Master's Thesis

Modeling the Non-Equilibrium Allosteric Response of Protein Domains

Fabian Rudolf

Supervisor
Prof. Dr. Gerhard Stock

Albert-Ludwigs-Universität Freiburg
Faculty of Mathematics and Physics
Institute of Physics
May 2025

universität freiburg

Erstgutachter	Prof. Dr. Gerhard Stock
Zweitgutachterin	Prof. Dr. Tanja Schilling
Datum	02.05.2025

Document Setup and Tools

This thesis was typeset with LaTeX, using the *Clean Thesis* style developed by Ricardo Langner. Analyses were mostly performed using Python 3, with essential libraries including NumPy, SciPy, and Matplotlib for data processing, statistical analysis, and visualization. Large language models, such as DeepL, were used as an aid in text production and translation. The protein visualizations were created using PyMOL.

Abstract

Allosteric regulation is a vital process in living organisms, enabling communication between distant regions in proteins. Despite its importance, molecular mechanisms underlying these processes remain unclear. The protein domains of the PDZ family are ideal model systems for studying allostery, as they are known for exhibiting allosteric transitions and are small enough to be studied through molecular dynamics simulations. In this work the second PDZ domain with an azobenzene photoswitch – designed to mimic ligand binding and trigger allosteric transitions – is studied. The analysis is based on a molecular dynamics simulation dataset from previous studies in the Stock group at the University of Freiburg. This work focuses on refining feature selection and identifies localized groups of dynamically collaborating contacts, referred to as contact clusters. A timescale analysis based on these clusters provides a mechanistic explanation for the experimentally observed dynamic timescales, spanning nanoseconds to microseconds. The opening of the binding pocket on a nanosecond timescale triggers further structural changes, including the unwinding of a helix distant from the binding site on a microsecond timescale – a change that can be considered an allosteric response. Furthermore, a comparison with three similar photoswitched PDZ systems shows that certain protein regions are dynamically important across the systems, even though the systems differ in their sequence and structure. These shared dynamic regions may represent key elements of more general allosteric pathways, providing insights into the broader principles of protein allostery.

Contents

1 Introduction **1**

2 Theory and Methods **5**
2.1 Systems: PDZ Domains 5
2.2 Molecular Dynamics Simulations 6
2.3 Simulation Details 8
2.3.1 PDZ2S 8
2.3.2 Other PDZ Systems 9
2.4 Internal Coordinates 9
2.5 MoSAIC Feature Selection 10
2.6 Principal Component Analysis 11
2.7 Timescale Analysis: Dynamic Content 13
2.8 Root Mean Square Deviation 14
2.9 Free Energy 14

3 PDZ2S: Data Processing and Assessment **16**
3.1 Available Datasets 16
3.2 Collective Variables: Internal Coordinates 17
3.2.1 Error Measures 18
3.3 Assessing Data Quality 19
3.3.1 Equilibrium Trajectories 19
3.3.2 Stability of Secondary Structures 21
3.4 Feature Selection: MoSAIC 22
3.4.1 Similarity Measure 22
3.4.2 MoSAIC Results: Cluster Identification 26
3.4.3 MoSAIC Results: Discussion 28
3.5 Resulting Datasets 29

4 Non-Equilibrium Response of PDZ2S **30**

4.1 Non-Equilibrium Photoinduced Response of Contact Clusters 30
4.1.1 Local Response to Photoswitching 32
4.1.2 Long Distance Response 34
4.1.3 Timescale Analysis 38
4.2 Free Energy Landscape Analysis of the Cis to Trans Transition 44
4.2.1 Principal Component Analysis 44
4.2.2 Defining Cis and Trans in Equilibrium 45
4.2.3 Non-Equilibrium Cis to Trans Transition 49

5 Comparison of Various Members of the PDZ Family **51**
5.1 Systems at Hand 51
5.1.1 Dataset Description 53
5.2 Comparison of Contact Clusters in PDZ Systems 53
5.2.1 Consistency of MoSAIC Clusters Across Systems 53
5.2.2 Cluster Comparisons Across PDZ Systems 54
5.2.3 Interpretation and Conclusion 56
5.3 Comparison of Dynamics in PDZ Systems 57
5.3.1 Introduction and Methodological Notes 57
5.3.2 PDZ3 Comparison (PDZ3L5 vs. PDZ3L6) 58
5.3.3 PDZ2 Comparison (PDZ2S vs. PDZ2L) 59
5.3.4 Discussion of Similarity in Dynamics 60

6 Conclusion and Outlook **61**

Bibliography **65**

A Appendix **71**
A.1 Concerning Chapter 3 71
A.2 Concerning Chapter 4 72
A.3 Concerning Chapter 5 74

1 Introduction

> *Allostery is the second secret of life*
>
> **— Jacques Monod**

Proteins as Fundamental Biomolecules

Proteins are fundamental to nearly all biological processes, serving, among others, as catalysts (e.g. enzymes), signal transducers (e.g. receptors), and structural components (e.g. the cytoskeleton) [1]. Proteins are polymeric molecules, consisting of amino acids linked in a chain. The composition of this chain defines a protein's primary structure. In these chains, recurring structural building blocks – like helices and sheets – are formed, which are called secondary structures. Further, the amino acid chains fold into complex 3D structures in its so-called tertiary structure consisting of the recurring secondary structures. The interplay of several proteins bonded non-covalently leads to the formation of functional complexes. This is referred to as a protein's quaternary structure (illustrated in Figure 1.1)[1]. It has long been believed that when the primary structure of the protein is known, the 3D structure and with this the function of the protein would be defined. This view was supported by early milestones such as the first fully resolved protein structure, that of myoglobin, studied by Kendrew et al. [2] using X-ray crystallography and NMR spectroscopy. By now, atomically detailed structures of more than 100,000 proteins are known. While a protein's structure undeniably plays an important role in its functionality, a protein is not confined to having just one structure. In fact, the dynamics of a protein play an equally important role in defining a protein's function [1]. Understanding the dynamics will be a major goal of this work.

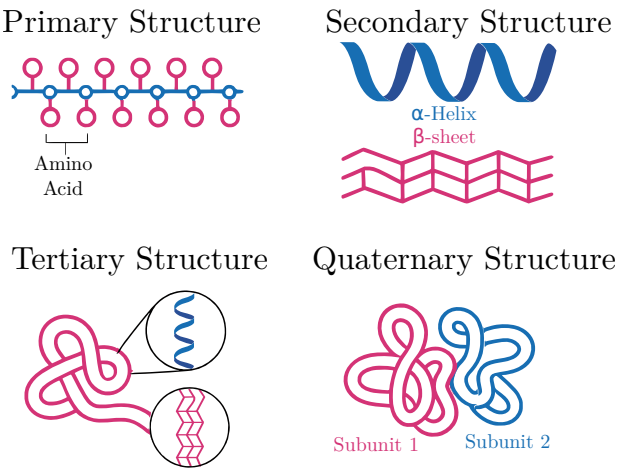

Fig. 1.1.: The figure shows the four levels of protein structure. A protein's primary structure is its amino acid chain. The secondary structure is given with stable arrangements of the chain in the form of e.g. helices or sheets. The tertiary structure describes the folded conformation and the quaternary structure describes the functional interplay of several proteins [1].

Allostery: A Key Mechanism of Protein Function

A key mechanism of proteins is seen in allostery: Allostery describes the processes in which a ligand binding to one site of a protein creates a conformational change at a distant site of the same protein, as illustrated in Figure 1.2. This allows for the regulation of an activity [3]. This is, for example, seen in the classical example of hemoglobin, which carries the oxygen in blood. In total, one hemoglobin protein is able to bind up to four oxygen molecules. When one oxygen molecule binds to it, a conformational change is induced at the other binding sites, increasing its affinity for binding further oxygen molecules, leading to a cooperativity effect that enables efficient transport of oxygen in the blood.

Despite being of such high importance for living organisms, the mechanisms behind allostery still lack a general, predictive atomic description [4]. Classical models treated allostery as a simple switch between two distinct conformations. However, modern approaches describe it as an ensemble of switching states, where regulation emerges from dynamic fluctuations rather than a single structural change. There are several allosteric phenomena seen on different ranges of disorder, from intrinsically disordered proteins over local unfolding, backbone and side-chain dynamics to rigid body motions [5].

The allosteric phenomena observed for the so-called PDZ (PSD-95/discs large/ZO-1) domains – the subject of this work – fall into the category of side-chain dynamics [5].

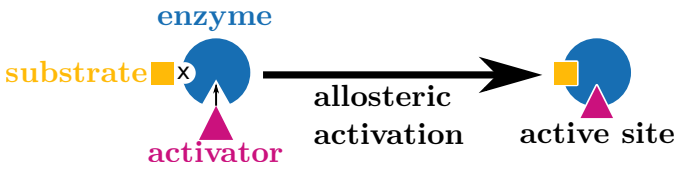

Fig. 1.2.: Schematic illustration of conformational change during the process of allosteric regulation. The substrate is only bound, when the activator is bound to the enzyme.

The PDZ domains are protein domains on the smallest side of still observable allosteric phenomena and thus a popular subject. More information on the PDZ domains analyzed in this work are given in the next chapter.

The Need for Molecular Dynamics Simulations

To gain insight into a protein's structure there are valuable experimental approaches like NMR spectroscopy or X-ray crystallography and theoretical approaches like homology modeling [6, 7]. However, they usually only offer static pictures of a protein's conformation and cannot deliver high time resolutions to keep track of a protein's dynamics [8, 9]. Dynamics can be resolved with ultra-fast IR spectroscopy experiments; however, this technique lacks the spatial resolution of the above-mentioned experimental methods [10].
Molecular Dynamics (MD) simulations aim to offer both in the form of *in silico* experiments. They can capture dynamic transitions at atomic resolution, using a computational approach with femtosecond time resolution [11]. Thus, they are often used for studying complex dynamic systems like proteins.

Non-Equilibrium Molecular Dynamics

Life is a non-equilibrium phenomenon, and so is allostery. The experimental and theoretical focus in allostery research is however often placed upon the equilibrium starting and end states of the allosteric transition [12]. The allosteric transition is hereby not directly observed. To observe the non-equilibrium response of an event triggering an allosteric response (e.g. a ligand binding), non-equilibrium simulations are needed. For this work, non-equilibrium means perturbing the system to start out in a non-equilibrium state and then letting it relax under equilibrium conditions. There are other non-equilibrium

simulation methods, like pulling MD simulations, where applied forces keep the system constantly in a non-equilibrium situation [13].

Research Gap and this Work

Despite the importance of allosteric regulation, only few studies focus on how these transitions behave under **non-equilibrium conditions**. Suitable systems to study allostery on are given with the **PDZ domains**, as they are small in size and yet still **examples of allosteric communication** [14]. In the Stock group at the University of Freiburg, the **second PDZ domain** (PDZ2) was studied with an **azobenzene photoswitch** linked across the binding pocket, which can mimic the binding of a ligand upon switching and thus force the cause for an allosteric transition. For this, molecular dynamics simulations have been performed by Buchenberg et al. [15, 16] and an analogous experimental study exists from Buchli et al. [17].
The **MD dataset for PDZ2** will be further analyzed in this work, following several previous works that have been done in this research group [15, 16, 18, 19, 20, 21]. Prior research on PDZ2 has established its potential for allosteric communication, but details on the **pathways, timescales**, **and mechanisms** remain unclear. It has been shown that the dataset is challenging to analyze and that a careful and judicious selection process on the data is needed.
This work aims to:

1. **Refine data processing and feature selection** to extract meaningful dynamic signals.
2. **Characterize allosteric response pathways** in PDZ2 using non-equilibrium MD.
3. **Compare photoswitched PDZ2 with other photoswitched PDZ systems** to identify general allosteric patterns.

To achieve these goals, advanced feature selection methods are introduced in chapter 2, providing a more structured approach to identifying relevant dynamic features. This is followed by a detailed data processing and assessment in chapter 3 ensuring the reliability of the analyzed trajectories. The core analysis, focusing on the dynamics and response pathways of PDZ2, is presented in chapter 4. Finally, in chapter 5, the findings from PDZ2 are compared with other PDZ systems to explore broader allosteric patterns and mechanistic insights.

2 Theory and Methods

2.1 Systems: PDZ Domains

The goal of this work is to delve further into allosteric communication. The main system analyzed in this work is the second PDZ domain (PDZ2) of the human tyrosine phosphatase (hPTP1E), which regulates several receptor-mediated signal transduction pathways, like cell growth and apoptosis in breast cancer [22, 23, 24]. The related third PDZ domain is known for featuring small-scale allosteric communication between its binding pocket and the distant α_3-helix. Removing the helix reduces the ligand affinity at the binding pocket by 21-fold, while ligand binding can speed up the thermal isomerization rate of a photoswitch applied at the α_3-helix [14, 25, 26, 27].
In the related PDZ2 domain, allostery is less clearly observed, but signs of it have been found [16]. PDZ domains are generally of interest in allostery research with molecular dynamics as they are known to exhibit allosteric effects and are small enough to be feasible to simulate.
The PDZ2 domain features 97 residues with two α-helices, and six β-strands. A binding groove for peptide ligands exists between the β_2-strand and the α_2-helix. Buchli et al. [17] have covalently linked an azobenzene photoswitch at residues 22 and 77 of the PDZ2 domain, spanning across this binding pocket. They performed ultrafast infrared experiments on this photoswitchable PDZ2 domain (which will be called PDZ2S here, *S* for switch). A visualization of the system is given in Figure 2.1 and a list of secondary structures is given in Table 2.1. The photoswitch can adopt either the *cis* or *trans* conformation. In the *trans* state, the azobenzene spans a larger distance, positioning the C_α-atoms of residues 22 and 77 similarly far apart as when a ligand is bound. In the *cis* conformation, this distance corresponds to the ligand-free state [28].

Residues	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

Tab. 2.1.: Locations of secondary structures in PDZ2.

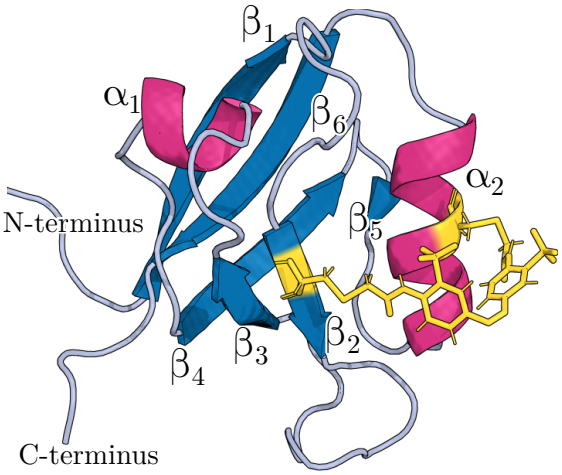

Fig. 2.1.: Cartoon representation of the PDZ2 domain. The azobenzene photoswitch at residues 22 and 77 is shown in yellow.

2.2 Molecular Dynamics Simulations

To perform an MD simulation, a starting structure with the positions of all atoms in the system must be given. This can be determined experimentally, e.g. using NMR spectroscopy [28]. Knowing these starting positions, the classical forces acting on all atoms can be calculated, and the equations of motion can be iteratively integrated over short time steps. With this, the atomic positions can be updated at every time step, before reevaluating the forces again [29]. In theory, solving the electronic Schrödinger equation would be necessary to accurately determine the forces. However, for larger systems such as proteins, this is computationally intractable. Therefore, MD simulations work with classical force fields to approximate interactions between atoms. The total potential energy of the system with atomic position vectors $\boldsymbol{R_1}, \boldsymbol{R_2}, ..., \boldsymbol{R_N}$ can be described as

$$V(\boldsymbol{R_1}, \boldsymbol{R_2}, ..., \boldsymbol{R_N}) = V_{\text{bonded}} + V_{\text{non-bonded}}, \tag{2.1}$$

where V_{bonded} accounts for bonded interactions and $V_{\text{non-bonded}}$ accounts for non-bonded interactions. The bonded contributions are viewed as the sum of bond and angle contributions modeled with harmonic expressions, and a term describing the torsional energy:

$$V_{\text{bonded}} = V_{\text{bonds}} + V_{\text{angles}} + V_{\text{torsions}}. \tag{2.2}$$

The non-bonded contributions are described with a van der Waals term and a Coulomb term:

$$V_{\text{non-bonded}} = V_{\text{vdW}} + V_{\text{Coulomb}}. \tag{2.3}$$

A typical functional form of the potential energy function is

$$\begin{aligned} V = &\sum_{\text{bonds}} \frac{K_{b_i}}{2}(b_i - b_{0_i})^2 + \sum_{\text{angles}} \frac{K_{\theta_i}}{2}(\theta_i - \theta_{0_i})^2 + \sum_{\text{torsions}} \sum_{n} K_{\phi_i}^{(n)} \left[1 + \cos(n\phi_i - \gamma^{(n)})\right] \\ &+ \sum_{\text{non-bonded } i,j} \left\{ \underbrace{4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{R_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{R_{ij}}\right)^{6}\right]}_{\text{van der Waals}} + \underbrace{\frac{1}{4\pi\epsilon_0}\frac{q_i q_j}{R_{ij}}}_{\text{Coulomb (electrostatic)}} \right\}, \end{aligned} \tag{2.4}$$

where [30]:

- b_i is the bond length between two atoms, and b_{0_i} is its equilibrium value.
- K_{b_i} is the bond force constant.
- θ_i is the bond angle, with equilibrium value θ_{0_i}.
- K_{θ_i} is the angle force constant.
- ϕ_i is the dihedral angle, with force constant $K_{\phi_i}^{(n)}$ and shift $\gamma^{(n)}$.
- σ_{ij} and ϵ_{ij} are Lennard-Jones parameters describing van der Waals interactions.
- q_i and q_j are atomic partial charges, and ϵ_0 is the permittivity of free space.
- R_{ij} is the distance between atoms i and j.

The listed parameters are typically determined from experiments. Several force fields with similar setups exist, and each has its specific strengths and weaknesses for different kinds of molecular systems.
The above-mentioned calculations are carried out for a molecule placed in a simulation

box with periodic boundary conditions, that is usually filled with a solvent like water. MD simulations reach time spans of several microseconds within reasonable wall clock times, which is sufficient to describe many dynamics of biological interest [31]. The need for long simulation times arises particularly for rare conformational changes. The transition barrier crossing probability is described by the Boltzmann factor

$$P \sim e^{-\Delta G/k_B T} \tag{2.5}$$

at temperature T with Boltzmann constant k_B. The probability of reaching the transition state decreases exponentially with the barrier height ΔG. This means that simulations tend to sample local minima most of the time, requiring these long simulation times to achieve good sampling of barriers.
Typically, molecular dynamics simulations are performed for systems in equilibrium. However, most biological processes also occur out of equilibrium, like the allosteric responses studied here. Therefore, also non-equilibrium simulations will be used in this work.

2.3 Simulation Details

2.3.1 PDZ2S

The PDZ2S simulations examined in this work comprise equilibrium (EQ) and non-equilibrium (NEQ) trajectories. All simulations were performed by S. Buchenberg with GROMACS using the Amber ff99SB*-ILDN force field and TIP3P water [32, 33, 34]. Frames were saved every $20\,\mathrm{ps}$. The EQ dataset originally contained 7 trajectories of $2.5\,\mu\mathrm{s}$ each for both cis and trans conformations [15]. Due to a lack in convergence (as also seen in later results in this work), 6 of these were extended to $10\,\mu\mathrm{s}$ by A. Gulzar.
The NEQ dataset consists of 100 short trajectories of $1\,\mu\mathrm{s}$ each. The system was forced from *cis* to *trans* following the potential-energy surface switching approach described in [35], and from there it was simulated under equilibrium conditions. This can be understood as a forceful stretching of the photoswitch visualized in Figure 2.1, leading to an opening of the binding pocket. The initial conditions were sampled from the EQ simulations. Due to poor convergence (most trajectories did not reach the *trans* state), 20 randomly selected trajectories were extended to $10\,\mu\mathrm{s}$ by Buchenberg et al. [16].

2.3.2 Other PDZ Systems

PDZ2L

The simulations for PDZ2L were carried out using GROMACS v2016 and the Amber99SB*-ILDN force field with TIP3P water. Frames are saved every $20\,\mathrm{ps}$. 100 non-equilibrium trajectories of $1\,\mu\mathrm{s}$ exist, of which 20 were extended to $10\,\mu\mathrm{s}$.

PDZ3L5

For the PDZ3L5 simulations, GROMACS v2020 was used, again with the Amber ff99SB*-ILDN force field and TIP3P water. Frames are saved every $20\,\mathrm{ps}$. 116 non-equilibrium trajectories exist, of which 90 are of $1\,\mu\mathrm{s}$ length and 22 are of $10\,\mu\mathrm{s}$ length.

PDZ3L6

For PDZ3L6, the same GROMACS and force field settings are used as for PDZ3L5. Coordinates are sampled with a time step of $200\,\mathrm{ps}$ however and there are 89 trajectories of $1\,\mu\mathrm{s}$ length and 10 of $10\,\mu\mathrm{s}$ length.

2.4 Internal Coordinates

In MD simulations, the dynamics of a protein are given in terms of 3N Cartesian coordinates. However, the Cartesian coordinate description of the protein is not suitable for analyzing the protein's behavior, as the overall protein rotation in the solvent and the internal motion will be mixed.
The analysis should focus on the internal motions of the protein, only. Therefore, internal coordinates should be used. Typically, dihedral angles or inter-residual distances are chosen as internal coordinates [36]. The dihedral angles ϕ_n and ψ_n of the protein backbone would be used to describe the proteins' conformation using local measures along the amino acid chain. In this work, more focus shall be laid upon side-chain dynamics and thus inter-residue contact distances will be majorly used as internal coordinates. Contact distances focus on the idea that the side chains of the amino acids are able to form non-covalent bonds, which are not necessarily represented in backbone information. It

was shown that contact distances between protein residues also capture protein dynamics more efficiently than dihedral angles, as fewer principal components are required to describe the cumulative fluctuations [37].
A representation of the difference between the often-used backbone C_α distance [37] and the contact distance of the same residues is given in Figure 2.2. It is important to note that the term *contact distance* does not imply that a bond is formed between the two residues at every point in time. It rather describes the instantaneous minimum distance between heavy atoms of two residues, that are able to form a contact.

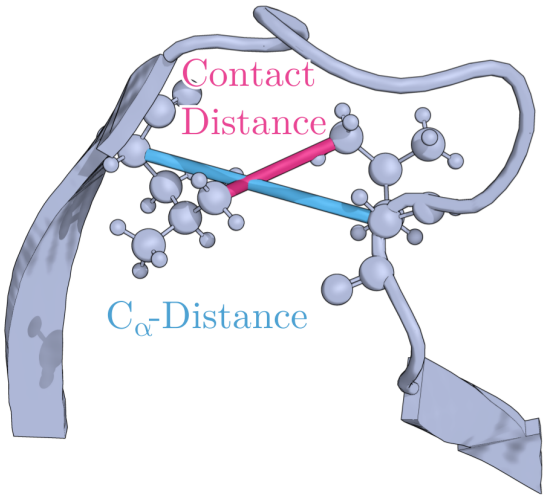

Fig. 2.2.: Contact distance (magenta) and C_α distance (blue) for the same residue pair. The residues form a non-covalent bond, where at least one heavy atom from each residue is in close proximity, stabilizing the interaction. This results in a shorter contact distance compared to the C_α distance, which measures the separation between backbone atoms.

2.5 MoSAIC Feature Selection

A good choice of collective variables is essential for the success of a subsequent analysis of molecular dynamics data [36]. Collective variables are a low-dimension set of observables describing processes of interest. In a protein MD dataset typically multiple processes occur, of which only some are relevant to the analysis. The goal of the MoSAIC algorithm by Diez et al. [38] is to identify these relevant processes and group variables that describe the same underlying motion. The algorithm is based on the idea that relevant variables show related motions, whereas randomly fluctuating variables or variables remaining constant

are considered irrelevant. These irrelevant variables (or features) can be considered as contributing to noise in the overall dataset and should thus be discarded. Applying such a filtering step before dimensionality reduction techniques like PCA is essential, as high-variance but meaningless fluctuations – such as randomly changing contact distances – might otherwise be considered "important" by variance-based methods like PCA [38]. The MoSAIC algorithm aims to find groups of variables that show related motion. To achieve this, each input variable will be viewed as a node in a graph, connected to the other nodes with some form of similarity measure as the edge weight. Diez et al. recommend measures like the linear Pearson correlation or the non-linear mutual information measure. Using the Leiden community detection algorithm, communities of strongly linked nodes in the graph are found. For this, the constant Potts model (CPM) is used as an objective function Φ_{CPM}, for which the community partitioning is iteratively optimized [38]:

$$\Phi_{\text{CPM}} = \sum_{c} \left[e_c - \gamma \binom{n_c}{2} \right]. \tag{2.6}$$

Here, n_c represents the number of nodes in cluster c, e_c refers to the sum of the edge weights in a cluster and $\binom{n_c}{2}$ stands for the number of possible edges within c. The resolution parameter γ indicates the minimum average correlation (or mutual information) for each resulting cluster of variables. Individual correlations in a cluster can be below γ if it helps for a better overall partitioning. The result of this method then consists of sets of correlated variables (clusters) and a set of variables considered noise.

