single trajectories and combining them into one single transition matrix, we can predict long-term dynamics via multiple multiplications of the resulting matrix with itself. This provides predictions at intervals of the lag time τ_{lag}.

In contrast to the MSM predictions, a new transition matrix is estimated for every MD data point with

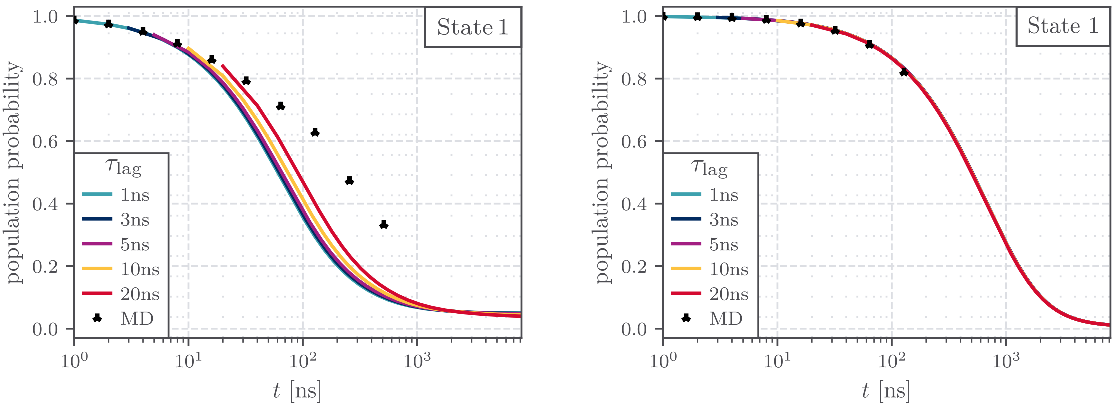

Figure 4.2.: Chapman-Kolmogorov test for microstate 1. *Left*: The trajectory after coring with a coring time of $\tau_{\text{cor}} = 30$ frames $= 0.6$ ns, which corresponds to the heuristic of the $W_1(t)$ test. *Right:* Trajectory after coring with a coring time of $\tau_{\text{cor}} = 500$ frames $= 10$ ns. The accordance of the MSM-predictions with the MD-predictions points is better compared to the coring time of $\tau_{\text{cor}} = 30$ frames. For technical reasons, the calculation of the MD data is suspended once the resulting transition matrix is not full dimensional anymore, i.e. the lag time is so long that single microstates are completely skipped by the sliding window.

the corresponding lag time τ_{lag}. Comparing both tells us whether the transition matrix, once estimated for the MSM, can predict the system's long-term dynamics by using short-term dynamics only.

One can see, that for all lag times a large discrepancy between the predictions from MD data and the MSM is present which demonstrated that the microstate partitioning does not behave Markov. A. Weber proposed to increase the coring time until a good agreement between the MSM- and the MD-predictions is provided [66]. This is the case for $\tau_{\text{cor}} \approx 500$ frames $= 10$ ns (see Fig. 4.2, *right side*). However, one can see that not only the MSM-, but also the MD-predictions are drastically altered which means that the coring process has a major impact on the microstate trajectory. 20.8% of the frames are changed when using a coring time $\tau_{\text{cor}} = 500$ frames compared to the uncored trajectory which represents a major intervention into the microstate trajectory. In order to understand how the coring mechanism is working and what might have caused these major differences between both, cored and uncored trajectory, let us consider the following example of a microstate trajectory and a fictive coring time of $\tau_{\text{cor}} = 10$ frames. Here and in the following, different microstates $i \in \Omega$ are represented by numbers.

$$\ldots, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, \underbrace{2, 2, 2, 2, 2, 2, 2, 2, \mathbf{4}, 2, 2, 2, 2, 2, 2}_{\text{Assigned to microstate 1}}, 3, 3, 3, 3, 3, 3, 3, \ldots$$

Even though the middle part of the trajectory underlined by the bracket clearly belongs to microstate 2, one small fluctuation (in form of microstate 4) is sufficient for the algorithm to not find a stable core in this region which leads to a complete discard of microstate 2. Consequently, the whole part is assigned to microstate 1 as it represents the last stable core.

This is indeed a huge problem in PDZ2. To illustrate this, it was found that altered sequences are often significantly longer than specified by the coring time τ_{cor} and extend up to 45.3 times the specified coring time τ_{cor} (see Fig. 4.4). On the one hand, that means that information is lost over the course of microseconds while on the other hand this leads simultaneously to spurious transitions in the transition matrix. Both of these problems significantly reduce the quality of the resulting MSM and make its predictive value questionable. Since the uncored microstate trajectory comprises $\sim 1.2 \cdot 10^6$ transitions in a total of $\sim 2 \cdot 10^7$ frames, it stands to reason that large parts of the trajectory are affected by the coring process.

> Applying the coring time suggested by the $W_i(t)$-test does not yield Markovian results in the Chapman-Kolmogorov tests. For significantly higher coring times, the Chapman-Kolmogorov test indicates Markovian behaviour, but heavily affects the microstate trajectory which makes an accurate representation of the MD data no longer possible.

4.2. Iterative Dynamical Coring

Iterative coring represents a remedy for above mentioned problems and works well for arbitrary long coring times. The basic idea is to iteratively core the trajectory in order to avoid that individual, incorrectly assigned frames prevent the formation of stable cores. Therefore we start with an initial coring time of $\tau_{\text{cor}} = 2$ frames in the first iteration which we then iteratively increase by one frame in each subsequent iteration until the final desired coring time is reached. Again, a simple example illustrates its effect. In the following example we assume a final coring time of $\tau_{\text{cor}} = 5$ frames.

$$
\begin{array}{rl}
 & \ldots, 5, 5, 5, 5, 5, \mathbf{2}, \mathbf{2}, \mathbf{2}, \underline{\mathbf{1}}, \mathbf{2}, \mathbf{2}, \mathbf{2}, \mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{1}, \underline{\mathbf{2}}, \mathbf{1}, \mathbf{1}, 3, 3, 3, 3, 3, \ldots \\
 & \qquad\qquad\qquad\qquad\quad \downarrow \underline{\tau_{\text{cor}} = 2} \\
 & \ldots, 5, 5, 5, 5, 5, \mathbf{2}, \mathbf{2}, \mathbf{2}, \underline{\mathbf{2}}, \mathbf{2}, \mathbf{2}, \mathbf{2}, \mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{1}, \underline{\mathbf{1}}, \mathbf{1}, \mathbf{1}, 3, 3, 3, 3, 3, \ldots \\
 & \qquad\qquad\qquad\qquad\quad \downarrow \underline{\tau_{\text{cor}} = 3, 4, 5} \\
\textbf{icor:} & \ldots, 5, 5, 5, 5, 5, \mathbf{2}, \mathbf{2}, \mathbf{2}, \mathbf{2}, \mathbf{2}, \mathbf{2}, \mathbf{2}, \mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{1}, 3, 3, 3, 3, 3, \ldots \\
\textbf{ccor:} & \ldots, 5, 5, 5, 5, 5, \mathbf{5}, \mathbf{5}, \mathbf{5}, \mathbf{5}, \mathbf{5}, \mathbf{5}, \mathbf{5}, \mathbf{5}, \mathbf{5}, \mathbf{5}, \mathbf{5}, \mathbf{5}, \mathbf{5}, \mathbf{5}, 3, 3, 3, 3, 3, \ldots
\end{array}
$$

Here and in the following we denote the iterative method by the abbreviation “icor” while “ccor” is used to indicate the classical method. For the sake of clarity, we often refer only to the coring time—in this case a coring time provided with the superscript i, i.e. $\tau^{\text{i}}_{\text{cor}}$ denotes iterative coring while $\tau^{\text{c}}_{\text{cor}}$ refers to the classical method. We see that the iterative coring method performs much better in this example. While the classical non-iterative method can only resolve two microstates in the curse of this example, the iterative method resolves all four microstates as desired.

Benchmarks

In order to verify whether iterative coring actually performs better on real data, a number of tests are carried out in the following. In the last section it was already suggested that classical coring alters sequences which are much longer than the specified coring time. In order to be able to estimate the impact of the different coring methods, the probability of each frame to be in a sequence of a certain length is plotted in Fig. 4.3 for multiple coring times ranging from no coring ($\tau_{\text{cor}} = 1$ frame) up to $\tau_{\text{cor}} = 600$ frames. We can use this test to judge the "reliability" of a coring method: If we find that coring alters mostly sequences which are significantly longer than the specified coring time τ_{cor} and leaves sequences with a similar range than the coring time τ_{cor} unaffected, we can state that the coring method does not work as desired.

Fig. 4.3 shows the distribution for iterative coring on top and for the classical coring method below. The grey solid line indicates the shortest sequence length possible (length sequence = τ_{cor}) for the respective coring time. One immediately sees that the iterative coring approach follows the grey line tightly corresponding to the desired behaviour while the classical coring method modifies the trajectory in such a way that sequences with the length of the coring time normally do not "survive" the coring process. Especially for longer coring times, the discrepancy between the shortest possible and the actual shortest sequence length is large for the classical coring method (note that the x-axis is plotted logarithmic).

Two other things are worth mentioning: If uncored, the by far most occurring sequences are shorter than 30 frames, clearly corresponding to spurious state fluctuations, which demonstrates the necessity for coring in general. Once the trajectory is cored however, the probability of a frame to be located within a sequence of length l is highest for sequences with than $l > 1834 \cdot 30 \approx 55000$ frames, independent of how long the coring time is. This is interesting as it suggests that the great majority of the trajectories consists of very long, stable sequences which are interrupted by shorter sequences possibly occupied by transient transition states. While the above provides a holistic view of the distribution of sequences in the data, we now shift the focus on those sequences which are altered by the two different coring algorithms. On the left side, Fig. 4.4 shows the number of changed sequences as a function of the sequence length. While for classical coring sequences up to a length of 45.3 times the coring time $\tau^{\text{c}}_{\text{cor}} = 500\,\text{frames} = 10\,\text{ns}$ are changed, which corresponds to 22654 frames, the longest altered sequence for iterative coring is only 1.5 times the coring time $\tau^{\text{i}}_{\text{cor}} = 500\,\text{frames}$ (corresponding to 747 frames). For the here displayed coring time of $\tau_{\text{cor}} = 500\,\text{frames}$, the shortest sequence is 500 frames for the iterative method and 585 frames for the classical approach.

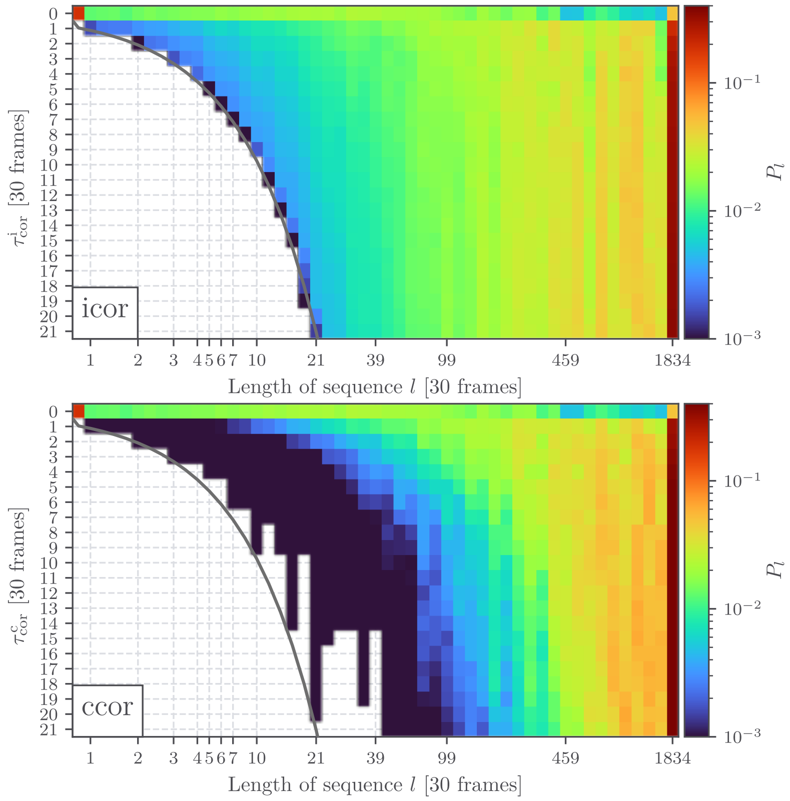

Figure 4.3.: *Top:* Iterative coring: The color indicates the probability P_l of a frame to be located in a sequence of the length l. ● The grey line indicates the shortest possible sequence length for the respective coring time τ_{cor}. *Bottom:* The same plot for the classically cored trajectory.

To better understand the plot on the right hand side of Fig. 4.4, we recall that the microstate trajectory is a projection of the high dimensional MD data onto a discrete, one-dimensional quantity. While the protein may already have significantly changed its conformation in the MD data, the microstate trajectory after coring may still pretend that little has changed since artifacts prohibit stable cores. We find that the microstate trajectory often fluctuates back and forth between two microstates which are kinetically closely connected and it could be that artifacts "freeze" the projection over an extended period of time by prohibiting stable cores while the MD data continues to describe conformational changes of the protein. If the MD data does not change significantly during this time, the resulting projection error would be comparatively small since MD data and microstate projection describe similar conformations of the protein. In this case we would most probably find that the cored trajectory agrees with the original, uncored trajectory shortly after the altered sequence is over as the uncored trajectory would fluctuate back into the state occupied in the cored trajectory due to their close dynamical connectivity. If, however, the MD data changes significantly while the microstate trajectory is still "frozen", the projection error would be large as these changes are not reflected within the frozen microstate trajectory.

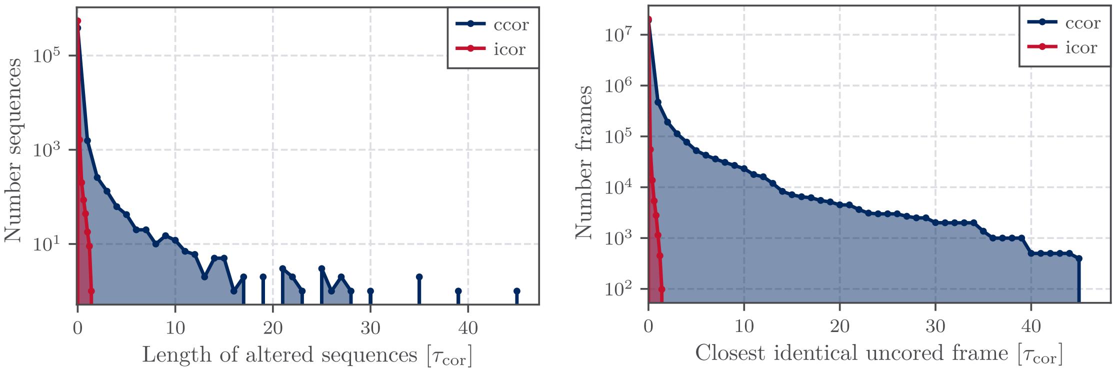

Figure 4.4.: $\tau_{\mathrm{cor}} = 500$ frames $= 10$ ns. *Left:* The length of altered sequences due to the different coring methods. *Right:* The distance between altered frames and the closest lying frame of the same state in the uncored trajectory.

In order to investigate this, the distance between an altered frame in the cored trajectory and a frame of the same state in the uncored trajectory is plotted (see Fig. 4.4, *right hand side*). If the above mentioned phenomenon applied, we would observe that distances to the closest identical uncored frame are only around half the length of the altered sequence since it would correspond to a $i \to j \to i$ transition. However, for our data, that is at least not the case for the longest altered sequences, as they can be individually identified by the plateaus in the plot at the right hand side, which agree in their length with the points in the plot on the left hand side. This means that the system already occupied another, kinetically not well connected, state in the MD data. As the distance between altered frames in the cored trajectory and the same state in the uncored trajectory is significantly higher for classical coring, the actual correct description from the MD data is substantially falsified over long periods of time for this coring method.

> All benchmarks underline the advantages of the iterative method over the classical coring approach as fluctuations are corrected before applying the final, desired coring time τ_{cor}. This ensures that trajectories are always cored reliable, no matter whether they are heavily affected by spurious transitions or not and makes *iterative dynamical coring* the better suited method for the correction of artifacts stemming from dimensionality reduction.

Applying Iterative Dynamical Coring

As the iterative coring method turned out superior, we repeat the $W_i(t)$-test in order to estimate a proper coring time $\tau^{\mathrm{i}}_{\mathrm{cor}}$. The plot can be seen in Fig. A.2 (Appendix on page 86). It is evident that

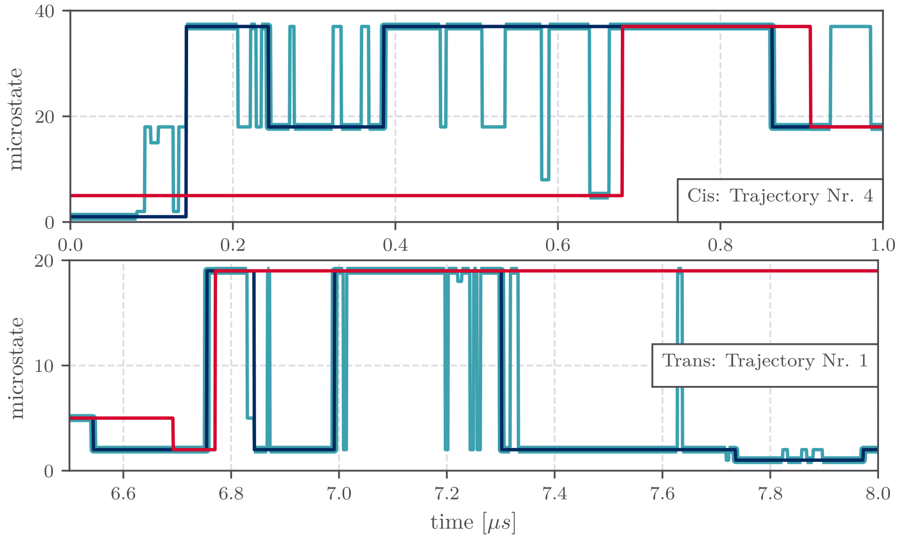

Figure 4.5.: Comparison of time traces of two trajectory excerpts (*top:* Fourth Cis trajectory, *bottom:* first Trans trajectory). ● the light blue line indicates the reference trajectory ($\tau^{\mathrm{i}}_{\mathrm{cor}} = 200$ frames), ● the dark blue line indicates the long iteratively cored trajectory ($\tau^{\mathrm{i}}_{\mathrm{cor}} = 3000$ frames), while the ● red line refers to the classically cored trajectory ($\tau^{\mathrm{c}}_{\mathrm{cor}} = 500$ frames).

now a coring time of $\tau^{\mathrm{i}}_{\mathrm{cor}} \approx 200$ frames must be applied instead of the prior $\tau^{\mathrm{c}}_{\mathrm{cor}} \approx 30$ frames for the classical method. This is remarkable but fits with the observations made in Fig. 4.4. Most probably, the difference stems from spurious assignments of frames to stable cores which consequently increases the metastability of the corresponding microstates when cored classically. As this effect hardly occurs for iterative coring, longer coring times are needed in order to obtain similar metastability.

We remember that coring eliminates all dynamics happening on timescales shorter than the applied coring time τ_{cor} regardless of the method used. So the question arises whether information on a very short timescale (30 frames $\leq t \leq$ 200 frames) might get lost through the application of higher coring times for the iterative method ($\tau^{\mathrm{i}}_{\mathrm{lag}} \approx 200$ frames) compared to shorter coring times for the classical method ($\tau^{\mathrm{c}}_{\mathrm{lag}} \approx 30$ frames). Yet, when we investigate the probability of a frame being located within a sequence which is shorter than $l = 210$ frames for the classically cored trajectory, we find it to be only 0.4 %, which is negligible compared to the benefits coming along with the iterative coring method (see Fig. 4.3).

Just as before, the trajectory needs to be cored longer in order to achieve good agreement between MSM and MD in the Chapman-Kolmogorov test. It turned out that a coring time of $\tau^{\mathrm{i}}_{\mathrm{cor}} = 3000$ frames = 60 ns results in high Markovianity. This coring time is significantly longer than the selected classical coring time of $\tau^{\mathrm{c}}_{\mathrm{cor}} = 500$ frames, but actually more transitions are preserved (799 vs. 473) and the overlap between the cored trajectories and a reference trajectory, which is cored with $\tau^{\mathrm{i}}_{\mathrm{cor}} = 200$ frames, is higher for iteratively cored trajectory, even though the coring time is much longer (see Tab. 4.1).

Table 4.1.: Comparison of the total transitions between the iteratively and classically cored model. For the agreement, the number of identical frames of both trajectories with the reference trajectory $\tau^{\mathrm{i}}_{\mathrm{cor}} = 200\,\mathrm{frames}$ were compared.

	Transitions [#]	Agreement [%]
ccor ($\tau^{\mathrm{c}}_{\mathrm{cor}} = 500\,\mathrm{frames}$)	473	83.01
icor ($\tau^{\mathrm{i}}_{\mathrm{cor}} = 3000\,\mathrm{frames}$)	799	93.71

Figure 4.5 shows two exemplary excerpts of different trajectories which demonstrate that the iteratively cored trajectory, even though the coring time is six times larger, follows the timetrace of the reference trajectory much closer than the classically cored trajectory. Both examples show that the classically cored trajectory occupies spurious microstates over a time interval of several of hundreds of nanoseconds which causes the microstate trajectory to loose track of the actual dynamics of the protein captured in the MD data. Differently, the iteratively cored trajectory represents a more “coarse-grained” version of the reference trajectory, efficiently eliminating highly frequent fluctuations between microstates.

It is necessary to core the reference trajectory as well, as the representation would be very fuzzy otherwise. This explains why the classically cored trajectory is e.g. not able to find a stable core between 7.4 μs and 7.6 μs, even though the reference clearly inhibits a well defined plateau there. Sticking with a coring time of $\tau^{\mathrm{i}}_{\mathrm{cor}} = 3000\,\mathrm{frames}$, one mirostate was identified as a trap state due to reassignments during the coring process, effectively reducing the number of microstates to 56.

