Theoretical Analysis of Biomolecular Systems:
Computational Simulations, Core-set Markov State Models, Clustering, Molecular Docking

Inaugural-Dissertation

to obtain the academic degree

Doctor rerum naturalium (Dr. rer. nat.)

submitted to the Department of Biology, Chemistry and Pharmacy
of Freie Universität Berlin

by

Oliver Lemke

October, 2019

The research presented in this thesis was carried out during the period October 2015 to October 2019 at the institute of chemistry and biochemistry in the group of Prof. Dr. Bettina G. Keller.

First Reviewer:

Prof. Dr. Bettina G. Keller
Department of Biology, Chemistry and Pharmacy
Physical and Theoretical Chemistry
Freie Universität Berlin
Arnimallee 22, 14195 Berlin

Second Reviewer:

Dr. Jan P. Götze
Department of Biology, Chemistry and Pharmacy
Physical and Theoretical Chemistry
Freie Universität Berlin
Arnimallee 22, 14195 Berlin

Date of Disputation: 03.02.2020

Acknowledgments

First of all, I would like to thank Prof. Dr. Bettina Keller for offering me the position to write my dissertation in her group, for her support in the last 4 years and for the management of collaborations with other working groups. I am very grateful for the opportunity to work in this interesting field of research of biomolecular dynamics focusing on MD simulations, data analysis and method development.

I would like to acknowledge the support of my second supervisor Prof. Dr. Christian Freund. Further, I thank Dr. Jan Götze for helping me to expand my portfolio of knowledge towards QM/MM and QM calculations, constructive discussions and for taking the role as the second reviewer of this thesis.

I would like to thank the whole theoretical chemistry group of the FU Berlin and especially Prof. Dr. Beate Paulus for the last 4 years. Special thanks go to my former office members Stevan Aleksić, Francesca Vitalini, Luca Donati, Edoardo Fertitta, Matthias Berg and Jhon Fredy Perez-Torres and my collaborators Johannes Dietschreit, Jagna Witek and Dominik Schumacher for the very nice time and interesting discussions.

A lot of gratitude goes to Stefanie Kieninger, Jennifer Anders, Marius Wenz, Jan Joswig, Marco Manni and Simon Petry, who unfortunately joined the theoretical group at a too late point in time. I would like to thank you, for the very nice time drinking coffee, watching movies, barbecuing or playing board games.

I thank my sister Marilena Lemke and my parents Lothar Lemke and Christa Lemke for their support. Although you do not live close-by, it was always possible to talk to you. Further, I would like to thank Fabian Weber and Karl Stenzel for always lending me an ear and helping me a lot to relax after work, so that I could keep a clear head to write this thesis.

Last but not least, I would give a lot of gratitude to Nadine Taube, for her support in good times and in bad times. You solaced me, you laughed with me and you helped me whenever you could. There were a lot of discussions but in the end we always act in concert.

Index

List of Publications

- **Lemke, O.**; Keller, B. G. "Density-based cluster algorithms for the identification of core sets", *J. Chem. Phys.* **2016**, *145*, 164104; Included in section 4.

- Schumacher, D.; **Lemke, O.**; Helma, J.; Gerszonowicz, L.; Waller, V.; Stoschek, T.; Durkin, P. M.; Budisa, N.; Leonhardt, H.; Keller, B. G.; Hackenberger, C. P. R. "Broad substrate tolerance of tubulin tyrosine ligase enables one-step site-specific enzymatic protein labeling", *Chem. Sci.* **2017**, *8*, 3471–3478; Included in section 5 of this thesis.

- **Lemke, O.**; Keller, B. G. "Common Nearest Neighbor Clustering – A Benchmark", *Algorithms* **2018**, *11*, 19; Included in section 4 of this thesis.