2.6 Principal Component Analysis

The principal component analysis (PCA) is a technique used to reduce dimensionality. To perform abstract analyses like free energy landscape analyses, the high-dimensional molecular dynamics data must be projected onto a low-dimensional space. The dimensionality of this space must be high enough to resolve the number of conformational states while remaining sufficiently low to allow for statistical analysis. The curse of dimensionality illustrates the challenge: if we distribute e.g. 10^6 data points on a 10D grid with 10 bins in each dimension, we obtain 10^{10} bins, and most of them will be empty. This would make density-based analyses impractical [36]. The PDZ2S system viewed in this work has $3 \cdot 1496 = 4.488$ Cartesian coordinates, corresponding to the x, y, and z

positions of all 1496 atoms. While PCA is typically applied to internal instead of Cartesian coordinates, this number illustrates the high dimensionality of molecular dynamics data. The effective dimension of degrees of relevant physical processes is typically much smaller however, due to nonlinear couplings and resulting cooperative effects in the protein.
A computationally efficient, numerical approach for dimensionality reduction is offered with the PCA. It projects data onto a low number of dimensions, that explain the largest amount of variance possible. For the n input coordinates $r_1 ... r_n$, each describing a time series with m time steps, the method starts by calculating the covariance matrix with entries

$$\text{Cov}(r_i, r_j) = \langle (r_i - \langle r_i \rangle)(r_j - \langle r_j \rangle) \rangle = \langle r_i r_j \rangle - \langle r_i \rangle \langle r_j \rangle. \tag{2.7}$$

Here, $\langle ... \rangle$ denotes the ensemble average. To emphasize on correlated motion and also on small-amplitude motion, the correlation matrix $\boldsymbol{C}$ is calculated from the covariance matrix

$$C_{ij} = \text{Corr}(r_i, r_j) = \frac{Cov(r_i, r_j)}{\sigma_{r_i} \sigma_{r_j}}, \tag{2.8}$$

with σ_{r_i} being the standard deviation of the input coordinate r_i. The eigenvectors of the correlation matrix $v^{(i)}$ are resembling the modes of collective motion in the data. The corresponding eigenvalues λ_i are the amplitudes of this motion [39, 40]

$$\boldsymbol{C}\boldsymbol{v}^{(\boldsymbol{i})} = \lambda_i \boldsymbol{v}^{(\boldsymbol{i})}. \tag{2.9}$$

The eigenvectors are orthogonal, normalized and sorted in descending order of their eigenvalues. The first eigenvector corresponds to the direction of the largest variance in the data, the second to the second largest and so forth. This is also illustrated in Figure 2.3. The projection of the data $\boldsymbol{r} = (r_1 ... r_n)$ onto the eigenvectors results in the principal components x_i:

$$x_i = \boldsymbol{v}^{(\boldsymbol{i})} \cdot \boldsymbol{r} \tag{2.10}$$

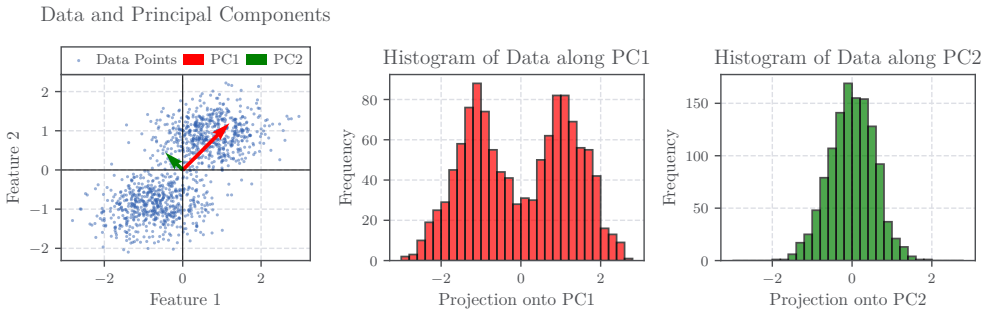

Fig. 2.3.: Illustration of a Principal Component Analysis (PCA). The left panel shows 2D data with two clusters, where the PCA finds the optimal axis (PC1, red) capturing the most variation. The middle panel projects the data onto PC1, reducing it to one dimension while preserving the cluster structure. In contrast, the right panel shows the projection onto PC2, which captures little meaningful variation, as seen in its Gaussian-shaped histogram. This highlights how PCA simplifies data complexity while retaining essential structure.

2.7 Timescale Analysis: Dynamic Content

Biomolecular processes occur on a wide spectrum of timescales. A timescale analysis can elucidate the relevant timescales, paired with localization of processes connected to these timescales. For this, a dynamic content at each timescale between $1\,\mathrm{ns}$ and $10\,\mu\mathrm{s}$ will be calculated both for the simulation data and for corresponding experimental data. To calculate the simulation's dynamic content, the first step is to model the time series of each ensemble-averaged contact distance with a multiexponential response function, following the procedure of Dorbath et al. [41]:

$$S(t) = \sum_k s_k e^{-t/\tau_k}. \tag{2.11}$$

Here, the time constants τ_k are equidistantly distributed on a logarithmic scale and the amplitudes s_k correspond to the timescales. By taking the square root of the sum of the squared amplitudes in a cluster of collective variables C_n, a cluster-n-resolved dynamic content D_n can be calculated for each timescale τ_k [12]:

$$D_n(\tau_k) = \sqrt{\sum_{i,j \in C_n} |s_k(i,j)|^2}. \tag{2.12}$$

From experimental time-resolved infrared spectroscopy, an overall dynamic content for the whole system can be calculated. For this, the amide-I band was probed, which reflects

backbone dynamics through coupled C=O vibrations [42]. The simulation data should be able to reproduce this, after combining the cluster-resolved dynamic content to an overall dynamic content as in Equation 2.13. The added benefit of the MD data is that they allow for an analysis showing which collective variables and clusters contribute to what timescales.

$$D(\tau_k) = \sqrt{\sum_n |D_n(\tau_k)|^2} \tag{2.13}$$

2.8 Root Mean Square Deviation

The root-mean-square deviation (RMSD) is commonly used in molecular dynamics to measure conformational differences between two structures or the change of a structure over time. It is calculated as the square root of the mean of the squared distances between the atoms of the two structures. For a structural change over time, the RMSD is calculated as

$$\text{RMSD}(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{r}_i(t) - \mathbf{r}_i^{\text{ref}})^2} \tag{2.14}$$

with N being the number of atoms, $\mathbf{r}_i(t)$ the position of atom i at time t and $\mathbf{r}_i^{\text{ref}}$ the position of atom i in the reference structure. A low RMSD indicates a high structural similarity [43].

2.9 Free Energy

The free energy quantifies the thermodynamic stability of different system states. To describe the connection between a protein's conformations, the physical image of an energy landscape can be drawn. The free energy landscape represents the possible conformations as regions in a multidimensional energy space. Conformational states are minima in this picture, separated through energy barriers. Thermal fluctuations then allow for the transitions over the barriers [44]. A schematic is given in Figure 2.4.
The energy in this landscape viewed is best described with the Gibbs free energy, encapsulating the entropy-enthalpy balance that governs processes like folding and allostery.

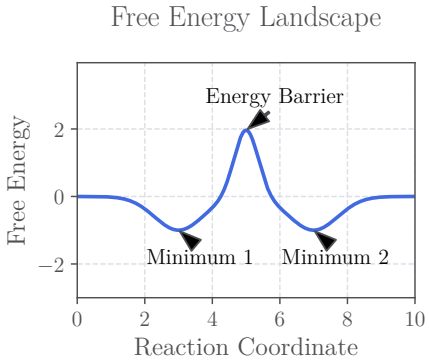

Fig. 2.4.: Illustration of an energy landscape along one reaction coordinate. Two energy minima indicating favorable conformations are seen. They are separated by an energy barrier, that is making thermal fluctuations between the minima less likely.

In folding, enthalpic gain from favorable interactions plays against entropic loss of disordered conformations. Protein dynamics span a wide range of timescales. Small energy barriers are connected to fast motions like bond vibrations on the picosecond scale, while larger energy barriers are connected to slow motions on the microsecond timescale, that involve conformational changes relevant for signalling [45].
In MD simulations, the free energy $\Delta G(\boldsymbol{r})$ can be calculated from the probability distribution $P(\boldsymbol{r})$ along a system's observable $\boldsymbol{r}$:

$$\Delta G(\boldsymbol{r}) = -k_B T \ln P(\boldsymbol{r}) \tag{2.15}$$

with k_B being the Boltzmann constant and T denoting the temperature [16]. The bin width of the histogram underlying $P(\boldsymbol{r})$ needs to be balanced between resolution and statistical noise. This free energy includes both the enthalpic and entropic contributions [44].

3 PDZ2S: Data Processing and Assessment

This chapter describes the preprocessing steps applied to the raw molecular dynamics simulation datasets to prepare them for the subsequent analysis. First, a set of internal coordinates is chosen to represent the system's dynamics. Next, noisy and irrelevant data are identified and excluded to ensure data quality. In addition, contact clusters are selected, which reduce the dimensionality of the data and help to focus the analysis on the decisive structural features. These steps ensure that the results of later analyses are both meaningful and robust.

3.1 Available Datasets

As presented before, the system to be analyzed is a photoswitchable PDZ2 domain. For this system, two sets of simulations have been generated by Buchenberg et al. [16] and Buchenberg et al. [15].

NEQ: $100 \times 1\,\mu\mathrm{s}$ short non-equilibrium trajectories are available, with 20 of these extended to $20 \times 10\,\mu\mathrm{s}$.

EQ: $6 \times 10\,\mu\mathrm{s}$ equilibrium trajectories exist for both cis and trans conformations, of originally 7 (one each was discarded in previous analyses [21]).

The original time step of the data is $dt = 20\,\mathrm{ps}$, which was filtered using a Gaussian kernel with a width of $2\,\mathrm{ns}$ and subsequently stridden, resulting in a final time step of $dt = 200\,\mathrm{ps}$ with $2\,\mathrm{ns}$ time resolution.

3.2 Collective Variables: Internal Coordinates

As explained in section 2.4, contact distances will be used to describe the internal motion of the system. To find a suitable set, the following calculations are performed on the NEQ long trajectories: For all residue pairs (i, j), with $j > i + 2$ (i.e. not first- or second-order neighbors) and $i, j \in \{1, 2, ..., 97\}$, the distance between the closest heavy atoms (i.e. non-hydrogen atoms)

$$d_{i,j}(t) = d_{j,i}(t) = \sqrt{(\vec{x}_i(t) - \vec{x}_j(t))^2} \tag{3.1}$$

of the residues is calculated, with the Cartesian coordinates $\vec{x}_i$ and $\vec{x}_j$ being defined as the coordinates of the heavy atoms of residues i and j with the shortest instantaneous distance.

If this distance is below $0.45\,\mathrm{nm}$ in at least $10\,\%$ of the simulation frames, the pair is considered to be able to form a contact and thus added to the set.
In general, an attempt should be made not to include irrelevant coordinates in the collective variables in order to maintain a good signal-to-noise ratio [36]. The restriction of $j > i+2$ is made to avoid trivial, neighboring contacts, but still allow for the inspection of helical bonds. In secondary structures such as α-helices, an $i \to i + 4$ bonding takes place. For 3_{10}-helices an $i \to i + 3$ pattern is seen [46]. Contact breaking of these structures is thus expected to be visible in the collective variables. The cutoff distance of $4.5\,\text{Å}$ is a typical value for contact distances in proteins [47].

In short, the collective variables are defined the following way:

- The distance between the closest heavy atoms of the residues is $d_{i,j} \leq 0.45\,\mathrm{nm}$.
- The residues are not first- or second-order neighbors.
- The distance criterion is satisfied in at least $10\,\%$ of the frames.

The set of inter-residue contact distances differs between data sets when using this definition. To ensure consistency in the analysis, the set of collective variables was defined based on the long non-equilibrium simulations and subsequently applied to all other data sets. Additional calculations on the other data sets confirmed that the resulting sets of collective variables were highly similar among the different data sets. Ultimately, 330 residue contact distances are considered, and they are displayed in Figure 3.1.

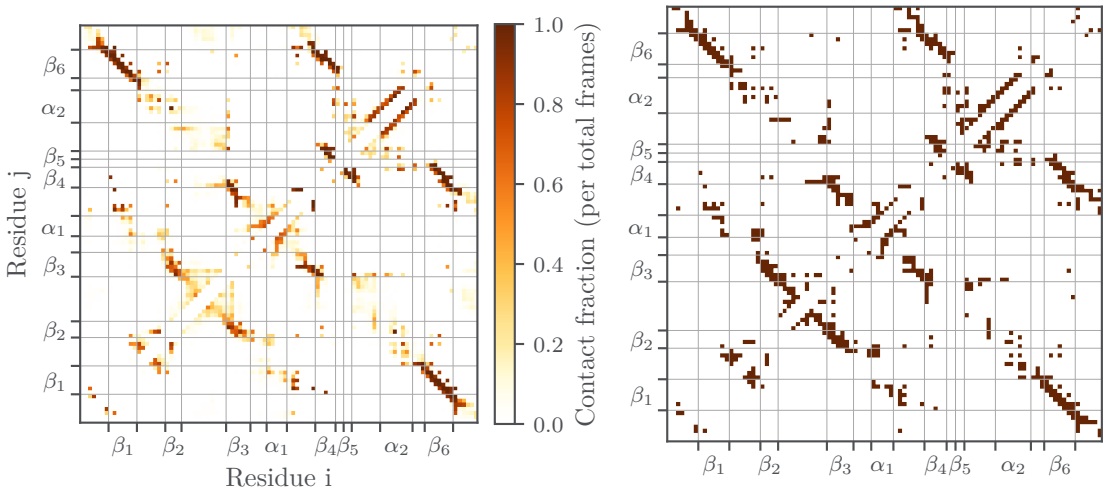

Fig. 3.1.: The heat map on the left shows the percentage of frames in which a residue pair forms a contact. The graph on the right shows only the 330 contacts that fulfill the 10% criterion and are thus selected as collective variables. Each pixel indicates a residue pair, starting from low to high residue index from the bottom left. The positions of secondary structures are marked with a grid. The matrices are symmetric.

Additionally, C_α-distances will be used as well at times. They are calculated as the distance between the C_α-atoms of two residues and focus more on describing the backbone motion than the heavy atom distances.

3.2.1 Error Measures

Commonly, when contact distances are used, they will be given as an ensemble mean. This averaged value comes with two error measures: the standard deviation σ (Equation 3.2) and the standard error of the mean (SEM, Equation 3.3). Both will be needed for different analyses. The standard deviation measures the spread of the data, while the standard error of the mean is informing about the precision of the mean value $\bar{x}$.
In the system analyzed here, individual contact distances exhibit highly heterogeneous dynamics. As a result, the ensemble mean must be interpreted with caution as many averages have large standard deviations. With low SEM values indicating high precision of the mean, this does not necessarily imply that the mean values themselves are occupied or more meaningful, particularly given the presence of non-Gaussian distributions.

Nevertheless, the SEM remains a useful indicator of the reliability of the calculated mean, that should mostly be used to identify trends in distance changes.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{3.2}$$

$$\text{SEM} = \frac{\sigma}{\sqrt{N}} \tag{3.3}$$

3.3 Assessing Data Quality

Previous studies on this dataset have reported low convergence and high noise in the data [16, 18]. A comparison with other photoswitched PDZ systems [48] suggests that this is likely due to the switch being attached at a more pivotal position of the protein – directly at the binding groove – rather than at more peripheral sites such as the α_3-helix. This has a significant impact on the protein's structure, and a more pronounced effect on the system can be expected. Therefore, a rigorous assessment of data quality is necessary to ensure that the analysis does not primarily capture destructive effects caused by the photoswitch, which are not of interest.

3.3.1 Equilibrium Trajectories

To interpret equilibrium molecular dynamics data, it must be checked first whether the trajectories have actually converged to equilibrium. A simple measure of convergence is the identification of a plateau after a relaxation period in the RMSD of the protein backbone [49]. The ensemble average C_α-atom RMSD between the starting points and all succeeding points of both the cis and trans trajectories is calculated and displayed in Figure 3.2. A relaxation period can be observed, with the largest changes happening in the first $3\,\mu$s and more stable values afterwards. A reasonable cutoff-value can thus be set at $3\,\mu$s. Data before that mark will be discarded.
During the later stages of the analysis, two of the available equilibrium trajectories (EQ cis 5 and EQ trans 1) were found to exhibit unusual behavior. Specifically, they appear to sample free energy regions that are more characteristic of the opposing photoswitch state. In Figure 3.3, the mean (time-averaged) RMSD of each trajectory was calculated relative to all trajectories within the same photoswitch state and compared to those in the opposing state. The difference between these means was analyzed to assess the

extent of deviation from the native state. Values below zero indicate that, on average, the trajectory is conformationally closer to the trajectories with opposing photoswitch state. This is indeed seen for the trajectories *EQ cis 5* and *EQ trans 1*. Combined with the doubts raised in a later performed free energy landscape analysis, these two trajectories are conservatively discarded as outliers.

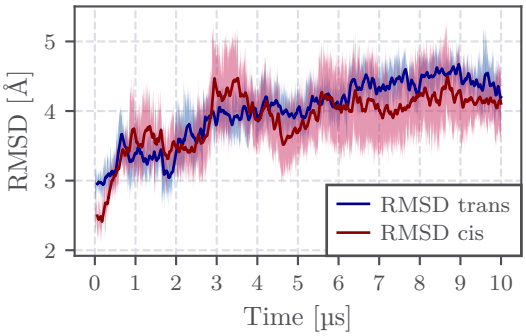

Fig. 3.2.: The ensemble-averaged RMSD of the C_α-atoms from their respective base structures was calculated for both the equilibrium *cis* and *trans* ensembles. A relaxation period is observed in the first approximately $3\,\mu$s. Filtered averages are given together with the standard error of the mean.

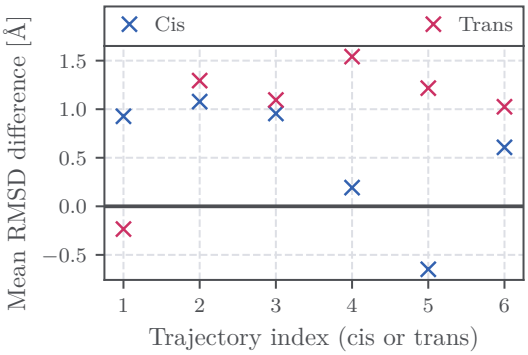

Fig. 3.3.: Difference between each trajectory's mean (over all frames) RMSD from all trajectories with same photoswitch state minus RMSD from all trajectories with opposing photoswitch state. Negative values indicate that, on average, a trajectory is conformationally more similar to the opposite state than to its own, which is unexpected. Thus, trajectories cis 5 and trans 1 appear as outliers.

3.3.2 Stability of Secondary Structures

The α-helices in the system exhibit varying stability throughout the simulations, likely influenced by the relatively drastic photoswitching, which introduces unnatural deformations. With the photoswitch being attached directly to the α_2-helix, deformations here will be viewed as unwanted. Trajectories that exhibit them will be removed as a precaution to not focus on secondary, photoswitch-specific dynamics. Structural changes in the more distant α_1-helix will be seen as more independent of direct photoswitch effects and thus those trajectories will be considered in the analysis.
To assess the stability of the α_2-helix, which is directly attached to the photoswitch, a DSSP (*database of secondary structure assignments*) analysis was performed on the trajectories. The DSSP assigns secondary structure classes to residues based on hydrogen-bonding patterns and geometric features [50]. For this analysis, the DSSP classification was calculated for each residue in the α_2-helix region (residues 74-81) across all frames. In most trajectories, the helix remains stable or loses its structure only briefly. To ensure data quality, trajectories are discarded if the combined proportion of α- and 3_{10}-helix assignments for the α_2-helix region residues falls below two thirds. Resulting graphs are given in the appendix, Figure A.2. To back this decision up, the RMSD from the corresponding protein database entry [28] of the C_α-atoms is calculated, additionally. Stable, low-fluctuation RMSD values around 0.5 Å are seen for intact helices. For the NEQ trajectories, the DSSP results match an equivalent cutoff of an average RMSD larger than 1 Å consistently. In the special case of the equilibrium cis trajectories, helical conformations are not as dominated by α-helix conformations, and 3_{10}-helical conformations play a larger role. Here, the RMSD reaches higher values while maintaining helical structure. Table 3.1 lists the trajectory indices for the 4 data sets, in which the α_2-helix loses its structure.

Data set	trajectories with broken α_2-helix	ratio of affected trajectories
NEQ short	1, 2, 13, 15, 16, 20, 26, 32, 38, 52, 55, 56, 59, 64, 68, 73, 77, 79, 80, 81, 86, 89, 92, 95, 96	25 %
NEQ long	1, 2, 12, 13, 17, 19	30 %
EQ cis	2, 6	33 %
EQ trans	-	0 %

Tab. 3.1.: Breakage of α_2-helices given for the 20 long and 100 short NEQ trajectories, as well as for the EQ data sets with 6 trajectories each. Also, the share of discarded trajectories from all trajectories in each data set is given.

The α_1-helix is also unfolding in some trajectories. For the shorter helices, like the α_1-helix (5 residues), the DSSP method does not prove to be as robust [51], so an average RMSD above 0.75 Å is set as the threshold of unfolding in comparison with inspection of protein visualizations. See Table 3.2 for the lists of impacted trajectories.

Data set	trajectories with broken α_1-helix	ratio of affected trajectories
NEQ short	7, 12, 19, 21, 26, 34, 35, 41, 54, 61, 68, 70, 74, 76, 83, 94	16 %
NEQ long	2, 3, 5, 6, 7, 10, 14, 17, 19	45 %
EQ cis	1, 2	33 %
EQ trans	-	0 %

Tab. 3.2.: Breakage of α_1-helices given for the 20 long and 100 short NEQ trajectories, as well as for the EQ data sets with 6 trajectories each. The breaking of α_1-helices is not seen as an artifact caused by the photoswitch and thus trajectories showing this behavior will be included in subsequent analyses.

In the EQ *trans* trajectories, the α-helices are more stable than in the EQ cis trajectories. No *trans* trajectories have to be discarded due to missing α-helix stability. The higher stability of the system in *trans* has been described previously [19].

3.4 Feature Selection: MoSAIC

A major advance in the investigation of this system compared to earlier investigations shall be the application of the MoSAIC feature selection, as introduced by Diez et al. [38]. This method is capable of finding the most relevant collective variables in a dataset and clusters collectively moving features together.

3.4.1 Similarity Measure

Linear Correlation

As explained in section 2.5, the MoSAIC method works on a matrix with similarity information of the collective variables. Typically, the covariance or the linear correlation are used as the similarity measure. The covariance is by design focussing on the largest motions (variances) (Equation 3.4). This leads to an underrepresentation of small-scale dynamics in the cluster analysis. It is thus not suitable for this system, where e.g., the unwinding of the α_1-helix is a major feature, which would overly draw attention to the α_1-helix dynamics and suppress others.

The linear correlation mitigates this problem by focussing on the relative motion of the collective variables, as for each feature (i.e. each contact distance), the values are normalized to values between -1 and 1. The covariance definition and various correlation definitions that will be discussed in the following are given for the mean free variables $\delta x(t) = x(t) - \langle x \rangle$:

$$\text{Cov} = \langle \delta x \delta y \rangle_{NT} \tag{3.4}$$

$$C^A_{xy} = \frac{\langle \delta x \delta y \rangle_{NT}}{\sigma_{x,NT}\sigma_{y,NT}} \tag{3.5}$$

$$C^B_{xy} = \left\langle \frac{\langle \delta x_n \delta y_n \rangle_T}{\sigma_{x_n,T}\sigma_{y_n,T}} \right\rangle_N \equiv \frac{1}{N}\sum_n^N C_{xy,n} \tag{3.6}$$

$$C^C_{xy} = \left\langle \left| \frac{\langle \delta x_n \delta y_n \rangle_T}{\sigma_{x_n,T}\sigma_{y_n,T}} \right| \right\rangle_N \equiv \frac{1}{N}\sum_n^N |C_{xy,n}| \tag{3.7}$$

In the typical case of several similarly set up MD simulations, resulting in an ensemble of molecular trajectories, the linear correlation can be calculated in different ways. The most straightforward way is to calculate the correlation for the whole set of trajectories as one concatenated trajectory (i.e. taking the standard deviation over all data points). This is represented in variant C^A_{xy} (Equation 3.5). However, this normalization over all data points at once suffers heavily from outlier trajectories. In the case of a major disruption that happens in just one trajectory, but affects several features, the correlation measure will see this as the deciding motion and will suppress all other relevant trajectories. This is visualized in Figure 3.4a, where in each trajectory, features 1 and 2 are clearly correlated, while feature 3 is not. Yet, due to the major disruption seen in trajectory 3 (e.g. a major disassembly of protein structure), the correlation C^A_{xy} will be high between the three features, as seen in Figure 3.4b (a). This is especially problematic in the system viewed here, where the photoswitching is known to have a destructive impact on some trajectories.