Figure 4.6 shows the Chapman-Kolmogorov test for two exemplary microstates. The highly populated microstate 1 (*left*) provides a comparison of the iteratively with the classically cored trajectory (see Fig. 4.2) whereas a lower populated microstate 43 (*right*) serves as an example for microstates which do not score as well in the Chapman-Kolmogorov test. The agreement between MSM- and MD-predictions is excellent for microstate 1 and the MD reference is by far not so much altered as it was for the classical coring with $\tau^{\mathrm{c}}_{\mathrm{cor}} = 500\,\mathrm{frames}$. In contrast, the agreement between MSM- and MD-predictions starts to differ at $t \approx 2 \cdot 10^2$ ns for microstate 43. In order to understand this phenomenon, the transitions in and out of microstate 43 as a function of the lagtime τ_{lag} are plotted in the lower part of Fig. 4.6. One can see that as soon as the predictions of both, MD and MSM, start to deviate strongly from each other, the number of transitions drops rapidly. It stands to reason that a decrease of transitions implies that short sequences at the end of a single trajectory are not registered anymore by the sliding window approach and that primarily long sequences remain which then leads to an increased metastability of the MD predictions. This is an artifact stemming from working with multiple short trajectories instead of one large one as the sliding window approach leads to a significant loss in data, mostly affecting short sequences for long lag times τ_{lag}.

Additionally, the appearance of the microstates in the different trajectories may also play a role. Microstate 43 for example appears in seven trajectories, of which five are short (1.1 μs). Microstate 1 in

contrast appears in 45 different trajectories and 19 of them are long trajectories (10 μs). Since the length of single sequences in short trajectories is more limited to the upper end owed to the comparatively short length of the trajectories, the MD predictions for microstates with an imbalanced occurrence in short and long trajectories tend to overestimate the metastability of a microstate for long lag times. This assumption is further supported by the fact that pure Cis or Trans microstates generally perform well in the Chapman-Kolmogorov tests while only appearing in equally long 10 μs trajectories.
Consequently, the MD data reference in Chapman-Kolmogorov tests for microstates which feature a similar course of their transitions as a function of the lag time as microstate 43 and which do occur in both, short and long trajectories, might not be reliable for all lag times. As more and more data—predominately short sequences—is ignored and largely long sequences remain, the life time for microstates in MD is consequently overestimated for long lag times.

> The trajectory was cored using iterative coring and a coring time of $\tau^{\mathrm{i}}_{\mathrm{cor}} = 3000$ frames = 60 ns which behaved more reliable compared to the practice of coring with $\tau^{\mathrm{c}}_{\mathrm{cor}} = 500$ frames.
> Besides, it was illustrated that working with multiple trajectories cause problems as
>
> 1. the application of the sliding window approach leads to a considerable loss in data.
> 2. they could falsify the MD reference in the Chapman-Kolmogorov test if they are of different length.

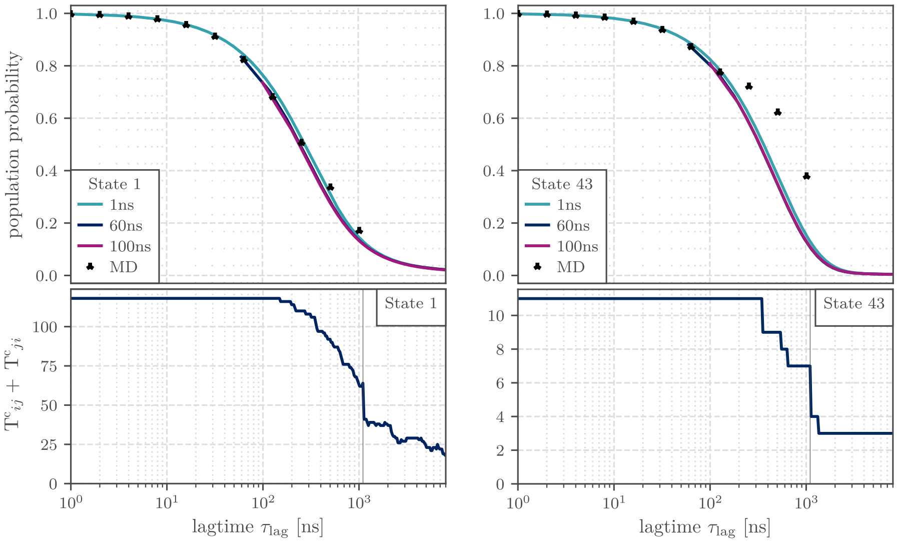

Figure 4.6.: *Top:* Chapman-Kolmogorov tests for the microstate 1 (*left*) and 43 (*right*). *Bottom:* Transitions [in and out of a microstate, see Eq. (2.15)] of the trajectories as a function of the lagtime τ_{lag}. The small grey line indicates the end of the 80 short nonequilibrium trajectories.

4.3. Classification

In the last section we obtained the cored microstate trajectory which is the basis for the following construction of the MSM. It is instructive to classify the microstates into different categories in order to facilitate the interpretation. As the ultimate goal is to describe the Cis→Trans transition, those two conformations should represent the first two categories. Additionally, microstates containing both Cis and Trans frames are expected to occur as there is a small region in the free energy landscape in which both conformations overlap—these microstates are called ambivalent microstates. Finally we need states which are neither Cis nor Trans but describe conformations which are occupied in the course of the transition. Those states are referred to as nonequilibrium microstates.

It is apparent to use the six Cis trajectories for the classification of the Cis microstates, the six Trans trajectories to classify Trans microstates and the 92 nonequilibrium trajectories for the classification of the nonequilibrium microstates. However, an examination of the equilibrium trajectories shows that they are not at all yet equilibrated but instead still undergo substantial conformational changes. Figure 4.7 shows the equilibrium time evolution of three distances covering different regions of the protein revealing a high overlap in the Cis and Trans conformation in the first ~2–3 μs. As the evolution of the distances slowly levels out afterwards, we consider only the last 7 μs as equilibrium and reassign the first three microseconds as nonequilibrium. This procedure turned out to be very valuable as it reduced the number of ambivalent microstates from 10 to 2, indicating a superior separation between Cis and Trans microstates. Nevertheless, We must point out that this consequently would mean that the PCA (see Sec. 3.3) would have to be redone without the first three microseconds of the equilibrium trajectories. However, as this procedure is very time consuming due to the subsequent clustering and coring studies, this was not done for data generation 1, but only for the second generation (see Ch. 5).

Classification scheme · Due to the four timers higher number of nonequilibrium frames compared to equilibrium frames (16.200.100 frames vs. 4.000.012 frames), we require nonequilibrium microstates to consist at least of 95% nonequilibrium frames. The high percentage of 95% is addition-

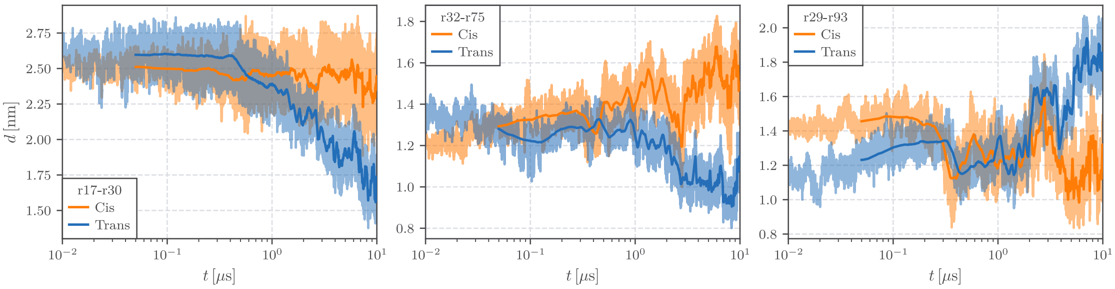

Figure 4.7.: Three exemplary distances between the residues 17and 30 (*left*), 32 and 75 (*middle*) and 29and 93 (*right*). ● orange denotes the distance in the Cis trajectories, and ● blue in the Trans trajectories. The solid lines represents a symmetric running mean over 500 frames.

ally justified because nonequilibrium frames usually dominate the microstate composition as they can, besides their high number, additionally adopt equilibrium conformations as well when the trajectory either starts in a Cis microstate or describes the arrival at a Trans microstate. All other microstates, which were not yet classified as nonequilibrium, are classified as Cis (Trans) if they feature at least 2% Cis (Trans) frames and less than 2% Trans (Cis) frames. In the case that both, Cis and Trans, are represented with more than 2% of the frames, the microstate is classified as ambivalent. This classification was primarily introduced by in Ref. [66] and is adopted here to be able to compare models. As a matter of fact, the classification scheme works well in the sense that a great majority of the microstates are robust in their classification, i.e. small changes in the classification criteria would not result in a significant change of the classification.

Figure 4.8 shows the free energy landscape projection on the first two PCs x_1 and x_2. The color of the box around the microstate's number indicates the classification. A deep color indicates pure microstates, i.e. they feature a small percentage of nonequilibrium frames. In contrast, pale colors indicate that they are dominated by nonequilibrium frames. When possible, we will stick with this continuous color scale, as it allows more precise statements about the composition of the microstates. The corresponding table, which contains the most important information on all microstates, can be found in Tab. 4.2.

The observation made in Fig. 3.6 that particularly x_1 splits Cis and Trans is supported by the arrangement of the microstates along x_1. Trans microstates are exclusively located at negative values of x_1 while Cis microstates are primarily characterized by positive values. In total, 16 Cis , 12 Trans , 26 nonequilibrium and 2 ambivalent microstates were identified.

For the purpose of illustrating the quality of the resulting microstates, 100 randomly selected overlays of the protein's conformation in six microstates 4, 22, 34, 39, 41 and 48 are exemplary shown in order to depict nonequilibrium, ambivalent, Cis and Trans microstates. It is notable that the different microstates exhibit varying levels of structural variability. So, for example the frames of the Trans microstate 48 are almost perfectly aligned in all regions of the protein while the Cis microstate 41 shows deviations in the $\beta_5\alpha_2$-loop. The frames of the third equilibrium microstate depicted, Cis microstate 34, are well aligned in all metastable structures (α-helices in red and β-sheets in blue) but the loop regions are more disordered compared to the other two. One difference that immediately stands out is the difference in the opening distance of the binding pocket between the α_2-helix and the β_2-sheet (see Fig. 2.1); the azobenzene photoswitch forces both structures to move apart from each other in the Trans state. Interestingly, one sees for the ambivalent microstate 39 that only the lower part of the β_2-sheet, where the azobenzene switch is located, is forced away from the α_2-helix, while the upper part lies close to the α_2-helix. This demonstrates what constitutes an ambivalent microstate and suggests that they might be an artifact stemming from stress within the protein induced by the azobenzene photoswitch. We can further speculate that deformation in the lower part of the α_2-helix is a result of the stress as well. For nonequilibrium microstates, the variance within the overlays is usually higher as almost all parts of the protein are strongly unraveled. In contrast to all other microstates, even some of the β-sheets which are usually tightly aligned now appear to be disordered. After all, this is however not really surprising as the nonequilibrium microstates are expected to amalgamate those microstates which are occupied by the system in its very dynamical phase of the Cis→Trans transition. However, it is important to note that microstates with higher microstate-numbers naturally tend to become tighter aligned as they contain less frames, which we e.g. can observe for microstate 4 in comparison to 22.

> Classifying the microstates resulted in 16 Cis, 12 Trans, 26 nonequilibrium and 2 ambivalent microstates. Removing the first three microseconds of the equilibrium trajectories from the Cis/Trans classification greatly reduced the number of classified ambivalent microstates from 10 down to 2. This is crucial since it is not clear what these constitute, apart from artifacts issued by the azobenzene photoswitch.

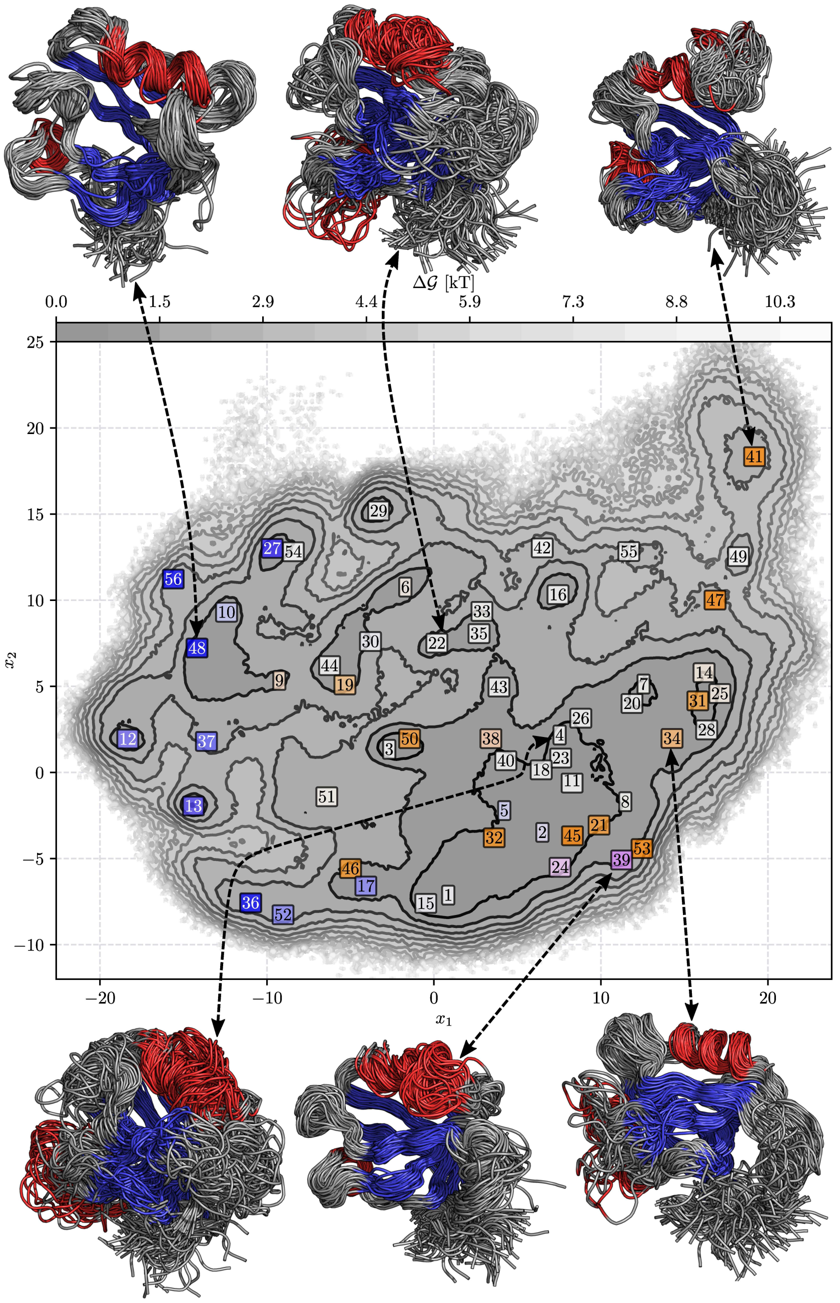

Figure 4.8.: Free energy landscape of the PDZ2 domain along its first two principal components. The microstate trajectory was cored with a coring time of $\tau^{\mathrm{i}}_{\mathrm{cor}} = 3000$ frames = 60 ns. ● Blue and ● orange microstates correspond to pure Trans or Cis microstates, respectively. The paler they get, the higher their Neq share is, which is why ○ white corresponds to pure Neq microstates. ● Pink microstates (e.g. microstate 24 and 39) denote ambivalent microstates, which feature both Trans and Cis frames. Overlays of 100 randomly selected frames are shown for 6 representative microstates.

Table 4.2.: Population of states with a final coring time of $\tau_{\mathrm{cor}}^{i} = 3000$ frames. States were ordered according to their population in the trajectory after coring. Coloring: ● Cis, ● Trans, ○ Neq and ● ambivalent. $\sum_j^i$ denotes the cumulative population up to the state i. “in” and “out” represent the number of inwards or outwards transitions.

MS_i	Pop [%]	$\sum_j^i \mathrm{Pop}_j$	% Neq	% Cis	% Trans	in	out	MS_i	Pop [%]	$\sum_j^i \mathrm{Pop}_j$	% Neq	% Cis	% Trans	in	out
1	6.80	6.8	99.1	0.0	0.9	42	76	29	1.37	79.3	100.0	0.0	0.0	4	4
2	5.34	12.1	89.1	1.1	9.8	68	67	30	1.34	80.6	98.8	0.0	1.2	9	7
3	4.65	16.8	100.0	0.0	0.0	20	17	31	1.32	81.9	26.6	73.4	0.0	15	14
4	4.21	21.0	99.3	0.7	0.0	43	38	32	1.22	83.2	9.7	90.3	0.0	9	9
5	4.19	25.2	85.0	0.0	15.0	64	65	33	1.10	84.3	98.6	1.4	0.0	7	14
6	4.15	29.3	93.4	6.6	0.0	21	13	34	1.09	85.4	43.2	56.8	0.0	11	12
7	4.13	33.5	98.7	1.3	0.0	37	34	35	1.09	86.4	100.0	0.0	0.0	6	5
8	3.34	36.8	98.0	2.0	0.0	12	16	36	1.00	87.4	0.0	0.0	100.0	1	1
9	3.22	40.0	80.7	18.0	1.3	32	26	37	1.00	88.4	37.9	0.0	62.1	7	6
10	3.07	43.1	81.9	0.0	18.1	14	10	38	0.98	89.4	69.5	28.6	1.9	8	12
11	2.62	45.7	100.0	0.0	0.0	36	33	39	0.97	90.4	54.8	13.6	31.6	7	9
12	2.54	48.2	44.6	0.0	55.4	6	2	40	0.87	91.3	100.0	0.0	0.0	5	4
13	2.51	50.8	31.8	0.0	68.2	11	12	41	0.85	92.1	0.0	100.0	0.0	2	2
14	2.37	53.1	90.4	9.6	0.0	18	20	42	0.78	92.9	100.0	0.0	0.0	3	3
15	2.37	55.5	98.5	0.0	1.5	20	19	43	0.77	93.7	100.0	0.0	0.0	5	6
16	2.37	57.9	100.0	0.0	0.0	14	14	44	0.69	94.4	100.0	0.0	0.0	5	4
17	2.19	60.1	55.5	0.0	44.5	16	16	45	0.66	95.0	0.0	100.0	0.0	5	4
18	1.81	61.9	100.0	0.0	0.0	29	27	46	0.63	95.6	15.8	84.2	0.0	6	6
19	1.77	63.6	52.3	47.7	0.0	14	12	47	0.56	96.2	0.0	100.0	0.0	4	3
20	1.77	65.4	100.0	0.0	0.0	14	14	48	0.54	96.7	0.0	0.0	100.0	1	1
21	1.73	67.1	20.2	79.8	0.0	18	19	49	0.46	97.2	100.0	0.0	0.0	2	2
22	1.66	68.8	100.0	0.0	0.0	16	15	50	0.46	97.7	19.4	80.6	0.0	3	2
23	1.59	70.4	100.0	0.0	0.0	19	18	51	0.44	98.1	96.3	3.7	0.0	3	3
24	1.57	72.0	81.3	7.0	11.7	13	16	52	0.44	98.5	53.9	0.0	46.1	3	3
25	1.56	73.5	92.0	8.0	0.0	19	18	53	0.43	99.0	0.0	100.0	0.0	4	3
26	1.55	75.1	100.0	0.0	0.0	12	12	54	0.43	99.4	100.0	0.0	0.0	6	5
27	1.43	76.5	16.7	0.0	83.3	3	3	55	0.32	99.7	100.0	0.0	0.0	2	2
28	1.40	77.9	100.0	0.0	0.0	21	17	56	0.28	100.0	0.0	0.0	100.0	1	1

4.4. Constructing and Interpreting the MSM

We have now completed all the preliminary steps (steps 1–4 in the workflow, see Sec. 2.7) and perform step 5, which is the construction of the MSM. Beforehand, there are a few things which need to be considered:

- *The concatenation limits*: As we are working with 112 trajectories, the algorithm for determining the transition count matrix must work in such a way that spurious transitions introduced by the concatenation of the end and the start of another trajectory are deducted. This might sound obvious, but is worth to be mentioned, as the introduction of 111 completely random transitions into a total of 799 transitions (see Tab. 4.1) would significantly undermine the validity of the MSM.

- *The lag time τ_{lag}:* We need to chose an appropriate lag time τ_{lag} to define the length of the sliding window. As all dynamics have already been removed on timescales smaller than $\tau_{\mathrm{cor}}^{\mathrm{i}} = 3000\,\mathrm{frames} = 60\,\mathrm{ns}$, lag times smaller than τ_{cor} do not make sense. On the other side, the lag time should not be chosen too large in order to maximize the temporal resolution, but must be large enough that the implied timescales are converged.
  Figure 4.9 shows the first implied timescale on the left hand side and one can see that they are not yet completely converged but still increase slightly, which is a typical behaviour for systems of biological interest. Since the Chapman-Kolmogorov test yield good results for a lag time of $\tau_{\mathrm{lag}} = 60\,\mathrm{ns}$, we apply this lag time.

- *(Non-)reversibility:* In equilibrium, it is often common for Markov state modeling to demand "detailed balance" or "reversibility" (compare Sec. 2.6), which should prevent the permanent production of work. We however aim to describe a nonequilibrium allosteric transition which was triggered by applying considerable external stress through the conformational change of the azobenzene photoswitch. This might allow certain inter-state transitions to occur only in one direction. As we do not want to introduce spurious backward transitions into the transition matrix, we therefore refrain from enforcing detailed balance.

Under consideration of the above mentioned points, a non-reversible MSM with a lag time of $\tau_{\mathrm{lag}} = 60\,\mathrm{ns}$ is built. The implied timescales and the corresponding transition matrix can be found in Fig. 4.9. One can see that the implied timescales [see Eq. (2.21)] are nearly constant in a time interval up to $\tau_{\mathrm{lag}} = 100\,\mathrm{ns}$ before they increase. This is most likely due to coring. As more and more intermediate states are skipped by an increasing sliding window, the implied timescales indicate that processes are becoming slower.