- Hormann, J.; Malina, J.; **Lemke, O.**; Hülsey, M. J.; Wedepohl, S.; Potthoff, J.; Schmidt, C.; Ott, I.; Keller, B. G.; Brabec, V.; Kulak, N. "Multiply Intercalator-Substituted Cu(II) Cyclen Complexes as DNA Condensers and DNA/RNA Synthesis Inhibitors", *Inorg. Chem.* **2018**, *57*, 5004–5012; Included in section 5 of this thesis.

- Witek, J.; Wang, S.; Schroeder, B.; Lingwood, R.; Dounas, A.; Roth, H.-J.; Fouché, M.; Blatter, M.; **Lemke, O.**; Keller, B.; Riniker, S. "Rationalization of the Membrane Permeability Differences in a Series of Analogue Cyclic Decapeptides", *J. Chem. Inf. Model.* **2019**, *59*, 294–308; Not included in this thesis.

- **Lemke, O.**; Götze, J. P. "On the Stability of the Water-Soluble Chlorophyll-binding Protein (WSCP) Studied by Molecular Dynamics Simulations" *submitted*; included in section 3 of this thesis.

Contribution to the Publications

- *"Density-based cluster algorithms for the identification of core sets"*

  I implemented all three examined cluster algorithms and performed all clustering steps. I introduced an hierarchical clustering approach and constructed the core-set Markov state models for all analyzed systems. I simulated the β-hairpin peptide and generated the trajectory for the two-dimensional potential. The simulation data for the alanine dipeptide were provided by Francesca Vitalini. I contributed to the writing of the paper and was involved in the creation of all figures with the exception of figures 2, 3f, 4e, 5 and 6a.

- *"Broad substrate tolerance of tubulin tyrosine ligase enables one-step site-specific enzymatic protein labeling"*

  I performed the computational studies, including the docking experiments and the molecular dynamics simulations for the protein-ligand complexes. I wrote the section "Computational studies" and created figure 2. For the supporting information, I created figure 4 and wrote the method sections 2.2.12 and 2.2.13.

- *"Common Nearest Neighbor Clustering – A Benchmark"*

  I performed all experiments and analyses, set up the simulation in the triple-well potential and constructed the core-set Markov state model. I was involved in writing the paper and created all figures with the exception of figure 1, 4a-f and 8d. The simulation data for figure 2 were provided by Jan O. Joswig. The Common-Nearest-Neighbor algorithm is available at *GitHub*.

- *"Multiply Intercalator-Substituted Cu(II) Cyclen Complexes as DNA Condensers and DNA/RNA Synthesis Inhibitors"*

  I proposed the *bis*-intercalating binding mode explaining the increased cytotoxicity. I conducted investigations regarding *bis*-intercalation and performed the molecular dynamics optimizations. I provided figure 6. For the supporting information, I wrote the method section "Molecular Modeling" and section S-14. I created figures S14.1 to S14.7 and analyzed the out-of-plane angle of the keto-moiety of the anthraquinone.

- *"Rationalization of the Membrane Permeability Differences in a Series of Analogue Cyclic Decapeptides"*

  I provided the implementation for the cluster algorithm and the core-set Markov state model and contributed to the interpretation of the outcome. I helped to fix problems that occurred during the analysis.

- *"On the Stability of the Water-Soluble Chlorophyll-binding Protein (WSCP) Studied by Molecular Dynamics Simulations"*

  I set up all simulations and performed all analyses. I participated in writing the paper and created all figures. I wrote the supporting information and provided all corresponding figures.

List of Abbreviations

a.u.	Atomic Units
AQ	Anthraquinone
ATP	Adenosine Triphosphate
Chl	Chlorophyll
CI	Centroid Index
CNN	Common-Nearest-Neighbor (clustering)
Crt	Carotenoid
CsA	Cyclosporine A
CsE	Cyclosporine E
cs-MSM	Core-set Markov State Model
CT-DNA	Calf Thymus Deoxyribonucliec Acid
DBSCAN	Density-Based Spatial Clustering of Application with Noise
DFT	Density Functional Theory
DNA	Deoxyribonucleic Acid
DSSP	Dictionary of Secondary Structure of Proteins
EB	Ethidium Bromide
FDA	Force Distribution Analysis
GBP	Green Fluorescent Protein Binding Nanobody
GFP	Green Fluorescent Protein
GGA	Generalized Gradient Approximation
HOMO	Highest Occupied Molecular Orbital
ITS	Implied Timescale
JP	Jarvis-Patrick (clustering)
kM	k-Means (clustering)
LD	Linear Dichromism
LDA	Local-Density Approximation
LGA	Lamarckian Genetic Algorithm
LSDA	Local Spin-Density Approximation
LUMO	Lowest Unoccupied Molecular Orbital