By normalizing the correlation for each trajectory separately and then averaging over all trajectories, the effect of outliers can be mitigated. This approach is represented by variant C^B_{xy}, where $C_{xy,n}$ describes the correlation seen over one trajectory n. However, this method fails in the case of anticorrelations. Anticorrelation occurs when two features are correlated, but with opposite signs: an increase in one feature is coupled to a decrease in the other. An illustrative example is given by features 1 and 2 in Figure 3.4b, which are correlated in trajectories 1, 2, and 3, but anticorrelated in trajectory 4. Although their motion is clearly connected, the averaged value of C^B_{xy} becomes misleadingly low

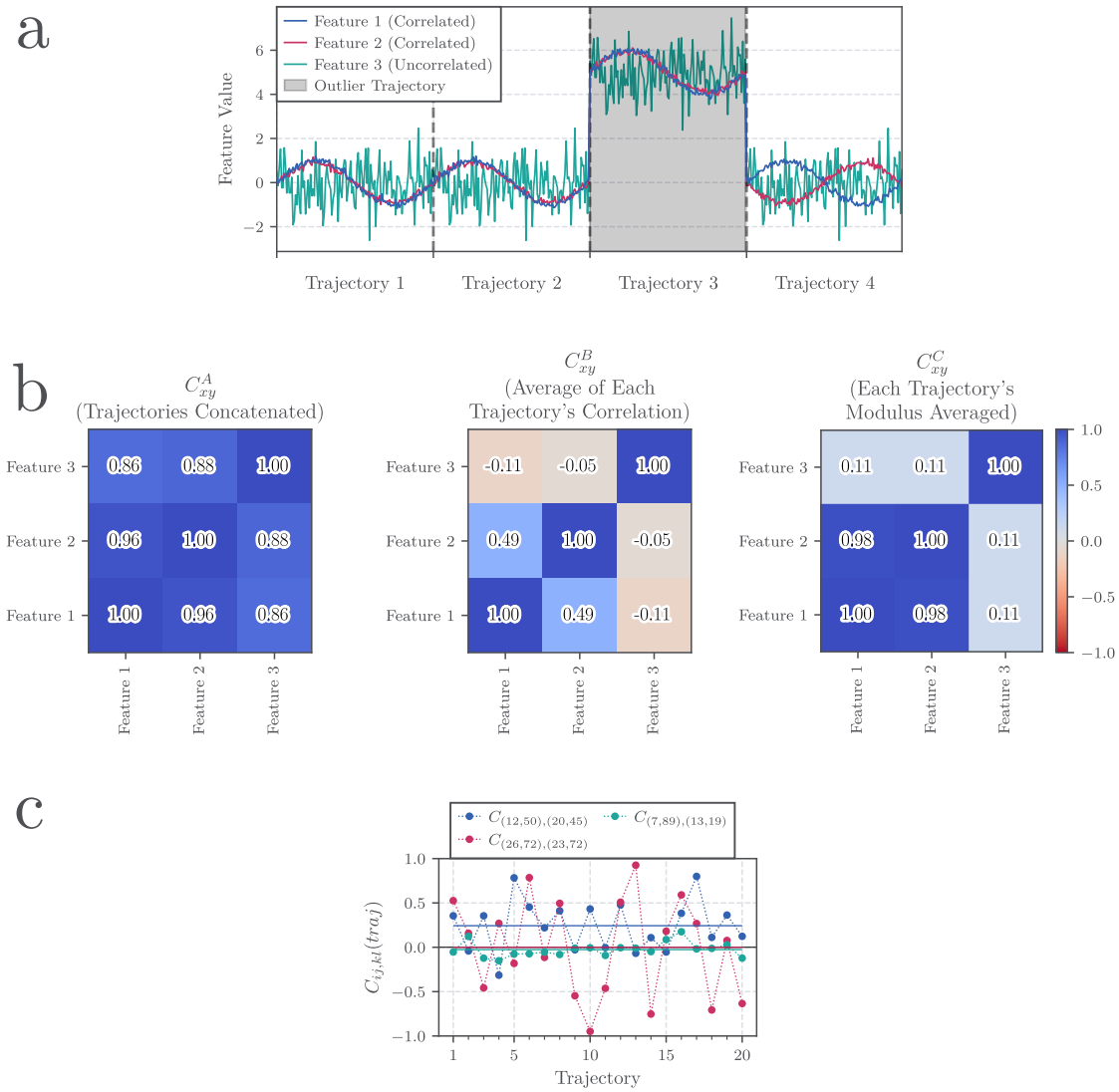

Fig. 3.4.: **a:** Illustrative time traces of three features in 4 trajectories. Features 1 and 2 are correlated, with an anticorrelation in trajectory 4. Feature 3 is uncorrelated. Trajectory 3 illustrates an artifactual trajectory where all features are affected (e.g. unexpected unfolding). **b:** Corresponding correlation matrices for the definitions C^A_{xy}, C^B_{xy}, and C^C_{xy}. C^A_{xy} sees correlation in all features due to the outlier trajectory. C^B_{xy} ignores the anticorrelation. Only C^C_{xy} correctly identifies a high correlation between features 1 and 2, but not 3. **c:** PDZ2S example of correlation cancelling: For three pairs of contact distances the correlations $C_{xy,n}$ are calculated for each of the 20 trajectories. For the pair shown in blue, the correlations are positive for most trajectories. For the pair shown in green, mostly around zero. This is also represented in their averages, shown as full lines. For the pair shown in red, however, either high correlations or high anti-correlations are seen in most trajectories. The average of those values according to C^B_{xy} leads to a possibly misleading mean correlation of around zero, indicating why measure C^C_{xy} should be used.

for features 1 and 2, as shown in Figure 3.4b. This exact effect is also observed in the PDZ2S system analyzed in this work, with an example given in Figure 3.4c. For instance, the averaged correlation value for $C_{(26,72),(23,72)}$ is close to zero, even though the interaction between these two contact distances is clearly relevant – being correlated in some trajectories and anticorrelated in others.
To address this issue at the cost of losing sign information, variant C^C_{xy} is introduced. Here, the absolute value of the correlation is calculated for each trajectory before subsequently averaging over the trajectories. As a result, the averaged correlation value becomes a measure of the strength of interaction between the two features, regardless of its direction. As illustrated in Figure 3.4c, the correlation between features 1 and 2 is high, while correlations involving feature 3 remain low. This matches the intuitive understanding.
When using the linear correlation as the similarity measure for MoSAIC, the variant C^C_{xy} should always be used as the definition of correlation. The other variants are prone to unexpected errors.

Mutual Information

A similarity measure that is not as prone to these errors is the normalized mutual information (NMI) introduced by Nagel et al. [52]. Unlike the linear correlation, mutual information is not requiring a normal distribution in the underlying stochastic variable. This makes mutual information the better motivated similarity measure for analyzing the PDZ2S simulations, as the given non-equilibrium data in the form of contact distances is not expected to be normally distributed. Also, mutual information finds non-linear correlations, e.g. for multidimensional data like Cartesian coordinates. The unbounded mutual information is defined for two random variables X and Y with the Kullback-Leibler divergence of the joint probability distributions $p(x, y)$ and the product of the marginal distributions $p_x p_y$:

$$I(X,Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right). \tag{3.8}$$

Nagel et al. introduced a normalized version of this mutual information (NMI), see [52] for details. The entropy calculations needed to evaluate the NMI result in considerably larger computational costs than for the Pearson correlation, however. For the dataset used in this work, an NMI calculation was feasible. It also proved to be the more stable similarity measure compared to C^C_{xy} and thus NMI is used as the similarity measure for the MoSAIC clustering in this work. All data points were assumed to be part of one ensemble, so the NMI was calculated over all trajectories (i.e. concatenated).

3.4.2 MoSAIC Results: Cluster Identification

Using MoSAIC clustering, the number of reaction coordinates can be reduced from 330 to 153. These coordinates are sorted in meaningful clusters of collective motion. The parameter for cluster resolution γ and the minimum cluster size were determined through a grid search, yielding $\gamma = 0.1$ and a minimum cluster size of 8 coordinates as the best options. This means, the minimum average NMI for each cluster is 0.1. It should be noted that no definitive criterion exists for identifying an optimal choice, and the parameter choice is somewhat flexible. The resulting clustered similarity matrix and a visualization of the clusters found for PDZ2S are seen in Figure 3.5. Table 3.3 lists the cardinalities of the clusters and Table 3.4 lists the contact distances that constitute the clusters. The labeling of the protein does not follow any biological principle and goes clockwise around the binding pocket.

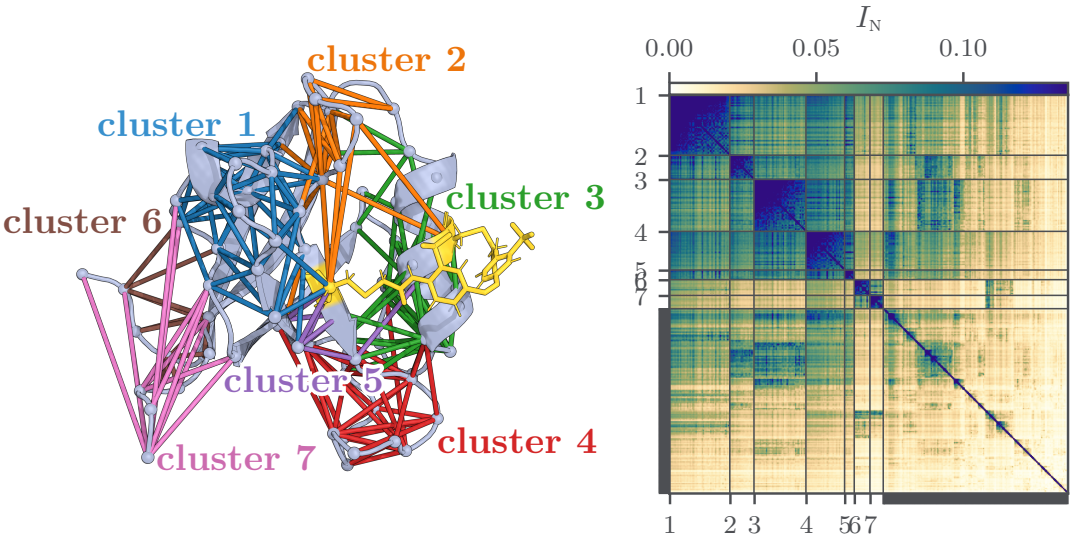

Fig. 3.5.: Left: Visualization of the clusters found using MoSAIC on normalized mutual information data for the long trajectories with resolution parameter $\gamma = 0.1$. Right: Resulting clustered similarity matrix (330×330). Each pixel represents the mutual information between two residue pairs. The average mutual information of the residue pairs in a cluster is larger than γ. Of the 330 contact distances, $46\,\%$ are not assigned to clusters and are considered noise (lower right entries).

Cluster	C1	C2	C3	C4	C5	C6	C7
Number of distances assigned	50	20	43	32	8	13	11

Tab. 3.3.: Number of residue pairs assigned to each cluster.

Cluster	1	2	3	4	5	6	7
Contacts	$d_{9,52}$,	$d_{12,19}$,	$d_{23,72}$,	$d_{24,35}$,	$d_{22,37}$,	$d_{2,6}$,	$d_{1,97}$,
	$d_{10,50}$,	$d_{13,18}$,	$d_{24,72}$,	$d_{24,36}$,	$d_{22,38}$,	$d_{2,8}$,	$d_{2,97}$,
	$d_{10,52}$,	$d_{13,19}$,	$d_{25,72}$,	$d_{24,37}$,	$d_{22,39}$,	$d_{2,52}$,	$d_{51,97}$,
	$d_{11,50}$,	$d_{14,17}$,	$d_{35,70}$,	$d_{25,28}$,	$d_{23,37}$,	$d_{2,92}$,	$d_{52,97}$,
	$d_{11,52}$,	$d_{14,18}$,	$d_{36,70}$,	$d_{25,35}$,	$d_{23,38}$,	$d_{3,6}$,	$d_{54,95}$,
	$d_{12,46}$,	$d_{14,19}$,	$d_{36,71}$,	$d_{25,36}$,	$d_{23,39}$,	$d_{3,92}$,	$d_{54,96}$,
	$d_{12,47}$,	$d_{14,20}$,	$d_{36,75}$,	$d_{25,37}$,	$d_{24,39}$,	$d_{4,92}$,	$d_{54,97}$,
	$d_{12,49}$,	$d_{15,18}$,	$d_{36,76}$,	$d_{26,29}$,	$d_{24,77}$	$d_{4,93}$,	$d_{56,95}$,
	$d_{12,50}$,	$d_{15,19}$,	$d_{62,75}$,	$d_{26,31}$,		$d_{4,94}$,	$d_{56,97}$,
	$d_{13,46}$,	$d_{15,83}$,	$d_{62,82}$,	$d_{26,34}$,		$d_{4,95}$,	$d_{92,95}$,
	$d_{19,45}$,	$d_{16,19}$,	$d_{63,78}$,	$d_{26,35}$,		$d_{4,96}$,	$d_{92,97}$
	$d_{19,46}$,	$d_{16,83}$,	$d_{65,71}$,	$d_{26,36}$,		$d_{5,92}$,	
	$d_{19,47}$,	$d_{17,22}$,	$d_{65,74}$,	$d_{26,37}$,		$d_{5,93}$	
	$d_{20,40}$,	$d_{17,77}$,	$d_{65,75}$,	$d_{26,72}$,			
	$d_{20,41}$,	$d_{18,22}$,	$d_{66,70}$,	$d_{27,30}$,			
	$d_{20,42}$,	$d_{19,22}$,	$d_{66,71}$,	$d_{27,31}$,			
	$d_{20,45}$,	$d_{19,77}$,	$d_{66,72}$,	$d_{27,33}$,			
	$d_{20,46}$,	$d_{19,86}$,	$d_{66,75}$,	$d_{27,34}$,			
	$d_{20,47}$,	$d_{19,88}$,	$d_{67,70}$,	$d_{27,72}$,			
	$d_{21,39}$,	$d_{21,38}$	$d_{67,72}$,	$d_{28,31}$,			
	$d_{21,40}$,		$d_{67,75}$,	$d_{28,32}$,			
	$d_{21,46}$,		$d_{67,76}$,	$d_{28,33}$,			
	$d_{21,47}$,		$d_{67,78}$,	$d_{28,34}$,			
	$d_{22,40}$,		$d_{67,79}$,	$d_{28,37}$,			
	$d_{22,42}$,		$d_{68,71}$,	$d_{28,67}$,			
	$d_{38,41}$,		$d_{68,72}$,	$d_{29,32}$,			
	$d_{40,53}$,		$d_{69,72}$,	$d_{31,34}$,			
	$d_{40,54}$,		$d_{70,73}$,	$d_{31,35}$,			
	$d_{41,47}$,		$d_{70,74}$,	$d_{31,37}$,			
	$d_{41,48}$,		$d_{70,75}$,	$d_{31,58}$,			
	$d_{41,53}$,		$d_{71,74}$,	$d_{33,37}$,			
	$d_{41,54}$,		$d_{71,75}$,	$d_{34,37}$			
	$d_{42,45}$,		$d_{72,75}$,				
	$d_{42,48}$,		$d_{72,76}$,				
	$d_{43,48}$,		$d_{73,77}$,				
	$d_{44,48}$,		$d_{74,77}$,				
	$d_{44,49}$,		$d_{74,78}$,				
	$d_{45,48}$,		$d_{75,78}$,				
	$d_{45,49}$,		$d_{75,79}$,				
	$d_{46,49}$,		$d_{76,80}$,				
	$d_{46,50}$,		$d_{78,82}$,				
	$d_{46,53}$,		$d_{82,85}$,				
	$d_{47,50}$,		$d_{82,86}$				
	$d_{47,51}$,						
	$d_{47,52}$,						
	$d_{47,53}$,						
	$d_{48,51}$,						
	$d_{49,52}$,						
	$d_{50,53}$,						
	$d_{51,54}$						

Tab. 3.4.: PDZ2S contact distances attributed to clusters. Secondary structure elements and important loops of the protein are color-coded: β_1, $\beta_1\beta_2$, β_2, $\beta_2\beta_3$, β_3, $\beta_3\alpha_1$, α_1, $\alpha_1\beta_4$, β_4, β_5, $\beta_5\alpha_2$, α_2, $\alpha_2\beta_6$, β_6

3.4.3 MoSAIC Results: Discussion

Large portions of the protein are covered by this resulting set of contact distances. Cluster C1 covers the α_1-helix and the surrounding area impacted by its dynamics. Cluster C2 covers the flexible $\beta_1\beta_2$-loop, which has been shown to play a role in important conformational changes of the protein before [16]. Cluster C3 covers the region around the α_2-helix, which is expected to be directly affected by photoswitching. Cluster C4 covers the $\beta_2\beta_3$-loop, which also plays a role in important conformational changes [16]. Cluster C5 links the beta strands β_2 and β_3 and contains a contact distance across the binding pocket to the α_2-helix.
Clusters C6 and C7 represent the dynamics in the N- and C-termini of the protein domain. While exhibiting high mutual information internally, these two clusters seem uncoupled from the rest of the protein in terms of instantaneous dynamics, as seen with the striking low NMI lines across the matrix in Figure 3.5. This is not surprising, as the loose ends of the amino acid chain inherently don't feature stabilizing secondary structures and are free to move rather unpredictably, compared to the rest of the protein. The coordinates of these two clusters are thus not of high interest and the work will focus on the dynamics of clusters C1-C5.
The resulting clusters cover all regions of the protein where internal motion is expected. Residues of the $beta$-strands β_1, β_4 and β_6 are rarely found in the clusters, despite being heavily represented in the initial set of distances that form contacts (as seen in Figure 3.1). This indicates their role as a stabilizing scaffolding in the protein domain, that does not show major contribution to the domain's dynamics.
Compared to other systems [38], and especially the similarly set up PDZ3 analysis by Ali et al. [48], where distinct clusters with low correlations between separate clusters were observed, high cross-correlations between clusters C1-C5 are seen here. With NMI and correlation being instantaneous measures, this indicates that there is a high general response to photoswitching in the system, vastly across the protein.

Stability of the Results

The stability of these results is quite high. Due to the non-deterministic nature of the underlying Leiden community detection algorithm small clustering variations are seen (~ 1 feature per cluster), but these are negligible. With the several simulation datasets at hand, several databases could be chosen for the clustering. Here, the long NEQ

trajectories were used, as the focus of the upcoming analysis shall be laid upon NEQ dynamics and as these are expected to be more converged than the short NEQ trajectories. For comparison, MoSAIC clusterings have also been performed on the above listed other datasets and with generally similar pictures drawn, the results are stable [1].
An additional observation worth noting is that MoSAIC does not require high time resolutions to gain stable clustering results. A comparison of results for different time resolutions is given in the appendix in Figure A.1.

3.5 Resulting Datasets

In summary, for the following analyses on PDZ2S the following datasets will be used, if not mentioned otherwise. The datasets feature a time resolution of $2\,\mathrm{ns}$, due to the Gaussian filtering applied.

NEQ: Of the originally 100 short trajectories, $75 \times 1\,\mu\mathrm{s}$ are selected, and of the original 20 long, $14 \times 10\,\mu\mathrm{s}$ trajectories with intact α_2-helices are selected.

EQ: Of the original trajectories, the first $3\,\mu\mathrm{s}$ are treated as relaxation periods and are therefore discarded. Of the $6 \times 10\,\mu\mathrm{s}$ cis EQ trajectories, 3 are selected and 5 of the $6 \times 10\,\mu\mathrm{s}$ trans EQ trajectories with intact α_2-helices are selected.

MoSAIC: Using the MoSAIC feature selection, clusters 1 to 5 are selected. They cover the relevant regions in the protein and show cross-correlations among each other.

[1] The clusterings for the long and short trajectories are also quite similar, which sounds unintuitive at first, but can be explained when looking at the starting positions of the trajectories. Some short trajectories seem to start out in a *cis*, some in a *trans* position, leading to a similar bandwidth of conformations like the long trajectories that start in *cis* and end in *trans*.

4 Non-Equilibrium Response of PDZ2S

This chapter focuses on the non-equilibrium response of the PDZ2S system to the photoinduced opening of the binding pocket, which mimics ligand binding to the protein domain. Using the datasets and contact clusters selected in the previous chapter, the dynamic response is analyzed with the aim of characterizing potential allosteric communication pathways. The timescales on which key structural processes occur will be analyzed and compared to experimental observations. Then, the equilibrium simulations of the *cis* and *trans* states are examined to identify their defining characteristics, and the non-equilibrium datasets are analyzed for transition paths between these states.

4.1 Non-Equilibrium Photoinduced Response of Contact Clusters

This analysis aims to identify the response to switching the system from *cis* to *trans*, in a cluster-resolved manner. For this, a combination of the short and long non-equilibrium trajectories is taken at hand. The side chain dynamics will be described by the breaking and formation of non-covalent bonds (contacts) over time and by the change in ensemble-averaged contact distances.

For each cluster C_n, the net change in formed contacts $K_n(t)$ and number of formed contacts normalized to the number S_n of residue pairs p in each cluster, $K_n^{\text{norm}}(t)$, are given. The ensemble-averaged number of contacts per cluster at time t is defined as

$$\tilde{K}_n(t) = \left\langle \sum_{p \in C_n} \Theta(d_{\text{cut}} - d_p(t)) \right\rangle_N . \tag{4.1}$$

Here, $\langle \dots \rangle_N$ denotes the ensemble average over the N trajectories. A contact is considered formed if the contact distance d_p for the residues pair is below the threshold of

$d_{\text{cut}} = 4.5\,\text{Å}$, as determined by the Heaviside function Θ. From this, the net change in contacts is given by $K_n(t) = \tilde{K}_n(t) - \tilde{K}_n(0)$, and the normalized change is defined as

$$K_n^{\text{norm}}(t) = \left\langle \frac{1}{S_n} \sum_{p \in C_n} \Theta(d_{\text{cut}} - d_p(t)) \right\rangle_N . \tag{4.2}$$

In other words, the quantity $K_n^{\text{norm}}(t)$ describes the fraction of residue pairs within a cluster that are forming a contact at time t. As a reminder: all residue pairs in the clusters are in principle capable of forming contacts – though not necessarily simultaneously – and have formed a contact at some point during the simulation, according to the applied contact definition (see section 3.2). As a third measure, the absolute change in contact distances, averaged over the cluster components, is defined as

$$r_n(t) = \frac{1}{S_n} \sum_{i,j \in C_n} |\langle \Delta r_{ij}(t) \rangle_N| . \tag{4.3}$$

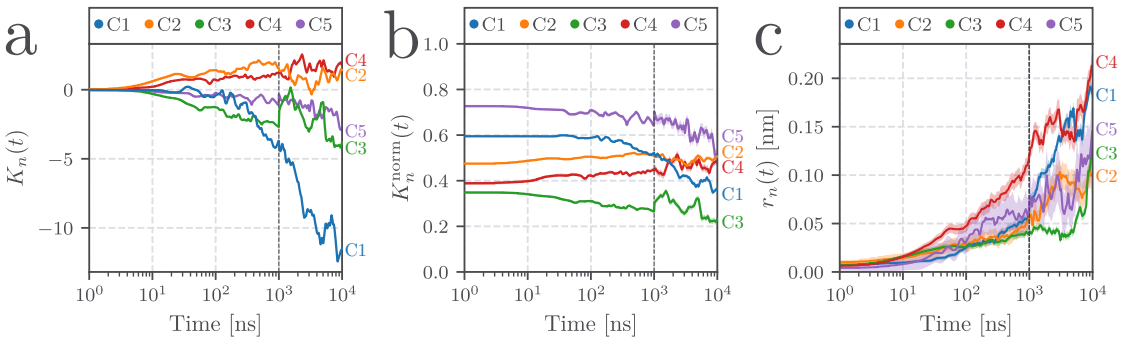

Fig. 4.1.: **a**: Net change in the number of formed contacts per cluster, K_n. **b**: Normalized number of formed contacts per cluster K_n^{norm}. **c**: Absolute ensemble-averaged change in contact distances r_n. All panels: Short and long non-equilibrium trajectories. The end of the short trajectories is marked with a dotted vertical line. Shaded areas indicate the standard error of the mean.

Figure 4.1 shows the response of these three measures after the photoswitching. The net change in the number of formed contacts per cluster $K_n(t)$ is shown in the left graph, and the relative number of contacts formed $K_n^{\text{norm}}(t)$ is shown in the center. The right panel shows the distance change $r_n(t)$. All three quantities are ensemble averages over the combined dataset of the 75 trajectories of $1\,\mu\text{s}$ length and the 14 extended $10\,\mu\text{s}$ long non-equilibrium trajectories. At a time of $1\,\mu\text{s}$, the shorter trajectories end, and the averages are from that point on computed solely from the smaller ensemble of longer trajectories. This reduction in ensemble size results in discontinuities and increased fluctuations beyond the vertical dotted line at $1\,\mu\text{s}$.

As a general remark on the quantity of absolute contact distances: by construction, this measure is always positive or zero. For large ensembles of equilibrium simulations, it is expected to converge to zero, since no net change in distances occurs. For the small ensemble of equilibrium trajectories analyzed in this work, however, values up to $0.5\,\mathrm{nm}$ are reached over $10\,\mu\mathrm{s}$, providing a reference scale for the level of fluctuations expected under equilibrium conditions. The significantly larger increases observed in Figure 4.1c therefore exceed this baseline and are indicative of a genuine non-equilibrium response rather than statistical noise.
The three quantities will be analyzed for the clusters C3 and C5 right at the photoswitch first, and for the more distant clusters C1, C2, and C4 in the section thereafter.