For the first three implied timescales we get for the selected lag time of 60 ns:

	t_1	t_2	t_3
Implied timescale [μs]	10.02	5.85	3.75

To better understand those values and which processes they describe, we will consider the corresponding eigenvectors shown in Fig. 4.11. On the very top, the population of the microstates found in the data is shown as a reference. Below the stationary distribution as it is predicted by the MSM is show in descending order, the first eigenvector and the second eigenvector belonging to the first and second implied timescale. The colors indicate, as before, the classification of the microstates (deep orange or blue correspond to pure Cis or Trans microstates respectively, while a paler color represents nonequilibrium microstates). We can clearly see a difference between the population in the data and the MSM predictions. Particularly the Trans microstates are more highly populated in the MSM while the Cis and the nonequilibrium microstates play a smaller role, especially compared to the MD data. This is a hint that the MSM indeed describes the allosteric transition, as one expects that the azobenzene photoswitch drives the protein in the Trans conformation.

For a better evaluation of the first two eigenvectors associated with the first two implied time scales, we use a graphical representation of the transition matrix, the so called network-representation, which is depicted in Fig. 4.10. The arrangement of network nodes, which represent the microstates, was optimized by the ForceAtlas2-algorithm [70]. In an iterative process the nodes are rearranged as a repulsive force is categorically ascribed to all nodes. Counteracting this is an attractive force resulting from the entries of the transition matrix, which can be calculated by the product of the matrix entry and the stationary distribution $P_i^{\mathrm{eq}} T_{ij}$. Shortly speaking, the connections between the nodes, called edges, are acting as springs which pull the repelling nodes into their final arrangement.

Regarding Fig. 4.10, two smaller clusters of strongly interacting Cis microstates are visible and we

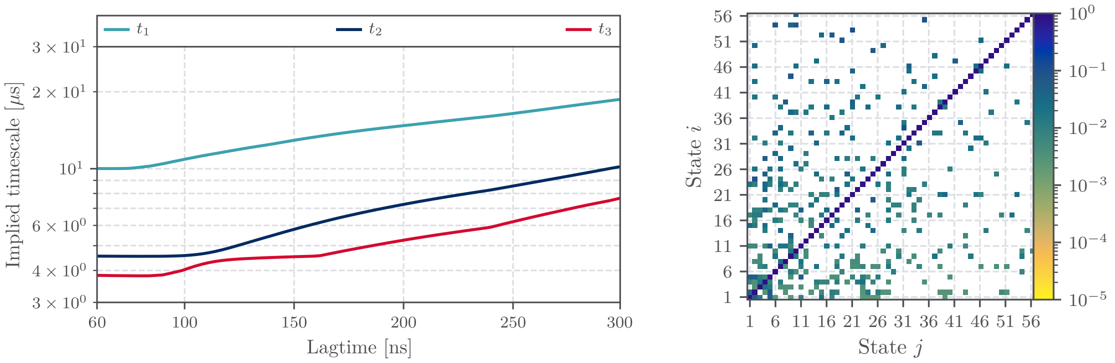

Figure 4.9.: *Left:* The first three implied timescales t_1–t_3 of the PDZ2-MSM based on equilibrium and nonequilibrium data for $\tau_{\mathrm{lag}} = 60$ ns. *Right:* The transition matrix of the corresponding MSM (line-normalized).

notice that one of them (around Cis microstate 45) is kinetically closer connected to the Trans cluster located on the right side of the network. About the Trans microstates, we can say that they are mostly located within one cluster which is strongly interconnected with the exception of the isolated cluster of the microstates 17, 36 and 52.

The majority of the nonequilibrium microstates is located between the second Cis cluster containing the microstates 31, 34, 41 and 47 and the Trans clusters. Only a few nonequilibrium microstates are located between the other Cis cluster and the Trans cluster. It stands to reason that the transitions from the Cis microstates 32, 45 and 46 to the Trans microstates, as described in the first eigenvector, are much more effective and happen more frequently due to their better kinetic connection.

The fact that most of the nonequilibrium microstates are located between the second Cis cluster and Trans microstates could have several reasons. It may be that the pathways starting from the Cis microstates of the second cluster are extremely ineffective as they include multiple highly disordered nonequilibrium microstates, while the pathways starting from the other Cis cluster around microstate 45 are the dominating ones, because their intermediate nonequilibrium microstates are less disturbed. Another reason could be that the projection of the nonequilibrium trajectories onto the eigenvectors of the equilibrium PCA delivers erroneous results as the first 3 μs of the equilibrium trajectories are not yet equilibrated. In Ch. 5, a PCA only on the remaining 7μs is performed in order to investigate whether this could be an artifact originating from an inaccurate PCA.

A third reason might be that the simulated nonequilibrium trajectories are biased due to incorrect simulation seeds (the Cis trajectories, which serve as simulation seeds, were not yet equilibrated). In fact, an analysis of the nonequilibrium trajectories yielded that in this projection from the PCA only 21 of 100 nonequilibrium trajectories increase their share of Trans from the first to the last frame. Therefore, we compare the microstates at the beginning and end of the trajectory and check to what percentage they consist of Cis or Trans frames, we also call this "transness". While only 17.5 % of the short nonequilibrium trajectories increase their transness, at least 35 % of the long nonequilibrium trajectories do. This indicates that the length of the nonequilibrium trajectories might also play a decisive role, and that they might still be too short for effectively describing the allosteric transition.

Coming back to the first eigenvector, we see in Fig. 4.11 that this eigenvector describes the flux from all parts of the network towards the tightly packed cluster in the upper right, which mainly consists of Trans microstates. This means, the process happening on the slowest timescale ($t_1 = 10\,\mu$s) describes the Cis→Trans transition. We must, however, mention the Trans microstates 17, 36 and 52, which do not belong to this cluster. The first two of them are also very notable in the second eigenvector as they contribute with large values. It looks like this eigenvector describes a drain, mainly from these three microstates and Trans microstate 27 towards Trans microstate 12, indicating that these three microstates and microstate 27 are relatively unstable. This is also consistent with the stationary distribution predicted by the MSM.

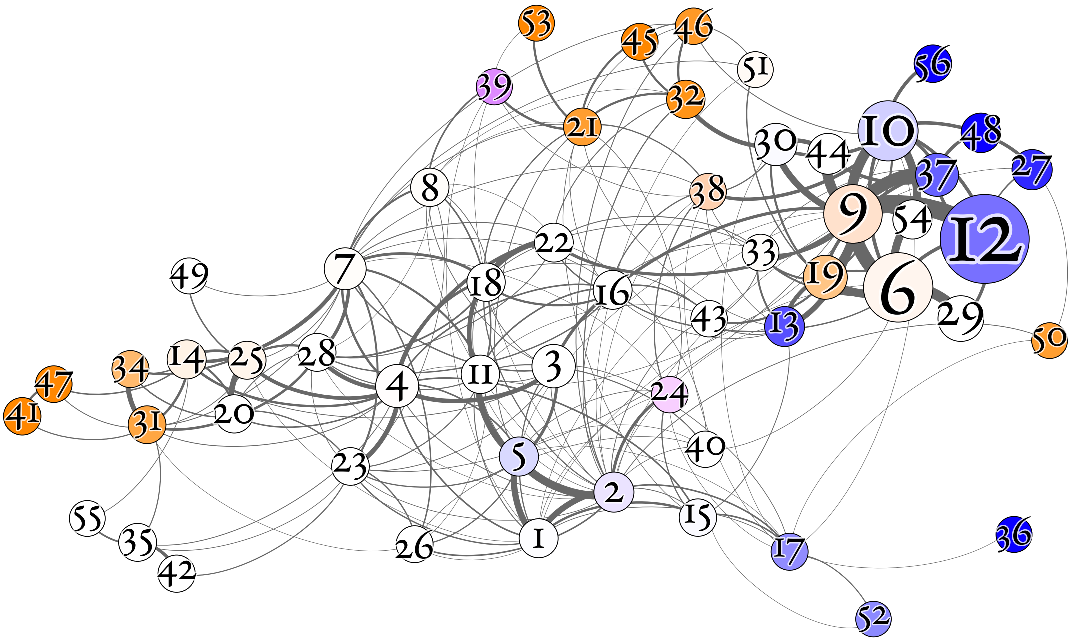

FIGURE 4.10.: Network representation of the $\tau^{\mathrm{i}}_{\mathrm{cor}} = 60\,\mathrm{ns}$ MSM. The edges (connections between the nodes) are weighted by $P^{\mathrm{eq}}_{ij}T_{ij}$ and the size of the nodes (circles indicating the microstates) is proportional to their stationary distribution. Cis microstates are marked in orange ●, Trans microstates in blue ●, Neq in white ○ and the ambivalent microstates in violet ●.

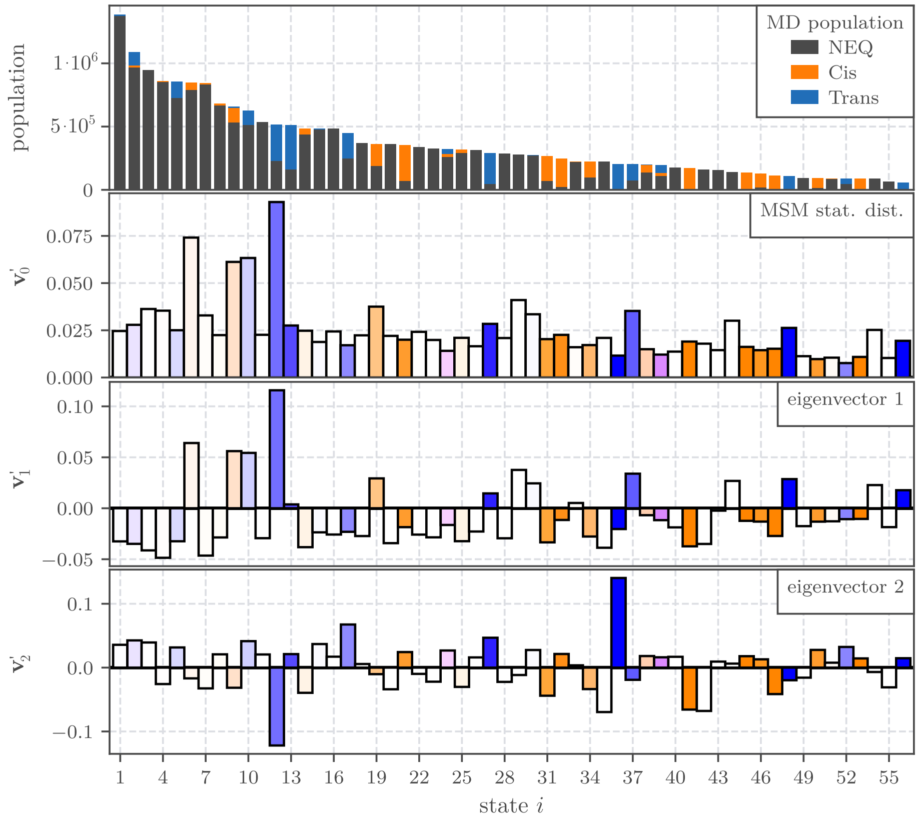

FIGURE 4.11.: *Top:* The population of the microstates found in the data. We define $v'_j \equiv \mathrm{sgn}(v_{j,i})\sqrt{|v_{j,i}|}$. *Upper middle:* The stationary distribution predicted by the MSM. *Lower middle:* The first eigenvector associated with the first implied timescale of 10.02 μs. *Bottom:* The second eigenvector with a corresponding implied timescale of 5.85 μs. The colors represent the usual scheme from Tab. 4.2.

Global Dynamics

Before we continue to discuss the dynamic in the system in detail in the next subsection, mostly by referring to Markov chain Monte Carlo (MCMC) pathways predictions, we shift the focus on the global dynamics of the system. As we classified all microstates as Cis, Trans, ambivalent (Amb) or nonequilibrium (Neq) in Sec. 4.3, it is now possible to describe "global dynamics" using the MSM predictions about the evolution of the system. For this purpose we consider the initial conditions given through the initial population in the nonequilibrium data and combine them in a normalized state vector, i.e. a population vector, which we subsequently propagate by multiplying it with the transition matrix. Thus, the shares in the various conformations can be tracked over time, which is shown on the right hand side of Fig. 4.12. To better understand these findings, the plot on the left of Fig. 4.12 shows the temporal development in the nonequilibrium trajectories.

The small difference between MSM-predictions and nonequilibrium data in the initial configuration results from the first matrix multiplication but the MSM is robust in that a randomly chosen, but normalized initial vector only affects the first few hundreds nanoseconds until the differences level off. We see that the MSM approximately reproduces the dynamics described in the nonequilibrium trajectories till ~2 μs, but then continues to predict dynamics up to ~60μs which correspond to six times the length of the simulated MD data. Even changes on shorter timescales, such as the temporally higher values for Cis compared to Trans in the range of ~3–6 μs are resolved by the MSM. In contrast to the trend visible in the nonequilibrium data, the nonequilibrium share in the MSM constantly decreases with simultaneous increase of the equilibrium shares with Trans ending up as the dominating conformation.

Concerning the nonequilibrium trajectories, we notice that they barely start in Cis microstates and keep the Cis level more or less in the course of the trajectories. For Trans, the picture is similar with a minimal trend towards a rise in the Trans share towards the end of the trajectories. For the nonequilibrium trajectories we can observe an opposite behaviour, i.e. a slow reduction towards the end of the trajectories. It should be mentioned here that all trajectories only show the expected trend at their end for times $t \geq 8$ μs. Recalling the problem of overlap between Cis and Trans mentioned in Sec. 4.3, it might be that the yet not equilibrated equilibrium trajectories serve as a biased simulation seed for the nonequilibrium trajectories. For many distances, we found heavy overlap between the Cis and Trans conformations within the first ~3 μs of the equilibrium simulations which could explain why we see a higher Trans than Cis share at the beginning of the nonequilibrium trajectories. Alongside the above mentioned problem that the trajectories show the expected trends only towards their very ends, which indicates again that they mostly have not reached Trans conformations yet, the validity of the nonequilibrium trajectories seems to be limited by suboptimal simulation seeds and length. After all, it is remarkable that the MSM still predicts the right trend for Trans, nonequilibrium and ambivalent.

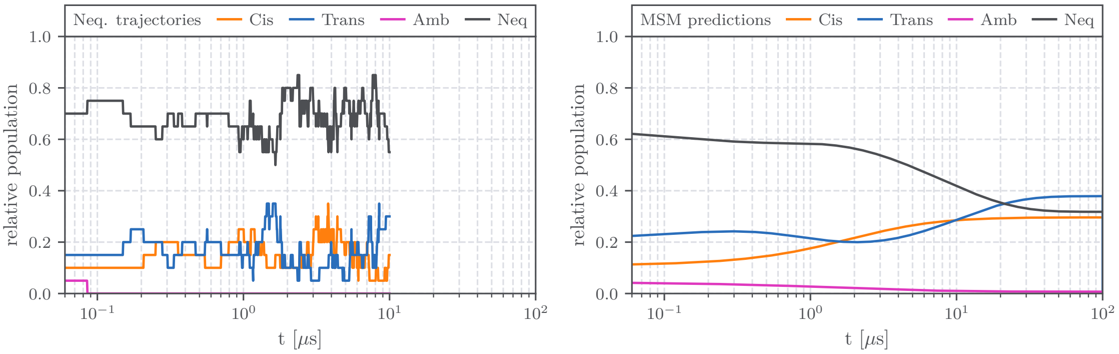

Figure 4.12.: *Left:* Mean time evolution of the 20 long nonequilibrium trajectories in shares of Cis, Trans, Amb and Neq. *Right:* Predictions of the share evolution by the MSM.

Most Important Pathways

Another way to extract information from the transition matrix is to determine the most important pathways by running a MCMC simulation on it. In order to initialize a MCMC, we define some initial and final microstates with which we seek describing the allosteric transition. To do so, we choose the pure Cis microstates 41, 45, 47 and 53 as initial microstates and for final microstates the Trans microstates 27, 36, 48 and 56. The simulated MCMC trajectory is 10^{12} frames long and was simulated using *MSMPathfinder* [71]. The 15 most sampled pathways are shown in Tab. 4.3. The most sampled pathway does only account for ~1.6 % of all events and the 15 most sampled pathways combined account for only 8.2 % of all events. Due to the enormous combinatoric possibilities of 56 microstates this low number is, however, not very surprising.

The MCMC also offers insights into the timescales involved. For this purpose we impose a maximum permitted length t_{max} on each pathway and investigate how many pathways shorter than t_{max} reach the Trans region. The quantity obtained this way is the cumulative distribution function (CFD). Initially, no pathway will make it to the Trans region, but with increasing t_{max}, more and more pathways will arrive there until every pathway which reaches the Trans region is shorter than t_{max}. By differentiating the CDF we obtain the probability density function (PDF) for the length of a pathway and can this way determine the most frequent pathway length which describe the Cis→Trans transition. However, we are mostly not interested in the length of the single pathways but rather in the timescales which are most dominant in the allosteric transition, which is why the PDF is multiplied by the time, i.e. $t \cdot \text{PDF}(t)$. In Fig. 4.13, all three curves are shown: The CDF, PDF and $t_{\text{max}} PDF$. Furthermore, the first implied timescale t_1 is shown. The good agreement between the first implied timescale t_1 and the maximum of $t_{\text{max}} \cdot \text{PDF}(t_{\text{max}})$, which indicates the dominant timescales involved, is remarkable. Since we explicitly specified the initial and final microstate in such a way that the pathways describe the

Cis→Trans transition, this underlines the assumption made before that the first eigenvector describes the allosteric transition.

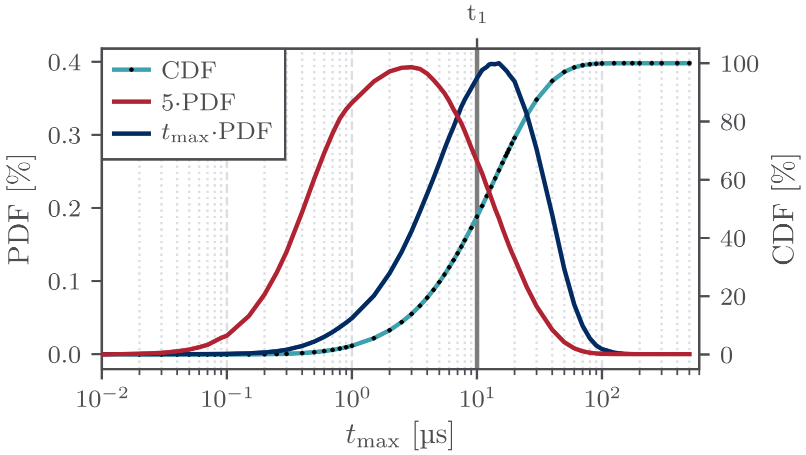

Figure 4.13.: ● Cumulative distribution function of the probability of all pathways being captured in less than $t_{\max}$ steps. The small black points indicate the actual computed values for a length limit $t_{\max}$. ● The derivative of the CDF, the probability density function (PDF), indicates the probability of a certain pathway length. ● For a better assessment of the actual impact of the single pathways, the PDF was weighted with the length of the pathways. ● the first implied timescale t_1.

In order to interpret the individual pathways, we can investigate some properties such as e.g. compactness or overlays of their conformation. Here, we exemplary show and analyze two pathways. Since the most and second most important pathway only differ in their final microstate, we analyze the most and third important pathway instead (pathway 1 and 3).

For the examination of the compactness of a microstate, 2000 frames in each conformation were randomly chosen and the root-mean-square deviation of atomic distances (RMSD) was calculated. Figure 4.14 on page 62 shows the evolution of the RMSD for each microstate along pathway 1 and 3 in a rain-cloud representation (see caption of the figure for a detailed explanation). Apart from microstate 46, which often follows microstate 45 and is slightly more aligned, an order-disorder-order behavior can be observed for the 15 most sampled paths. While Cis and Trans microstates are generally well aligned, nonequilibrium microstates usually exhibit higher disorder. Furthermore, we can see that the overlays of nonequilibrium microstates appearing in the MCMC pathways are usually better aligned than other nonequilibrium microstates like e.g. the most populated one. This indicates that a certain amount of disorder may be prerequisite to allow conformational changes from Cis→Trans, but that efficient pathways, which are frequently sampled, do not contain such highly variable microstates.

The corresponding overlays for the pathways 1 and 3 are shown in the figures 4.15 on page 63. This representation allows us to see which parts of the protein undergo larger changes during the allosteric transition. While the β-sheets are generally very stable and well aligned, the α_2-helix, which is shown on top in red, is the most affected metastable structure of the protein when it comes to conformational heterogeneity. Often already slightly distorted from the very beginning, the disorder increases in the course of the pathway until the α_2-helix becomes strongly aligned in the final Trans microstates. This might be due to stress induced by the azobenzene photoswitch, spanned from the β_2-sheet to the α_2-helix, which diminishes over the 10 μs of the nonequilibrium trajectories. Also noteworthy is the $\beta_2\beta_3$-loop, which is marked in yellow. In all analyzed pathways, a lot of dynamics can be observed in this area which can be explained by the fact that this loop contains relatively few contacts which leads to a large conformational heterogeneity. Still it is well aligned in the final Trans microstates and in particular microstate 48 sticks out as its $\beta_2\beta_3$-loop has a strongly aligned, but twofold arrangement. The second most affected loop region is the $\beta_1\beta_2$-loop, which is firmly aligned at the start and the end but disordered in between. Both loops, $\beta_1\beta_2$ and $\beta_2\beta_3$ seem to be distant at first but then approach each other during the disordered nonequilibrium microstates before they are very close in the final Trans microstates.

> A MSM based on the data selected in Ch. 3 was constructed in this section. The eigenvector associated with the first implied timescale of the resulting transition matrix describes the allosteric transition from Cis to Trans at a timescale of $t_1 \approx 10\,\mu$s. The predictions of timescales involved by the MCMC agree well with those of the MSM, which again emphasizes that this eigenvector indeed describes the Cis→Trans transition.
> Analyzing the pathway predictions of the MCMC, we found that they feature an order-disorder-order behaviour, where the nonequilibrium microstates inhibit the highest disorder. However, nonequilibrium microstates featuring even more disorder also exist, but they do not occur in the most important pathways. As those nonequilibrium microstates are located between a second Cis cluster, which is far away from the Trans microstates, and the Trans cluster, it stands to reason that a high conformational heterogeneity of transition states may hinder effective transition.