MCMC	Markov-Chain Monte-Carlo (sampling)
MD	Molecular Dynamics
MM	Molecular Mechanics
MSM	Markov State Model
MO	Molecular Orbital
NMR	Nuclear Magnetic Resonance (spectroscopy)
PCA	Principle Component Analysis
PCCA	Perron-Cluster Cluster Analysis
PCNA	Proliferating-Cell-Nuclear-Antigen
PCR	Polymerase Chain Reaction
PCF	Point Charge Field
PES	Potential Energy Surface
QM	Quantum Mechanics
QM/MM	Quantum Mechanics/Molecular Mechanics
RMSD	Root-Mean-Square Deviation
RNA	Ribonucleic Acid
SANS	Small-Angle Neutron Scattering
SAXS	Small-Angle X-ray Scattering
SD	Standard Deviation
TD-DFT	Time-Dependent Density Functional Theory
TIC	Time-lagged Independent Component
TICA	Time-lagged Independent Component Analysis
TTL	Tubulin Tyrosine Ligase
UPLC-MS	Ultra Performance Liquid Chromatography Mass-Spectrometry
UV/Vis	Ultraviolet-Visible (spectroscopy)
WSCP	Water-Soluble Chlorophyll-binding Protein

Abstract

The analysis of the structural and the dynamical behavior of biomolecules is very important to understand their biological function, stability or physico-chemical properties. In this thesis, it is highlighted how different theoretical methods to characterize the aforementioned structural and dynamical properties can be used and combined, to obtain kinetic information or to detect biomolecule-ligand interactions.

The basis for most of the analyses, performed in the course of this work, are molecular dynamics simulations sampling the conformational space of the biomolecule of interest. Using molecular dynamics simulations, the remarkable stable water-soluble-binding-protein is examined first. On a theoretical basis, structural modifications that can influence the stability of the protein are discussed. Additionally, by combining the simulations with a QM/MM optimization scheme and quantum chemical calculations, spectroscopical properties can be investigated.

Markov State Models are applied frequently to capture the slow dynamics within simulation trajectories. They are based on a discretization of the conformational space. This discretization, however, introduces an error in the outcome of the analysis. The application of a core-set discretization can reduce this error. In this thesis, it is discussed how density-based cluster algorithms can be used to determine these core sets, and the application on linear and cyclic peptides is highlighted. The performance of a promising cluster algorithm is investigated and error sources in the construction of the Markov models are discussed.

Finally, it is shown how molecular docking combined with molecular dynamics simulations can be used to determine the binding behavior of ligands towards biomolecules. In this context, the important interactions within the active site of an enzyme, and different binding modes of DNA intercalators are identified.

Zusammenfassung

Die Analyse der strukturellen und dynamischen Eigenschaften von Biomolekülen ist wichtig, um ihre biologische Funktion, ihre Stabilität oder ihre physikalisch-chemischen Eigenschaften zu verstehen. In dieser Arbeit wird gezeigt, wie unter Verwendung und Kombination verschiedener theoretischer Methoden, die strukturelle und dynamische Eigenschaften charakterisieren können, kinetische Information erhalten oder Biomolekül-Ligand-Wechselwirkungen detektiert werden können.