4.1.1 Local Response to Photoswitching

The clusters 3 ($\approx \alpha_2$-helix) and 5 ($\approx \beta_2\beta_3$-sheet) will be used to describe the local response to photoswitching, as they are positioned at the binding pocket.

Cluster 3 The number of contacts formed in cluster 3 (Figure 4.1a, b) decreases approximately logarithmically over time from $34\,\%$ to $23\,\%$, corresponding to a reduction of four contacts in total. Contrary to the steady logarithmic decrease over the whole simulation time in the number of contacts, the average change in contact distances only shows a small logarithmic increase of $0.5\,$Å up to $5\,\mu\mathrm{s}$, followed by a sudden increase up to $1.2\,$Å afterwards (Figure 4.1c). The late, sudden jump seen for the distance changes does not seem to be linked to instantaneous contact changes, but indicates that some contacts had to be broken to allow for the later distance change.

Looking at representative distances in Figure 4.2b-c, this process can be seen as a restructuring of the non-covalent bonds between the α_2-helix and the region of the β_5-strand with its surrounding residues. Figure 4.2b shows the breaking of the previously stable bond between residue 25 in the $\beta_2\beta_3$-loop and the residue 72 close to the α_2-helix ($d_{25,72}$), starting at around $5\,\mu\mathrm{s}$. The α_2-helix moves away from its initial position towards the direction of its ASN81-end, parallelly to the β_2-strand. It is forming new stabilizing contacts at around $5\,\mu\mathrm{s}$ with the end of the β_4-strand ($d_{62,75}$, Figure 4.2d) and the loops around the β_5-strand, seen e.g. with $d_{67,72}$. The process is visualized in Figure 4.2a.

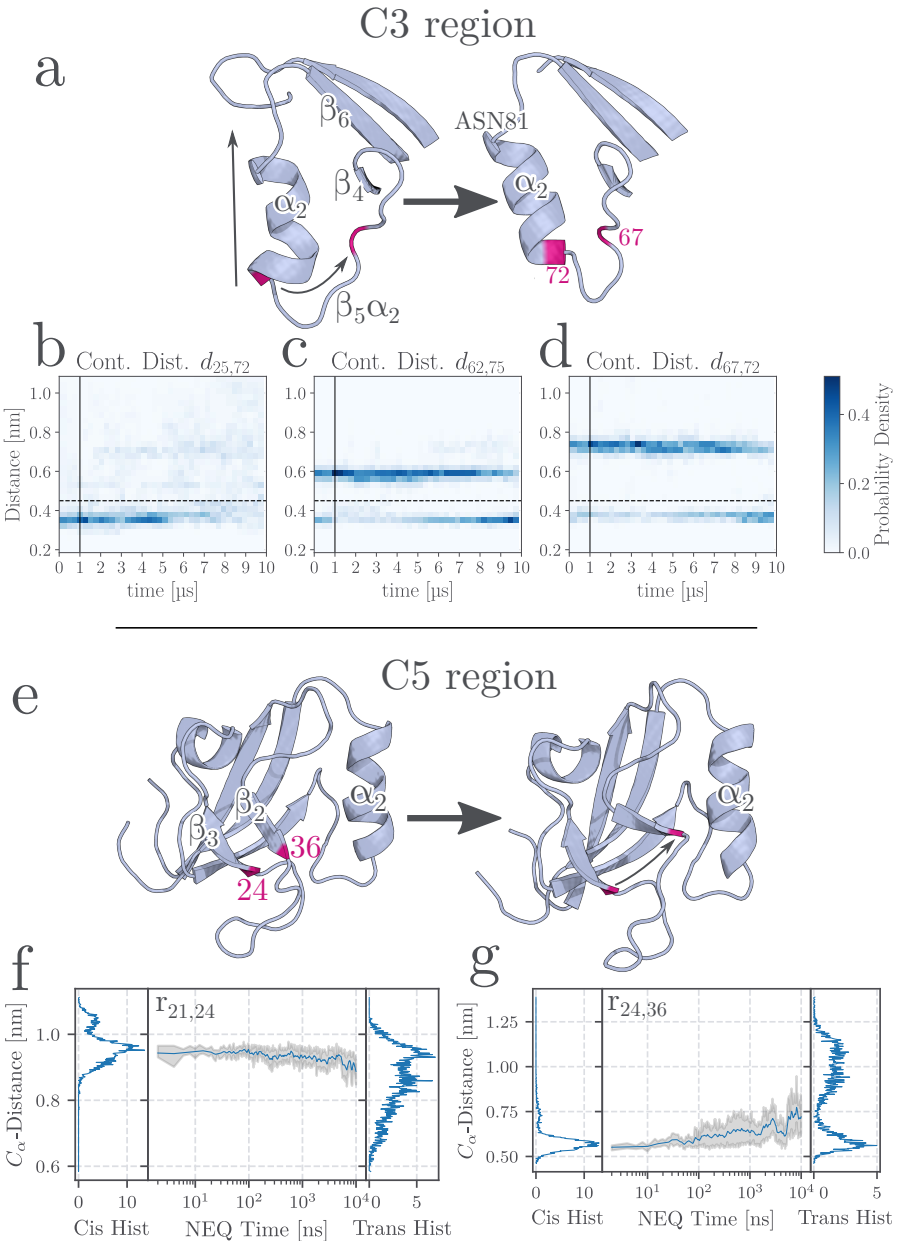

Fig. 4.2.: Local non-equilibrium responses in C3 and C5. a: Visualization of the rearrangement in the C3 region. The α_2-helix shifts "up" towards its ASN81 end, and the $\beta_5\alpha_2$-loop forms new stabilizing bonds. **b, c, d**: Probability densities of observed contact distances across the non-equilibrium trajectories. The $0.45\,\mathrm{nm}$ contact definition line is marked horizontally, the end of the short non-equilibrium trajectories is marked at $1\,\mu\mathrm{s}$, vertically. Starting at $5\,\mu\mathrm{s}$, the $d_{25,72}$-bond breaks, while the $d_{62,75}$- and $d_{67,72}$-bonds are formed. **e**: Visualization of β_2 and β_3 drifting apart at one end as β_2 bends. **f, g**: *Cis* and *trans* equilibrium distributions and non-equilibrium time evolution of $r_{21,24}$ and $r_{24,36}$ C_α-distances, showing β_2 bending, and β_2 and β_3 drifting apart, respectively. In the equilibrium histograms it becomes apparent that the bonded β-sheet is the favored conformation in the *cis* equilibrium simulations, while *trans* shows a significant occurrence of a broken β-sheet. Both ensemble averages shift from the typically *cis*-sampled distances towards more typically *trans*-sampled distances.

Cluster 5: Bending of the β_2-strand For cluster 5, three out of eight possible contacts break over the course of $10\,\mu\mathrm{s}$ (Figure 4.1a). Similar to the cluster's average change in contact distances of $1.4\,\text{Å}$ in total (Figure 4.1c), this occurs in a stretched-exponential-like manner. The contacts breaking here are β-sheet bonds between β_2 and β_3. This is connected to an outward bending of the β_2-strand on the microsecond timescale, which becomes evident when examining the C_α-distances $r_{24,36}$ and $r_{21,24}$ given in Figures 4.2f and g. The first distance describes the spacing between the two ends of the β_2-strand and shows a decrease over time, indicating the bending. The latter reflects the increasing distance between the end of the β_2-strand and the beginning of the β_3-strand, which are initially in contact but separate over time. Bonds between β-strands have been shown to break under mechanical forces [53], and the directly attached photoswitch could be the cause of this. The process is illustrated in Figure 4.2e.

4.1.2 Long Distance Response

In this section, the focus shall be laid upon the more distant response sites at cluster 1 around the α_1-helix, cluster 2 at the $\beta_1\beta_2$-loop and cluster 4 around the $\beta_2\beta_3$-loop, as well as the response site at the C-terminus, that was found by Buchenberg et al. [16].

Cluster 1 The percentage of formed contacts in cluster 1 drops from $60\,\%$ to only $36\,\%$ over the course of the simulations (Figure 4.1b). This process of breaking contacts starts from $100\,\mathrm{ns}$ on and happens exponentially in time. Both in relative and absolute change in number of contacts this is the strongest response seen among the clusters. Conjointly, the average change in contact distances reaches about $2\,\text{Å}$ per residue pair during simulation time in a stretched exponential manner (Figure 4.1c).
The underlying process of these changes is an unfolding of the α_1-helix. This can be seen representatively in the interhelical contact distance $d_{47,50}$ (Figure 4.3a), which, after starting out as a stable contact, fulfills the condition of a contact in less than half of the ensemble at the end of the simulation time, matching the results presented in Table 3.2. Additionally, contacts of the α_1-helix are lost with the $\beta_1\beta_2$-loop ($d_{20,41}$, Figure 4.3b) and the $\beta_3\alpha_1$-loop ($d_{42,48}$, Figure 4.3c). This shows that there is a high likelihood of the helix unfolding during the course of the non-equilibrium response and that the whole region of cluster 1 is rebuilt into a more flexible loop-like structure.
This long distance response from the binding pocket to the α_1-helix is also seen in form of log-periodic oscillations in contact distances like $d_{23,51}$, see Figure 4.4. These findings

suggest that the processes in the binding pocket and the distant α_1-helix are likely coupled via staircase shaped free energy barriers [41]. The log-periodic oscillations found here[1] appear with a period of $\tau_{log} = 1.49$ orders of magnitude. As introduced by Dorbath et al. [41], this is to be interpreted as a period in logarithmic time: whereas a linear time period τ characterizes time series such as $\cos\left(\frac{2\pi}{\tau}t\right)$, a logarithmic period refers to time series of the form $\cos\left(\frac{2\pi}{\tau_{log}}\ln(t)\right)$.

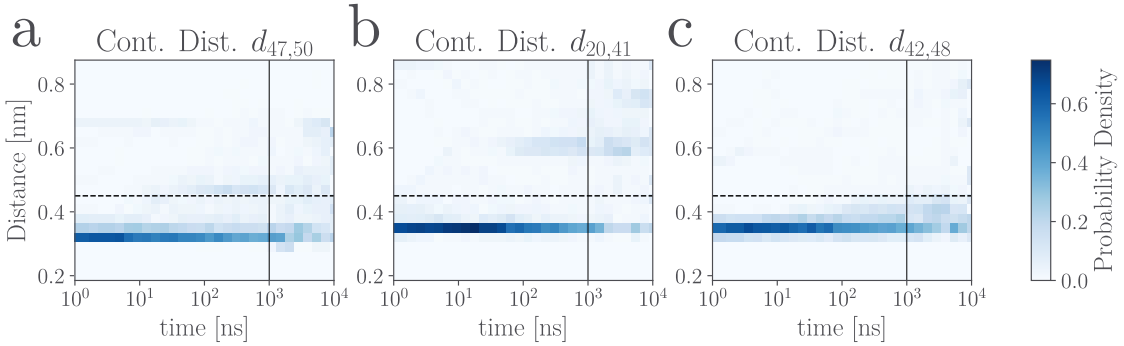

Fig. 4.3.: **Cluster 1:** Probability densities of contact distances across the non-equilibrium trajectories. The $0.45\,\mathrm{nm}$ contact definition line is marked horizontally, the end of the short non-equilibrium trajectories is marked at $1\,\mu$s, vertically. $d_{47,50}$ is a distance inside the α_1-helix, while $d_{20,41}$ and $d_{42,48}$ describe bonds between the α_1-helix and the $\beta_1\beta_2$-loop and $\beta_3\alpha_1$-loop, respectively. All three contact distances typically start out with a non-covalent bond ($d < 0.45\,\mathrm{nm}$), but lose their bonded nature over the course of the simulation.

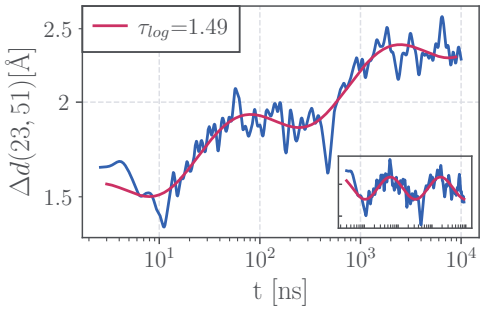

Fig. 4.4.: Log-oscillations seen in $d_{23,51}$, reaching from β_2-strand to α_1-helix region. In blue, the ensemble-averaged C_α-distance is given, in red a log-oscillation fit with period $\tau_{log} = 1.49$ is given. The inset shows the same data, adjusted for its linear increase to clearly point out the log-oscillatory behavior.

[1]The 14 long NEQ trajectories have been used as the underlying dataset. Fourteen trajectories is a small ensemble size for such an analysis. Yet, the oscillations are still visible for the shorter dataset with 75 trajectories up to the trajectory end at $1\,\mu$s, indicating that these are genuine dynamics and not just noise.

Cluster 2 Cluster 2 does not show changes this large, as it only forms between one and two additional contacts over the simulation time (Figure 4.1a). The average contact distance change increases exponentially to a modest 1 Å within the first $3\,\mu$s, after which it plateaus, indicating a stable conformation beyond $3\,\mu$s (Figure 4.1c).

Cluster 4 Cluster 4 shows a more dynamic behavior than cluster 2: While the number of formed contacts remains relatively stable – gaining two additional contacts over the course of the $10\,\mu$s simulation – the average contact distance changes considerably. It exceeds 2.2 Å per residue, which is the largest change seen among all clusters (Figure 4.1c). This reflects a considerable restructuring in cluster 4, which is linked to the transitions observed in neighboring clusters 3 and 5. In cluster 4, effects of the rearrangement of the α_2-helix – previously described for cluster 3 and illustrated in Figure 4.2a – are clearly visible. In line with the rearrangement of the α_2-helix, the previously described breaking of the $d_{25,72}$ bond (associated to C3) is replaced by the formation of the $d_{27,72}$ bond (associated to C4, Figure 4.5), which better accommodates the new position of the α_2-helix.

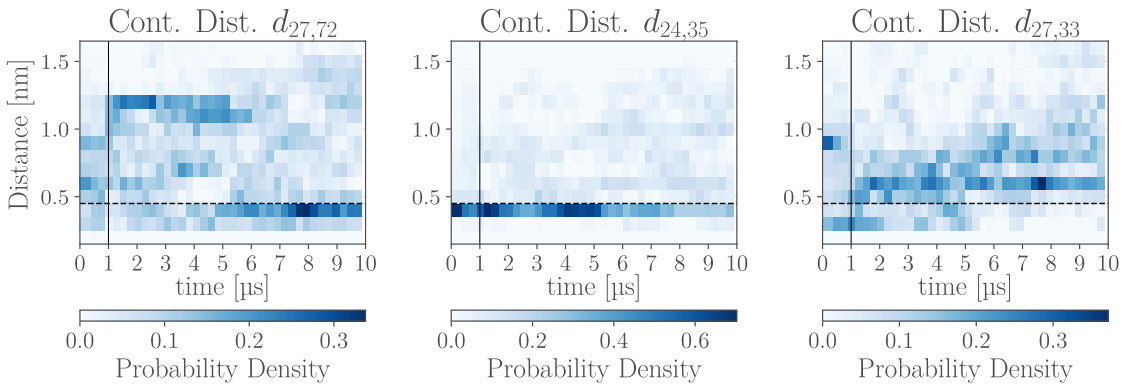

Fig. 4.5.: Cluster 4: Probability densities of contact distances across the non-equilibrium trajectories. The 0.45 nm contact definition line is marked horizontally, the end of the short non-equilibrium trajectories is marked at $1\,\mu$s, vertically. The distance $d_{27,72}$ connects the $\beta_2\beta_3$-loop to a residue near the α_2-helix, where a non-covalent bond forms after $5\,\mu$s. The distance $d_{24,35}$ lies at the interface between the $\beta_2\beta_3$-sheet and the $\beta_2\beta_3$-loop. It responds to the separation of the β-sheet by breaking its non-covalent bond starting at $5\,\mu$s. The distance $d_{27,33}$ is located within the $\beta_2\beta_3$-loop itself and increases continuously over time.

The effects of the β-sheet separation in cluster 5 are also visible in cluster 4. This is prominently reflected with the steady decrease in bonds in the cluster 4 contact

distance $d_{24,35}$ (Figure 4.5) close to the $\beta_2\beta_3$-sheet. Neither the formation nor the breaking of contacts is dominant here, however. Instead, many contact distances of non-bonded residue pairs are continuously increasing, likely caused by the β_2- and β_3-strands departing. For instance, the distance $d_{27,33}$ (Figure 4.5) is continuously increasing, after the contact distance surpassed the bonding criterion of $0.45\,\mathrm{nm}$.

C-Terminal Response Comparing PDZ2 to the well-analyzed PDZ3, where allosteric communication was found between the binding pocket and the α_3-helix at the C-terminus of the protein domain, it is natural to ask whether there is also communication to the C-terminal end of the protein domain in PDZ2. From the MoSAIC analysis in section 3.4 it was already visible that the clusters 6 and 7 at the N- and C-termini show very little correlation to the rest of the protein, due to their unbound and independent nature.
Buchenberg et al. [16] previously suggested a non-equilibrium response in the C-terminal distance $d_{92,97}$ after the photoswitching was done[2]. This distance is revisited in Figure 4.6. The mean values do not show significant changes from its starting value on, considering the standard error of the mean (Figure 4.6a). The average distance holds limited significance in this context anyway, as becomes apparent when looking at the clearly non-Gaussian distribution in the free energy (Figure 4.6b) and in distance time-traces with resolved probability densities (pFigure 4.6c). The C-terminus indeed predominantly exhibits random fluctuations with a few slightly preferred distances ($0.3\,\mathrm{nm}$ and $1\,\mathrm{nm}$), but hardly no overall trend.

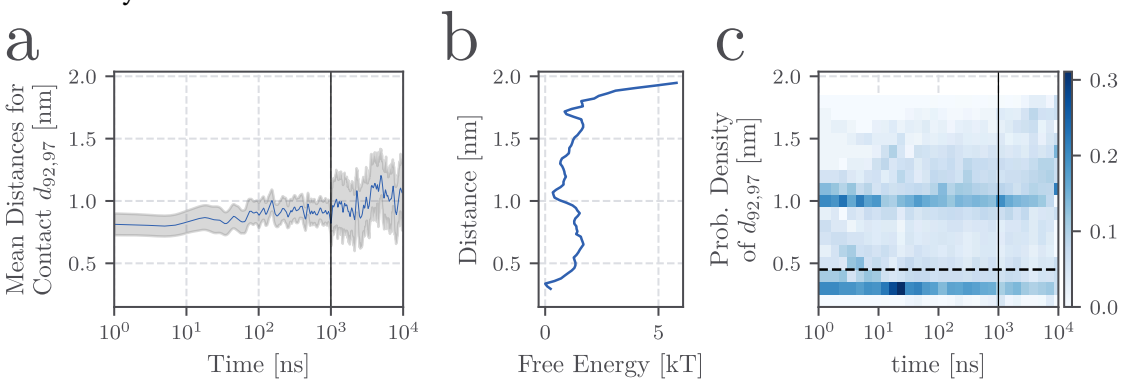

Fig. 4.6.: For the contact distance between residues 92 and 97, the mean of the time evolution is given with the standard error of the mean in gray in **a**. In **b**, the free energy landscape for the long trajectories is shown. In **c**, the time-traces with sampling probability densities are shown. A significant response to the photoswitching is hard to argue here. The black vertical lines indicate the dataset reduction after $1\,\mu\mathrm{s}$ in a and b.

[2]Note: The reference uses zero-based indexing (residues 91 and 96), whereas one-based indexing is used in this work.

There are several methodological differences between the analysis of the C-terminus in reference and in this work to note: here, contact distances are used, whereas the reference relied on C_α-distances. Additionally, the approximately $25\,\%$ of trajectories featuring a breaking of the α_2-helix are excluded here, as they were deemed unrepresentative of wild-type PDZ2 (i.e. PDZ2 without a photoswitch) behavior. Also, here, the ensemble averaged distance values are placed on a scale more representative of the range of fluctuations, which becomes apparent when inspecting the free energy and probability density along this distance in Figure 4.6b and c. The averaged distance increases by $0.1\,\mathrm{nm}$ in the better sampled first $1\,\mu\mathrm{s}$, but this remains negligible in a high-fluctuation environment, where distances in a range between $0.3\,\mathrm{nm}$ and $1.7\,\mathrm{nm}$ are frequently observed. Notably, this increase stays within the bounds of the standard error of the mean and can thus hardly be considered a response to the *cis* to *trans* transition of the photoswitch.

Local and Long Distance Response Summary Local non-equilibrium responses are seen with a rearrangement of the α_2-helix and a destabilization of the bond between the β_2- and β_3-strands. The $\beta_2\beta_3$-loop is affected by this and restructures. A more long-range, allosteric response is seen in the unfolding of the α_1-helix. Overall, contact changes and large-scale restructuring both locally and at distant sites are caused by the *cis* to *trans* transition of the photoswitch, indicating allosteric propagation across the protein.

4.1.3 Timescale Analysis

To understand the biomolecular timescales relevant for this system, it is instructive to compare the experimental data with the simulation data. This helps to identify which dynamic processes contribute to the overall behavior of the system and how these contributions differ between structural regions. The analysis was performed following the dynamic content calculation procedure described in section 2.7.

The simulation data was analyzed based on the contact distances of clusters 1 to 5, using the 14 long non-equilibrium trajectories introduced in the previous chapters. The dynamic content in the simulations is shown in Figure 4.7 b and d (overall and cluster resolved dynamic content, respectively). While an analysis using all C_α-distances in the PDZ2S dataset has been performed already [12], the present analysis offers a refined, cluster-resolved view.

This cluster resolved view is the key advantage of using simulation data, featuring its ability to break down the overall dynamics observed experimentally into contributions from specific regions and structural processes. This enables a mechanistic interpretation of the peaks observed in the experiment. When a shared timescale appears in both, the simulation data can help locate the structural regions responsible for the signal in the experiment. A direct comparison of absolute dynamic content values between experiment and simulation is not meaningful, however. In experiment and simulation, different observables are recorded (spectroscopic time traces vs. contact distance time traces). From these, different projections of the dynamic content in the protein are calculated, making their amplitudes hard to compare – therefore only the timescales will be compared.

The experimental dynamic content for the non-equilibrium PDZ2S system, shown in Figure 4.7a, features peaks at $0.6\,\mathrm{ns}$ (I), $10\,\mathrm{ns}$ (II), $100\,\mathrm{ns}$ (III) and at $4\,\mu\mathrm{s}$ (IV). The experimental data is taken from [12]. In the following subsections, the experimentally observed peaks will be discussed in detail. Using the simulation data, we aim to explain the molecular origin of these timescales, beginning with the longest timescale, the $4\,\mu\mathrm{s}$ peak.

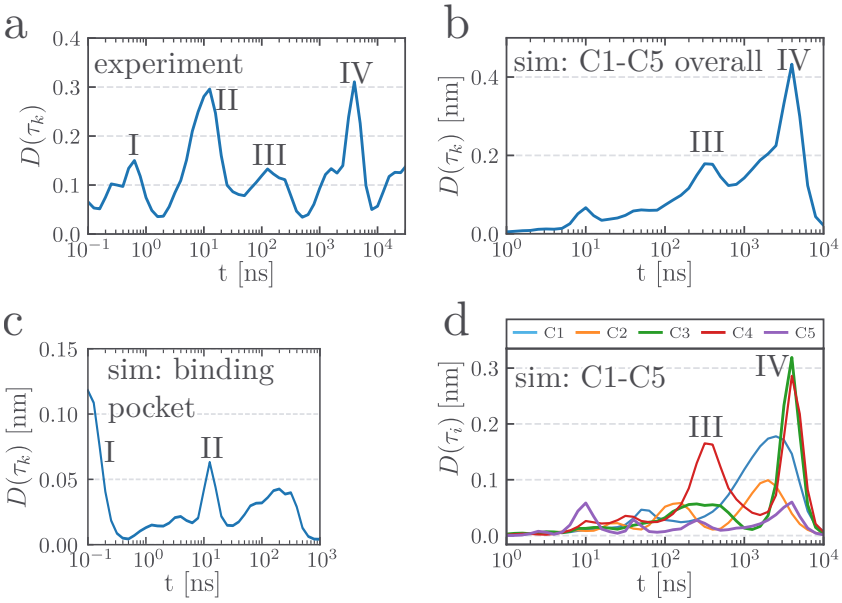

Fig. 4.7.: Timescale analysis and dynamic content results. **a**: Experimental data from [12]. Four major peaks are visible. **b:** Simulation results for the overall dynamic content in contact distance clusters 1 to 5. The peaks III and IV can also be identified here. The peak at $10\,\mathrm{ns}$ does not prove to be statistically significant. **c**: Dynamic content of all C_α pairs of the β_2-strand and the α_2-helix. Here, the peaks I and II seen in experiment can be identified. **d:** The result of b, but cluster-resolved. Peak III can be attributed to dynamics in C4, peak IV to dynamics in C1, C3 and C4.