Table 4.3.: Most sampled pathways of the $\tau_{\mathrm{cor}}^{\mathrm{i}} = 3000\,\mathrm{frames} = 60\,\mathrm{ns}$ model. MCMC with a length of 10^{12} steps was performed. Initial states were the pure Cis states 41, 45, 47and 53. The final states were the Trans states 27, 36, 48 and 56.

i	Pathway$_i$	Count [%]	$\sum^i$ Counts [%]
1	45 → 46 → 10 → 56	1.642	1.642
2	45 → 46 → 10 → 27	1.639	3.281
3	45 → 32 → 30 → 9 → 37 → 48	0.621	3.902
4	53 → 21 → 13 → 9 → 37 → 48	0.472	4.373
5	45 → 21 → 13 → 9 → 37 → 48	0.451	4.825
6	45 → 32 → 46 → 10 → 56	0.423	5.248
7	45 → 32 → 46 → 10 → 27	0.423	5.671
8	45 → 32 → 30 → 44 → 37 → 48	0.343	6.013
9	45 → 46 → 10 → 9 → 37 → 48	0.339	6.353
10	45 → 46 → 2 → 50 → 27	0.333	6.686
11	45 → 46 → 2 → 27	0.332	7.018
12	45 → 46 → 2 → 27	0.309	7.327
13	45 → 32 → 30 → 9 → 10 → 27	0.304	7.632
14	45 → 32 → 30 → 9 → 10 → 56	0.304	7.936
15	45 → 32 → 2 → 50 → 27	0.244	8.179

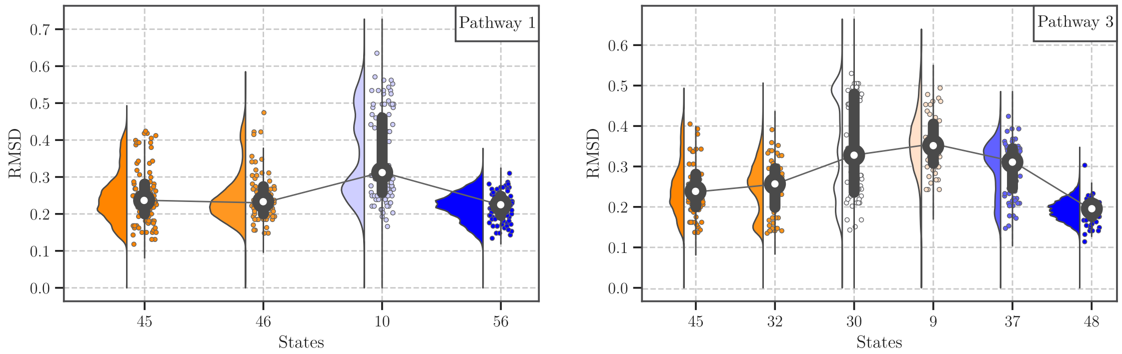

Figure 4.14.: The evolution of the "compactness" of microstates during two of the most sampled pathways. For the calculation of compactness, 2000 frames were randomly selected for each microstate and the root-mean-square deviation of atomic distances (RMSD) was determined.
For every microstate, one can see a rain-cloud plot: On the left the probability density is shown. The white point in the bar indicates the median of the distribution. The thick grey bar shows the second quartile from the bottom to the median and the third quartile from the median up to the top end of the thick grey bar. The smaller underlying grey line shows the interval which contains 95% of all data points. 100 randomly selected data points are shown in order to illustrate the PDF. The colors correspond to the usual Cis/Trans colors which are used in Tab. 4.2.

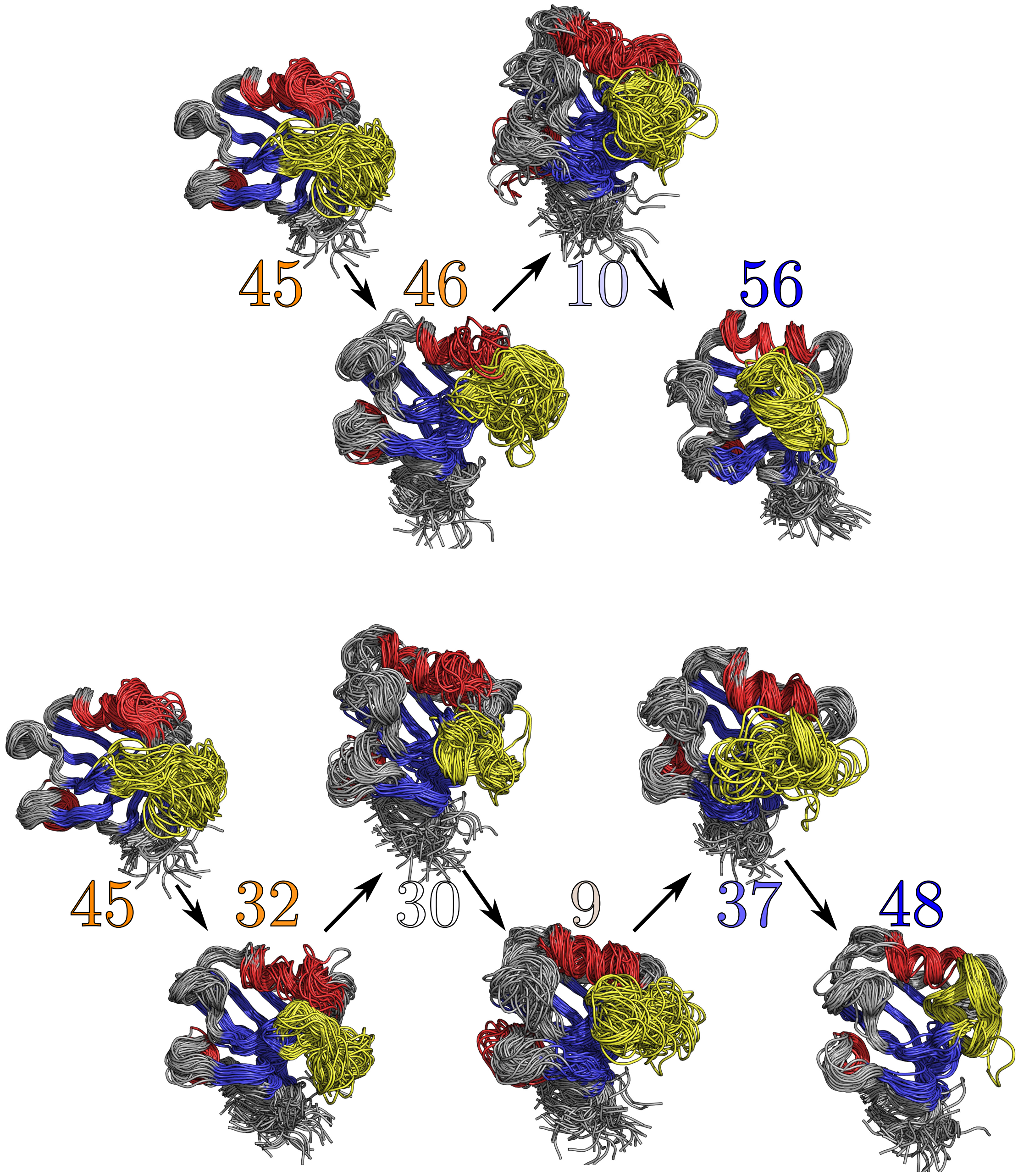

Figure 4.15.: *Top:* The most probable pathway. *Bottom:* The third most probable pathway. 100 randomly selected overlays are plotted for each microstate occuring in the pathway. While the β-sheets (marked in ●) are relatively closely aligned, in particular the α_2-helix (marked in ● red) and the $\beta_2\beta_3$-loop (marked in ○ yellow) seem to undergo major changes.

5. Data Generation 2: MSM Based on Machine Learned Coordinates

Statisticians, like artists, have the bad habit of falling in love with their models.

George E.P. Box

In the last chapter we constructed a MSM which is able to describe the Cis→Trans allosteric transition, but it has its limits. For the underlying set of data, we saw that the first three microseconds of the equilibrium trajectories were not yet equilibrated and that they heavily overlapped which is why we discard them in the following. By building a second MSM based on newly chosen coordinates, we want to check to what extent the predictions of the MSM are reliable and reproducible when improving the reliability of the data. Furthermore, we examine whether the discard of the first 3 μs of the equilibrium trajectories leads to a more clear-cut separation between Cis and Trans and hence to a preciser predictions about nonequilibrium states involved in the allosteric transition.

Instead of choosing important contact-distances by hand, we are now using a method based on supervised machine learning in order to identify a small set of essential contact-distances (see Sec. 5.1). Besides facilitating the construction of a second MSM, the results from the decision tree used by the machine learning approach are also interesting for making assumptions about the complexity of the system or pointing out important mechanisms. We find that this resulting set of "essential coordinates" is—due to the collective motion of many contact-distances—still correlated which is why we subsequently perform a correlation analysis of the remaining distances (see Sec. 5.2). This way we reduce the initial set of 429 contact-coordinates down to 54, on which we eventually construct a second MSM after PCA and clustering. Since most of the steps necessary for the construction of a MSM should look familiar by now, we will refrain from discussing them in great detail, but instead refer to the chapters 3 and 4. The MSM will be coarse grained by dynamical lumping, a newly introduced technique (see Sec. 5.5), in order to allow statistically sound statements about the allosteric transition (see Sec. 5.6).

5.1. Feature Importance

Brandt and Sittel developed a machine learning decision tree based on XGBoost [72] which assigns unknown MD data points—given by contact-distances in our case—to the microstate they most probably belong to [45]. As we will see in the following, it iteratively improves the prediction of assigning the MD structures to a specific microstate by creating a decision tree for each microstate.
70% of the MD data, represented by the contact-distances from Fig. 3.2 were used to train the model to assign a single frame $\boldsymbol{r} = (r_1, ..., r_n)$ to the resulting microstates from clustering obtained in Sec. 3.4 and the remaining 30% were retained for testing the accuracy. On the basis of the test set, the model's accuracy can be estimated by counting how many of the frames are correctly assigned $\varepsilon = \frac{N_\text{correct}}{N}$ in respect to the total number of frames N (multiclass error, N_correct is the number of correctly assigned frames). Furthermore, the accuracy of the classification for one particular microstate i is calculated as well: $\varepsilon_i = \frac{N_{i,\text{correct}}}{N_i} - \frac{N_{i,\text{incorrect}}}{N}$, where $N_{i,\text{correct}}$ denotes the frames correctly assigned to microstate i, while $N_{i,\text{incorrect}}$ stands for the erroneously assigned frames. Latter penalizes miscalculation and prevents the formation of one "trash"-state which contains all MD frames whose classification is not entirely clear. By optimizing a loss function containing the number of erroneously classified frames and the tree size the model is adjusted. By holding the tree size as small as possible, overfitting is prevented.
We measure the importance of different coordinates for the microstates classification by evaluating how impactful a move in the decision tree is regarding that contact-distance. Sorting all contact-distances by their importance, we either discard the most or the least important contact-distance successively and reiterate the procedure by retraining the model for the remaining contact-distances:

- *Removing the most important contact-distance (RMI).* We iteratively remove the most important coordinate in order to access "hidden" important contact-distances. As the dynamics of PDZ2 is highly correlated in terms of contact-distances, collective motion related to the most important contact- distance may overshadow other important, but independent contact-distances in PCA. Therefore, the most important contact-distance is removed in each iteration to clear the way for other important and uncorrelated ones.

- *Removing the least important contact-distance (RLI).* In order to filter out all nonessential contact-distances from the data set, the least important contact-distances is discarded in each iteration.

While we identify the most important contact-distances by iteratively removing the most important one from the set and therefore enable to assess its overall importance independent of the other coordinates, we use RLI to greatly reduce the number of total contact-distances by filtering out all nonessential contact distances [45]. Matthias Post set up the algorithm on the basis of the model constructed in Ch. 4 for both approaches (RMI and RLI) in order to combine their predictions. Figure 5.1 shows

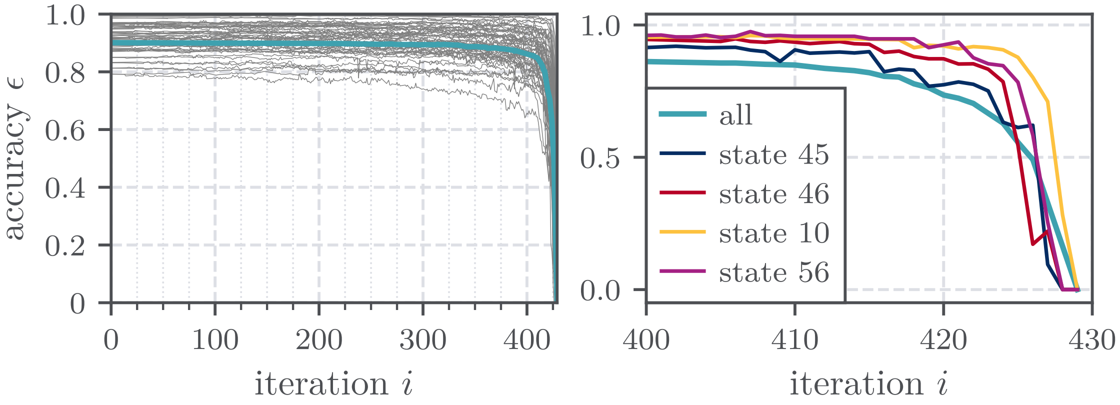

Figure 5.1.: Accuracy ε evolving by removing the least important coordinate. *Left:* ● the mean for all microstates, ● all microstates individually. *Right:* The microstates of the most sampled MCMC pathway (see Sec. 4.4). Data provided by Matthias Post.

the evolving accuracy ε by removing the least important contact-distance for all microstates on the left and for the microstates of the most important pathway of the model based on data generation 1 on the right (see Fig. 4.15). One can see, that the microstates along the most important pathway can still be well resolved using only 10 contact-distances or less. For the microstate classification in general, even after ~380 discarded contact-distances there is no real drop in accuracy visible which suggests that the dynamics of PDZ2 is indeed governed by the collective motion of relatively few distances and that many contact-distances are equally able to discriminate the microstates. This means that one can greatly reduce the number of contact-distances and still end up with an accurate description of the system. Furthermore, the model can in particular be trained to discriminate states of major interest such as, e.g., the microstates along the most probable pathway. Combining the most important contact-distances resulting from both approaches and the most important contact-distances for discriminating the microstates along the most sampled pathway, the 429 initial contact-distances (see Fig. 3.2) can be reduced down to 77 (RMI ∪ RLI ∪ Pathway 1).

The two most essential contact-distances are $r_{19,27}$ and $r_{64,74}$, which describe the relative motion between the $\beta_1\beta_2$-loop and the $\beta_2\beta_3$-loop and the relative distance between the $\beta_4\beta_5$-loop and the α_2-helix, respectively. The second most important contact-distance, $r_{64,74}$ was expectable as the azobenzene photoswitch lies very close-by and forces this coordinate to change significantly upon the Cis → Trans transition. Also the most important contact-distance $r_{19,27}$ agrees perfectly with the observations made by inspecting the most important pathways resulting from the MCMC simulation. There, we found that the distance between the $\beta_1\beta_2$ and the $\beta_2\beta_3$-loop decreases during the Cis→Trans transition. However, one must mention here, that the algorithm used the MSM from Ch. 4 as a reference for evaluating the importance of each contact-distance in the set of the 429 already manually preselected contact-distances. Therefore, we could only achieve a reduction of the number of contact-distances and an identification of the most important ones. It would have been even more useful to use all possible input coordinates as input, which is due to the enormous computational effort not feasible.

> *...what the machine learned.*
>
> We trained a machine learning algorithm which is based on decision trees with the initial contact-coordinates and the clustering from Ch. 3 and this way were able to reduce the number of contact-distances down to 77 which are considered to be particularly essential. The predictions for the most important contact-distances fit very well with the observations from the MCMC simulations for the MSM based on data generation 1.

5.2. Correlation Analysis

Despite removing the least important coordinates in the last section, the 77 contact-distances yielded by the machine learning algorithm are still not completely uncorrelated but require further reduction. To this end, the correlation matrix of those 77 contact-distances was calculated and is shown on the left hand side in Fig. 5.2. Still, 36 off-diagonal elements representing strongly correlated contact-distances, i.e. feature a correlation of higher than 80%, can be identified. While most of the highly

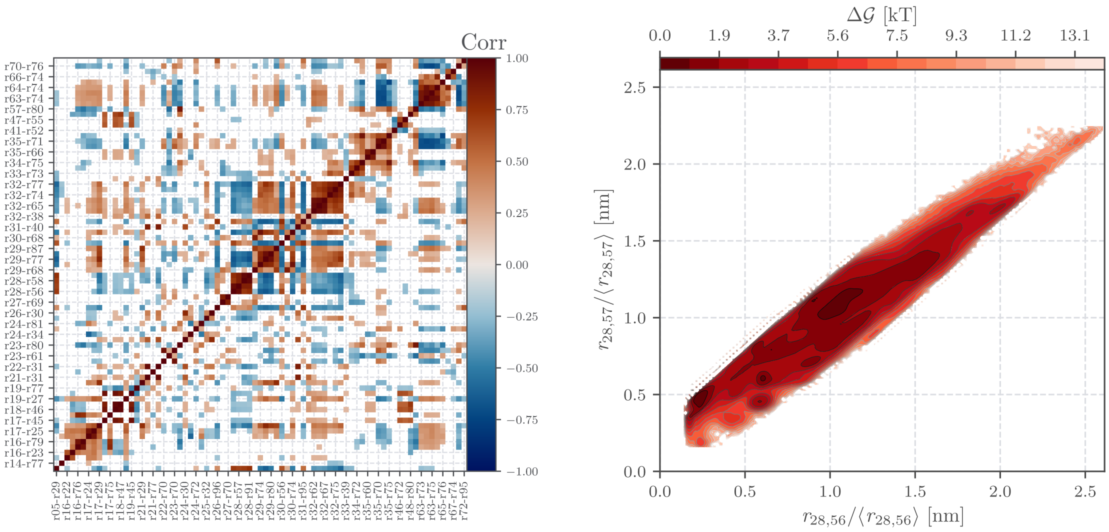

Figure 5.2.: *Left:* Correlation matrix of the 77 contact-distances resulting from the XGBoost decision tree studies. Only values $|\mathrm{Corr}(r_{i,j}, r_{k,l})| \geq 0.2$ are shown. *Right:* Free energy projection onto the contact-distance $r_{28,56}$ and $r_{28,57}$.

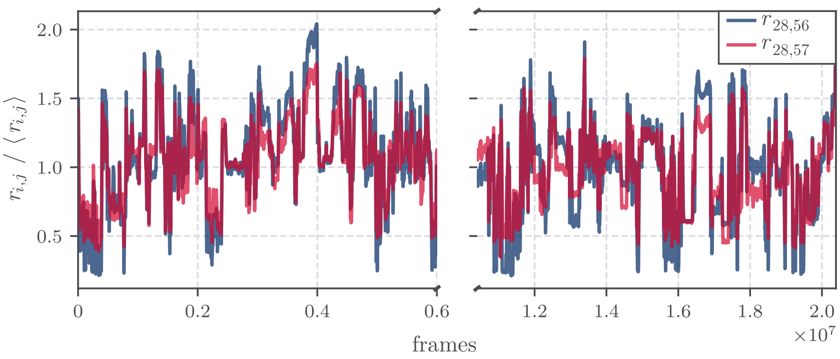

Figure 5.3.: Time trace of the contact-distances $r_{28,56}$ and $r_{28,57}$. Their absolute values have been divided by their mean value in order to facilitate comparability. The gap in the middle indicates the range of the short 1.1 μs nonequilibrium trajectories which are not shown for the sake of clarity.

correlated distances lie close to each other and are represented by entries close to the diagonal (consider e.g. distances $r_{28,56}$ and $r_{28,57}$), other entries are located far away from the diagonal (e.g. $r_{23,80}$ and $r_{57,80}$) and represent contact-distances whose correlation is not immediately obvious.

Before we discard highly correlated contact-distances, we need to confirm whether they indeed contain almost identical information. For this purpose, the free energy [see Eq. (2.11)] is plotted along those contact-distances that could qualify to be sufficiently represented by only one contact-distance. This is shown exemplary for the distances $r_{28,56}$ and $r_{28,57}$ on the right hand side of Fig. 5.2. The free energy along both contact-distances is approximately diagonal which indicates that both coordinates behave almost identical. If the shape would differ significantly from the diagonal shape, this would indicate that one coordinate contains information which cannot be resolved by the other coordinate. To illustrate that two coordinates which are characterized by a diagonal shape of their free energy do indeed contain almost identical information, the time trace (all 10 μs trajectories concatenated) of both contact-distances is shown in Fig. 5.3—the time traces of both distances are almost identical. Both time traces were first normalize them by dividing by their mean value in order to facilitate the comparison. By repeating this procedure for all pairs of strongly correlated contact-distances, the initial 77 contact-distances could be reduced to 54.

> The correlation analysis of the set of contact-distances yielded by the XGBoost decision tree approach allowed to further reduce the number of input coordinates by 30%. Instead of 77 contact-distances, only 54 were retained.