Als Basis für die meisten in dieser Arbeit verwendeten Analysen dienen moleküldynamische Simulationen, in denen der Konformationsraum des untersuchten Biomoleküls gesampelt wird. Im ersten Teil dieser Arbeit wird mit Hilfe der moleküldynamischen Simulationen die hervorzuhebende Stabilität des wasserlöslichen, Chlorophyll-bindenden Proteins erforscht. Auf theoretischer Ebene werden Modifikationen, welche die Stabilität des Proteins beeinflussen können, diskutiert. In Kombination mit einem QM/MM-Optimierungsschema und quantenchemischen Berechnungen werden zudem spektroskopische Eigenschaften untersucht.

Markov Modelle werden häufig verwendet, um die langsamen Dynamiken innerhalb der Simulationstrajektorien einzufangen. Die Modelle basieren auf einer Diskretisierung des Konformationsraumes. Diese Diskretisierung führt jedoch zu einem Fehler in der Analyse. Unter Verwendung einer *Core-Set*-Diskretisierung ist eine Reduzierung des Fehlers möglich. In dieser Arbeit wird diskutiert, wie man mit Hilfe von dichte-basierten Clusteralgorithmen diese *Core-Sets* bestimmen kann und deren Anwendung an linearen und cyclischen Peptiden aufgezeigt. Des Weiteren wird die Leistungsfähigkeit eines vielversprechenden Clusteralgorithmus untersucht und Fehlerquellen in der Konstruktion von Markov Modellen diskutiert.

Abschließend wird gezeigt, wie Dockinguntersuchungen in Kombination mit moleküldynamischen Simulationen genutzt werden können, um das Bindungsverhalten von Liganden zu Biomolekülen zu bestimmen. Wichtige Wechselwirkungen im aktiven Zentrum eines Proteins werden identifiziert und die Bindungsarten von DNS-Interkalatoren untersucht.

1 Introduction

Analyzing properties of biomolecular systems on a theoretical level has become quite important in the last few decades. Frequently applied techniques in this context are computational simulations [1]. They can be based on a classic mechanical level [2], sometimes incorporating quantum mechanical effects [3]. The simulations can be used to model the dynamics of the biomolecule in order to understand its structural and dynamical properties [4] as well as its function [5], and to interpret experimental findings [6]. Tools such as Markov State Models [7–14], clustering [15–17] and other methods accounting for intramolecular stability of the biomolecule [18–20], provide insights into the biomolecular metastable conformations, kinetics and further dynamical properties. Besides computational simulations, other techniques such as binding mode prediction of biomolecule-ligand interactions using molecular docking [21–23], or theoretical spectroscopy [24, 25] are performed frequently.

The aim of this section is to provide a general overview over the analyzed systems and the methods used in this thesis. For the latter, the focus lies on applications, advantages and drawbacks. The corresponding theory behind these methods is discussed in section 2.

1.1 Investigated Biomolecules

1.1.1 Proteins and Peptides

Proteins are important biomolecules, which can be found within intra- and extracellular environments. Their function is quite diverse. They can, for example, efficiently and selectively catalyze chemical reactions in cells, act as transmitters, or provide structural properties for the cell's stability and flexibility. Proteins are therefore an important research topic regarding diseases [26–29], catalysis [30–32], and the understanding of cell properties [33, 34].

Proteins, and their smaller versions peptides, consist of building blocks, called amino acids, forming a polymer. Each protein has a primary sequence of these amino acids, determining its overall structure and function. Two amino acids are connected via a peptide bond that is formed by a condensation reaction (see figure 1.1). There exist 20 canonical amino acids in all living creatures. Each amino acid consists of a backbone and a side chain, where the latter varies for every amino acid and features different bulkiness, polarity and functionality.

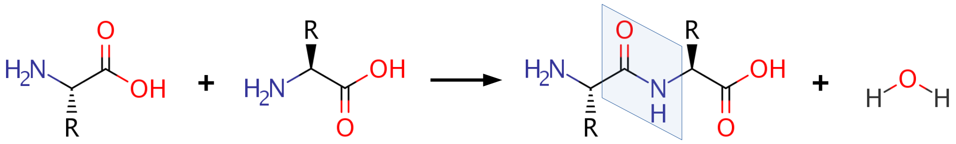


Figure 1.1: Reaction of two amino acids to a dipeptide. The amino acids in the dipeptide are connected via a peptide bond.