Peak IV: α_1, α_2 and $\beta_2\beta_3$ Regions

The experimentally observed $4\,\mu$s peak is clearly reflected in the simulation's dynamic content, where the cluster-resolved analysis (Figure 4.7d) localizes this timescale predominantly in clusters 3 and 4. The peak in dynamics in cluster 1 at $2.5\,\mu$s can likely also be attributed to the experimental peak at $4\,\mu$s, considering the expected variability in timescale between experiment and the approximating simulation.
The dynamics at this timescale can in fact be traced back to the processes already described for clusters 1, 3 and 4 in the previous section, as they all appeared on a $2\,\mu$s to $5\,\mu$s timescale: The α_1-helix unwinds on this timescale, with many bonds breaking in its proximity (cluster 1). The α_2-helix rearranges in a restructuring of the non-covalent bonds between the α_2-helix itself and the region of the β_5-strand (cluster 3). Finally, the $\beta_2\beta_3$-loop restructures (cluster 4) (see the previously described Figures 4.3, 4.2a and 4.5, respectively).
Among these three processes, the unwinding of the α_1-helix represents a distal structural response and can be interpreted as an allosteric effect of the *cis* to *trans* transition of the photoswitch.

Peak III: $\beta_2\beta_3$-flip

A likely counterpart to the experimental $100\,\mathrm{ns}$ peak appears at $300\,\mathrm{ns}$ in the simulation. This can be traced back to the other significant peak in cluster 4: A different reordering of the $\beta_2\beta_3$-loop can be seen at this timescale, involving 22 of its 32 contact distances. The loop orients itself from a rather weak association with the scaffolding strands β_2, β_3 and β_4 to the region β_5/α_2, as displayed in Figure 4.8. The process is visualized in Figure 4.8a and three representative contact distances illustrate this transition in Figure 4.8b.
Initially, the most likely starting position of the loop's outer segment (residues 28-31) is close to the β_3-strand, as indicated by the contact $d_{31,37}$ between the loop and the β_3-strand. This contact then breaks, and the loop flips towards the β_5/α_2-region, decreasing the distance to the $\beta_5\alpha_2$-loop $d_{28,67}$. The final position is stabilized by a new contact in the outer loop, seen in $d_{28,31}$.
Loop dynamics tend to be sensitive to the chosen the force field [34]. In this case, the now-outdated force field *Amber99* was used. This could explain the difference in timescale between experiment and simulation ($100\,\mathrm{ns}$ vs. $300\,\mathrm{ns}$) for the loop-region.

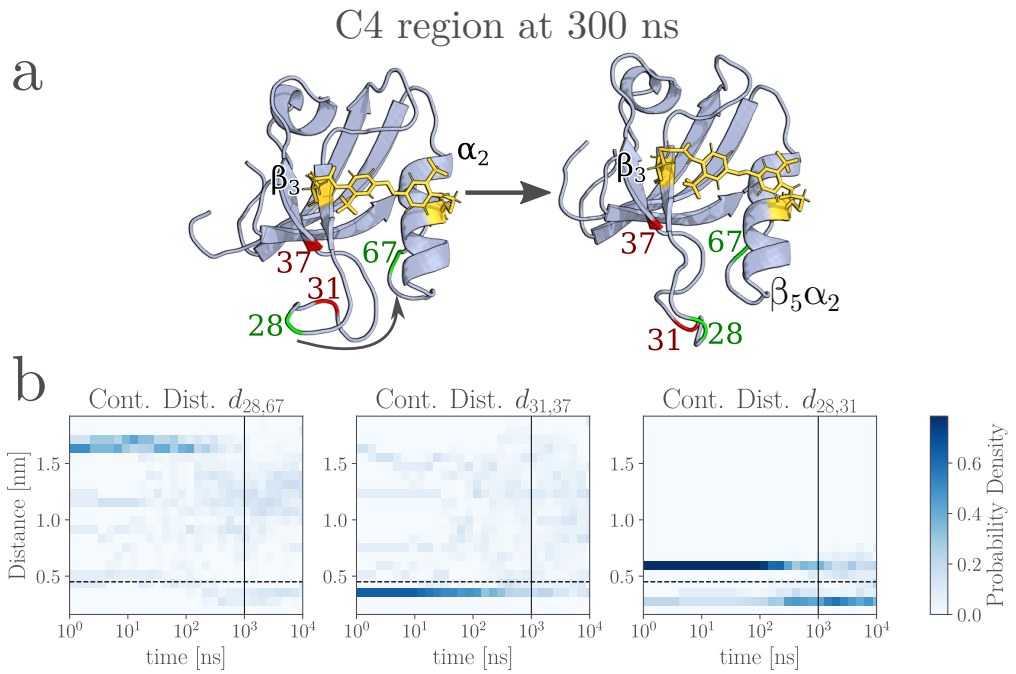

Fig. 4.8.: **Peak III: $\beta_2\beta_3$-flip.** **a:** Reordering happening in cluster 4 at the 300 ns timescale. The outer part of the loop ($\approx$ res. 28-31) flips over from the β_3 side of the protein to the α_2 side. The $d_{31,37}$ bond breaks and residues 28 and 67 approach each other. A new bond is formed in $d_{28,31}$. **b:** Probability densities of the mentioned contact distances across the non-equilibrium trajectories. The 0.45 nm contact definition line is marked horizontally, the end of the short non-equilibrium trajectories is marked at 1 μs, vertically.

Peaks I and II: Binding Pocket

Note on methodology The second-largest peak seen in experiment, the 10 ns peak, is not nearly as prominently pronounced in the contact distance simulation data in Figure 4.7b and d. In fact, even though a peak is visible at that timescale for cluster 5, a careful investigation of the underlying data shows that this peak is likely not significant: This timescale analysis was performed solely on the 14 elongated non-equilibrium trajectories, which represent a relatively small sample size. The combination of the long and short trajectory datasets introduces artifacts and could not be used for the timescale analysis, as explained in Figure A.3 in the appendix. Hence, the long trajectories have been used for calculating the dynamic content, as they allow to still represent long timescale dynamics. However, the peak at a 10 ns timescale can be better understood using the better-sampled short trajectory dataset, as it lies well inside its time boundaries. When using those trajectories, the 10 ns peak no longer

appears in the equivalent contact distance timescale analysis of cluster 5, indicating that the small peak observed in Figure 4.7b is likely a statistical outlier.
The question arises why the clear $10\,\mathrm{ns}$ experimental peak is not as prominently visible in the simulation data. The answer may lie in the nature of the internal coordinates chosen. An inherently important region of the system is its binding pocket. Yet, only one contact distance spanning it ($d_{24,77}$) is found in the collective variables of clusters 1 to 5. Six further contacts ($d_{21,79}$, $d_{22,76}$, $d_{22,77}$, $d_{23,72}$, $d_{23,79}$, $d_{24,72}$) exist in the full set of the 330 contact distances, but were identified as noise in the MoSAIC clustering process. These contact distances indeed do not exhibit relevant dynamics on the $10\,\mathrm{ns}$, except for $d_{22,76}$ which does not show any contact breaking or formation that could be considered relevant, however. Given that the photoswitch already enforces a covalent bond across the binding pocket, the dynamics in this region may be more closely tied to backbone motions. Instead of contact distances, C_α-distances are better able to describe these.

Performing a timescale analysis over the C_α-distances of the binding pocket (all residue pairs between the β_2-strand and the α_2-helix) reveals a dynamic content peak at $10\,\mathrm{ns}$ (Figure 4.7c). This suggests that the experimentally observed short-timescale response is likely associated with opening motions of the binding pocket. To maximize the time resolution (up to $20\,\mathrm{ps}$) no temporal filtering was applied on the underlying distances traces for this analysis.
In addition to the $10\,\mathrm{ns}$ peak, two further peaks emerge in this analysis. A first peak is seen at a timescale of $0.1\,\mathrm{ns}$, which reflects a faster component of binding pocket dynamics, and it could be attributed to the $0.6\,\mathrm{ns}$ peak (I) seen in experiment. While the timescales differ between experiment and simulation, both may reflect early motions localized in the binding pocket. Deviations on these timescales can be expected close to the photoswitch, given the limited validation of the force field for the azobenzene photoswitch. Finally, a broader peak is seen in this analysis around $200\,\mathrm{ns}$. It is likely connected to the similar peak seen in cluster 4, which is in proximity and can explain the dynamics at that timescale.

These findings support a two-step model of binding pocket opening, occurring on distinct timescales $0.1\,\mathrm{ns}$ and $10\,\mathrm{ns}$, as seen in the representative C_α distance time trace between residues 24 and 77 (Figure 4.9). Importantly, this two-step behavior does not seem to be mediated by a "zip-lock"-like sequential breaking or forming of contacts, as no such patterns were observed among the 330 contact distances examined.

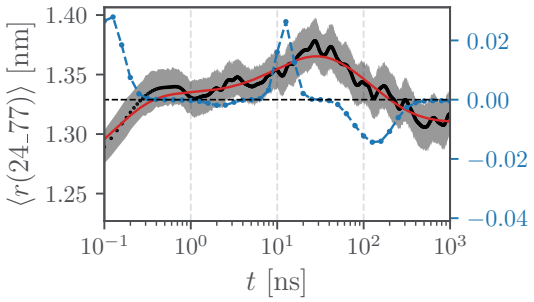

Fig. 4.9.: C_α-distance between residues 24 and 77, spanning the binding pocket from the β_2-strand to the α_2-helix. The black line shows the ensemble average over the short non-equilibrium trajectories at a time resolution of $20\,\mathrm{ps}$, with the standard error of the mean given in gray. The fitted multiexponential function is shown in red, and the corresponding fitting amplitudes s_k (cf. section 2.7) for each timescale are displayed in blue. An initial pronounced $0.04\,\mathrm{nm}$ opening of the binding pocket occurs on a $0.1\,\mathrm{ns}$ timescale, followed by a further $0.04\,\mathrm{nm}$ opening on a slower $10\,\mathrm{ns}$ timescale. The subsequent decrease in distance does not reflect a closing of the binding pocket, but instead results from a lateral shift of the α_2-helix along the pocket.

While this analysis was specifically targeted at identifying the fast timescale dynamic content in the binding pocket, it is likely that such dynamics are not observed elsewhere. In the standard, filtered analysis (shown in Figure 4.7b and d), the experimental timescale at $0.6\,\mathrm{ns}$ cannot be expected to appear, due to the limited time resolution of $2\,\mathrm{ns}$ (see section 3.5). However, even when analyzing all 330 contact distances without applying a filter, this timescale appears in only a single distance: $d_{22,76}$. But even for this distance, the change at that timescale does not reflect contact formation or breaking, as the distance decreases but remains outside from typical contact range (see Figure A.4 in the appendix). No similar feature is seen in any of the 330 other contact distances across the protein. Thus, the $0.6\,\mathrm{ns}$ dynamic appears to be localized nowhere else but the binding pocket.

Timescale Analysis Summary

In summary, four important dynamic timescales were found in the experimental data at $0.6\,\mathrm{ns}$, $10\,\mathrm{ns}$, $100\,\mathrm{ns}$ and $4\,\mu\mathrm{s}$. Corresponding peaks in the simulation appear at $0.1\,\mathrm{ns}$, $10\,\mathrm{ns}$, $300\,\mathrm{ns}$ and $4\,\mu\mathrm{s}$. The experimental peaks can thus be linked with specific internal processes: The early experimental peaks at $0.6\,\mathrm{ns}$ and $10\,\mathrm{ns}$ can be linked to opening motions of the binding pocket, which are visible in its backbone structure. On a timescale of hundreds of nanoseconds, the reorientation of the $\beta_2\beta_3$-loop from the β_3-strand toward

the α_2-helix occurs. The experimental $4\,\mu\mathrm{s}$ peak can be explained by restructurings around the binding pocket – specifically involving the β_2-strand and the α_2-helix – as well as by the distant unwinding of the α_1-helix. Among these processes, the α_1-unwinding and the associated restructuring of the region around the α_1-helix can be considered as an allosteric process, while the other explained dynamics primarily describe direct reactions in the photoswitch region.

4.2 Free Energy Landscape Analysis of the Cis to Trans Transition

A primary goal of the analysis is to understand how the conformational shift from the *cis*-conformation to the *trans*-conformation of PDZ2S is done. The three different simulation setups should help to understand this transition: The *cis* and *trans* equilibrium simulations aim to clarify typically occupied *cis* and *trans* conformations, while the non-equilibrium simulations aim to resolve the transition from *cis* to *trans*. To identify the relationships between these three simulation sets, a low-dimensional free energy landscape (FEL) analysis shall be performed. For this, a PCA will be carried out. The PCA will capture the dynamics with the most variance. For all following analyses, the contact distances of clusters 1 to 5 will be used, if not mentioned otherwise.

4.2.1 Principal Component Analysis

The PCA is performed on the non-equilibrium data, as the goal is to reduce the dimensionality of the dataset that samples the transition. The *cis* and *trans* equilibrium data will be projected onto the same principal components, to allow for comparability of conformations between equilibrium *cis*, equilibrium *trans* and non-equilibrium.
Between the two dataset options – 14 trajectories of $10\,\mu\mathrm{s}$ length and 75 trajectories of $1\,\mu\mathrm{s}$ length – the longer trajectories are chosen. They constitute more data points and, due to their extended simulation times, offer a higher likelihood of relaxing into a *trans* equilibrium state.
As input features to the PCA, the normalized and filtered (with a $2\,\mathrm{ns}$ Gaussian kernel) contact distances of contact clusters 1 to 5 are used. This selection already reduces noise

in the input features and shifts attention to important, collaborative dynamics.
The resulting free energies, cluster contributions, cumulated covariance, lifetime, and entropy statistics of the PCA are displayed in Figure 4.10. The first four principal components explain $56\,\%$ of the variance ($24\,\%$, $17\,\%$, $8\,\%$, and $7\,\%$, respectively). This suggests that a substantial portion of the proteins conformational changes can be described with just a few collective motions.
While the dynamics along the principal components are complex, an approximate and abstract structural interpretation can be made with the help of Figure 4.10b. It shows the contribution of each contact distance to the eigenvectors, grouped by clusters. The red lines indicate the average of the modulus of the eigenvector entries per cluster, to indicate the importance of a cluster to an eigenvector. For example, in the fourth eigenvector, contact distances from cluster four have a particularly strong influence. A very abstract interpretation of dynamics along PC4 can thus be made by looking at this cluster 4: High values of PC4 reflect the completed restructuring described for cluster 4 earlier in Figure 4.5. Next, the third principal component is heavily influenced by dynamics in cluster 2: low values reflect a stronger binding of the $\beta_1\beta_2$-loop to the β_1- and β_4-strand, higher values indicate more flexibility in the loop. For the second and first PC, many dynamics come into play, but major contributions are seen from cluster 1 here, which shows the α_1-helix unwinding along increasing PC1 and PC2 values.

4.2.2 Defining Cis and Trans in Equilibrium

From previous works it is known that actual convergence of the non-equilibrium trajectories to *trans* states is rarely seen [18]. To find these transitions from *cis* to *trans*, a definition of the *cis* and *trans* conformations is needed first. For this, the definitions will be based on FEL projections of the *cis* and *trans* equilibrium trajectories onto the previously described non-equilibrium principal components. The free energies are calculated using the binning method described in section 2.9.
Figure 4.11 shows the distributions of the equilibrium trajectories projected onto the first four principal components. While some PCs, like the third PC, barely separate *cis* from *trans*, others, like the first and the fourth PC, work as good discriminators. This is also quantitatively described by the Wasserstein distance (also known as Earth Mover's Distance), which measures the dissimilarity between two probability distributions [54]. The fourth principal component has the largest Wasserstein distance here, marking it as the best discriminator, followed by the first principal component.

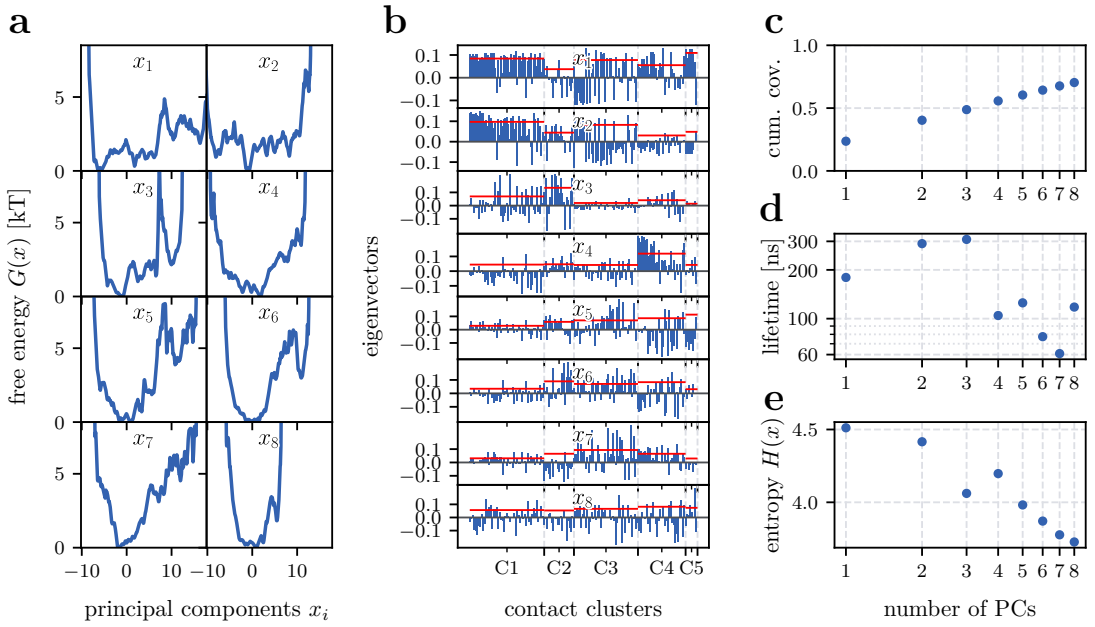

Fig. 4.10.: (a) Free energies along the first 8 principal components. (b) Contribution of each contact distance to each eigenvector. The red lines indicate the mean modulus of the eigenvector entries, to indicate the importance of a cluster to an eigenvector. (c) shows the cumulative sum of the eigenvalues. (d) shows the lifetime of the autocorrelation functions for each eigenvector. (e) shows the entropy of the free energy for each eigenvector.

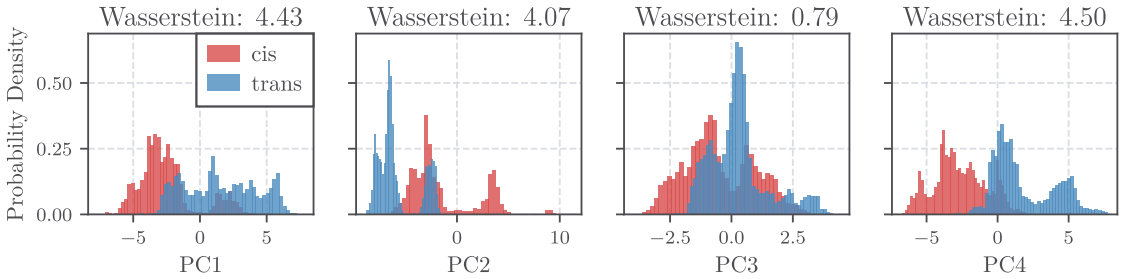

Fig. 4.11.: Distributions of the equilibrium trajectories projected onto the first four principal components. The Wasserstein distance between the *cis* and *trans* distributions is given, with high values indicating greater dissimilarity between distributions and thus between the conformational ensembles of the two states. The first and the fourth principal components separate *cis* and *trans* the best.

A two-dimensional free energy landscape of the equilibrium trajectories along these first and fourth non-equilibrium principal components is given in Figure 4.12a. The *cis* and *trans* regions (red and blue, respectively) can be clearly separated in this two-dimensional projection. Panels b and c of this figure will be discussed in subsection 4.2.3. One has to keep in mind, that this representation with just two principal components describing $31\,\%$ of the variance is prone to projection errors, i.e., a conformation might appear to lie in the *cis* region within this 2D projection but could, in fact, be better associated with a transitional state when additional dimensions are considered [36].

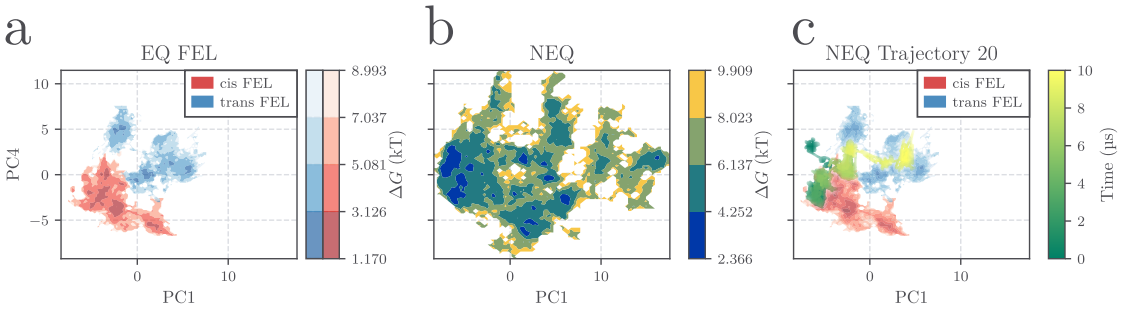

Fig. 4.12.: Free energy landscape projections along principal components 1 and 4, with consistent axes across the figures. **a:** Free energy landscapes of the equilibrium *cis* (red shades) and *trans* (blue shades) simulations. Distinct, non-overlapping minima regions exist for both states in this projection. **b:** Free energy landscape for the 14 long non-equilibrium trajectories, showing sampling of both *cis* and *trans* regions, as well as intermediate regions. The minima sampled for PC1 values larger than 10 correspond to a complete disintegration of the α_1-helix region, which is not part of the equilibrium minima in panel a. **c:** A representative non-equilibrium trajectory (index 20) is overlaid onto panel a as a time-trace, revealing that the transition from *cis* to *trans* minima occurs at $9.2\,\mu$s.

A conformation will only be regarded as *cis*- or *trans*-typical if it lies in the respective region across all first four principal components. Conformations that fulfill this for all four principal components are considered *cis* or *trans*, respectively. Higher-order principal components would only contribute marginally to this separation, as shown in the appendix (Figure A.5).

The system with the photoswitch attached in its *cis*-form is frustrated. The first two principal components already suggest that the *cis* states are not ergodically sampled, as there are unconnected regions along the most important principal components, when revisiting Figure 4.11. The *cis* projections are based on three equilibrium simulations only, and all mostly sample one independent minimum, each. The system features multiple

co-existing states rather than a single stable conformation. Buchenberg et al. [16] already showed that the *cis* conformations consist of several conformational states. The frustrated nature of the *cis* system is also visible in simpler metrics. It can already be seen when looking at the distribution of the width of the binding pocket (C_α-distance) in equilibrium in Figure 4.13. The distribution varies considerably across *cis* trajectories, with each trajectory favoring a different minimum between $1.15\,\mathrm{nm}$ and $1.40\,\mathrm{nm}$. In contrast, all *trans* trajectories predominantly sample around a consistent C_α distance of $1.45\,\mathrm{nm}$, with the single exception of *trans* trajectory 2.

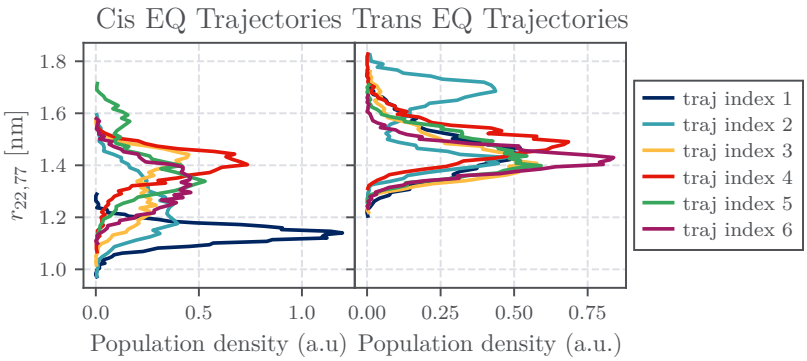

Fig. 4.13.: Population densities of the C_α-distances $r_{22,77}$ (residues of the photoswitch) in the equilibrium trajectories. Each *cis* trajectory favors its own minimum between $1.15\,\mathrm{nm}$ and $1.4\,\mathrm{nm}$. The *trans* trajectories predominantly sample around a consistent C_α distance of $1.45\,\mathrm{nm}$, with the single exception of *trans* trajectory 2. This is an indication of the frustrated nature of the system in its *cis* conformations.

The classification rules of *cis* and *trans* introduced here are based on the distributions shown above and rely only on three out of the original six *cis* equilibrium trajectories. As discussed in section 3.3, this subset was chosen carefully but cannot be fully verified. If more *cis* equilibrium trajectories were available, it is likely – given the diversity and frustrated nature already seen among the three trajectories – that additional minima would be sampled. This leads to the conclusion that the definition of *cis* conformations used here should be considered tentative. Not having a clear picture of *cis* conformations makes the effort of clearly identifying transitions from *cis* to *trans* in non-equilibrium simulations – which is the focus of the next section – more difficult. Nonetheless, the *cis* equilibrium simulations serve solely as sources for the starting points of the non-equilibrium trajectories. During the non-equilibrium simulations, the photoswitch is kept in the *trans* configuration. Therefore, the system is always driven from its starting positions toward a *trans*-like configuration.