5.3. PCA, Clustering and Coring

The set of 54 contact-distances, which was obtained in the last section, is used to repeat the steps 2–4 of the workflow (see Sec. 2.7). As in Sec. 3.3, the PCA was first performed only on the equilibrium data in order to calculate reaction coordinates that maximize the variance between the Cis and Trans conformations of the protein. In a subsequent step, the nonequilibrium data was projected onto the eigenvectors of this PCA. The tests which were performed in Sec. 3.3 show better results for the PCs x_1–x_6 compared to the x_1–x_5 and x_7, used for data generation 1 (see Fig. A.5–A.8 on the pages 89-91 for of the the cumulative fluctuations, the temporal evolution along the PCs, the ACF plots and the grid representation of the two-dimensional free energy projections) Since the number of input coordinates was greatly reduced, we find that the first 6 PCs now cover almost 70% of the overall variance within the set of input coordinates, which is about 10% more than before. Also due to a smaller number of input coordinates, we fine a smaller clustering and lumping radii, which decreased compared to the ones obtained in Sec. 3.4 from $R = d_{\text{lump}} = 0.94$ to $R = d_{\text{lump}} = 0.34$ by a factor of $\frac{0.34}{0.94} = 0.36$. This corresponds approximately to the reduction which was expected due to geometrical considerations [$\sqrt{\frac{429}{54}} = 0.38$, see Eq. (2.12)].
We decided to cluster 52 microstates, of which none was detected as trap state. There are two possible reasons for the disappearance of trap states: Firstly, the PCs are different now and the MD data is projected slightly different onto the microstate space which could prevent the formation of a trap state. Secondly, the PCs now used for clustering could describe additional or different dynamics that are essential for the trajectory to leave the trap state. Just as before, the resulting microstates trajectory was noise-corrected and iteratively cored with $\tau^{\text{i}}_{\text{cor}} = 60\,\text{ns}$.

Comparison of the Clustering in Both Models

We further analyze how the new PCA and clustering affect the definition of the microstates. With almost the same number (52 vs. 56) of microstates it is convenient to compare how each MD frame is projected onto the discrete microstate space Ω. By doing this, we check whether certain microstates exist which represent universal, essential conformations of the protein and are therefore equally present in both microstates trajectories. Fig. 5.4 shows a comparison of both trajectories where we compare each frame one by one and map their connection for each microstate. It is apparent that not a single microstate is present which appears identically in both trajectories. On the one hand, we see for some microstates, that it is more likely that they are rather uniquely assigned to another microstate from the respective other trajectory. If we study the graph from top to bottom, these are—to name some examples—the microstates 10 → 15, 19 →, 28 → 27, 29 → 36, 44 → 46 and more. On the other hand there are microstates which are assigned to a multitude of other microstates, e.g. for data generation

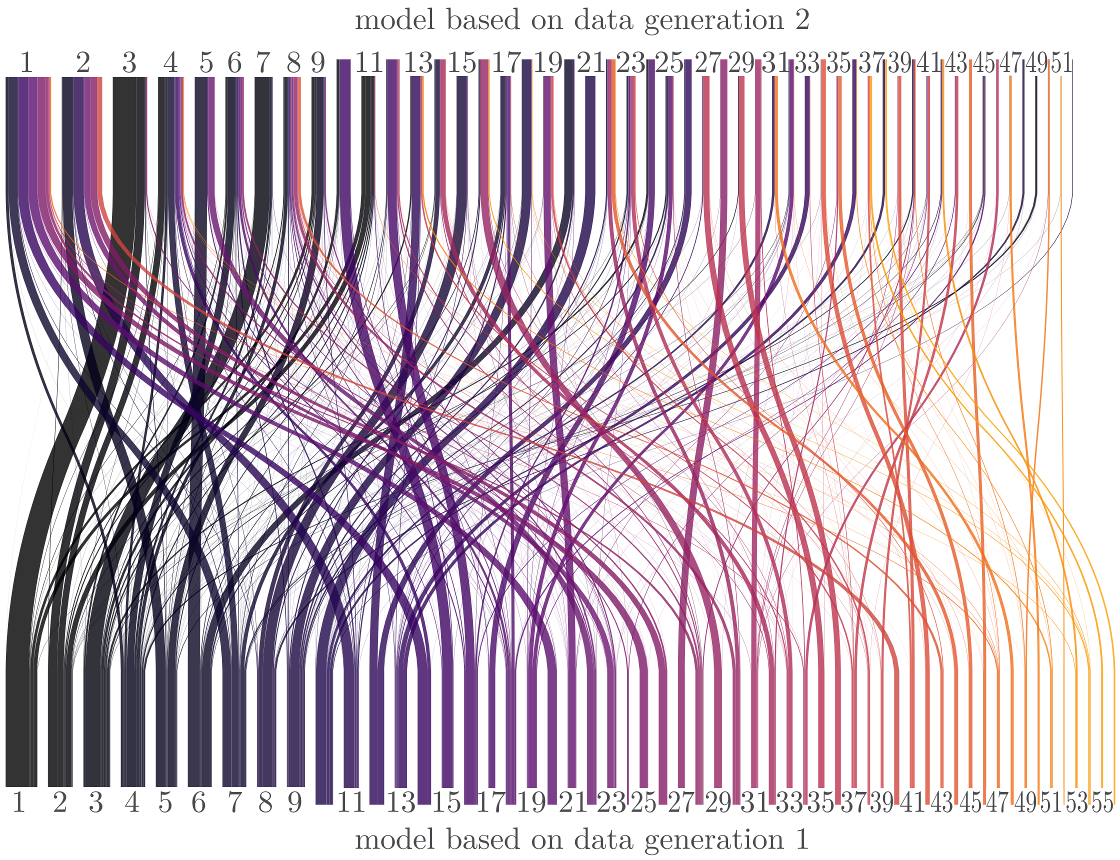

Figure 5.4.: Comparison of the microstates which were found for data generation 1 (*bottom*) and data generation 2 (*top*). Since data generation 2 discards them, the frames of the first 3 μs of all equilibrium trajectories had to be deleted in data generation 1 as well, which is why Cis and Trans microstates are sometimes smaller than subsequent nonequilibrium microstate. The colors do not specify characteristics of the microstates and only serve to facilitate the assignment.

2 the microstates 1, 2, 4, 8, 23 or 40. For data generation 1, we identify, for example the microstates 2 to 7, 9, 18 and 19, 38, 54 and many more. Interestingly, we find predominantly equilibrium microstates (Cis and Trans, see tables A.2 and 4.2) to be relatively uniquely assigned, while nonequilibrium microstates are often not uniquely mapped but assigned to a great number of microstates in the other trajectory. This is most probably due to two reasons: In the first place, we use exclusively the equilibrium data to calculate the PCA correlation matrix in order to prioritize a high resolution for the Cis and Trans initial/final microstates. Secondly, we saw in the last chapter that Cis and Trans conformations feature a lower heterogeneity within their microstates while nonequilibrium conformations are often more variable. Consequently, equilibrium microstates are characterized by a lower free energy minima with better defined microstate borders. In contrast, nonequilibrium microstates require more volume in

the conformational space which causes border regions to overlap and MD frames located within this regions are easily assigned to a different microstate once the definition of the PCs changes slightly.

> A PCA was performed on data generation 2 and the PCs x_1–x_6 were selected for the subsequent clustering. This resulted 52 microstates which were again cored with a coring time of $\tau_{\text{cor}}^{\text{i}} = 60\,\text{ns}$. Comparing the resulting microstate trajectory with the microstate trajectory of data generation 1, we find that equilibrium microstates are more likely to be relatively uniquely mapped while nonequilibrium microstates usually get widely spread mapped.

5.4. Construction of the MSM

Again, the MSM was set up with a lag-time of $\tau_{\text{lag}} = 60\,\text{ns}$. Similar to before, Trans microstates are peculiarly strongly represented in the stationary distribution and the first eigenvector splits Cis and Trans apart from few exceptions (see Fig. 5.5). In contrast to before, the first implied timescale decreases from 10.0 μs to 7.5 μs (see Fig. A.9). One reason for this might be the fact that we did not identify any trap state in data generation 2. Reassigning the trap state in data generation 1 to the microstate visited before yields a highly metastable microstate which could ultimately prolong the dynamics.
We consider Fig. 5.6 for the representation of the network and we do not pay attention to the colored markings yet. In contrast to data generation 1, an improved separation of Cis and Trans is evident. Thus, the majority of non-equilibrium microstates now lies between the pure Cis microstates and the large cluster of strongly interacting microstates—mostly Trans microstates—on the right while we find a cluster of Cis microstates, i.e. 32, 45 and 46, to be closely located to the Trans cluster in data generation 1. This is a strong indication that the machine-preselected and shortened input coordinates

Table 5.1.: Most sampled pathways of the MSM based on data generation 2. A MCMC-simulation with 10^{12} steps was performed. Initial states were the pure Cis states 35, 44 and 46. The final states were the Trans states 28, 33, 36 and 52.

i	Pathway$_i$	Count [%]	$\sum$ Counts [%]
1	35 → 19 → 47 → 25 → 18 → 28	0.252	0.252
2	46 → 30 → 11 → 18 → 28	0.242	0.494
3	46 → 30 → 1 → 25 → 18 → 28	0.206	0.700
4	46 → 30 → 1 → 37 → 27 → 23 → 36	0.122	0.822
5	44 → 14 → 3 → 43 → 33	0.116	0.938
6	35 → 14 → 3 → 43 → 33	0.102	1.041

in data generation 2 result in a clearer separation between Cis and Trans.
Again, we consult a MCMC-simulation in order to calculate the most important pathways and thus identify mechanisms which play a crucial role. For the initial states we chose the pure Cis microstates 35, 44 and 46 and as final state the relatively pure Trans microstates 28, 33 and 52 (all Trans microstates with a Trans share higher than 88.5%). The results of the MCMC simulation are depicted in Tab. 5.1. Just like predicted by the MCMC simulation based on data generation 1, these pathways again feature an order-disorder-order behaviour (see Fig. A.10 on page 93). A closer look at Tab. 5.1 reveals that the contributions of the most important pathways is very small. Compared to data generation 1, we expected the numbers to be smaller, since we are now faced with much more nonequilibrium microstates which are located in between the Cis and Trans cluster which ultimately increases the combinatorial possibilities significantly. So, the first six most important pathways only account for little more than 1% of all sampled pathways. This is still not a small amount compared to the enormous amount of combinatorial possibilities, but it is difficult to make statistically relevant statements about the underlying mechanisms of the Cis→Trans transition.

> The MSM based on data generation 2 splits the Cis and Trans clusters in such a way, that the transition from the initial Cis to the final Trans microstates is dynamically well described by the nonequilibrium microstates in between. This suggests that the PCA projection calculated only on the last 7μs of the equilibrium data is more reasonable than for the PCA projection which was calculated for data generation 1. However, this also leads to significantly smaller contributions of the most important pathways which describe the Cis→Trans transition to the overall MCMC trajectories, since those nonequilibrium microstates now appear as possible nodes. This makes it considerably more difficult to make statistically verifiable statements.

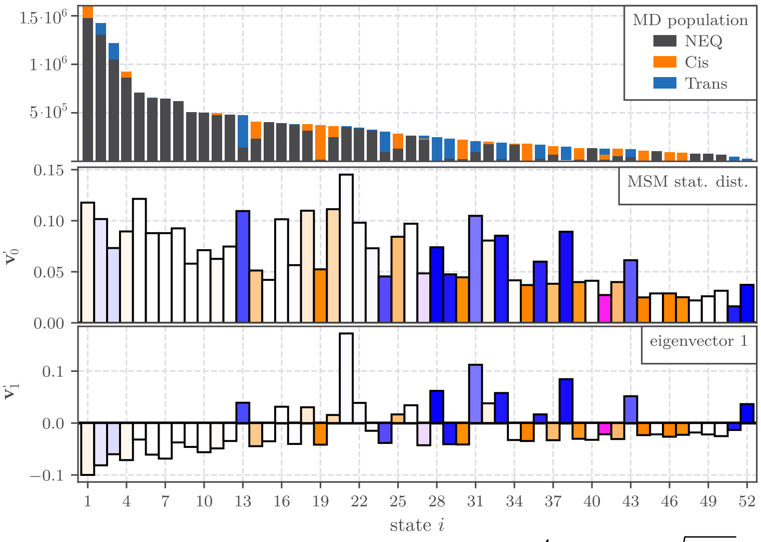

Figure 5.5.: *Top:* Population of the states found in the data. We define $v'_j \equiv \mathrm{sgn}(v_{j,i})\sqrt{|v_{j,i}|}$. *Middle:* Stationary distribution predicted by the MSM. *Bottom:* The first eigenvector which indicates the slowest occurring process.

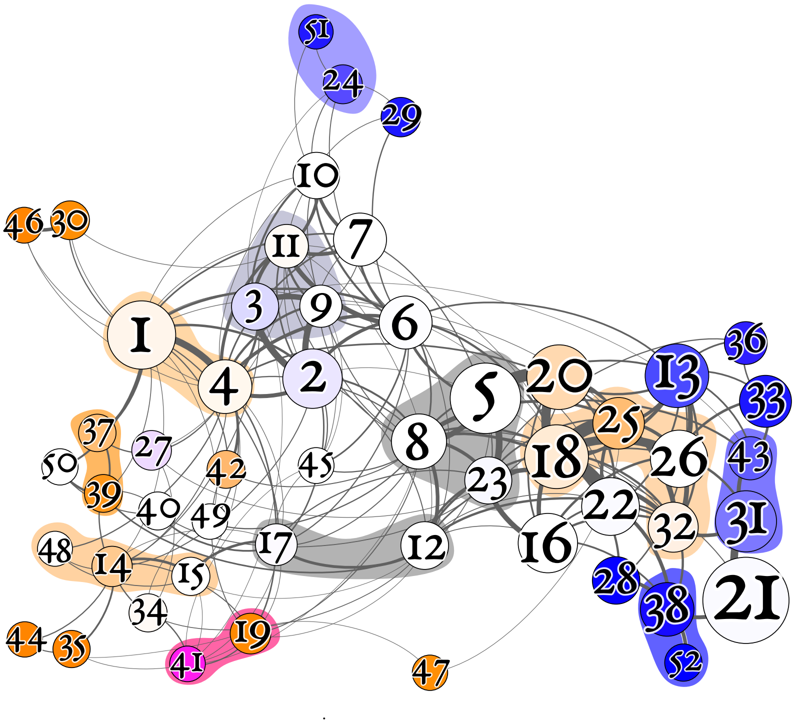

Figure 5.6.: Network representation of the $\tau^{\mathrm{i}}_{\mathrm{cor}} = 60$ ns MSM based on data generation 2. The edges are weighted by $P^{\mathrm{eq}}_{ij}T_{ij}$ and the size of the nodes is proportional to their stationary distribution. Cis microstates are marked in orange ●, Trans microstates in blue ●, Neq in white ○ and the ambivalent microstates in violet ●. Cluster of microstates which are marked by a colored area are lumped due to the dynamical lumping algorithm (see Sec. 5.5).

5.5. Dynamical Lumping

In order to increase the relevance of the most important pathways which result from the MCMC-simulation by increasing their contribution to the MCMC trajectory, we introduce a method called *dynamical lumping*. The idea is to accept small losses in the structural homogeneity of the microstates by lumping kinetically close microstates and thus to increase their relevance, to consequently make pathways along these lumped microstates more relevant. It lumps microstates according to their kinetic connectivity and is inspired by the *Most Probable Path* algorithm [73]. In contrast to this approach, however, dynamical lumping does not take the population of a microstate into account, but lumps them only according to their transition rates T_{ij}. For a system like PDZ2, which is actively driven out of its equilibrium conformation, this has proven to be advantageous, since very metastable microstates are usually not the highest populated states in the data and are therefore systematically penalized by the *Most Probable Path* algorithm. The application of the *Most Probable Path* algorithm would therefore merge states particularly into highly populated nonequilibrium states, which are expected to have a short lifetime for PDZ2.

Application on PDZ2 · The exact implementation is explained in form of a pseudocode in Algorithm 1 (see Appendix, page 94). Applied on the microstate trajectory and the model of the last section, the dynamical lumping algorithm lumps the microstates according to the dendrogram shown in Fig. 5.7.

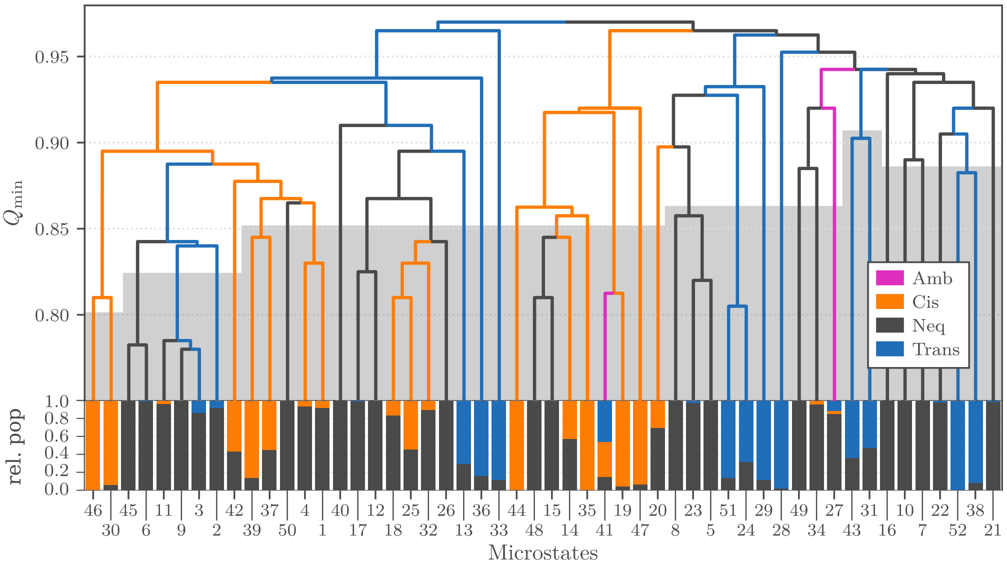

Figure 5.7.: *Top:* Dendrogram which indicates the metastability of the microstates. If two lines are merged, it means that the left microstate is lumped into the right microstate. The grey area indicates the selected Q_{min}−threshold under which all microstates are merged. Lines above indicate the microstates which remained in the final trajectory after lumping. *Bottom:* Relative state population of each microstate.

Microstates which are dynamically well connected are positioned side by side, where the microstate on the right always possesses a higher metastability. Once the metastability of the left microstate falls below the required metastability Q_{min}, we see a horizontal line connecting both microstates and the left microstate is lumped into the right one. Here we can freely decide which value of Q_{min} we choose for different microstates—so we have, for example, left pure Cis or Trans microstates untouched, as we want to preserve a high structural resolution in the initial and final states. An exception to this is the pure Trans microstate 52, which is kinetically so closely connected to the Trans microstate 38. Microstate 38 is nearly a pure Trans microstate as well and is therefore considered as final microstate and its much higher connectivity to intermediate nonequilibrium microstates prevents any MCMC trajectory from reaching microstate 52. The final state lumping is indicated by the grey area in Fig. 5.7. All microstates which are connected by a horizontal line within this area are lumped together. In order to better understand whether such a lumping is reasonable, the lower part of the plot shows the microstate's relative population in shares of Cis, Trans and nonequilibrium. This way, we make sure that we do not accidentally lump Cis microstates with Trans microstates or Trans microstates with Cis microstates, respectively. The clusters of actually lumped microstates are indicated by the colored areas in the network representation shown in Fig. 5.6.

> The *dynamical lumping* algorithm allows to merge microstates reliably according to their metastability. We applied the *dynamical lumping* algorithm on the microstate trajectory based on data generation 2 and this way reduced the number of microstates from 52 to 35.

5.6. Coarse-Grained MSM

By means of the dynamical lumping algorithm, the total number of microstates was reduced from 52 down to 35. The dynamical network representation of the resulting MSM with $\tau_{\text{lag}} = 60\,\text{ns}$ is shown in Fig. 5.8. The eigenvector of the first timescale still splits Cis and Trans microstates suggesting that it is still describing the allosteric transition from Cis→Trans (see Fig. A.12). In order to verify whether dynamical lumping has led to statistically more reliable statements, another MCMC simulation with 10^{12} steps is performed and the most important pathways are shown in Tab. 5.2.

The most important pathways are now sampled much more frequently than before (compare Tab. 5.1). As a result, the most important five pathways now account for 3.7% instead of only 1.0% as before lumping. More remarkable, however, is the fact that the microstate sequence $6 \rightarrow 7 \rightarrow 5$ is strikingly frequent—not only in the 10 most important pathways, where this sequence occurs six times. It can therefore be considered as the most important intermediate path. Consequently, this makes this sequence an excellent subject for the study of the allosteric transition in PDZ2.

We consider Fig. 5.9 on page 80 for an overlay representation of the contributing microstates including two initial and final microstates (pathways 2–5). In accordance with the MSM based on data generation

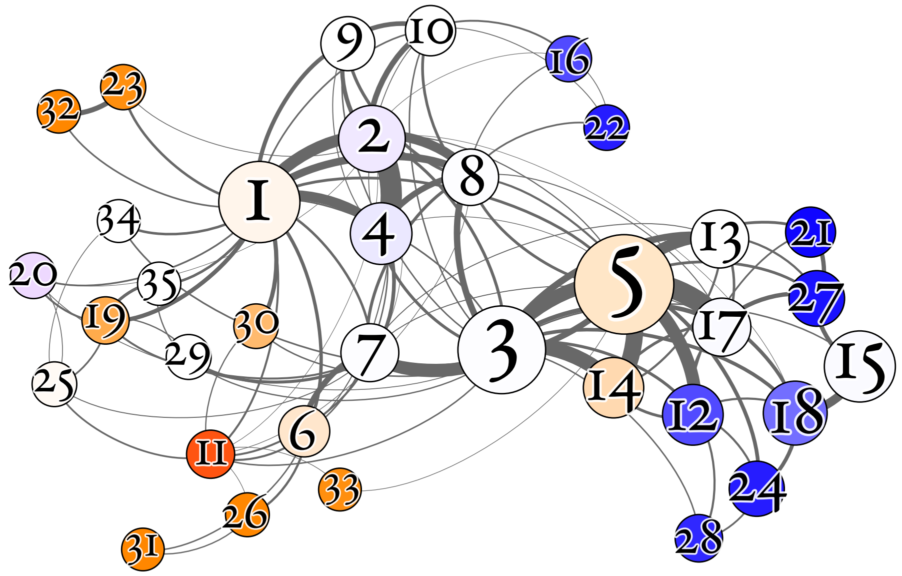

Figure 5.8.: Network representation of the resulting MSM of the microstate trajectory obtained by dynamical lumping according to Fig. 5.7.