The sequence of amino acids is called primary structure. The spatial arrangement of this chain, also called protein folding, is achieved by interactions of the amino acids. Hydrogen bonding of the backbone forms the secondary structure. The main motives are α-helical structures and extended structures (β-sheets) linked by turns, loops or random coils. The structural arrangement of an α-helix and a β-sheet is presented in figure 1.2.

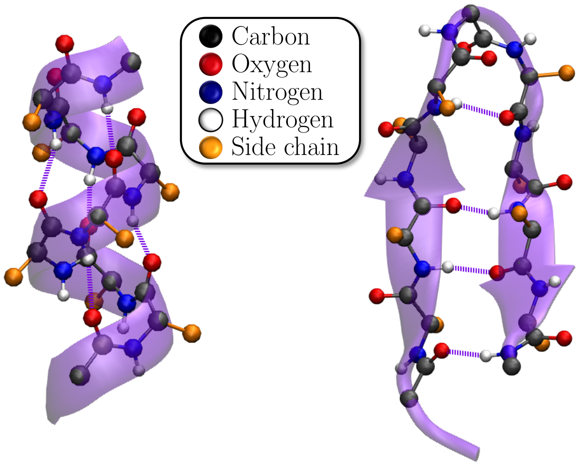


Figure 1.2: Main secondary structure motives of proteins: α-helix (left) and β-sheet (right). Backbone hydrogen bonds are highlighted.

Each secondary structure motive can be classified by analyzing the dihedral angles of the backbone. The backbone of an amino acid residue in a peptide chain has three torsion angles ω, ϕ and ψ (figure 1.3). As the peptide bond features a partial double-bond character, ω is usually assumed to be fixed while the angles ϕ and ψ can rotate freely. The only exception is the amino acid proline, where the rotation around ϕ is limited. The combination of ϕ and ψ can be displayed using a Ramachandran plot [35], where each combination accounts for a specific secondary structure arrangement (figure 1.3).

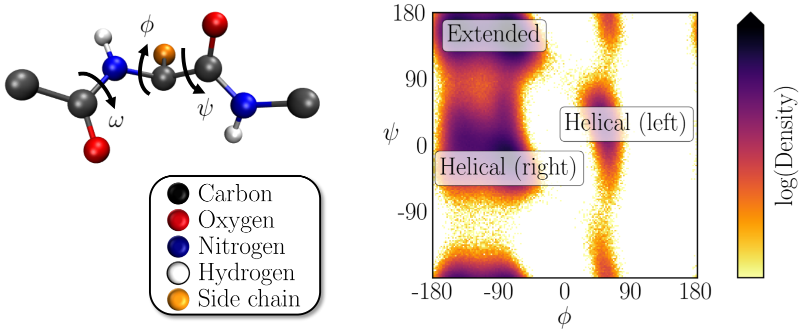


Figure 1.3: Dihedral angles ω, ϕ and ψ of the amino acid backbone (left). Distribution of angles ϕ and ψ of alanine dipeptide depicted in a Ramachandran plot (right). Corresponding secondary structures are highlighted. As the angle ω is usually fixed due to a partial double bond character, it is not displayed in the plot.

In addition to the backbone interactions, interactions of the side chains form the tertiary structure of the protein. These interactions include ionic interactions by charged residues, hydrogen bonding by polar amino acids or hydrophobic interactions by nonpolar amino acids between side chain and/or backbone moieties. Another possible interaction is the formation of a covalent disulfide bridge between two cysteines. For some proteins, an interaction of multiple folded proteins is observed forming the quarternary structure.