4.2.3 Non-Equilibrium Cis to Trans Transition

Identifying Transitioning Trajectories

With the definition of *cis* and *trans* approximately given, one can search for non-equilibrium trajectories that sample the transition from one state to the other. Figure 4.12b shows the free energy landscape projection of the non-equilibrium simulations on the first and fourth principal components, representatively. In this projection, the non-equilibrium simulations clearly sample both *cis* and *trans* conformations, as well as connecting regions in between, when compared to panel a. This is a good indicator that the transition between the specified definitions of *cis* and *trans* conformations is captured in the non-equilibrium data. Continuing with the abstract structural interpretation introduced in subsection 4.2.1, PC1 spans configurations from a stable α_1-helix to its complete disintegration at large PC1 values, while increasing PC4 values correspond to a transition from the initial cluster 4 structure to the restructured conformation. Using the free energy landscapes of the first four principal components, we can analyse what pathways the non-equilibrium trajectories follow. An example is given for the non-equilibrium trajectory with index 20 in Figure 4.12c. The *cis* and *trans* FEL minima of the equilibrium trajectories are displayed again (cf. panel a) and the time trace of sampled conformations of trajectory 20 is laid upon it. It shows that trajectory 20 samples mostly conformations in the *cis* minima up to a time mark of $9.2\,\mu$s, where a transition to *trans* regions takes place. This trajectory is one of two, where such a clear transition is seen for the first four principal components. The other trajectory is the long non-equilibrium trajectory with index 11, which transitions at $2.7\,\mu$s. This also means that most non-equilibrium trajectories don't equilibrate from a *cis* conformation to a *trans* conformation in the simulation time as expected.

Analyzing the Transition

From the FEL analysis above, the transitioning points in time were inferred for the trajectories in question. The question arises what structural changes at these transitioning points occur. The major conformational changes[3] at $2.7\,\mu$s in trajectory 11 and $9.2\,\mu$s in trajectory 20 occur in cluster 5, characterized by the sudden bending of the β_2-strand

[3] This is determined by performing a PCA on each cluster's set of contact distances. A major jump in the time trace of its first components at the respective transitioning time marks is only seen for cluster five. For the other clusters, the changes at those points in time are marginal.

and its detachment from the β_3-strand, as previously described in Figures 4.2f and g. This conformational change was already discussed and visualized in the context of the ensemble-averaged dynamics in Figure 4.2e. In the histograms in Figures 4.2f and g it also becomes apparent, that the bonded $\beta_2\beta_3$-sheet is the favored conformation in the *cis* equilibrium simulations, while the *trans* equilibrium simulations show a significant occurrence of a broken β-sheet.

Despite being visible so clearly as the identifying mechanism in the *cis* to *trans* transition, the breaking of the bond between β_2 and β_3 is not seen in the literature for the *trans* state. The NMR based PDZ2S-*cis* and -*trans* structures found by Buchli et al. [28] both still show a connected β-sheet, with the C_α-distance between the end of the β_2-strand and the beginning of the β_3-strand of $r_{24,37} = 5.4\,\text{Å}$ for the *cis* and $r_{24,37} = 5.5\,\text{Å}$ for the *trans* structure. In contrast, in the non-equilibrium simulation a much larger separation of $r_{24,37} = 8.2\,\text{Å}$ is reached. Therefore, while this process is clearly seen as the discriminating feature in the *cis* to *trans* transition based on the combined analysis of equilibrium and non-equilibrium data, it likely does not occur naturally in wild-type PDZ2. It is rather a direct consequence of the usage of the photoswitch, spreading the binding pocket.

5 Comparison of Various Members of the PDZ Family

In the chapters above, the idea of finding allosteric signaling was presented by introducing the concept of a protein consisting of rigid parts and connecting hinges on the example of PDZ2S. The clusters found using MoSAIC can hereby be counted as the flexible parts, i.e. as the hinges. This conceptual framework is general enough that it should be applicable to a class of proteins. Moreover, one assumption is that when comparable clusters are found in related proteins, their dynamics should be similar.
The second PDZ domain considered in this work belongs to the PDZ family, a protein family consisting of more than 200 similarly constructed domains. The structural similarity lies mainly in the secondary structures of the PDZ domains, with most domains having the same setup of the α_1- and α_2-helix as well as the six β-strands also found in PDZ2. As an exception in secondary structure, PDZ3 contains a third α-helix at its C-terminus. The primary structures between the PDZ domains differ more. A BLAST comparison [55, 56] of the primary sequences of PDZ2S and PDZ3 analyzed in the following chapter reveals $37\,\%$ identical amino acid residues and $60\,\%$ positives (i.e., identical residues and biochemically similar substitutions). For some PDZ domains it is known to show allosteric signaling from the binding pocket to sites distant from it, like seen with the α_1-helix in PDZ2S and the α_3-helix in PDZ3 [14, 57]. The question that arises is to what extent these findings can be generalized to the broader PDZ family.

5.1 Systems at Hand

Available systems in the Stock group at the University of Freiburg are 4 non-equilibrium simulation sets of photo-switched PDZ domains. Two of them are simulations of PDZ3, two of PDZ2. All of them essentially aim to simulate a ligand binding (PDZ2S) or unbinding (all others), with different strategies:

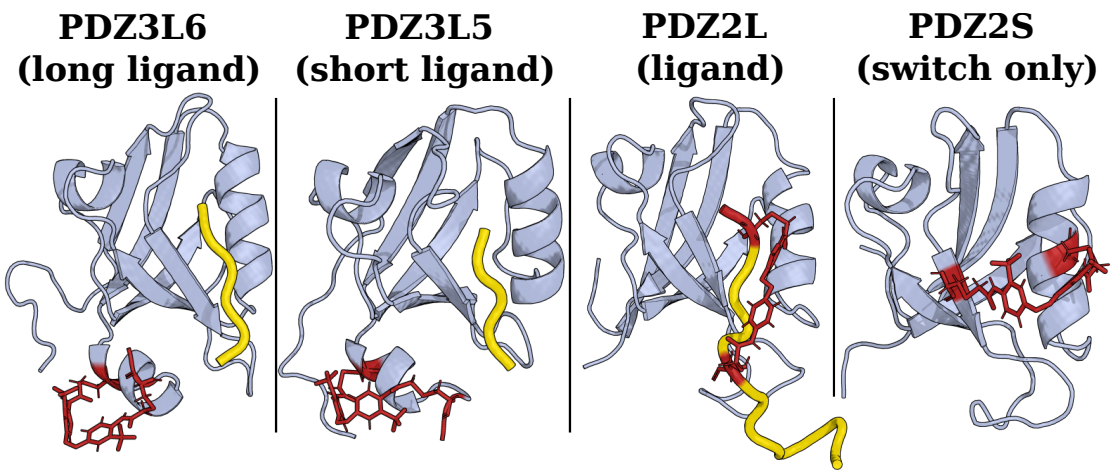

Fig. 5.1.: The four different setups of PDZ non-equilibrium simulations are illustrated with ligands in yellow and photoswitches in red. For PDZ3, the switch is breaking the α_3-helix, for PDZ2L it is squeezing the ligand out of its pocket, and for PDZ2S it is mimicking the ligand itself.

PDZ2S: One of the PDZ2 simulations is the one widely discussed in this work, with an azobenzene photoswitch (PDZ2*S*, with *S* for switch) spanning across the binding pocket. It mimics the binding of a ligand, with the *cis*-state representing the free domain and the *trans*-state representing the bound state. The switch undergoes a photoinduced change from *cis* to *trans*. It is the only simulation set without a peptide ligand.

PDZ2L: Another PDZ2 simulation features explicitly a 16 residue long peptide ligand (PDZ2*L*, with *L* for ligand) in the binding pocket with a photoswitch attached to it. In the non-equilibrium simulation, the switch undergoes a photoinduced change from *trans* to *cis*, squeezing the ligand out of the binding groove.

PDZ3L6: In the PDZ3 simulations, the photoswitch is attached to the α_3-helix. When it is opened from *cis* to *trans*, it breaks the helical structure of α_3, resembling the removal of the helix, which was shown to reduce ligand binding affinity. The simulation features a peptide ligand with a length of 6 residues in the binding pocket.

PDZ3L5: The other PDZ3 simulation set is identical, except that the ligand in this system is only 5 residues long. Again a photoswitch stretches the α_3-helix.

The four different setups are illustrated in Figure 5.1. In addition to these, equilibrium

wild-type (WT) simulations exist for both PDZ2 and PDZ3. Wild-type means that no photoswitches are attached, and only a ligand is bound to the protein.

5.1.1 Dataset Description

For PDZ2S, the results from the previous analyses are directly used. For the remaining three systems, the foundational work on the data was carefully conducted by Adnan Gulzar, Emanuel Dorbath and Ahmed Ali, and all data processing steps were carried out under their supervision and have been adopted in this thesis without modification.
The three additional systems have datasets of the following size:
For PDZ2L, 100 non-equilibrium trajectories exist for *trans* to *cis*, of which 80 trajectories are of $1\,\mu$s length and 20 are of $10\,\mu$s length.
For PDZ3L6, 99 non-equilibrium trajectories exist for *cis* to *trans*, of which 89 are of $1\,\mu$s length and 10 are of $10\,\mu$s length.
For PDZ3L5, 116 non-equilibrium trajectories exist for *cis* to *trans*, of which 90 are of $1\,\mu$s length and 22 are of $10\,\mu$s length.
For PDZ2 wild-type six equilibrium simulations of $1\,\mu$s each exist.
For PDZ3 wild-type four equilibrium simulations of $1\,\mu$s each exist.

5.2 Comparison of Contact Clusters in PDZ Systems

5.2.1 Consistency of MoSAIC Clusters Across Systems

MoSAIC clustering identifies groups of contact distances that show related motion. If similar MoSAIC clusters appear for all systems, a robust picture of protein regions with system-overarching importance for dynamics is gained. This may reduce the need for extensive non-equilibrium simulations, provided the primary goal is to detect regions of allosteric importance. However, this comparative approach relies on the ability to robustly find MoSAIC clusters for a single system first, which is limited by two challenges:

Resolution Parameter Selection There is no clear criterion on how to choose the resolution parameter γ for MoSAIC's Leiden community detection algorithm. There are data-driven methods like the silhouette score, which help to get in the right direction,

but are often different to the intuitively chosen resolution parameter. A range of γ values can yield valid results, leading to clustering solutions that vary in structural resolution.

Algorithmic Variability Even with an optimal choice of γ, the Leiden algorithm is not fully deterministic, meaning that minor variations in clustering results can occur across different runs.

With these limitations in mind, the following sections will show that clustering patterns do remain quite consistent across systems, allowing for a comparative analysis across different systems.

5.2.2 Cluster Comparisons Across PDZ Systems

MoSAIC clusterings have been performed for the above-mentioned PDZ2 and PDZ3 systems. They are visualized in Figure 5.2 and an association of structural regions to the clusters is given in Table 5.1. The clustering parameters and detailed lists of the residue pairs attributed to the clusters can be found in the appendix (Figure A.7, Tables A.1, A.2, A.3, A.4, A.5). Clusters in the same regions of the proteins will be given the same name. For PDZ3L5, the clustering was previously published by Ali et al. [48]. Since a different naming scheme is used here, Table 5.1 provides a comparison of both schemes.

Cluster	Region	Ali et al. [48]
C1	α_1	C7
C2	$\beta_1\beta_2$	C6
C3	β_2 / α_2 / ligand	C4
C4	$\beta_2\beta_3$ / β_5	C3
C5	β_2 / β_3	–
C6	$\alpha_2\beta_6$ / $\beta_1\beta_2$	C5
C7	α_3 / ligand	C1
C8	α_3 / β_3 / $\beta_2\beta_3$	C2
C9	α_2 / β_2 / $\beta_1\beta_2$	–

Tab. 5.1.: The clusters 1 to 9 have been found at least for some of the six viewed systems. The clusters found in several systems that link the same regions (like C7 linking the α_3-helix and the ligand) are given the same name. For reference, the naming scheme used by Ali et al. [48] for PDZ3L5 is given in the last column.

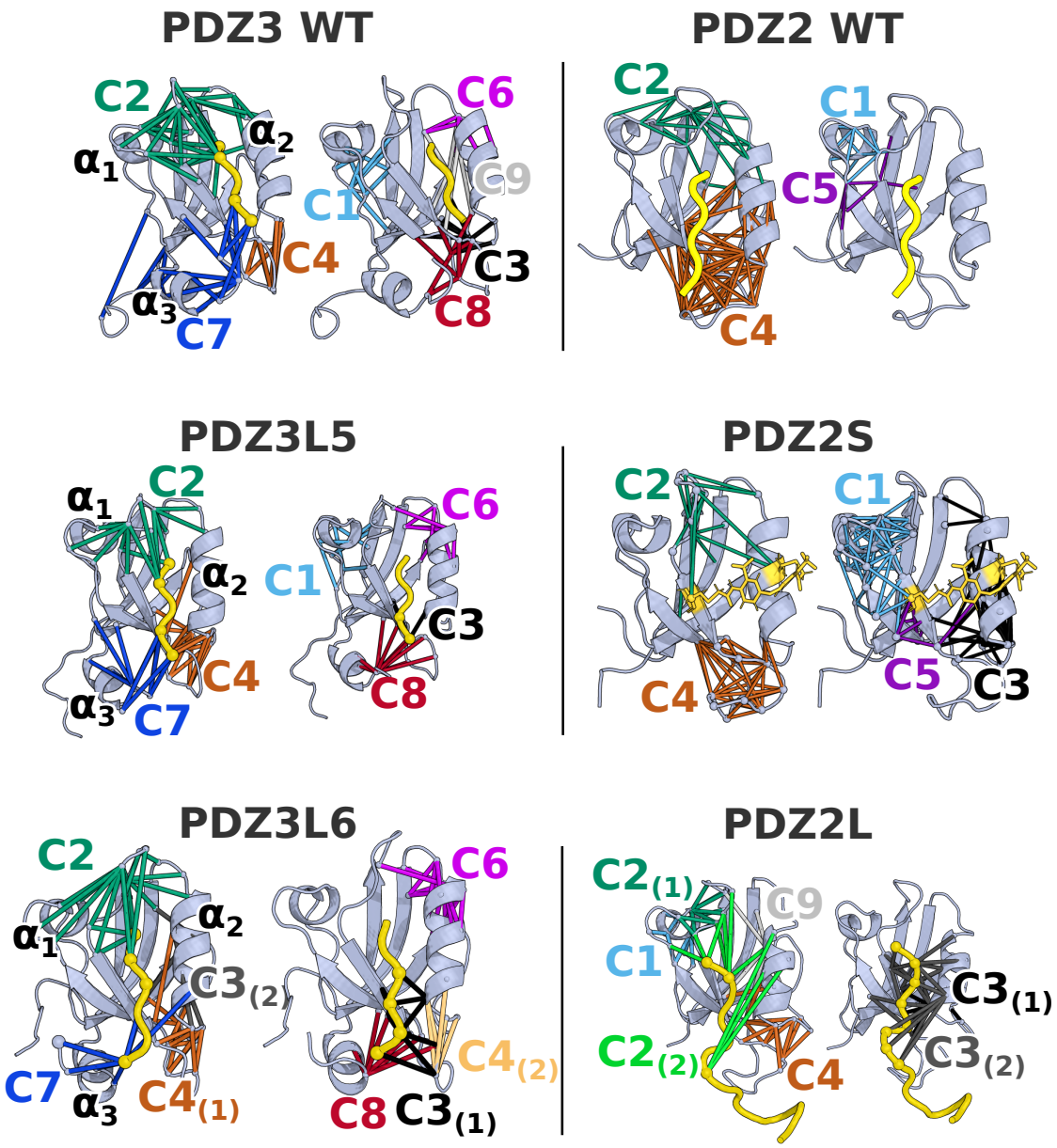

Fig. 5.2.: MoSAIC clustering results for the 6 compared systems. Note: The cluster names are consistent for PDZ2S compared to previous chapters, while colors were changed in this chapter for better comparability. Resolution parameters: $\gamma^{\text{corr}}_{\text{PDZ3WT}} = 0.35$, $\gamma^{\text{corr}}_{\text{PDZ3L5}} = 0.5$, $\gamma^{\text{corr}}_{\text{PDZ3L6}} = 0.4$, $\gamma^{\text{NMI}}_{\text{PDZ2WT}} = 0.05$, $\gamma^{\text{NMI}}_{\text{PDZ2S}} = 0.1$, $\gamma^{\text{corr}}_{\text{PDZ2L}} = 0.6$.

PDZ3 Systems For the PDZ3 systems, MoSAIC clustering identifies **eight clusters**. A comparison of the closely related long and short ligand **PDZ3** systems reveals:

- Clusters **C2, C4, C6, and C8** are highly similar.
- Clusters **C3 and C7** exhibit differences considering the contacts to the ligands, which are of different size in the two systems.
- Cluster **C1** appears exclusively in the short-ligand system.

For the wild-type equilibrium PDZ3 system (without photoswitching), clusters **C1, C2, C3, C4, and C6** remain largely conserved, while **C7 and C8** differ from the photoswitched non-equilibrium systems. These two clusters are connected to the α_3-helix, and thus it can be expected that the photoswitching dynamics at this helix influence the cluster formation. Additionally, a new minor cluster, **C9**, is observed for the wild type system. Overall, a large overlap between the PDZ3 systems is seen.

PDZ2 Systems For the PDZ2 systems the yielded clusters are more diverse. Nevertheless, there are key similarities to the PDZ3 systems:

- Clusters **C1, C2, and C4** appear across all PDZ2 systems, consistent with findings for PDZ3.
- Clusters **C3 and C5** are present but vary across PDZ2 variants.
- Clusters **C7 and C8**, observed in PDZ3, are absent. This is expected due to their association with the PDZ3-specific α_3-helix.
- **C3** is present in both non-equilibrium systems (PDZ2S, PDZ2L) but absent in wild-type PDZ2.
- The minor cluster **C9** is exclusively found in PDZ2L.

5.2.3 Interpretation and Conclusion

Overall, there is a **strong similarity** between clusters identified across different systems. Notably, the following clusters appear consistently:

- **C2** ($\beta_1\beta_2$ loop)

- **C3** (α_2/binding pocket)
- **C4** ($\beta_2\beta_3$ loop)

Clusters **C1** (α_1) and **C6** ($\alpha_2\beta_6$-loop and $\beta_1\beta_2$-loop) also show partial consistency.

It is important to acknowledge that a perfect correspondence of clusters across the systems is unrealistic. Again, PDZ2 and PDZ3 share a very similar secondary structure, but the underlying primary structure differs substantially [57], leading to differences in side chain dynamics. Furthermore, cluster-specific differences are expected:

- **C7 and C8** are linked to the PDZ3-specific α_3-helix and therefore do not appear in PDZ2 systems.
- **C5 and C9** are very small in size. Smaller clusters are more likely to emerge in the MoSAIC analysis due to system- or even trajectory-specific dynamics and are therefore less likely to play a major role in overall protein dynamics.

C6 on the other hand could be expected to be present across all systems, but it is absent in PDZ2S. Despite these discrepancies, the **overall agreement between clusters is remarkably strong** across the systems. This suggests that these regions are:

1. Functionally relevant for protein dynamics.
2. Largely confined as independently correlated regions.

5.3 Comparison of Dynamics in PDZ Systems

5.3.1 Introduction and Methodological Notes

Now that structurally similar clusters across the different PDZ systems are identified, the goal is to compare their dynamic characteristics. For PDZ2S, the analysis of dynamics used the quantities of *changes in number of contacts*, *changes in contact distances*, and the associated dynamic contents and *timescales*. However, this comparative analysis will focus exclusively on the dynamic content retrieved from the timescale analysis.
The metric of contact changes is limited in its comparability. The count of formed contacts is naturally very dependent on the very set of residue pairs in a cluster, which do vary between the systems looked at here. Their time evolutions will thus not be used

for comparison here. The distance change offers similar insights into dynamics as the dynamic content, and for sake of overview only the well-defined dynamic content will be used. The results, resolved by clusters, are presented in Figure 5.3. The result for PDZ3L5 was published by Ali et al. [48].

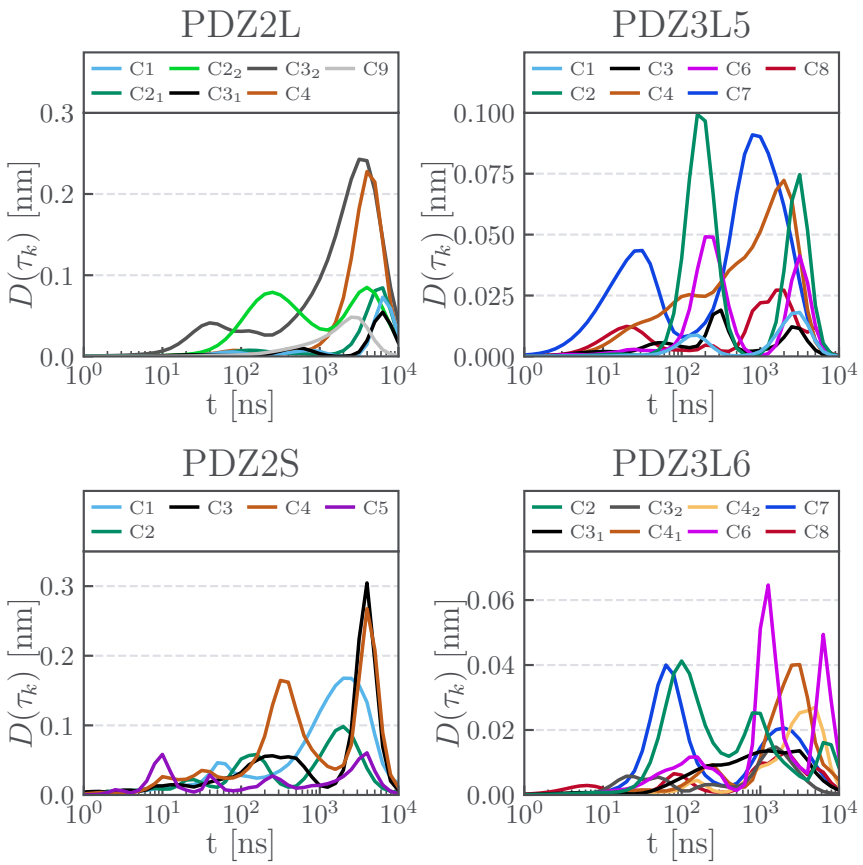

Fig. 5.3.: Dynamic contents for the 4 non-equilibrium systems, resolved for each cluster. Each subplot shows the dynamic content for the individual clusters of the four systems: PDZ2L, PDZ2S, PDZ3L5 and PDZ36. Notably, cluster 2 shows a recurring pattern across several systems, pointing to generalizable dynamics in the $\beta_1\beta_2$-loop. Differences between PDZ3L5 and PDZ3L6 highlight how minor variations in the system's setup can lead to large dynamic differences.

5.3.2 PDZ3 Comparison (PDZ3L5 vs. PDZ3L6)

First, the two PDZ3 systems will be compared, as they are the most similar pair of systems in this comparison. They only differ in the length of the ligand, having 5 or 6 residues.

For both systems, the first response seen is with cluster 7, which is a plausible and expected result, as some residues of the switched α_3-helix are part of the cluster. Cluster 7 responds slower in PDZ3L6 than in PDZ3L5, which can be explained by the extra residue in the ligand, that is added at the end of the ligand closer to the α_3-helix. This residue can lead to more stabilizing contacts between ligand and helix, hence slowing down the response.
A distant response is seen in cluster 2, with two major dynamic content peaks for PDZ3L5. One is at a $100\,\mathrm{ns}$ timescale, the other on the microsecond timescale. Interestingly, they both follow up to the peaks in cluster 7, roughly with a delay of one order of magnitude in time. Ali et al. [48] have identified this response as an allosteric transition.
For PDZ3L6, cluster 2 also follows up the first peak seen for cluster 7 with a major peak. This response is very fast however - the dynamic content peaks are almost fully overlapping on the logarithmic time axis. A logarithmic time axis more accurately reflects the underlying physical processes in this case, assuming allosteric transitions involve the crossing of exponential energy barriers. Further thorough checking on the PDZ3L6 system would be needed to confirm this as an allosteric response.
The two clusters described exhibit the largest dynamic content responses. Cluster 6 also shows some prominent peaks, appearing on similar timescales as cluster 2. Given their spatial proximity, this is to be expected, and similar conclusions can be drawn. Another strong dynamic response is seen in cluster 4, which shows one major peak around the $2\,\mu\mathrm{s}$ mark. Also, further dynamic features appear in the microsecond regime, but they are difficult to interpret – especially considering the drastically reduced ensemble size from $1\,\mu\mathrm{s}$ onward. More simulation data would be needed in this timescale region to improve reliability.
To conclude, similarities between C2, C4, C6 and C7 are found for the two PDZ3 domain systems.