Table 5.2.: The most important pathways predicted by a MCMC simulation of the length of 10^{12} steps. The sequence 6 → 7 → 5 appears in many pathways.

i	Pathway_i	Count [%]	$\sum$ Counts [%]
1	32 → 23 → 1 → 8 → 22	1.062	1.062
2	31 → 6 → 7 → 5 → 27	0.698	1.760
3	31 → 6 → 7 → 5 → 21	0.696	2.456
4	26 → 6 → 7 → 5 → 27	0.619	3.075
5	26 → 6 → 7 → 5 → 21	0.618	3.694
⋮	⋮	⋮	⋮
9	31 → 6 → 7 → 5 → 13 → 27	0.358	5.556
10	26 → 6 → 7 → 5 → 13 → 27	0.319	5.874
⋮	⋮	⋮	⋮
13	31 → 6 → 7 → 3 → 28 → 24	0.260	6.721
⋮	⋮	⋮	⋮
29	31 → 6 → 7 → 3 → 5 → 21	0.189	10.44
⋮	⋮	⋮	⋮

1, one can again observe a clear order-disorder-order behaviour. This time we refrain from showing another RMSD plot, because the order-disorder-order is clearly visible to the naked eye in Fig. 5.9. Also the $\beta_1\beta_2$- and $\beta_2\beta_3$-loop regions, in which XGBoost detected one of the most important contact-distances for the Cis→Trans transition, behaves similarly as before: At the beginning still far away, both loops approach each other along the path until they are very close. And also the already observed ordering of the α_2-helix is reflected along the most important intermediate path and can be seen in more detail in the grey area in Fig. 5.9: slightly disordered in the Cis microstates as 3_{10}-helix, the α_2-helix becomes highly disordered in the initial intermediate nonequilibrium microstate and then realigns itself more and more towards the final Trans microstates, where it ends up highly ordered. This might be due to stress induced by the azobenzene photoswitch which is connected to the α_2-helix and is switched to its Trans-conformation at the beginning of the nonequilibrium trajectories pushing the α_2-helix away from the β_2-sheet. The data shows that the β-sheets are more stable than the α-helices. The stress induced by the photoswitch looks for the weaker point which results in temporally induced deformations within the α_2-helix until the stress in the system disperses.

Furthermore, we observed in the last chapter, that highly disordered intermediate microstates do not participate in effective pathways. This is not the case here, as microstate 1—representing the microstate with the highest disorder (see Fig. A.11)—already appears in the most important pathway. Considering the dynamical network representation of this model in Fig. 5.8, we see that the second highest disordered microstate 3 is located in between microstate 7 and microstate 5 and is connected strongly to both of them. Nonetheless, the sequence $6 \rightarrow 7 \rightarrow 3$ does not appear until pathway 27, which is besides the addition of the highly disorder microstate 3, identical to the third most important pathway. Microstate 3 is in fact the secondly highest disordered microstate overall (see Fig. A.11). Even though one node more is required for this pathway and the probability for this pathway is therefore reduced, it is remarkable that the direct jump from microstate 7to 5 is favored and—despite less connectivity—almost four times more sampled. We cautiously mention here that this might be yet another sign that highly disordered microstates are preferably avoided in allosteric transitions for PDZ2—that is, if it is possible (see microstate 1).

However, the intermediate nonequilibrium and the final Trans microstates also reveal a small disadvantage of the lumping method, namely a lower conformational resolution of the microstates, which is evident when comparing Fig. 5.9 with pathways of the MSM before lumping in Fig. 4.15. As several microstates, which are initially clustered according to a high conformational homogeneity, are now lumped together, the conformational resolution suffers. Nevertheless and despite the coarser model, we were still able to make precise observations for which we can now provide a statistically more relevant foundation—a rewarding tradeoff.

Applying the dynamical lumping method, we were able to lump the microstates in order to find a coarse-grained model which still describes the allosteric transition from Cis to Trans conformation. This lumping procedure allowed to identify a sequence of intermediate states which occurs in a large variety of pathways which describe the Cis→Trans transition. Analyzing this sequence in detail revealed that the $\beta_1\beta_2$–$\beta_2\beta_3$ region plays a decisive role in the allosteric transition for PDZ2, which is in line with the predictions of the XGBoost decision tree and the MSM based on data generation 1 constructed in Ch. 4. Moreover, it was found that the α_2-helix, which is linked to the azobenzene photoswitch, also undergoes conformational changes during the allosteric transition. These are manifested by the fact that the α_2-helix always ends up being stable despite featuring a larger conformational heterogeneity along the intermediate microstates. The order-disorder-order behaviour seems to emerge as an universal principle in the allosteric transition of PDZ2, as it was observed in both models independently of the input coordinates and the projection of the MD data.

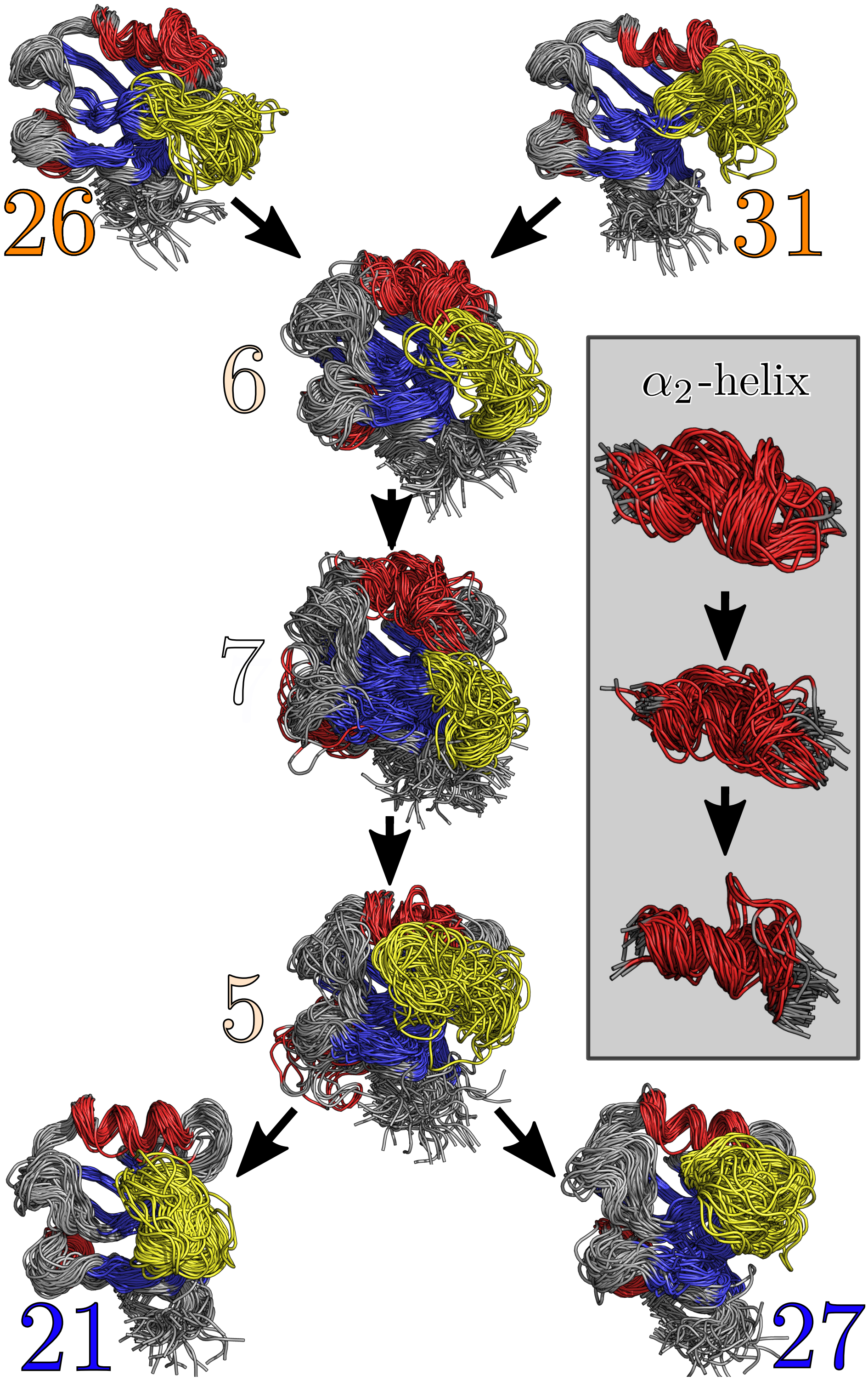

Figure 5.9.: 100 overlays per state of the states appearing along the most important pathways. The microstate sequence 6 → 7 → 5 occurs with striking frequency. The structures that are shown in the grey box are overlays of the α_2-helix for the states 6, 7 and 5 and are shown separately for a better visualization.

6. Conclusion and Outlook

Das ist ein weites Feld.

Effie Briest

This thesis addressed the complete transition mechanisms involved in the Cis→Trans allosteric transition of the photoswitchable PDZ2 protein. To this end, we employed a state-of-the-art six-fold workflow onto a vast ensemble of equilibrium and nonequilibrium trajectories in order to shed light on the underlying mechanisms that govern the allosteric transition in the protein and to explore whether they are of dynamical or conformational nature.
In the following we summarise our findings and discuss possible future projects.

Summary

We extracted the internal motion of the protein in Ch. 3 for obtaining microstates representating the protein's metastable conformations. To do so, different input-coordinates such as C_α-distance or contact-distance have been examined for their suitability. We found that contact-distances are more appropriate as fast side chain dynamics appear to play a crucial role in the allosteric transition. This is in line with other investigations [29]. Depending on their ability to discriminate Cis and Trans conformation, we chose a set of 429 contact-distances and performed an equilibrium principal component analysis on this data. In a subsequent step, the nonequilibrium data was projected onto the resulting eigenvectors of the equilibrium principal component analysis. The resulting free energy projections indicated a clear separation between Cis and Trans conformations and six suitable principal components were selected for the consequent clustering. For this purpose, we applied the clustering method by Sittel [49] in order to cut the microstates along their free energy barriers. Nevertheless, those barriers are sparsely sampled and artifacts due to dimensionality reduction are inevitable which often lead to a misclassification of frames at the states borders.
Chapter 4 therefore considers ways to correct these misclassifications via the application of dynamical coring [26]. Dynamical coring requires the trajectory of the microstate to spend at least a minimum coring time τ_{cor} in a new state and discards changes in the microstate trajectory which happen on shorter timescales. τ_{cor} is the only input parameter of dynamical coring which can be estimated by demanding

a mono exponential decay of the microstate population probabilities. However, for PDZ2 we find that the coring time suggested by this heuristic must be further extended in order to obtain a Markov description of the system's dynamics in terms of the microstates trajectory. Applying dynamical coring on these timescales alters the microstate trajectory significantly and introduces spurious artifacts which contradict the molecular dynamics data. We therefore introduce an iterative coring method which does not cause these artifacts and, thus, yields a significantly improved representation of the actual molecular dynamics data in terms of the microstate trajectory.

Based on the cored microstate trajectory we constructed a Markov state model which is able to describe the allosteric transition from the Cis to the Trans conformations of the protein. The resulting Markov state model is capable of describing the nonequilibrium transition as its stationary distribution differs from the distribution found in the MD data in that Trans states are higher populated. However, a large number of the nonequilibrium microstates were not projected between the Cis and Trans clusters and consequently did not contribute to a detailed description of the allosteric transition. In addition, it was found that the first 3 μs of the equilibrium trajectories were not yet fully equilibrated and that Cis and Trans heavily overlap during this time.

In order to investigate whether the positions of the nonequilibrium microstates are stemming from a projection error or rather are a result of biased simulation seeds, the above mentioned first three microseconds of each equilibrium trajectory were discarded. We also used a machine learning algorithm to reduce the number of input coordinates and retained only those contact-distances which are crucial for a description of the allosteric transition in terms of our previous Markov state model. Combining both, the shortened equilibrium trajectories and the smaller number of contact-distances, we projected the molecular dynamics data onto a new microstate trajectory in Ch. 5. In this projection, the nonequilibrium states are able to provide a detailed picture of the allosteric transition from Cis to Trans:

Important Transition Regions · We found that the allosteric transition of PDZ2 is governed by an order-disorder-order principle which is primarily fueled by the very flexible $\beta_2\beta_3$- and $\beta_1\beta_2$-loop. But also the more stable α-helices were not unaffected—in this respect we detected an initial deformation in the α_2-helix once the azobenzene photoswitch mimicked the docking of a ligand by pushing the α_2-helix and the β_2-sheet away. Subsequently, a structural reorganization process in the course of the allosteric transition leads to a highly arranged α_2-helix. Whether this also occurs in the wild-type PDZ2 without photoswitch or rather is an artifact due to the design of the photoswitch, which may introduce too much internal stress, and to what extent this affects the conclusions here, is a question that remains open for further studies.

Allostery · Finally, we to put the findings of our research into perspective on the discussion about allostery. As mentioned at the beginning of this thesis, there is an ongoing debate on whether allostery is either being governed by underlying dynamical or conformational processes. We analyzed vast data on both allosteric equilibrium-conformations (Cis and Trans) as well as the nonequilibrium transition between them. Using a spatially high-resolution method for dimensionality-reduction, namely principal component analyis, we obtained reaction coordinates which well separate Cis and Trans. In both, the free energy landscape and the clustered microstates, we saw distinct differences between the Cis and Trans conformation. Furthermore, we were able to make precise predictions about the allosteric transition by analyzing the nonequilibrium pathways. Also by means of this investigation we were able to identify significant changes in the conformation of the protein in the course of the allosteric transition. On the other hand, the strong disorder in the intermediate states indicates that dynamical processes also play a role and that, although all states can be clearly distinguished by their conformation, particularly nonequilibrium states are subject to strong dynamics. To conclude, it seems difficult to describe allostery exclusively as an clear-cut dynamical or structural process but it rather appears to be an interplay of both.

Outlook

The central pillar of Markov state modeling is the separation of timescales between slow interstate transitions and fast intrastate fluctuations. We saw that artifacts from dimensionality reduction significantly affect the quality of this separation, but with *dynamical iterative coring* we have a powerful tool at hand to correct those artifacts.

We consider the question why coring times are necessary which are higher than the ones suggested by theoretical considerations in order to find and better understand possible drawbacks in our working pipeline. In the end, it all boils down to the same cause. Barriers are overseen which leads to a wrong connectivity of the microstates. Reasons therefore can be manifold. For example, it may be that the projection onto a lower dimensionality merges two different metastable conformations into one microstate—despite being separated by a free energy barrier. If both, originally separated metastable conformations feature substantially different connectivities to a third microstate, this would consequently result in two timescales for the same transition which cannot be covered by MSM. Only by using significantly longer coring times processes occurring on shorter time scales are eliminated and the system consequently looses its memory. While this is necessary to describe the dynamics in terms of a Markov state model, a lot of information on shorter time scales gets lost, even though it was obtained with an immense computational effort. One possible remedy for reducing artifacts introduced by dimensionality reduction could be the application of modern non-linear, machine learning dimensionality reduction methods, such as e.g. VAMPnets [74] or auto-encoder based methods [51, 75]. Similar or identical effects can also result from the clustering process. Recently it was shown that the method used here, namely robust density clustering [49], has problems with clusters which are poorly separated

and partly overlap [76]. In Ref. [76], Westerlund and Delemotte proposed a clustering algorithm called InfleCS which is based on a Gaussian mixture free energy estimator and that is able to distinguish multiple, overlapping clusters with high precision.
Another problem, closely related to dimensionality reduction, is the preselection of the right data which is to be clustered. As mentioned before, one requisite for the construction of a Markov state model is the separation of timescales between the fast intrastate fluctuations and the slow interstate transitions. We argued that these slow interstate motion is resolved by the first few principal components and that fluctuations are aggregated in the remaining principal components. In an ideal world, these essential principal components are unmistakably separated from the rest by vast differences in the decay of their autocorrelation function. However, for PDZ2 the selection of a few principal componentents was not a matter of course and discarded principal components could still contain valuable information. Rodriguez and Laio therefore proposed a method which clusters the data without the preceding reduction of dimensionality [77]. This way it can be ensured that all important data is taken into account.
While most of the methods mentioned here are still in their infancy and require more work to yield valuable results apart from toy models, there is one further point concerning simulations, which could possibly be worth mentioned. We have seen that the conformational change of the azobenzene photoswitch significantly affects the stability of the α_2-helix. Prominently linked across the binding groove, the photoswitch energetically imposes a considerable structural change on PDZ2 upon activation. By the investigation of a slightly different PDZ domain, Petit et al. showed that less invasive modification in N-terminus of the protein reduces the affinity of binding a ligand by 21-fold [29]. Preparing and simulating such a less artificial construct could lead to an improved accordance with the real wild-type system.

To conclude, the here constructed Markov state model is capable of describing the allosteric transition—an intrinsic nonequilibrium process. Nonequilibrium processes drive many of the very fundamental processes that occur in living systems—from gene transcription to photosynthesis—and unraveling them leads to an ever better understanding of life itself. In the words of Feynman, with which we started the thesis:

> *"...if we were to name the most powerful assumption of all, which leads one on and on in an attempt to understand life, it is that all things are made of atoms, and that everything that living things do can be understood in terms of the jigglings and wigglings of atoms."*

A task like tailor-made for molecular dynamics simulations and their analysis—including Markov state models.

A.

Appendix

> *Nature uses only the longest threads to weave her patterns, so each small piece of her fabric reveals the organization of the entire tapestry.*
>
> Richard P. Feynman

Contribution of Single PCs to the Overall Variance

The contribution of some selected PCs to the overall variance in the data set of data generation 1 is shown in the table below. The PCs x_1–x_5 and $x_7$yield an effective description of the dynamics as they contribute with 56.2 % to the overall variance. The remaining 423 PCs cover 43.8% of the variance in the data.

Table A.1.: Data generation 1: Contribution of some PCs (x_i) to the overall variance of the system and the cumulative sum of it.

x_i	λ_i	$\sum_{j=1}^{i} \lambda_j / (\sum_k \lambda_k)$
1	0.2593	0.2593
2	0.0997	0.3590
3	0.0731	0.4315
4	0.0607	0.4928
5	0.0456	0.5284
6	0.0344	0.5628
7	0.0334	0.5962
8	0.0305	0.6267
9	0.0233	0.6500
10	0.0209	0.6709
⋮	⋮	⋮
20	0.0077	0.8008
⋮	⋮	⋮
44	0.0025	0.9004
⋮	⋮	⋮
78	0.0009	0.9503
⋮	⋮	⋮
192	0.0001	0.9900
⋮	⋮	⋮
429	$2.9 \cdot 10^{-6}$	1.0000

Which Contact-Distances to Choose?

$r_{29,94}$ on the left shows an example for an contact-distance which well discriminates the Cis and Trans conformation of the protein and is therefore retained. In contrast, $r_{2,87}$ hardly resolves any differences at all and is therefore discareded.

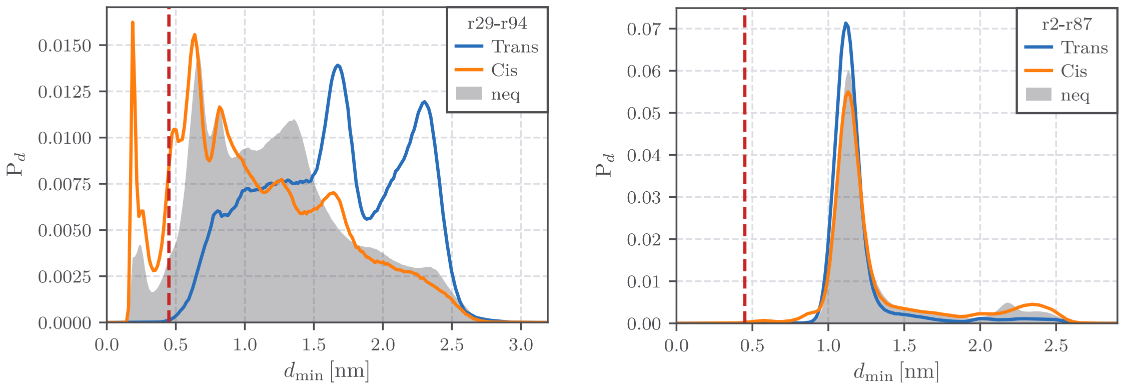

Figure A.1.: Probability distribution for Cis (● orange), Trans (● blue) and nonequilibrium (● grey). *Left:* Example of a distance which clearly shows a major difference between the Cis- and Trans-conformation. *Right:* Example of a distance in which there is hardly no difference between Cis- and Trans-conformation. Therefore, the distance between the residues 29 and 94 on the left is retained while the distance between the residues 2 and 87 on the right is discarded.

$W_1(t)$ for Iterative Coring

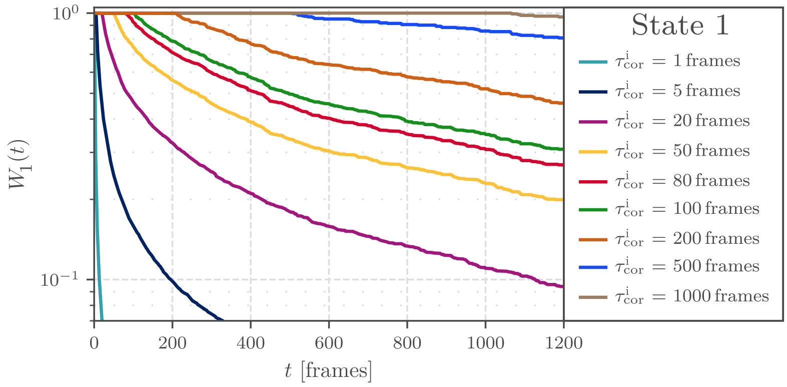

Figure A.2.: State 1: Population probability W_i as a function of time t for different coring times $\tau^{\mathrm{i}}_{\mathrm{cor}}$ (iterative method). In order to remove the strong initial decay, a coring time of $\tau_{\mathrm{cor}} \approx 200$ frames $= 4$ ns seems appropriate. Back-transitions are not included in this plot.

Regions in the PCs Sampled by Single Equilibrium Trajectories

We plotted the equilibrium trajectories (Cis and Trans) individually in order to see whether trajectories of the same conformation sample similar regions. A high overlap (in all dimensions) would indicates converged equilibrium simulations. This one-dimensional representation suggests that trajectories describing the same conformation seem to sample similar values which suggests that equilibrium simulations are largely converged (as mentioned above, this is only a vague suggestion as we are limited to one dimension).