The structure of a peptide is not entirely limited to the folding of the linear amino acid chain. It can also incorporate covalent bonds between different amino acids forming a cyclic peptide. This peptide can either be connected via its backbone connecting both termini, as in cyclosporines [36, 37], and/or via its side chains, as in amanitin [38]. Also, a connection between a terminus and a side chain is possible [39]. The cyclization of a peptide can improve its stability [40, 41], can limit its conformational space [42] and/or can increase its membrane permeability. The capability of passive membrane diffusion differs for every cyclic peptide and is a recent field of research [43–47].

Proteins catalyzing chemical reactions are called enzymes and are classified with respect to their reaction type. Ligases, for example, form covalent bonds which can be split by hydrolases catalyzing hydrolysis, or oxidoreductases catalyzing redox reactions. For the reaction the substrate(s) have to bind to the active center of the molecule. This active center consists of a binding site and a catalytic site [48]. Due to the constitution of the binding site, enzymes are highly specific with respect to the bound substrate (see figure 1.4). For the binding of the substrate, different mechanisms were proposed. Emil Fischer proposed a lock-and-key binding mechanism where the substrate has to fit perfectly into the active center. Since this mechanism is too simplistic, two other mechanisms were introduced and are currently used: the induced-fit and the conformational-selection mechanism. In both mechanisms, a rearrangement of the protein is permitted, either upon or before binding of the substrate [49–52]. Some enzymes need additional cofactors such as metal ions, or coenzymes such as adenosine triphosphate, to fulfill their function.

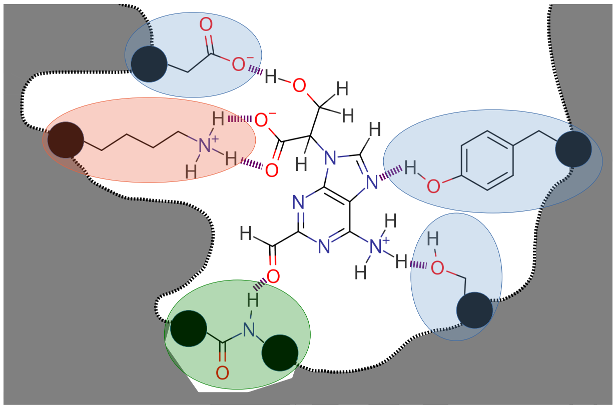


Figure 1.4: Depiction of a ligand binding to the active site of a protein. Side chain hydrogen bonding (blue), backbone hydrogen bonding (green) and charge-charge interactions (orange) are highlighted.

To observe the structural arrangement of proteins, different experimental techniques can be used. Large proteins can be investigated using cryo-electron microscopy, however, the resolution is limited [53, 54]. To get a deeper insight into the protein constitution, X-ray crystallography [55] or nuclear magnetic resonance (NMR) [56, 57] measurements can be performed. Average size and shape can be determined using small-angle X-ray scattering (SAXS) and small-angle neutron scattering (SANS) [58–60]. Given a protein structure, dynamical information can be obtained from molecular dynamics (MD) simulations [2].

1.1.2 Deoxyribonucleic Acid

The deoxyribonucleic acid (DNA) encodes the information for the synthesis of proteins. Similar to proteins, the DNA consists of small building blocks. These building blocks are called nucleotides and consist of a nucleobase, the sugar deoxyribose and a phosphate-moiety. The sugar and the phosphate constitute an alternating chain forming the negatively charged backbone of the DNA. The sequence of the nucleobases contains the information stored in the DNA. There exist four different nucleobases, which are highlighted in figure 1.5. DNA forms a double-stranded helical structure, with a small minor groove and a larger major groove. The helix is stabilized by hydrogen bonds between pairs of nucleobases (figure 1.5). These nucleobase pairs are Adenine-Thymine and Guanine-Cytosine. The complementary base pairing allows for easy replication of the DNA and error detection during the replication process.