5.3.3 PDZ2 Comparison (PDZ2S vs. PDZ2L)

For the PDZ2 systems generally larger dynamic contents are seen compared to the PDZ3 systems. While dynamic content peak height is not a very robust measure, also individual ensemble averaged distances show larger changes for the PDZ2 systems, when comparing similar distances over the systems. This means there are generally larger dynamics seen in the PDZ2 setups.
For PDZ2S, cluster 2 shows peaks at similar timescales (in the $300\,\mathrm{ns}$ and in the microsec-

onds regime) as the PDZ3 systems. For PDZ2L, these peaks are also visible for cluster $C2_2$. However, the contact distances in PDZ2L leading to this peak can be related to distances between the α_2-helix and the ligand, which are only *correlated* with the more typical cluster 2 distances, usually found in the $\beta_1\beta_2$ loop. These more typical distances don't show the mentioned peak for this system. So despite being attributed to cluster 2, these timescale peaks are not clearly correlated to the cluster 2 peaks seen in all other systems. Apart from this, only the peak of cluster 4, here at $\approx 3\,\mu$s instead of $\approx 2\,\mu$s, can be seen again in both PDZ2 systems. Other dynamic content timescales of the PDZ2 systems don't show mentionable similarities.

5.3.4 Discussion of Similarity in Dynamics

The only difference between the setups of the non-equilibrium simulations for PDZ3L5 and PDZ3L6 is one additional residue in the ligand. This is already causing strong differences in dynamics. For the even more distant PDZ2 systems, little overlap in the dynamics of the clusters is found. Nonetheless, cluster 2 shows a robust pattern of two characteristic peaks in at least 3, and with reservations 4 systems. This points to a recurring allosteric theme at the $\beta_1\beta_2$-region. Cluster 4 shows clear buildups to a congruent peak at around $2\,\mu$s – $3\,\mu$s over the systems, with PDZ2S showing an additional earlier peak, however.

Having two clusters with similar dynamics supports the interpretation that the clustering method reveals generally important regions for the PDZ family. The clusters showing these similarities also share one common trait: they are more spatially distant from the site of perturbation (i.e., the photoswitch site). This suggests that the more "setting-independent" clusters may share similar timescales. It is to be expected that clusters located close to the perturbation site experience greater impacts and therefore differ more from the corresponding cluster in another system. In contrast, clusters that are further away or less influenced by the photoswitch are more likely to exhibit similar timescales.

6 Conclusion and Outlook

In this thesis, the non-equilibrium allosteric behavior of PDZ protein domains is explored using molecular dynamics simulations. The central system is a PDZ2 domain modified with an azobenzene photoswitch, where the transition from the *cis* to the *trans* state imitates the effect of ligand binding. This transition is examined for a potential allosteric response. To uncover broader principles of allosteric communication in this family of protein domains, three additional similar photoswitchable PDZ domains are analyzed and compared, focusing on shared structural elements and dynamical response patterns.

Summary

To start the analysis on the PDZ2S system, the quality of the data was first assessed, and preprocessing steps were performed. Criteria for discarding individual trajectories from the dataset were developed based on structural stability, particularly focusing on the α_2-helix, which loses its integrity in several equilibrium and non-equilibrium trajectories. Affected trajectories were conservatively excluded from the analysis, as this is likely an unwanted effect of the photoswitch. The equilibrium trajectories were found to include an equilibration period of $3\,\mu$s each, which was discarded.
To describe the dynamics of the non-equilibrium simulations, 330 protein-internal coordinates in the form of inter-residue contact distances were chosen. Feature selection on these internal coordinates was performed using a Leiden-clustering-based method. This resulted in a reduced and clustered set of 153 contact distances, capturing the system's essential collective dynamic modes. The clusters were found for residue pairs in important loop regions, like the $\beta_1\beta_2$-loop and the $\beta_2\beta_3$-loop, which are known to play key roles in conformational changes [16]. Additional clusters include regions with cooperative dynamics, such as around the α_1-helix. Overall, the clusters span many regions of the protein, while largely excluding parts that suggest a stable structural scaffolding formed by the β_1, β_4, and β_6 strands – these are instead surrounded by flexible regions that enable dynamic processes potentially related to allostery.
With this abstract map of dynamically relevant clusters, the time-resolved response to the

photoswitch transition was analyzed for each structural cluster. Both local responses at the binding pocket and distal responses were observed, marked by significant changes in contact formation and inter-residue distances. Notably, the region around the α_1-helix shows a strong response on a microsecond timescale, including unwinding and restructuring. This is accompanied by log-periodic oscillations of the distance between the helix and the binding pocket, indicating structured transitions across energy barriers. Interestingly, the α_1-helix has previously been identified as a site of allosteric communication in PDZ domains [57], which aligns with its pronounced response observed here.

A timescale analysis was performed on the simulation data, aiming to further characterize the biomolecular processes and to complement earlier findings from experimental studies [12]. The four main timescales of dynamics identified in the experimental data – $0.6\,\mathrm{ns}$, $10\,\mathrm{ns}$, $100\,\mathrm{ns}$, and $4\,\mu\mathrm{s}$ – were successfully reproduced in the simulation and can even be explained mechanistically: the two faster timescales correspond to binding pocket opening dynamics, the $100\,\mathrm{ns}$ dynamics can be attributed to a reorientation of the $\beta_2\beta_3$-loop towards the α_2-helix, and the slower $4\,\mu\mathrm{s}$ timescale can be linked to a restructuring of the α_1-helix, a loosening of the $\beta_2\beta_3$ bond accompanied by a $\beta_2\beta_3$-loop restructuring, and a relocation of the α_2-helix.

An attempt was made to define *cis*-typical and *trans*-typical conformations based on the respective equilibrium simulations, using a free energy landscape analysis following PCA-based dimensionality reduction. To complement this, the non-equilibrium response was also projected into a reduced four-dimensional space, providing insight into how the starting and ending conformations of an allosteric transition – namely the *cis* and *trans* states – might be characterized. Defining *cis*-typical conformations proved challenging, with a wide range of adopted structures of the frustrated system making the picture somewhat vague. Still, the data points toward the separation of the previously bound β_2- and β_3-strands as the dominant structural change during the *cis* to *trans* transition, which is an unlikely process in wild-type PDZ2, however.

In a broader perspective, the final chapter aims to identify general patterns of dynamics across three additional photoswitched PDZ systems. All three systems – one more PDZ2 and two PDZ3 variants – feature a ligand that is brought to unbinding by the photoswitch. Interestingly, a contact clustering reveals that, despite differences in primary structure and the specific photoswitching strategies used, similar regions of cooperative dynamics emerge across all four systems. These spatially confined, distinct, and likely functionally relevant regions include clusters in the $\beta_1\beta_2$-loop, the $\beta_2\beta_3$-loop, and the α_2-helix/binding pocket region. We see that these regions act cooperatively across systems. However, their specific dynamics and timescales vary. Even minor structural changes – like the single

additional residue in the ligand – can significantly shift the dynamics within these clusters. Exceptions are clusters farther from the photoswitch or the differing ligand, such as cluster 2, which shows similar timescale dynamics across all systems. This consistency further supports its potential role in recurring internal processes involving the $\beta_1\beta_2$ region.

Outlook

This thesis focused on analyzing the response of PDZ domains to a photoswitch-induced perturbation. Several points arose during the analysis that point to important directions for future work.

Considering PDZ2 First, the photoswitched PDZ2S system needs to be better understood. So far, only three equilibrium trajectories could be used in this work. While each trajectory individually samples relatively homogeneous conformations, the conformations differ significantly between trajectories, making the dataset too small to confidently capture the ensemble of typical *cis* states. A larger number of equilibrium runs would be useful here. This would allow a clearer characterization of the equilibrium states and a better understanding of the *cis* to *trans* transition. This is essential for two reasons: first, to define what the final positions – the *cis* and *trans* states – actually look like; and second, to evaluate the reliability of the starting structures used in the non-equilibrium simulations. At present, these starting structures don't seem fully representative of typical *cis* conformations, which complicates interpretation of the transition in the non-equilibrium trajectories.
To better study wild-type PDZ2 under more natural conditions, a different and less artificial perturbation method than the photoswitch could be considered – such as pulling MD simulations, where external constraints are applied to open the binding pocket [13]. The photoswitch is elegant because it allows a precisely timed perturbation in experiments that can be simulated in an identical molecular dynamics setup. However, it also introduces significant artifacts, like the unnaturally abrupt opening of the binding pocket, that heavily impacts the proteins structure.

Allostery across systems In a broader picture, the comparison of the four systems suggests that there are recurring patterns of dynamics across the PDZ family. The processes in PDZ2S are extensively analyzed in this work, and PDZ3L5 has been studied

in detail by Ali et al. [48]. The other two systems require further investigation to better understand their specific dynamic behaviors. In particular, PDZ3L6 could be examined closer, as it may exhibit a response in cluster 2 that is coupled to a change in cluster 7 – similar to PDZ3L5, but with a faster response. However, it remains unclear whether this coupling truly exists in PDZ3L6. Clarifying whether such a communication pathway is present could shed light on more general principles of signal propagation. A deeper understanding of all four systems would allow for a more robust identification of shared or diverging patterns. Ultimately, the goal would be to develop a more general, mechanistic model in which changes in one structural region, or cluster, trigger changes in others – maybe mediated via secondary structures. This could provide insight into how signals propagate through the protein allosterically. So far, the analysis has mainly relied on correlation-based measures. A more intuitive and informative approach to this problem could make use of causality-based methods like transfer entropy [58] that can reveal actual causal influences between clusters. This could help uncover the underlying communication pathways within protein structures.

Bibliography

[1] Ivet Bahar, Robert L Jernigan, and Ken A Dill. *Protein actions: Principles and modeling*. Garland Science, 2017.

[2] J. C. Kendrew, G. Bodo, H. M. Dintzis, et al. "A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis". In: *Nature* 181.4610 (1958), pp. 662–666.

[3] Jingjing Guo and Huan-Xiang Zhou. "Protein Allostery and Conformational Dynamics". In: *Chemical Reviews* 116.11 (2016). PMID: 26876046, pp. 6503–6515.

[4] Shoshana J. Wodak, Emanuele Paci, Nikolay V. Dokholyan, et al. "Allostery in Its Many Disguises: From Theory to Applications". In: *Structure* 27.4 (2019), pp. 566–578.

[5] Hesam N. Motlagh, James O. Wrabl, Jing Li, and Vincent J. Hilser. "The ensemble nature of allostery". In: *Nature* 508.7496 (2014), pp. 331–339.

[6] Kurt Wüthrich. "Protein structure determination in solution by NMR spectroscopy." In: *Journal of Biological Chemistry* 265.36 (1990), pp. 22059–22062.

[7] Muhammed Tilahun Muhammed and Esin Aki-Yalcin. "Homology modeling in drug discovery: Overview, current applications, and future perspectives". In: *Chemical biology & drug design* 93.1 (2019), pp. 12–20.

[8] Jacob D Durrant and J Andrew McCammon. "Molecular dynamics simulations and drug discovery". In: *BMC Biology* 9.1 (2011).

[9] Samuel Hertig, Naomi R. Latorraca, and Ron O. Dror. "Revealing Atomic-Level Mechanisms of Protein Allostery with Molecular Dynamics Simulations". In: *PLOS Computational Biology* 12.6 (2016), pp. 1–16.

[10] Klemens L Koziol, Philip JM Johnson, Brigitte Stucki-Buchli, Steven A Waldauer, and Peter Hamm. "Fast infrared spectroscopy of protein dynamics: advancing sensitivity and selectivity". In: *Current Opinion in Structural Biology* 34 (2015). Carbohydrate-protein interactions Biophysical and molecular biological methods, pp. 1–6.

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

[12] Gerhard Stock and Peter Hamm. "A non-equilibrium approach to allosteric communication". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 373.1749 (2018), p. 20170187.

[13] Steffen Wolf, Benjamin Lickert, Simon Bray, and Gerhard Stock. “Multisecond ligand dissociation dynamics from atomistic simulations”. In: *Nature communications* 11.1 (2020), p. 2918.

[14] Chad M. Petit, Jun Zhang, Paul J. Sapienza, Ernesto J. Fuentes, and Andrew L. Lee. “Hidden dynamic allostery in a PDZ domain”. In: *Proceedings of the National Academy of Sciences* 106.43 (2009), pp. 18249–18254.

[15] Sebastian Buchenberg, Volker Knecht, Reto Walser, Peter Hamm, and Gerhard Stock. “Long-Range Conformational Transition of a Photoswitchable Allosteric Protein: Molecular Dynamics Simulation Study”. In: *The Journal of Physical Chemistry B* 118.47 (2014), pp. 13468–13476.

[16] Sebastian Buchenberg, Florian Sittel, and Gerhard Stock. “Time-resolved observation of protein allosteric communication”. In: *Proceedings of the National Academy of Sciences* 114.33 (2017), E6804–E6811.

[17] Brigitte Buchli, Steven A Waldauer, Reto Walser, et al. “Kinetic response of a photoperturbed allosteric protein”. In: *Proceedings of the National Academy of Sciences* 110.29 (2013), pp. 11725–11730.

[18] Georg Diez. “Markov Modeling of Nonequilibrium Biomolecular Data”. Master’s Thesis. Albert-Ludwigs-Universität Freiburg, 2020.

[19] Sebastian Buchenberg. “Energy and Signal Transport in Proteins: A Molecular Dynamics Simulation Study”. PhD Thesis. Albert-Ludwigs-Universität Freiburg, 2016.

[20] Leonard Franz. “Feature Selection for Allosteric Processes”. Bachelor’s Thesis. Albert-Ludwigs-Universität Freiburg, 2022.

[21] Anna Weber. “Markov State Modeling of an Allosteric Transition”. Master’s Thesis. Albert-Ludwigs-Universität Freiburg, 2019.

[22] Gilles Freiss and Françoise Vignon. “Protein tyrosine phosphatases and breast cancer”. In: *Critical Reviews in Oncology/Hematology* 52.1 (2004), pp. 9–17.

[23] Stoyan Milev, Saa Bjeli, Oleg Georgiev, and Ilian Jelesarov. “Energetics of Peptide Recognition by the Second PDZ Domain of Human Protein Tyrosine Phosphatase 1E”. In: *Biochemistry* 46.4 (2007), pp. 1064–1078.

[24] Ho-Jin Lee and Jie J Zheng. “PDZ domains and their binding partners: structure, specificity, and modification”. In: *Cell communication and Signaling* 8 (2010), pp. 1–18.

[25] Olga Bozovic, Brankica Jankovic, and Peter Hamm. “Sensing the allosteric force”. In: *Nature communications* 11.1 (2020), p. 5841.

[26] Olga Bozovic, Jeannette Ruf, Claudio Zanobini, et al. “The speed of allosteric signaling within a single-domain protein”. In: *The Journal of Physical Chemistry Letters* 12.17 (2021), pp. 4262–4267.

[27] Ahmed AAI Ali, Emanuel Dorbath, and Gerhard Stock. “Allosteric communication mediated by protein contact clusters: A dynamical model”. In: *Journal of Chemical Theory and Computation* 20.23 (2024), pp. 10731–10739.

[28] Brigitte Buchli, Steven A. Waldauer, Reto Walser, et al. “Kinetic response of a photoperturbed allosteric protein”. In: *Proceedings of the National Academy of Sciences* 110.29 (2013), pp. 11725–11730.

[29] Scott A Hollingsworth and Ron O Dror. “Molecular dynamics simulation for all”. In: *Neuron* 99.6 (2018), pp. 1129–1143.

[30] Wendy D. Cornell, Piotr Cieplak, Christopher I. Bayly, et al. “A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules”. In: *Journal of the American Chemical Society* 117.19 (1995), pp. 5179–5197.

[31] Franco Ormeño and Ignacio J General. “Convergence and equilibrium in molecular dynamics simulations”. In: *Communications Chemistry* 7.1 (2024), p. 26.

[32] Viktor Hornak, Robert Abel, Asim Okur, et al. “Comparison of multiple Amber force fields and development of improved protein backbone parameters”. In: *Proteins: Structure, Function, and Bioinformatics* 65.3 (2006), pp. 712–725.

[33] Robert B Best and Gerhard Hummer. “Optimized molecular dynamics force fields applied to the helix- coil transition of polypeptides”. In: *The journal of physical chemistry B* 113.26 (2009), pp. 9004–9015.

[34] Kresten Lindorff-Larsen, Stefano Piana, Kim Palmo, et al. “Improved side-chain torsion potentials for the Amber ff99SB protein force field”. In: *Proteins: Structure, Function, and Bioinformatics* 78.8 (2010), pp. 1950–1958.

[35] Phuong H Nguyen and Gerhard Stock. “Nonequilibrium molecular dynamics simulation of a photoswitchable peptide”. In: *Chemical physics* 323.1 (2006), pp. 36–44.

[36] Florian Sittel and Gerhard Stock. “Perspective: Identification of collective variables and metastable states of protein dynamics”. In: *The Journal of Chemical Physics* 149.15 (2018), p. 150901.

[37] Matthias Ernst, Florian Sittel, and Gerhard Stock. “Contact- and distance-based principal component analysis of protein dynamics”. In: *The Journal of Chemical Physics* 143.24 (2015), p. 244114.

[38] Georg Diez, Daniel Nagel, and Gerhard Stock. “Correlation-Based Feature Selection to Identify Functional Dynamics in Proteins”. In: *Journal of Chemical Theory and Computation* 18.8 (2022), pp. 5079–5088.

[39] Andrea Amadei, Antonius B. M. Linssen, and Herman J. C. Berendsen. “Essential dynamics of proteins”. In: *Proteins: Structure, Function, and Bioinformatics* 17.4 (1993), pp. 412–425.

[40]Yuguang Mu, Phuong H. Nguyen, and Gerhard Stock. “Energy landscape of a small peptide revealed by dihedral angle principal component analysis”. In: *Proteins: Structure, Function, and Bioinformatics* 58.1 (2005), pp. 45–52.

[41]Emanuel Dorbath, Adnan Gulzar, and Gerhard Stock. “Log-periodic oscillations as real-time signatures of hierarchical dynamics in proteins”. In: *The Journal of Chemical Physics* 160.7 (2024).

[42]Andreas Barth and Christian Zscherp. “What vibrations tell about proteins”. In: *Quarterly reviews of biophysics* 35.4 (2002), pp. 369–430.

[43]Irina Kufareva and Ruben Abagyan. “Methods of protein structure comparison”. In: *Homology modeling: Methods and protocols* (2012), pp. 231–257.

[44]David L Beveridge and Frank M Dicapua. “Free energy via molecular simulation: applications to chemical and biomolecular systems”. In: *Annual review of biophysics and biophysical chemistry* 18.1 (1989), pp. 431–492.

[45]Katherine Henzler-Wildman and Dorothee Kern. “Dynamic personalities of proteins”. In: *Nature* 450.7172 (2007), pp. 964–972.

[46]Roger Armen, Darwin O.V. Alonso, and Valerie Daggett. “The role of -, 310-, and -helix in helixcoil transitions”. In: *Protein Science* 12.6 (2003), pp. 1145–1157.

[47]Daniel Nagel, Sofia Sartore, and Gerhard Stock. “Selecting Features for Markov Modeling: A Case Study on HP35”. In: *Journal of Chemical Theory and Computation* 19.11 (2023), pp. 3391–3405.

[48]Ahmed A. A. I. Ali, Emanuel Dorbath, and Gerhard Stock. “Allosteric Communication Mediated by Protein Contact Clusters: AăDynamical Model”. In: *Journal of Chemical Theory and Computation* 20.23 (2024), pp. 10731–10739.

[49]Franco Ormeño and Ignacio J. General. “Convergence and equilibrium in molecular dynamics simulations”. In: *Communications Chemistry* 7.1 (2024).

[50]W. Kabsch and C. Sander. “Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features”. In: *Biopolymers* 22.12 (1983), pp. 2577–2637.

[51]Indraneel Majumdar, S. Sri Krishna, and Nick V. Grishin. “PALSSE: A program to delineate linear secondary structural elements from protein structures”. In: *BMC Bioinformatics* 6.1 (2005), p. 202.

[52]Daniel Nagel, Georg Diez, and Gerhard Stock. “Accurate estimation of the normalized mutual information of multidimensional data”. In: *The Journal of Chemical Physics* 161.5 (2024).

[53]David J Brockwell, Emanuele Paci, Rebecca C Zinober, et al. “Pulling geometry defines the mechanical resistance of a β-sheet protein”. In: *Nature Structural & Molecular Biology* 10.9 (2003), pp. 731–737.

[54]Victor M Panaretos and Yoav Zemel. “Statistical aspects of Wasserstein distances”. In: *Annual review of statistics and its application* 6.1 (2019), pp. 405–431.

[55] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, et al. “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs”. In: *Nucleic acids research* 25.17 (1997), pp. 3389–3402.

[56] Stephen F Altschul, John C Wootton, E Michael Gertz, et al. “Protein database searches using compositionally adjusted substitution matrices”. In: *The FEBS journal* 272.20 (2005), pp. 5101–5109.

[57] Amy O Stevens and Yi He. “Allosterism in the PDZ Family”. In: *International Journal of Molecular Sciences* 23.3 (2022), p. 1454.

[58] Aysima Hacisuleyman and Burak Erman. “Entropy transfer between residue pairs and allostery in proteins: quantifying allosteric communication in ubiquitin”. In: *PLoS computational biology* 13.1 (2017), e1005319.

Appendix

A

A.1 Concerning Chapter 3

The choice of the time step Δt does not play a large role on the clustering process of MoSAIC at $\gamma = 0.35$, as seen in Figure A.1. The figure shows mean cluster sizes and standard deviations over 50 MoSAIC runs for time steps chosen between a lower end of $200\,\mathrm{ps}$ and an upper end of $2\,\mu\mathrm{s}$. A large variance is only seen for the second-largest cluster. Apart from that, the results are very stable.

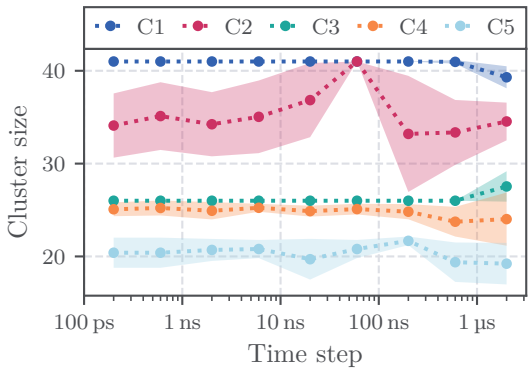

Fig. A.1.: At least for the PDZ2S non-equilibrium trajectories, performing MoSAIC clustering delivers stable results even for large time steps between frames. The dots mark average cluster sizes averaged over 50 cluster calculations per time step. The standard deviations are shown as filled areas.

The α_2-helix is not stable throughout the non-equilibrium simulations of PDZ2S. Figure A.2 shows the percentage over residues in the α_2-helix and over all frames in which helical structures are assigned by DSSP. In cis trajectory 1, almost all residues are in α-helix conformation for all frames, and in trajectory 5 over $60\,\%$ are. In the other trajectories, 3_{10}-helix configurations are seen more prominently.

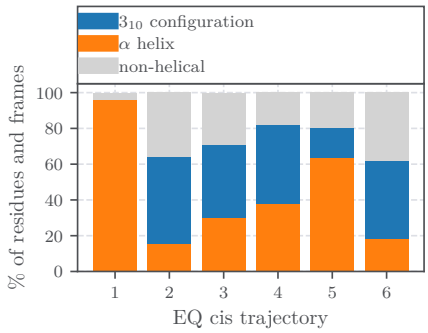

Fig. A.2.: The share of DSSP-assigned helical structures for the α_2-helix residues 74-81 over frames. 100 % share would mean that *all residues* have been in a helical structure for *all equilibrium frames*. Typically, the higher residue numbers 80 and 81 are seen less in a helical structure than the other residues.

A.2 Concerning Chapter 4

Figure A.3 shows the result for a timescale analysis of the same contact distance for a dataset of only the long NEQ trajectories on the left and the combined short and long dataset on the right. Combining the short trajectory dataset (with a larger sample size) and the long trajectory dataset (with a smaller sample size) is not feasible for this timescale analysis, as it would cause problems: Even minor jumps in the averaged time trace, caused by shifts in dataset size at 1 μs, introduce errors that render the identified timescales incomprehensible and inestimable. This is seen with the unreasonable fit and thus unreasonable peaks in the right image.

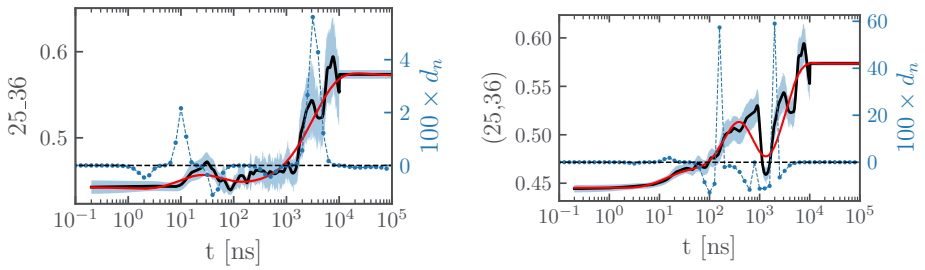

Fig. A.3.: Exemplary TSA result for contact distance $d_{25,36}$. Left: Long trajectories only. Right: Short and long combined. The black line is the ensemble averaged distance, with the error margin around it in light blue. The red line indicates the TSA multiexponential fit and its underlying amplitudes for each timescale are given as blue dots. The artifactual drop at 1 μs is heavily influencing the TSA result.