However, we see that the fifth Cis trajectory and the first Trans trajectory cover values in x_1 which are typically covered by trajectories of the respective other conformation. This explains why some of the Cis states are located closely to the Trans cluster in the network representations.

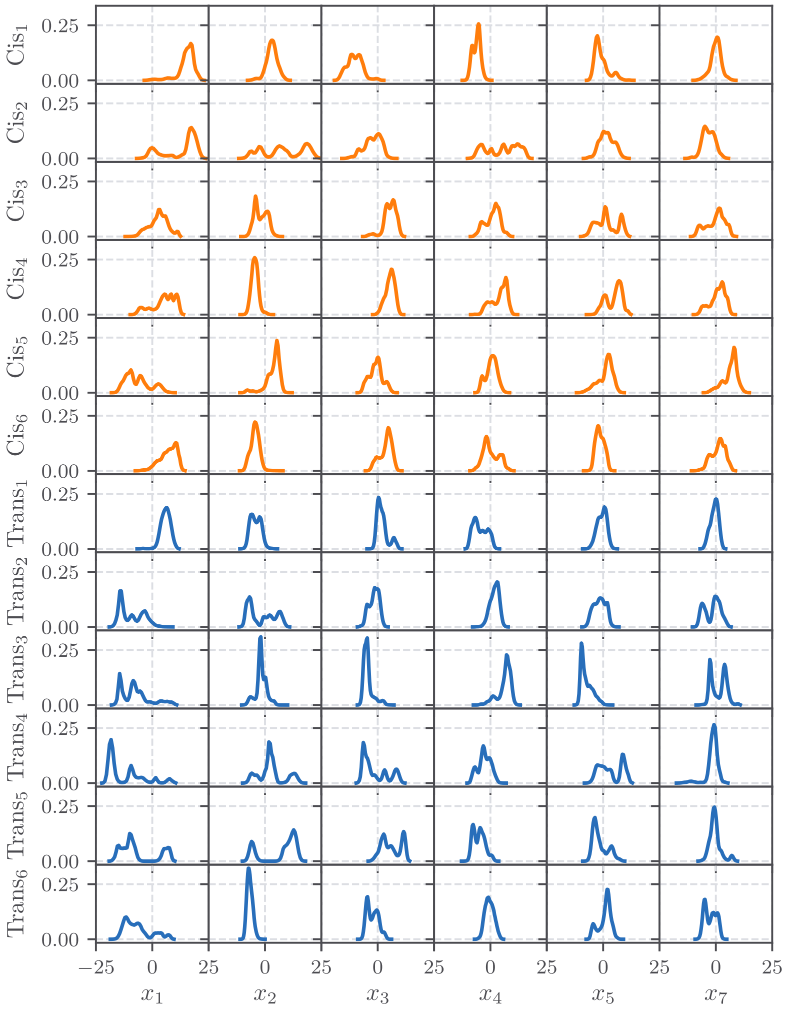

Figure A.3.: Normalized probability distribution of the equilibrium trajectories in PC $x_1 - x_5$ and x_7.

Impact of Coring on the Location of the Microstates

Coring can also shift the location of the microstates in the free energy landscape. Here we exemplary investigated the microstate 1 and 2 and plotted their geometrical position along the PCs x_1–x_5 and x_7. It is evident that the iterative approach respects the geometric location of the microstates more than the classical coring. In particular for state 1, classical coring shifts the position of the microstates considerably.

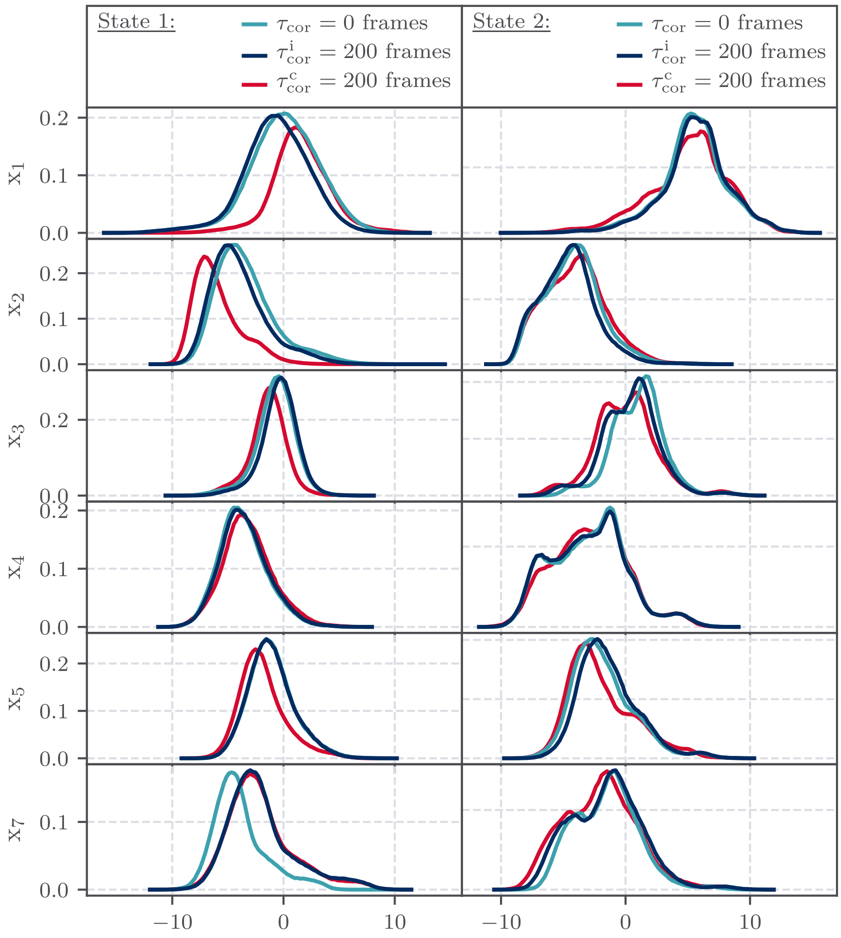

Figure A.4.: Probability density of state 1 (*right*) and state 2 (*left*) along the PCs. ● light blue indicates the location of the state in the uncored trajectory, ● dark blue indicates the location of the state in the iterative cored trajectory and ● red of the classical cored trajectory. The test was performed for a coring time of $\tau^{\mathrm{c/i}}_{\mathrm{cor}} = 200$ frames.

Data Generation 2: On the Selection of the PCs for Clustering

Just like for data generation 1 (compare Sec. 3.3), the same tests were performed for data generation 2 in order to select a suitable set of PCs which is subsequently clustered in order to identify the microstates representing the metastable conformations of the protein. This includes the eigenvalues of the PCs and the resulting cumulative flux (see Fig. A.5), the temporal evolution of the data along the PCs (see A.7), the analysis of the autocorrelation functions (see Fig. A.6) and the free energy projections onto the PCs (see Fig. A.8). The first 6 PCs contain 68 % of the variance present in the data set of data generation 2 and are chosen for clustering as they perform best on the tests performed for selecting a suitable set of PCs.

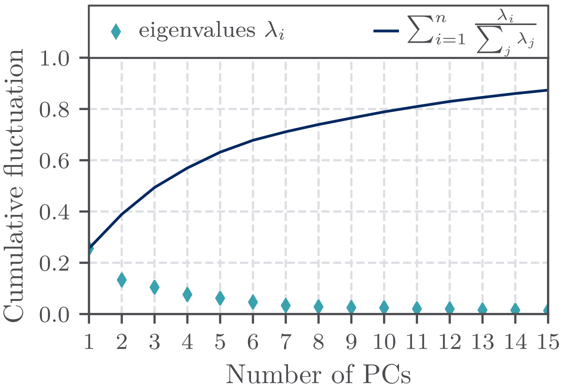

Figure A.5.: Data generation 2: Autocorrelation function for Cis (*left*) and Trans (*right*).

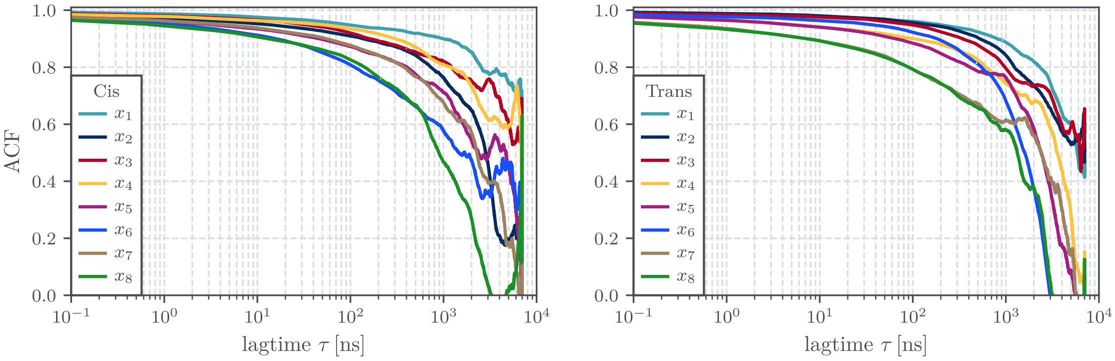

Figure A.6.: Data generation 2: Autocorrelation function for Cis (*left*) and Trans (*right*).

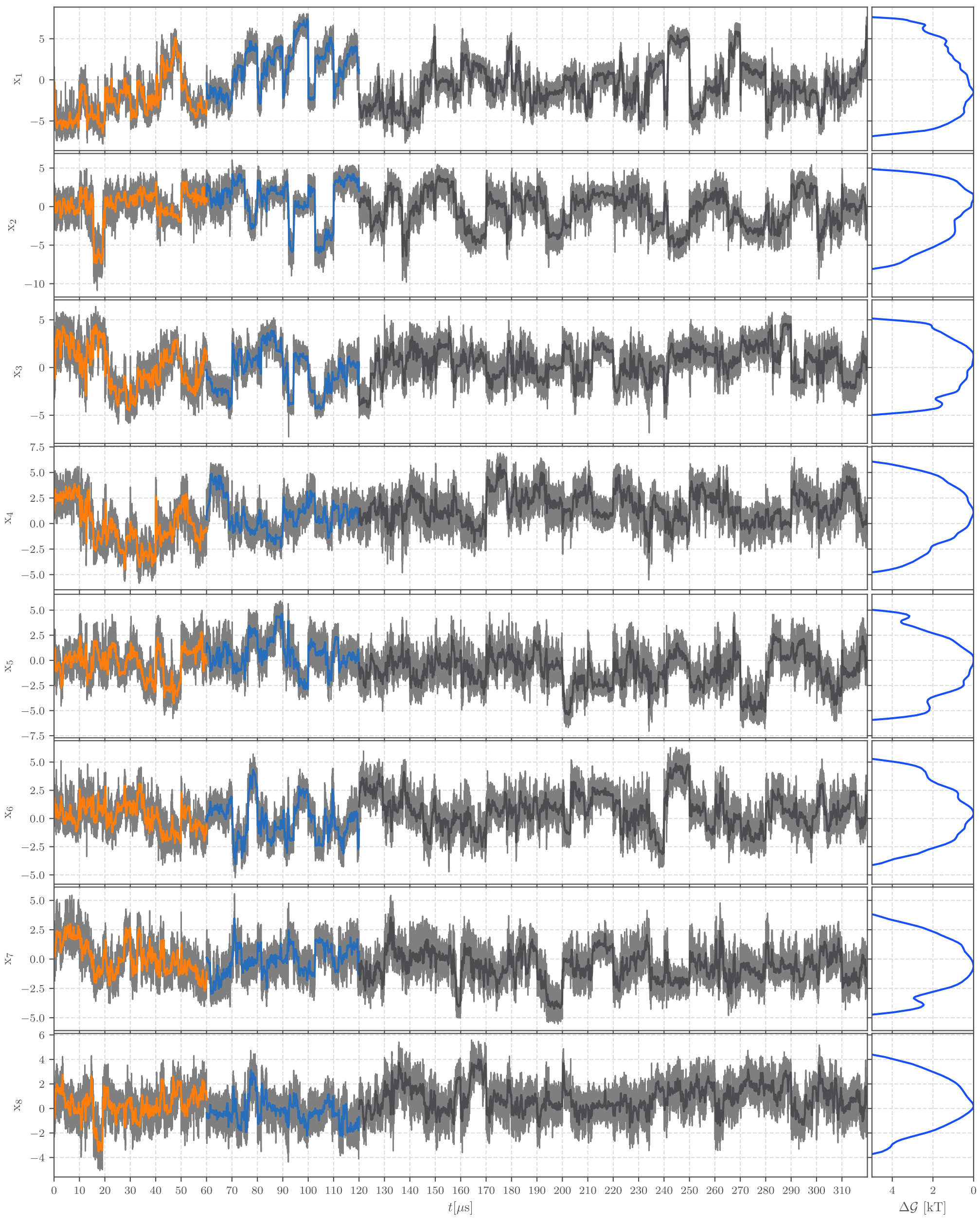

Figure A.7.: Data generation 2: *Left:* Temporal evolution along the first 8 PCs. The first three microseconds of the nonequilibrium data were discarded. Orange ● denotes the symmetric time average over 10^4 Cis frames, blue ● for Trans and grey ● for nonequilibrium frames. *Left:* The free energy projection on the corresponding PC.

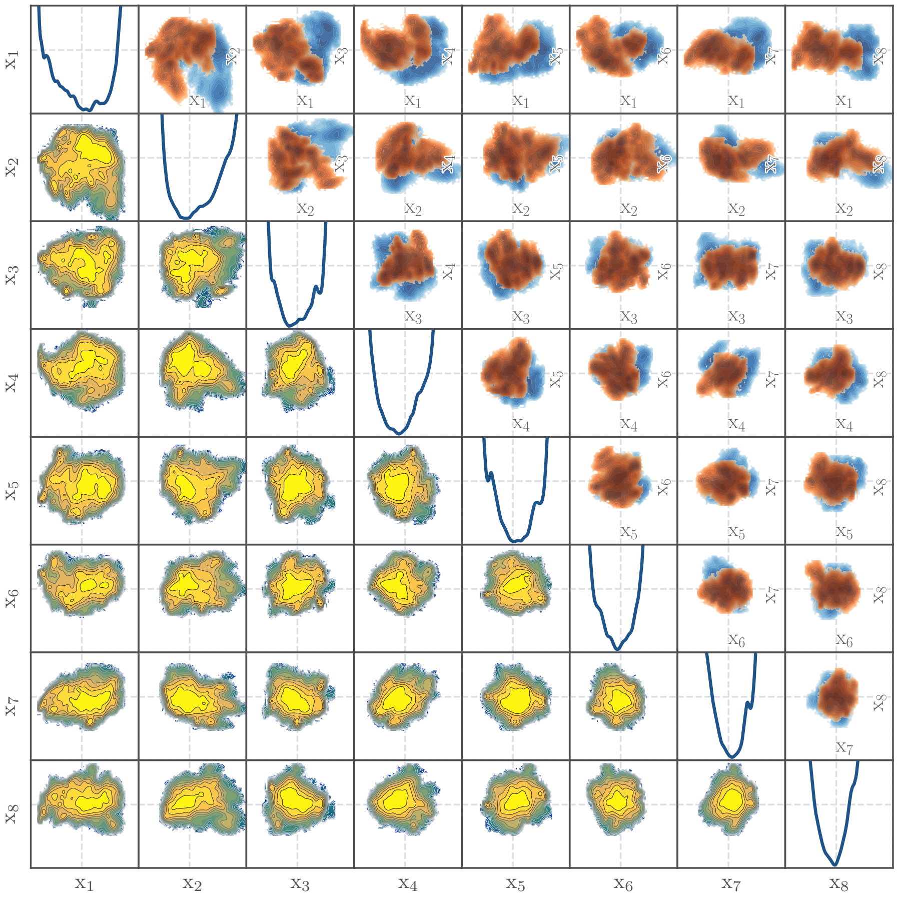

Figure A.8.: Grid representation of the free energy landscape along the PCs x_1–x_6. For a detailed description of this plot, see Fig. 3.6.

Implied Time Scales of the MSM Based on Data Generation 2

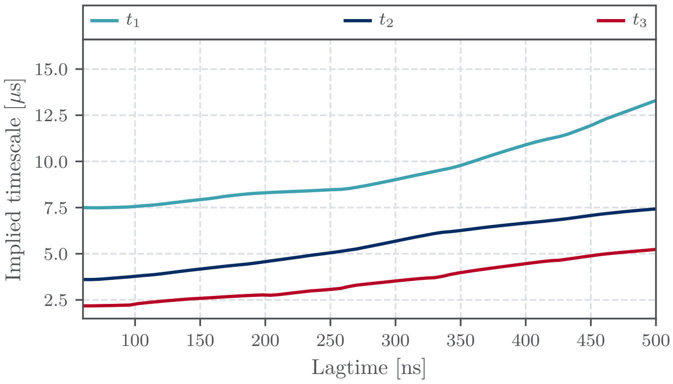

Figure A.9.: Data generation 2: The first three implied timescales as a function of the lag time τ_{lag} of the transition matrix.

Table A.2.: Data generation 2: Population of states with a final coring time of $\tau^{i}_{\text{cor}} = 3000$ frames. States were ordered according to their population in the trajectory after coring. Coloring: ● Cis, ● Trans, ● Neq and ● ambivalent.

MS_i	Pop [%]	$\sum_j^i \text{Pop}_j$	% Neq	% Cis	% Trans	in	out
1	8.62	8.6	92.1	7.9	0.0	32	34
2	7.66	16.3	91.8	0.0	8.2	55	54
3	6.54	22.8	86.0	0.0	14.0	38	65
4	4.96	27.8	93.7	6.3	0.0	54	52
5	3.79	31.6	100.0	0.0	0.0	17	17
6	3.53	35.1	99.3	0.0	0.7	39	34
7	3.46	38.6	100.0	0.0	0.0	17	14
8	3.32	41.9	100.0	0.0	0.0	28	29
9	2.74	44.6	100.0	0.0	0.0	37	36
10	2.70	47.3	100.0	0.0	0.0	21	18
11	2.67	50.0	96.5	3.5	0.0	36	35
12	2.58	52.6	100.0	0.0	0.0	19	21
13	2.56	55.1	29.7	0.0	70.3	10	10
14	2.18	57.3	57.2	42.8	0.0	21	19
15	2.18	59.5	100.0	0.0	0.0	13	21
16	2.12	61.6	100.0	0.0	0.0	8	8
17	2.06	63.7	99.0	0.0	1.0	23	22
18	2.06	65.7	83.6	16.4	0.0	30	23
19	1.99	67.7	4.2	95.8	0.0	9	10
20	1.96	69.7	69.3	30.7	0.0	17	12
21	1.94	71.6	98.4	0.0	1.6	4	3
22	1.87	73.5	98.1	0.0	1.9	12	11
23	1.76	75.2	97.4	0.0	2.6	17	19
24	1.64	76.9	31.6	0.0	68.4	6	7
25	1.53	78.4	45.9	54.1	0.0	16	16
26	1.42	79.8	100.0	0.0	0.0	12	9
27	1.42	81.2	85.0	3.2	11.7	5	5
28	1.34	82.6	2.4	0.0	97.6	2	4
29	1.25	83.8	11.7	0.0	88.3	3	3
30	1.21	85.0	6.2	93.8	0.0	8	8
31	1.11	86.1	47.1	0.0	52.9	5	4
32	1.08	87.2	89.4	10.6	0.0	14	10
33	1.03	88.3	11.5	0.0	88.5	3	2
34	0.98	89.2	96.0	4.0	0.0	5	5
35	0.98	90.2	0.0	100.0	0.0	6	5
36	0.91	91.1	16.1	0.0	83.9	2	2
37	0.84	92.0	45.0	55.0	0.0	7	7
38	0.82	92.8	8.5	0.0	91.5	6	4
39	0.73	93.5	14.1	85.9	0.0	8	7
40	0.71	94.2	100.0	0.0	0.0	5	4
41	0.70	94.9	15.1	38.8	46.1	6	8
42	0.69	95.6	43.3	56.7	0.0	7	5
43	0.67	96.3	36.2	0.0	63.8	4	4
44	0.58	96.9	0.0	100.0	0.0	4	5
45	0.55	97.4	100.0	0.0	0.0	7	7
46	0.51	97.9	0.0	100.0	0.0	6	6
47	0.47	98.4	6.8	93.2	0.0	1	1
48	0.42	98.8	100.0	0.0	0.0	5	5
49	0.41	99.2	100.0	0.0	0.0	3	3
50	0.37	99.6	100.0	0.0	0.0	4	3
51	0.25	99.9	14.2	0.0	85.8	2	3
52	0.14	100.0	0.0	0.0	100.0	1	1

Data Generation 2: Order and Disorder of the Microstates

Unlumped Model

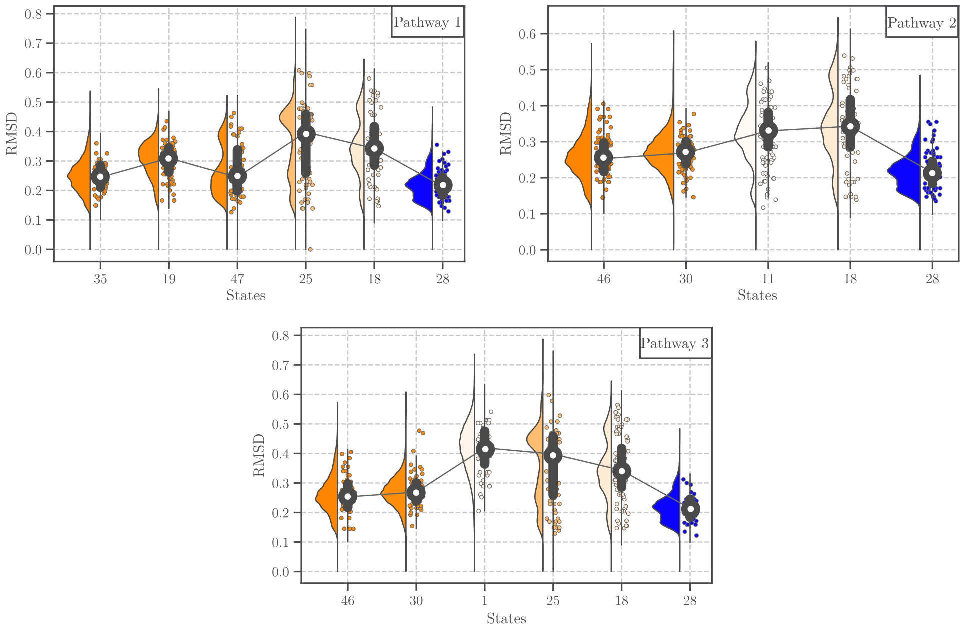

Figure A.10.: Microstate trajectory of data generation 2, not lumped. The three most important pathways predicted by the MSM and a MCMC simulation. Similar to Fig. 4.14—where one can find a detailed description of the plot—we can observe an order-disorder-order behaviour. In pathway 1, we can see that state 47 is more ordered than its predecessor. However, this is due to the fact that this state itself is a very pure Cis state. Only the very high requirements for initial states in the MCMC, i.e. a relative Cis population of 100% causes the states 35 and 19 to precede state 47.