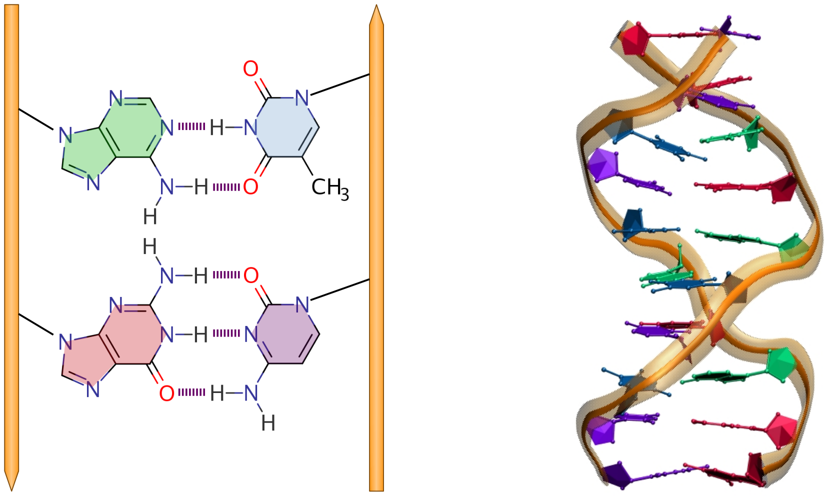


Figure 1.5: Nucleobase pairs (left): adenine (green) - thymine (blue) and guanine (red) - cytosine (purple) with antiparallel phosphate-sugar backbone (orange). Helical structure of the DNA (right, PDB:1ZEW [61]); The nucleosides (sugar and nucleobase) are shown explicitly.

There are three different types of double-stranded helical DNA: A-DNA, B-DNA and Z-DNA with B-DNA considered the main type of DNA in cells. B-DNA is a right-handed helix with nucleobases orientated orthogonal to the helical axis. The A-DNA features a broader diameter than the B-DNA and the nucleobases are not orthogonal anymore with respect to the helical axis. Contrary to the A- and B-DNA, the Z-DNA is a left-handed helical structure [62, 63]. Besides the helical arrangements of the DNA, there also exists circular plasmid DNA occurring in bacteria.

Due to the planar structure of the nucleobase pairs, it is possible for other planar, in the best case positively charged molecules, to be inserted between two nucleobase pairs. This insertion is called intercalation and results in a small unwinding of the DNA at the position of the intercalation. DNA-intercalators can be used as a fluorescent tag [64], as an inhibitor for DNA replication [65, 66], or as a carrier for metal ions to the sugar-phosphate backbone in order to cleave the DNA [67–69]. Thus, the design of DNA-intercalators is important e.g. as a anti-cancer drug [70, 71].

1.2 Computational Simulations

1.2.1 Molecular Dynamics Simulations

Molecular dynamics (MD) simulations are a powerful tool in the analysis of biomolecular properties. They are based on a propagation of an atomic structure with respect to Newton's equations of motion incorporating the forces acting on every single atom within the system. These interactions are typically stored in a force field, acting as a potential energy function for all interatomic interactions. A detailed description of the theory behind MD simulations is given in section 2.1.1.

Compared to other methods like quantum mechanics (QM), MD simulations are computationally cheap as every atom is treated as one particle. Additionally, no reevaluation of the force field parameters describing the simulated system is typically done during the simulation [72]. Incorporating all dynamical properties of the system such as the bond vibration requires propagation time steps in the femtosecond regime (10^{-15} seconds) or lower [73]. In a feasible amount of real time thus it is possible to gain total simulation lengths of microseconds (10^{-6} seconds) to milliseconds (10^{-3} seconds) [74]. However, the dynamics of some biomolecules are in the range of seconds or even longer [75]. If in the kinetics are also of interest, transitions between different conformations have to be sampled multiple times to obtain significant estimates of the relative populations and the transition rates.