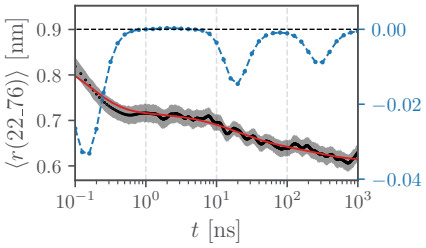

Fig. A.4.: Of all 330 contact distances in PDZ2S, $d_{22,76}$ is the only contact distance with significant timescale peaks at $0.1\,\mathrm{ns}$ and $10\,\mathrm{ns}$. There is no contact formed or broken at those timescales, however.

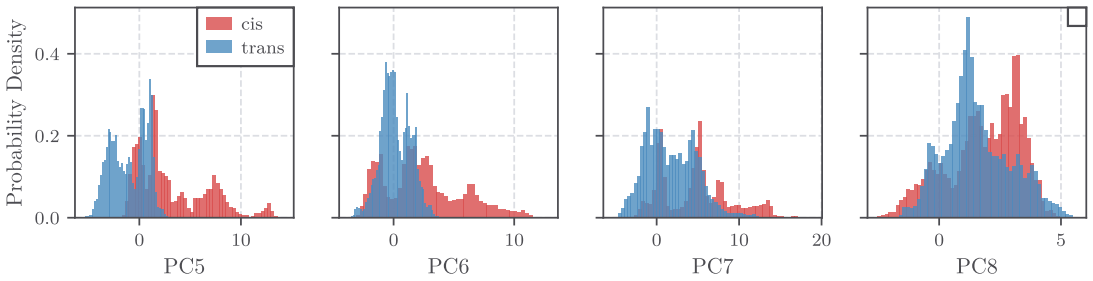

Fig. A.5.: Distributions of the equilibrium trajectories projected onto the fourth to eighth principal components. The explained variances of these NEQ PCs are $4\,\%$, $4\,\%$, $3\,\%$ and $2\,\%$, respectively, and thus they only represent minor structural changes in the data.

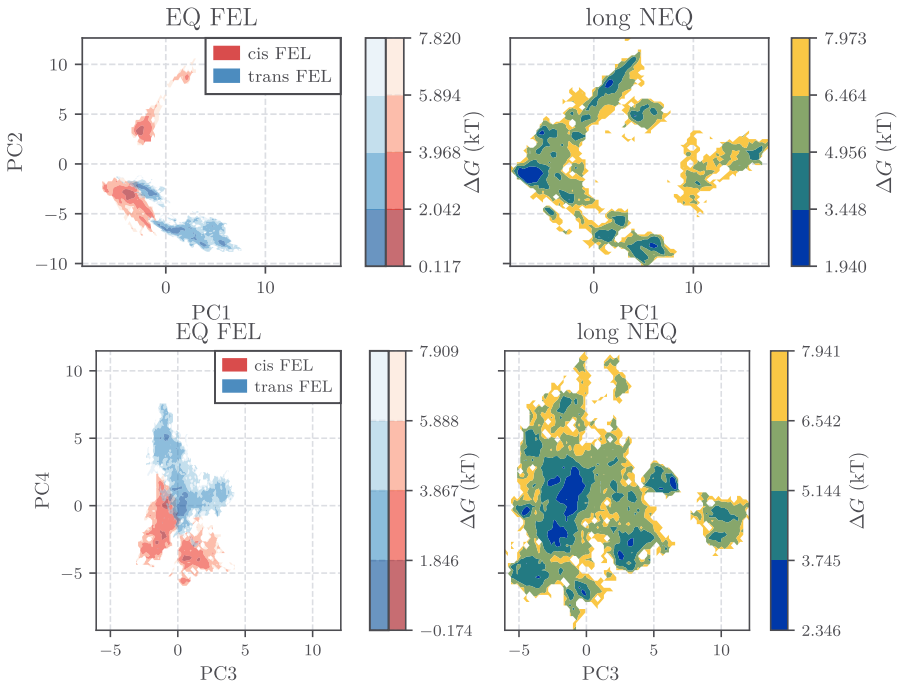

Fig. A.6.: Equilibrium and non-equilibrium free energy landscape projections along the first two (top) and third and fourth (bottom) non-equilibrium principal components. In the left panels the equilibrium cis (red shades) and trans (blue shades) EQ free energies are given. The right panels show the NEQ FELs. The EQ FELs are mostly well-separated in the projections and the NEQ FELs sample both cis and trans minima and transitioning regions. Also, the NEQ trajectories sample FEL regions clearly outside typical equilibrium minima, that are likely not transition regions, like the "island" for PC1 values larger than 10 describing fully disintegrated α_1-helices. EQ Trajectories used: *cis* 1,3,4; *trans* 2,3,4,5,6. NEQ trajectories: intact α_2, clusters 1-5.

A.3 Concerning Chapter 5

Figure A.7 shows the correlation/NMI matrices of the PDZ systems compared in chapter 5 and the tables thereafter show the exact contact distances attributed to MoSAIC clusters and the locations of secondary structures for the various systems analyzed in chapter 5.

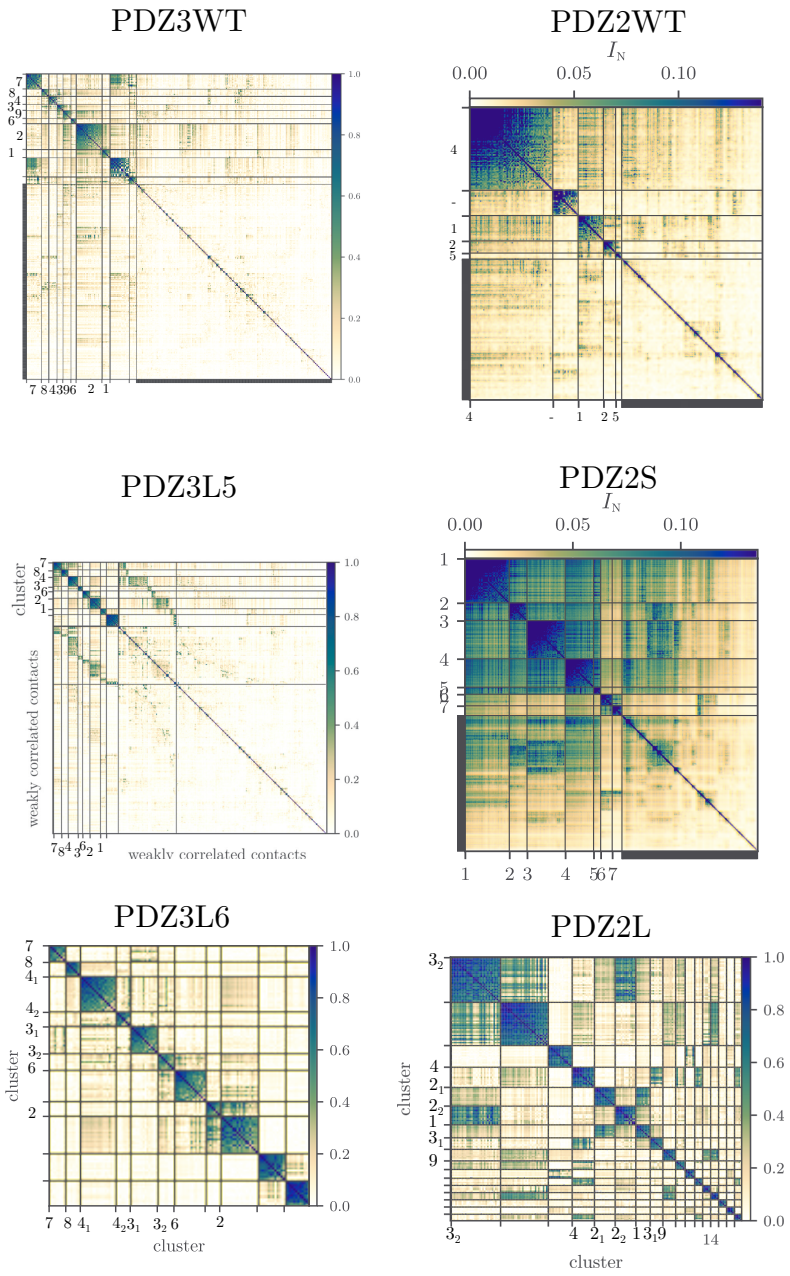

Fig. A.7.: MoSAIC matrices of the compared PDZ systems. For PDZ3L6 and PDZ2L only the non-noise parts of the matrices are given, for sake of better overview. $\gamma^{\text{corr}}_{\text{PDZ3WT}} = 0.35$, $\gamma^{\text{corr}}_{\text{PDZ3L5}} = 0.5$, $\gamma^{\text{corr}}_{\text{PDZ3L6}} = 0.4$, $\gamma^{\text{NMI}}_{\text{PDZ2WT}} = 0.05$, $\gamma^{\text{NMI}}_{\text{PDZ2S}} = 0.1$, $\gamma^{\text{corr}}_{\text{PDZ2L}} = 0.6$.

Cluster	1	2	2 (ctd.)	3	4	6	7	8	9
Contacts	$d_{47,52}$,	$d_{21,-4}$,	$d_{24,46}$,	$d_{59,68}$,	$d_{33,70}$,	$d_{78,82}$,	$d_{98,102}$,	$d_{29,35}$,	$d_{67,79}$,
	$d_{16,52}$,	$d_{21,25}$,	$d_{21,46}$,	$d_{35,69}$,	$d_{33,69}$,	$d_{82,86}$,	$d_{28,102}$,	$d_{28,35}$,	$d_{62,82}$,
	$d_{14,50}$,	$d_{20,-4}$,	$d_{80,83}$,	$d_{35,68}$,	$d_{33,68}$,	$d_{79,82}$,	$d_{102,-2}$,	$d_{30,35}$,	$d_{36,79}$,
	$d_{15,52}$,	$d_{18,-5}$,	$d_{25,42}$,	$d_{60,68}$,	$d_{33,71}$,	$d_{77,81}$,	$d_{8,94}$,	$d_{34,100}$,	$d_{63,78}$,
	$d_{25,53}$,	$d_{21,43}$,	$d_{17,22}$,	$d_{35,70}$,	$d_{32,70}$,	$d_{81,86}$,	$d_{95,102}$,	$d_{34,58}$,	$d_{62,86}$,
	$d_{47,51}$,	$d_{24,-5}$,	$d_{23,45}$,	$d_{58,68}$,	$d_{32,69}$,	$d_{63,82}$	$d_{8,92}$,	$d_{35,100}$,	$d_{59,62}$,
	$d_{50,53}$,	$d_{18,-4}$,	$d_{26,42}$,	$d_{61,68}$	$d_{34,69}$,		$d_{7,92}$,	$d_{35,67}$,	$d_{59,79}$,
	$d_{47,50}$,	$d_{22,-4}$,	$d_{18,81}$,		$d_{32,71}$,		$d_{103,-2}$,	$d_{30,34}$,	$d_{70,75}$,
	$d_{52,57}$,	$d_{21,24}$,	$d_{25,-5}$,		$d_{34,68}$,		$d_{97,101}$,	$d_{29,34}$	$d_{67,78}$,
	$d_{12,91}$	$d_{21,45}$,	$d_{19,46}$,		$d_{30,33}$		$d_{103,-1}$,		$d_{36,75}$
		$d_{18,82}$,	$d_{24,45}$,				$d_{5,54}$,		
		$d_{42,-4}$,	$d_{25,79}$,				$d_{99,102}$,		
		$d_{21,42}$,	$d_{25,47}$,				$d_{94,102}$,		
		$d_{24,-4}$,	$d_{27,-5}$				$d_{8,93}$,		
		$d_{18,83}$,					$d_{7,94}$,		
		$d_{18,86}$,					$d_{8,95}$,		
		$d_{19,22}$,					$d_{39,102}$,		
		$d_{79,83}$,					$d_{9,92}$,		
		$d_{18,21}$,					$d_{101,-1}$		
		$d_{19,84}$							

Tab. A.1.: PDZ3WT: Contact distances assigned to MoSAIC clusters.

Cluster	1	2	3	4	6	7	8
Contacts	$d_{47,52}$,	$d_{21,25}$,	$d_{28,-4}$,	$d_{35,69}$,	$d_{78,82}$,	$d_{102,-3}$,	$d_{29,100}$,
	$d_{47,51}$,	$d_{20,-1}$,	$d_{27,-3}$,	$d_{35,68}$,	$d_{62,82}$,	$d_{28,102}$,	$d_{37,100}$,
	$d_{16,52}$,	$d_{18,0}$,	$d_{29,-4}$,	$d_{33,69}$,	$d_{82,86}$,	$d_{39,102}$,	$d_{35,100}$,
	$d_{25,53}$,	$d_{21,43}$,	$d_{27,-4}$,	$d_{34,69}$,	$d_{79,82}$,	$d_{28,101}$,	$d_{30,100}$,
	$d_{15,52}$,	$d_{21,-1}$,	$d_{28,-3}$,	$d_{35,67}$,	$d_{63,82}$,	$d_{97,102}$,	$d_{28,100}$,
	$d_{14,50}$,	$d_{21,45}$,	$d_{72,-4}$	$d_{33,70}$,	$d_{81,86}$,	$d_{101,-4}$,	$d_{97,100}$,
	$d_{50,53}$,	$d_{21,24}$,		$d_{34,68}$,	$d_{77,81}$,	$d_{101,-3}$,	$d_{34,100}$,
	$d_{51,54}$	$d_{24,0}$,		$d_{58,68}$,	$d_{79,83}$,	$d_{94,102}$,	$d_{31,100}$,
		$d_{22,-1}$,		$d_{33,71}$,	$d_{18,81}$,	$d_{103,-4}$,	$d_{96,100}$
		$d_{18,-1}$,		$d_{35,70}$,	$d_{83,86}$	$d_{55,102}$,	
		$d_{21,42}$,		$d_{59,68}$,		$d_{103,-3}$	
		$d_{18,82}$,		$d_{33,68}$,			
		$d_{21,46}$,		$d_{29,71}$,			
		$d_{18,83}$,		$d_{36,79}$			
		$d_{18,86}$					

Tab. A.2.: PDZ3L5: Contact distances assigned to MoSAIC clusters.

Cluster	2	3_1	3_2	4_1	4_2	6	7	8
Contacts	$d_{21,25}$,	$d_{35,69}$,	$d_{70,75}$,	$d_{35,69}$,	$d_{31,72}$,	$d_{78,82}$,	$d_{102,-5}$,	$d_{29,100}$,
	$d_{20,-1}$,	$d_{33,69}$,	$d_{63,78}$,	$d_{33,69}$,	$d_{30,71}$,	$d_{62,82}$,	$d_{101,-5}$,	$d_{37,100}$,
	$d_{18,0}$,	$d_{35,68}$,	$d_{59,62}$,	$d_{35,68}$,	$d_{31,71}$,	$d_{82,86}$,	$d_{100,-5}$,	$d_{28,100}$,
	$d_{24,0}$,	$d_{34,69}$,	$d_{67,79}$,	$d_{34,69}$,	$d_{31,73}$,	$d_{63,82}$,	$d_{73,-5}$,	$d_{30,100}$,
	$d_{21,-1}$,	$d_{35,67}$,	$d_{78,86}$,	$d_{35,67}$,	$d_{31,70}$,	$d_{79,82}$,	$d_{103,-5}$,	$d_{35,100}$,
	$d_{21,24}$,	$d_{33,70}$,	$d_{67,74}$,	$d_{33,70}$,	$d_{30,72}$	$d_{81,86}$,	$d_{-5,-2}$,	$d_{97,100}$
	$d_{21,43}$,	$d_{58,68}$,	$d_{71,75}$	$d_{58,68}$,		$d_{77,81}$,	$d_{28,-5}$	
	$d_{18,-1}$,	$d_{34,68}$,		$d_{34,68}$,		$d_{18,81}$,		
	$d_{21,45}$,	$d_{33,71}$,		$d_{33,71}$,		$d_{77,82}$,		
	$d_{22,-1}$,	$d_{35,70}$,		$d_{35,70}$,		$d_{79,83}$,		
	$d_{21,42}$,	$d_{59,68}$,		$d_{59,68}$,		$d_{83,86}$,		
	$d_{18,82}$,	$d_{33,68}$,		$d_{33,68}$,		$d_{80,83}$,		
	$d_{18,83}$,	$d_{29,71}$,		$d_{29,71}$,		$d_{63,81}$,		
	$d_{21,46}$,	$d_{36,79}$,		$d_{36,79}$,		$d_{82,85}$		
	$d_{18,86}$,	$d_{32,69}$,		$d_{32,69}$,				
	$d_{25,0}$,	$d_{60,68}$		$d_{60,68}$				
	$d_{24,46}$							

Tab. A.3.: PDZ3L6: Contact distances assigned to MoSAIC clusters.

Cluster	1	2_1	2_2	3_1	3_2	3_2 (ctd.)	4	9
Contacts	$d_{41,47}$,	$d_{18,46}$,	$d_{17,-1}$,	$d_{23,71}$,	$d_{75,-2}$,	$d_{21,-1}$,	$d_{25,69}$,	$d_{35,78}$,
	$d_{40,47}$,	$d_{17,45}$,	$d_{16,-1}$,	$d_{24,71}$,	$d_{22,-3}$,	$d_{76,-1}$,	$d_{26,69}$,	$d_{22,78}$,
	$d_{42,47}$,	$d_{18,45}$,	$d_{20,-1}$,	$d_{22,71}$,	$d_{71,-2}$,	$d_{72,-5}$,	$d_{26,70}$,	$d_{66,78}$,
	$d_{40,46}$,	$d_{19,45}$,	$d_{18,-1}$,	$d_{22,75}$,	$d_{22,-2}$,	$d_{23,-4}$,	$d_{25,70}$,	$d_{78,85}$,
	$d_{44,48}$,	$d_{19,46}$,	$d_{78,-1}$,	$d_{35,74}$,	$d_{72,-2}$,	$d_{72,-6}$,	$d_{26,68}$,	$d_{13,78}$
	$d_{46,52}$,	$d_{20,46}$,	$d_{17,0}$,	$d_{35,71}$,	$d_{71,-3}$,	$d_{24,-4}$,	$d_{27,70}$,	
	$d_{46,51}$,	$d_{14,45}$,	$d_{20,0}$,	$d_{66,69}$	$d_{22,-4}$,	$d_{72,-1}$,	$d_{24,70}$,	
	$d_{44,47}$	$d_{18,44}$,	$d_{19,0}$,		$d_{71,-4}$,	$d_{21,0}$,	$d_{35,70}$,	
		$d_{13,45}$,	$d_{79,-6}$,		$d_{71,-5}$,	$d_{71,-6}$,	$d_{22,74}$,	
		$d_{19,44}$,	$d_{41,0}$,		$d_{21,-2}$,	$d_{20,-2}$,	$d_{26,67}$,	
		$d_{12,45}$,	$d_{76,-6}$,		$d_{21,-3}$,	$d_{79,-1}$,	$d_{35,69}$,	
		$d_{11,49}$	$d_{75,-6}$		$d_{38,-3}$,	$d_{71,-7}$,	$d_{35,75}$,	
					$d_{23,-3}$,	$d_{71,-1}$,	$d_{25,35}$	
					$d_{75,-1}$,	$d_{72,-7}$,		
					$d_{22,-1}$	$d_{38,-4}$		

Tab. A.4.: PDZ2L: Contact distances assigned to MoSAIC clusters.

Cluster	1	2	4	4 (ctd.)	4 (ctd.)	4 (ctd.)	5
Contacts	$d_{18,46}$,	$d_{12,17}$,	$d_{-5,23}$,	$d_{24,70}$,	$d_{31,36}$,	$d_{70,74}$,	$d_{11,20}$,
	$d_{19,40}$,	$d_{12,18}$,	$d_{-5,71}$,	$d_{24,71}$,	$d_{31,90}$,	$d_{71,75}$,	$d_{20,39}$,
	$d_{19,44}$,	$d_{12,45}$,	$d_{-4,22}$,	$d_{25,28}$,	$d_{32,36}$,	$d_{72,75}$,	$d_{20,40}$,
	$d_{19,45}$,	$d_{12,83}$,	$d_{-4,23}$,	$d_{25,30}$,	$d_{32,57}$,	$d_{72,76}$,	$d_{20,87}$,
	$d_{19,46}$,	$d_{13,16}$,	$d_{-4,24}$,	$d_{25,31}$,	$d_{33,57}$,	$d_{73,76}$,	$d_{40,53}$,
	$d_{41,44}$,	$d_{13,17}$,	$d_{-4,27}$,	$d_{25,32}$,	$d_{33,69}$,	$d_{73,77}$,	$d_{40,54}$
	$d_{41,47}$,	$d_{13,45}$,	$d_{-4,28}$,	$d_{25,33}$,	$d_{34,57}$,	$d_{74,77}$	
	$d_{44,47}$,	$d_{13,78}$,	$d_{-4,71}$,	$d_{25,34}$,	$d_{35,69}$,		
	$d_{44,48}$,	$d_{13,79}$,	$d_{-3,23}$,	$d_{25,35}$,	$d_{35,70}$,		
	$d_{46,49}$,	$d_{13,81}$,	$d_{-3,38}$,	$d_{25,69}$,	$d_{35,71}$,		
	$d_{46,50}$,	$d_{13,82}$,	$d_{-3,71}$,	$d_{25,70}$,	$d_{35,74}$,		
	$d_{46,51}$	$d_{14,43}$,	$d_{22,35}$,	$d_{25,71}$,	$d_{36,57}$,		
		$d_{14,44}$,	$d_{22,58}$,	$d_{26,33}$,	$d_{36,58}$,		
		$d_{14,45}$,	$d_{22,71}$,	$d_{26,68}$,	$d_{57,67}$,		
		$d_{16,79}$,	$d_{22,75}$,	$d_{26,69}$,	$d_{58,61}$,		
		$d_{17,45}$,	$d_{22,78}$,	$d_{26,70}$,	$d_{58,67}$,		
		$d_{18,44}$,	$d_{23,31}$,	$d_{27,70}$,	$d_{58,78}$,		
		$d_{18,45}$,	$d_{23,34}$,	$d_{27,71}$,	$d_{58,87}$,		
		$d_{18,87}$,	$d_{23,71}$,	$d_{28,32}$,	$d_{59,66}$,		
		$d_{61,81}$,	$d_{24,27}$,	$d_{28,33}$,	$d_{59,67}$,		
		$d_{61,87}$,	$d_{24,28}$,	$d_{28,71}$,	$d_{66,69}$,		
		$d_{66,77}$,	$d_{24,30}$,	$d_{30,33}$,	$d_{66,74}$,		
		$d_{77,81}$,	$d_{24,31}$,	$d_{30,34}$,	$d_{66,78}$,		
		$d_{78,81}$,	$d_{24,34}$,	$d_{30,36}$,	$d_{69,73}$,		
		$d_{81,85}$	$d_{24,35}$	$d_{31,34}$	$d_{69,74}$		

Tab. A.5.: PDZ2WT: Contact distances assigned to MoSAIC clusters.

Structure	β_1	β_2	β_3	α_1	β_4	β_5	α_2	β_6	α_3
PDZ2S	7–13	21–24	36–41	46–50	58–62	65–66	74–81	85–91	94–99
PDZ3	12–17	25–28	36–40	46–50	56–62	85–92	72–81	85–92	94–99
PDZ2L	6–12	20–24	35–40	45–49	57–61	64–65	71–80	84–90	

Tab. A.6.: Residues in secondary structures in PDZ2S, PDZ2L and PDZ3.

Acknowledgement

I truly want to thank Prof. Dr. Gerhard Stock for the opportunity to write my Master's thesis in his group. His door was quite literally always open, and he always took the time for discussions. His supervision was both genuinely engaged and encouraging, which was invaluable.

Many thanks also to the entire Stock group for the warm and welcoming atmosphere – I greatly appreciate being part of it. I am especially grateful to Georg Diez and Sofia Sartore from my office, who were always ready to help with any questions. Thanks also to Emanuel Dorbath and Ahmed Ali, my fellow companions in the world of PDZ domains.

Special thanks to Marko Weittenhiller for valuable discussions and plenty of good Mensa breaks, and to Svenja Koch – for her support, patience, and love throughout this time.

Erklärung gem. der geltenden Prüfungsordnung:

Hiermit versichere ich, dass

1. ich die eingereichte Masterarbeit bzw. bei einer Gruppenarbeit meinen Anteil entsprechend gekennzeichneten Anteil der Arbeit selbständig verfasst habe,
2. ich keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommenen Inhalte als solche kenntlich gemacht habe und
3. die eingereichte Masterarbeit weder vollständig noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahrens war oder ist.

_____ _____

Ort, Datum Unterschrift

Auszug Prüfungsordnung M.Sc.
Vom 19. August 2005 (Amtliche Bekanntmachungen Jg. 36, Nr. 46, S. 269–293) in der Fassung vom 25. September 2020 (Amtliche Bekanntmachungen Jg. 51, Nr. 66, S. 328–337)

§20 (8) Bei der Einreichung hat der/die Studierende schriftlich zu versichern, dass

1. er/sie die eingereichte Masterarbeit beziehungsweise bei einer Gruppenarbeit seinen/ihren entsprechend gekennzeichneten Anteil der Arbeit selbständig verfasst hat,
2. er/sie keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommenen Inhalte als solche kenntlich gemacht hat und
3. die eingereichte Masterarbeit weder vollständig noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahrens war oder ist.