Lumped Model

The 10 most populated microstates depicted here correspond to the lumped microstate trajectory which is described in Sec. 5.6.

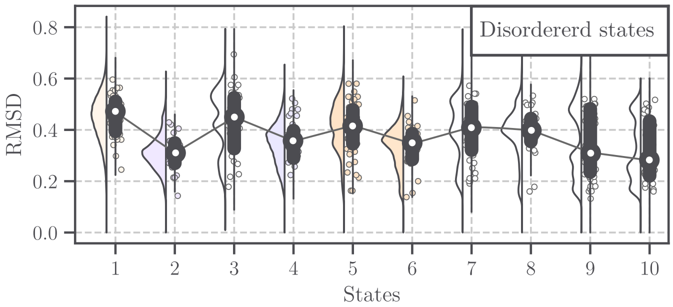

Figure A.11.: By dynamical lumping coarse grained microstate trajectory: The 10 most populated microstates and their conformational heterogeneity.

Pseudocode Dynamical Lumping

Algorithm 1: Dynamical Lumping for the lumping of multiple microstates according to their kinetic connectivity.

```
function dynlumping (trajectories, $\tau_{\text{lag}}$, $q_{\text{cut}}$);
Input    : trajectories, lag time $\tau_{\text{lag}}$ and cutoff value $q_{\text{cut}}$
Output : Lumped microstate-trajectory
 Find all states in microstates trajectory
Microstates: unique elements in all trajectories → {states} $\equiv \Omega$
Determine transition matrix $T_{i,j}$
Transition matrix: $T_{i,j}\ \forall i, j \in \Omega$
Loop through various $q_{min}$ values
while $q_{min} \in [0, 1] \leq q_{cut}$ do
    for $\tilde{i} \in \{states\}$ do
        if $q_{min} < T_{\tilde{i},\tilde{i}}$ then
            continue
        else
            Sort $T_{\tilde{i},j}$ in descending order
            for $\tilde{j} \in \{j \mid \max(\{T_{\tilde{i},j} \mid j \in \Omega\}), ..., \min(\{T_{\tilde{i},j} \mid j \in \Omega\})\}$ do
                Rule out wrong assignment (Self transition)
                if $\tilde{i} = \tilde{j}$ then
                    continue
                Rule out wrong assignment
                else if $q_{min} < T_{\tilde{j},\tilde{j}}$ then
                    continue
                Assign $\tilde{i}$ to $\tilde{j}$
                else
                    $\tilde{i} \to \tilde{j}$
```

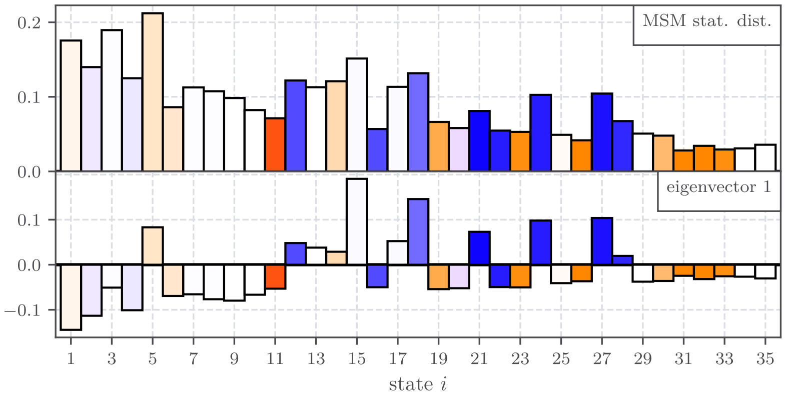

Figure A.12.: *Top:* Stationary distribution predicted by the MSM which was constructed onto the lumped microstate trajectory. *Bottom:* First eigenvector indicating the slowest process in the system.

Table A.3.: Population of state which resulted from the dynamical lumping process. States were ordered according to their population in the trajectory after lumping. Coloring: ● Cis, ● Trans, ● Neq and ● ambivalent.

MS_i	Pop [%]	$\sum_j^i \text{Pop}_j$	% Neq	% Cis	% Trans	in	out
1	13.58	13.6	92.7	7.3	0.0	65	65
2	11.95	25.5	91.6	0.8	7.6	58	83
3	8.88	34.4	99.5	0.0	0.5	41	44
4	7.66	42.1	91.8	0.0	8.2	55	54
5	6.09	48.2	79.0	21.0	0.0	44	30
6	4.78	52.9	80.5	19.5	0.0	18	24
7	4.64	57.6	99.6	0.0	0.4	34	35
8	4.07	61.7	99.4	0.0	0.6	42	37
9	3.46	65.1	100.0	0.0	0.0	17	14
10	2.70	67.8	100.0	0.0	0.0	21	18
11	2.69	70.5	7.1	81.0	12.0	10	13
12	2.56	73.1	29.7	0.0	70.3	10	10
13	2.12	75.2	100.0	0.0	0.0	8	8
14	1.96	77.1	69.3	30.7	0.0	17	12
15	1.94	79.1	98.4	0.0	1.6	4	3
16	1.88	81.0	29.3	0.0	70.7	5	7
17	1.87	82.8	98.1	0.0	1.9	12	11
18	1.78	84.6	43.0	0.0	57.0	6	5
19	1.58	86.2	30.6	69.4	0.0	11	10
20	1.42	87.6	85.0	3.2	11.7	5	5
21	1.34	88.9	2.4	0.0	97.6	2	4
22	1.25	90.2	11.7	0.0	88.3	3	3
23	1.21	91.4	6.2	93.8	0.0	8	8
24	1.03	92.4	11.5	0.0	88.5	3	2
25	0.98	93.4	96.0	4.0	0.0	5	5
26	0.98	94.4	0.0	100.0	0.0	6	5
27	0.95	95.3	7.3	0.0	92.7	5	3
28	0.91	96.3	16.1	0.0	83.9	2	2
29	0.71	97.0	100.0	0.0	0.0	5	4
30	0.69	97.7	43.3	56.7	0.0	7	5
31	0.58	98.2	0.0	100.0	0.0	4	5
32	0.51	98.7	0.0	100.0	0.0	6	6
33	0.47	99.2	6.8	93.2	0.0	1	1
34	0.41	99.6	100.0	0.0	0.0	3	3
35	0.37	100.0	100.0	0.0	0.0	4	3

B.

Erklärung

Hiermit versichere ich, dass ich die eingereichte Masterarbeit selbständig verfasst habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommenen Inhalte als solche kenntlich gemacht. Weiter versichere ich, dass die eingereichte Masterarbeit weder vollständig noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahrens war oder ist.

Ort und Datum

Georg Gabriel Diez

List of Figures

List of Tables

Bibliography

1. J. C. Kendrew, G. Bodo, H. M. Dintzis, R. Parrish, H. Wyckoff, and D. C. Phillips. “A three-dimensional model of the myoglobin molecule obtained by x-ray analysis”. *Nature* 181:4610, 1958, pp. 662–666.
2. A. Platt, H. C. Ross, S. Hankin, and R. J. Reece. “The insertion of two amino acids into a transcriptional inducer converts it into a galactokinase”. *Proceedings of the National Academy of Sciences* 97:7, 2000, pp. 3154–3159.
3. K. Dill, R. L. Jernigan, and I. Bahar. *Protein actions: Principles and modeling*. Garland Science, 2017.
4. C. I. Branden and J. Tooze. *Introduction to protein structure*. Garland Science, 2012.
5. C. M. Dobson, A. Šali, and M. Karplus. “Protein folding: a perspective from theory and experiment”. *Angewandte Chemie International Edition* 37:7, 1998, pp. 868–893.
6. K. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl. “The protein folding problem”. *Annu. Rev. Biophys.* 37, 2008, pp. 289–316.
7. K. F. Winklhofer, J. Tatzelt, and C. Haass. “The two faces of protein misfolding: gain-and loss-of-function in neurodegenerative diseases”. *The EMBO journal* 27:2, 2008, pp. 336–349.
8. M. D. Scott and J. Frydman. “Aberrant protein folding as the molecular basis of cancer”. In: *Protein misfolding and disease*. Springer, 2003, pp. 67–76.
9. S. M. Larson, C. D. Snow, M. Shirts, and V. S. Pande. “Folding@ Home and Genome@ Home: Using distributed computing to tackle previously intractable problems in computational biology”. *arXiv preprint arXiv:0901.0866*, 2009.
10. G. Stock and P. Hamm. “A non-equilibrium approach to allosteric communication”. *Philosophical Transactions of the Royal Society B: Biological Sciences* 373:1749, 2018, p. 20170187.
11. J. Liu and R. Nussinov. “Allostery: an overview of its history, concepts, methods, and applications”. *PLoS computational biology* 12:6, 2016.
12. R. Nussinov and C.-J. Tsai. “Allostery in disease and in drug discovery”. *Cell* 153:2, 2013, pp. 293–305.

13. A. Cooper and D. Dryden. “Allostery without conformational change”. *European Biophysics Journal* 11:2, 1984, pp. 103–109.

14. T. C. McLeish, T. Rodgers, and M. R. Wilson. “Allostery without conformation change: modelling protein dynamics at multiple scales”. *Physical biology* 10:5, 2013, p. 056004.

15. R. Nussinov and C.-J. Tsai. “Allostery without a conformational change? Revisiting the paradigm”. *Current opinion in structural biology* 30, 2015, pp. 17–24.

16. B. Buchli, S. A. Waldauer, R. Walser, M. L. Donten, R. Pfister, N. Blöchliger, S. Steiner, A. Caflisch, O. Zerbe, and P. Hamm. “Kinetic response of a photoperturbed allosteric protein”. *Proceedings of the National Academy of Sciences* 110:29, 2013, pp. 11725–11730.

17. S. Buchenberg, V. Knecht, R. Walser, P. Hamm, and G. Stock. “Long-range conformational transition of a photoswitchable allosteric protein: molecular dynamics simulation study”. *The Journal of Physical Chemistry B* 118:47, 2014, pp. 13468–13476.

18. G. R. Bowman, V. S. Pande, and F. Noé. *An introduction to Markov state models and their application to long timescale molecular simulation*. Vol. 797. Springer Science & Business Media, 2013.

19. J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé. “Markov models of molecular kinetics: Generation and validation”. *The Journal of chemical physics* 134:17, 2011, p. 174105.

20. J. D. Chodera and F. Noé. “Markov state models of biomolecular conformational dynamics”. *Current opinion in structural biology* 25, 2014, pp. 135–144.

21. F. Noé and E. Rosta. *Markov Models of Molecular Kinetics*. 2019.

22. F. Nüske, H. Wu, J.-H. Prinz, C. Wehmeyer, C. Clementi, and F. Noé. “Markov state models from short non-equilibrium simulations—Analysis and correction of estimation bias”. *The Journal of Chemical Physics* 146:9, 2017, p. 094104.

23. B. Reuter, K. Fackeldey, and M. Weber. “Generalized Markov modeling of nonreversible molecular kinetics”. *The Journal of chemical physics* 150:17, 2019, p. 174103.

24. A. Jain and G. Stock. “Hierarchical folding free energy landscape of HP35 revealed by most probable path clustering”. *The Journal of Physical Chemistry B* 118:28, 2014, pp. 7750–7760.

25. F. Sittel and G. Stock. “Perspective: Identification of collective variables and metastable states of protein dynamics”. *The Journal of Chemical Physics* 149:15, 2018, p. 150901.

26. D. Nagel, A. Weber, B. Lickert, and G. Stock. “Dynamical coring of Markov state models”. *The Journal of chemical physics* 150:9, 2019, p. 094111.

27. C.-J. Tsai, A. Del Sol, and R. Nussinov. "Allostery: absence of a change in shape does not imply that allostery is not at play". *Journal of molecular biology* 378:1, 2008, pp. 1–11.

28. B. Z. Harris and W. A. Lim. "Mechanism and role of PDZ domains in signaling complex assembly". *Journal of cell science* 114:18, 2001, pp. 3219–3231.

29. C. M. Petit, J. Zhang, P. J. Sapienza, E. J. Fuentes, and A. L. Lee. "Hidden dynamic allostery in a PDZ domain". *Proceedings of the National Academy of Sciences* 106:43, 2009, pp. 18249–18254.

30. C. N. Chi, A. Bach, K. Strømgaard, S. Gianni, and P. Jemth. "Ligand binding by PDZ domains". *Biofactors* 38:5, 2012, pp. 338–348.

31. J. Zhang, P. J. Sapienza, H. Ke, A. Chang, S. R. Hengel, H. Wang, G. N. Phillips Jr, and A. L. Lee. "Crystallographic and nuclear magnetic resonance evaluation of the impact of peptide binding to the second PDZ domain of protein tyrosine phosphatase 1E". *Biochemistry* 49:43, 2010, pp. 9280–9291.

32. M. Karplus. *Molecular dynamics simulations of biomolecules*. 2002.

33. S. A. Adcock and J. A. McCammon. "Molecular dynamics: survey of methods for simulating the activity of proteins". *Chemical reviews* 106:5, 2006, pp. 1589–1615.

34. D. E. Shaw, M. M. Deneroff, R. O. Dror, J. S. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson, K. J. Bowers, J. C. Chao, et al. "Anton, a special-purpose machine for molecular dynamics simulation". *Communications of the ACM* 51:7, 2008, pp. 91–97.

35. W. F. van Gunsteren, S. Billeter, A. Eising, P. Hünenberger, P. Krüger, A. Mark, W. Scott, and I. Tironi. "Biomolecular simulation: the GROMOS96 manual and user guide". *Vdf Hochschulverlag AG an der ETH Zürich, Zürich* 86, 1996.

36. D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. Berendsen. "GROMACS: fast, flexible, and free". *Journal of computational chemistry* 26:16, 2005, pp. 1701–1718.

37. W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. "A second generation force field for the simulation of proteins, nucleic acids, and organic molecules". *Journal of the American Chemical Society* 117:19, 1995, pp. 5179–5197.

38. V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. "Comparison of multiple Amber force fields and development of improved protein backbone parameters". *Proteins: Structure, Function, and Bioinformatics* 65:3, 2006, pp. 712–725.

39. R. B. Best and G. Hummer. "Optimized molecular dynamics force fields applied to the helix- coil transition of polypeptides". *The journal of physical chemistry B* 113:26, 2009, pp. 9004–9015.

40. K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw. "Improved side-chain torsion potentials for the Amber ff99SB protein force field". *Proteins: Structure, Function, and Bioinformatics* 78:8, 2010, pp. 1950–1958.

41. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. "Comparison of simple potential functions for simulating liquid water". *The Journal of chemical physics* 79:2, 1983, pp. 926–935.

42. S. Buchenberg, F. Sittel, and G. Stock. "Time-resolved observation of protein allosteric communication". *Proceedings of the National Academy of Sciences* 114:33, 2017, E6804–E6811.

43. M. Ernst, F. Sittel, and G. Stock. "Contact-and distance-based principal component analysis of protein dynamics". *The Journal of chemical physics* 143:24, 2015, 12B640_1.

44. R. B. Best, G. Hummer, and W. A. Eaton. "Native contacts determine protein folding mechanisms in atomistic simulations". *Proceedings of the National Academy of Sciences* 110:44, 2013, pp. 17874–17879.

45. S. Brandt, F. Sittel, M. Ernst, and G. Stock. "Machine learning of biomolecular reaction coordinates". *The journal of physical chemistry letters* 9:9, 2018, pp. 2144–2150.

46. T. Zhou and A. Caflisch. "Distribution of reciprocal of interatomic distances: A fast structural metric". *Journal of chemical theory and computation* 8:8, 2012, pp. 2930–2937.

47. Y. Mu, P. H. Nguyen, and G. Stock. "Energy landscape of a small peptide revealed by dihedral angle principal component analysis". *Proteins: Structure, Function, and Bioinformatics* 58:1, 2005, pp. 45–52.

48. A. Altis, M. Otten, P. H. Nguyen, R. Hegger, and G. Stock. "Construction of the free energy landscape of biomolecules via dihedral angle principal component analysis". *The Journal of chemical physics* 128:24, 2008, 06B620.

49. F. Sittel and G. Stock. "Robust density-based clustering to identify metastable conformational states of proteins". *Journal of chemical theory and computation* 12:5, 2016, pp. 2426–2435.

50. G. E. Hinton and R. R. Salakhutdinov. "Reducing the dimensionality of data with neural networks". *Science* 313:5786, 2006, pp. 504–507.

51. T. Lemke and C. Peter. "Encodermap: Dimensionality reduction and generation of molecule conformations". *Journal of chemical theory and computation* 15:2, 2019, pp. 1209–1215.

52. S. Wold, K. Esbensen, and P. Geladi. "Principal component analysis". *Chemometrics and intelligent laboratory systems* 2:1-3, 1987, pp. 37–52.

53. F. Sittel, T. Filk, and G. Stock. "Principal component analysis on a torus: Theory and application to protein dynamics". *The Journal of chemical physics* 147:24, 2017, p. 244101.

54. A. Altis, P. H. Nguyen, R. Hegger, and G. Stock. “Dihedral angle principal component analysis of molecular dynamics simulations”. *The Journal of chemical physics* 126:24, 2007, p. 244111.

55. A. Hyvärinen and E. Oja. “Independent component analysis: algorithms and applications”. *Neural networks* 13:4-5, 2000, pp. 411–430.

56. E. Facco, M. d’Errico, A. Rodriguez, and A. Laio. “Estimating the intrinsic dimension of datasets by a minimal neighborhood information”. *Scientific reports* 7:1, 2017, pp. 1–8.

57. R. Hegger, A. Altis, P. H. Nguyen, and G. Stock. “How complex is the dynamics of peptide folding?” *Physical review letters* 98:2, 2007, p. 028102.

58. C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard. “A direct approach to conformational dynamics based on hybrid Monte Carlo”. *Journal of Computational Physics* 151:1, 1999, pp. 146–168.

59. J. A. Hartigan and M. A. Wong. “Algorithm AS 136: A k-means clustering algorithm”. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28:1, 1979, pp. 100–108.

60. H. Frauenfelder, S. G. Sligar, and P. G. Wolynes. “The energy landscapes and motions of proteins”. *Science* 254:5038, 1991, pp. 1598–1603.

61. N.-V. Buchete and G. Hummer. “Coarse master equations for peptide folding dynamics”. *The Journal of Physical Chemistry B* 112:19, 2008, pp. 6057–6069.

62. K. Dill and S. Bromberg. *Molecular driving forces: statistical thermodynamics in biology, chemistry, physics, and nanoscience*. Garland Science, 2012.

63. W. C. Swope, J. W. Pitera, and F. Suits. “Describing protein folding kinetics by molecular dynamics simulations.” *The Journal of Physical Chemistry B* 108:21, 2004, pp. 6571–6581.

64. F. Sittel. “The secret life of proteins: Exploring molecular dynamics through statistics”. PhD thesis. Albert-Ludwigs-Universität Freiburg, 2018.

65. S. Ohnemus. *Markov State Modeling of a Photoswitchable PDZ2 Domain*. Albert-Ludwigs-Universität Freiburg, Bachelor's Thesis. 2018.

66. A. Weber. *Markov State Modeling of an Allosteric Transition*. Albert-Ludwigs-Universität Freiburg, Master's Thesis. 2019.

67. M. Ernst, S. Wolf, and G. Stock. “Identification and validation of reaction coordinates describing protein functional motion: Hierarchical dynamics of T4 lysozyme”. *Journal of chemical theory and computation* 13:10, 2017, pp. 5076–5088.

68. N. Kawashima, J. Gubernatis, and H. Evertz. “Loop algorithms for quantum simulations of fermion models on lattices”. *Physical Review B* 50:1, 1994, p. 136.

69. D. Bartle. *Sampling and Equilibration of Biomolecular Simulations*. Albert-Ludwigs-Universität Freiburg, Bachelor's Thesis. 2019.

70. M. Jacomy, T. Venturini, S. Heymann, and M. Bastian. “ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software”. *PloS one* 9:6, 2014.

71. D. Nagel, A. Weber, and G. Stock. “Identification of pathways in Markov state models”. *Manuscript in preparation*.

72. T. Chen and C. Guestrin. “Xgboost: A scalable tree boosting system”. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 785–794.

73. A. Jain and G. Stock. “Identifying metastable states of folding proteins”. *Journal of chemical theory and computation* 8:10, 2012, pp. 3810–3819.

74. A. Mardt, L. Pasquali, H. Wu, and F. Noé. “VAMPnets for deep learning of molecular kinetics”. *Nature communications* 9:1, 2018, pp. 1–11.

75. C. Song, F. Liu, Y. Huang, L. Wang, and T. Tan. “Auto-encoder based data clustering”. In: *Iberoamerican Congress on Pattern Recognition*. Springer. 2013, pp. 117–124.

76. A. M. Westerlund and L. Delemotte. “InfleCS: Clustering Free Energy Landscapes with Gaussian Mixtures”. *Journal of chemical theory and computation* 15:12, 2019, pp. 6752–6759.

77. A. Rodriguez and A. Laio. “Clustering by fast search and find of density peaks”. *Science* 344:6191, 2014, pp. 1492–1496.