To improve simulation times within the limitations of the available computational power, enhanced sampling techniques have been developed. These methods include, amongst others, replica exchange molecular dynamics [76–78], umbrella sampling [79, 80] and metadynamics [81–83]. All of these methods bias the kinetics and therefore can not be directly used to extract dynamical properties such as rate constants. For some of these techniques, reweighting schemes have been developed taking this bias into account and restoring the correct kinetics [84–86]. Another approach to reduce the computational costs is to treat the environment of the simulated system implicitly. However, this simplification often reduces the accuracy of the simulation [87]. Moving from all-atomistic simulations to coarse-grained simulations by grouping a set of atoms into one particle, also reduces the computational costs. However, this results in a worse resolution compared to all-atomistic simulations [88–90].

One big drawback of MD simulations is that they rely on the incorporated force field. With respect to the kinetics, it was shown that for small peptides the chosen force field, which is used to propagate the structure, highly influences the dynamic properties [91]. Another drawback of classical MD simulations, that emerges, is the treatment of every atom as a charged mass point. It is therefore hard to incorporate effects that involve electrons such as polarization or bond rearrangement. To account for polarization effects, special force fields have been developed [72, 92]. To include bond breaking such as chemical reactions or hydrogen rearrangement, techniques on a higher theory-level like QM/MM [93–95] have to be applied.

1.2.2 Quantum Mechanics/Molecular Mechanics

To reduce the drawbacks of classical MD simulations, quantum mechanics (QM) can be incorporated into the simulation. As most biomolecules are too large to be treated completely on a QM level, quantum mechanics and molecular mechanics (MM) can be combined [3, 93–98]. The resulting QM/MM method was first introduced in Ref. [95] and was rewarded with a Nobel prize in chemistry in 2013 (M. Karplus, M. Levitt and A. Warshel). In a QM/MM calculation, the simulated system is split into at least two parts: A QM part describing the reaction center, and an MM part accounting for the environment. With respect to enzymes the reaction center is often chosen to be the active site, whereas the environment is the rest of the protein. The splitting of the two parts, later referred to as layers, is possible since the effect of most chemical reactions in solution is often strongly localized. By considering the part that is not involved in the reaction as an environment, which is propagated on an MM level, allows for a dynamical electrostatic representation [94]. A detailed description of QM/MM can be found in section 2.1.2.

1.3 Analysis of the Computational Simulations

1.3.1 Markov State Models

To get an insight into the dynamics of the simulated system, a Markov State Model (MSM) can be constructed. An MSM is a stochastic model in which the MD trajectory is approximated as a stochastic jump-process between different discrete states. The key component of an MSM is a transition probability matrix $\mathbf{T}(\tau)$. The matrix elements T_{ij} represent the probability to observe a transition starting in state C_i and ending up in a state C_j after a certain time τ. This time τ is called the lag time and denotes a parameter of the constructed transition probability matrix. The lag time τ has to be chosen such that the analyzed dynamics are Markovian, i.e. they have no knowledge about the history of the jump-process and the transition probability towards any state only depends on the current state [11, 99]. If this applies, the transition matrix can be constructed on a large set of trajectories. This has the advantage that, although the total simulation length might not be reduced, multiple simulations can be run in parallel [100]. Nevertheless, finding an appropriate lag time often depends on the definition of the discrete states [101].

Projecting the continuous dynamics onto a set of discrete states causes a loss of information within these states and therefore a loss of Markovianity. To restore this Markovianity, the lag time has to be increased. Using a carefully optimized discretization, however, one can minimize this discretization error [11] as highlighted in figure 1.6. One way to achieve this, is the use of a very fine discretization [9, 102, 103]. Due to an exponential scaling of the micro states needed to maintain the fine discretization with respect to increasing dimensionality of the investigated conformational space, this type of discretization, however, gets impractical for high-dimensional systems. In addition, due to the finite simulation length, the statistical uncertainty of the model increases as the number of microstates increases [104]. Using a discretization that accounts for the slow dynamics instead of a fine grid, the increasing dimensionality can be handled for higher dimensional conformational spaces. This can be achieved either by clustering [15, 105, 106] or by using human-made borders [91]. Nonetheless, these states have to be chosen properly to reduce recrossing at the borders between the states [12].

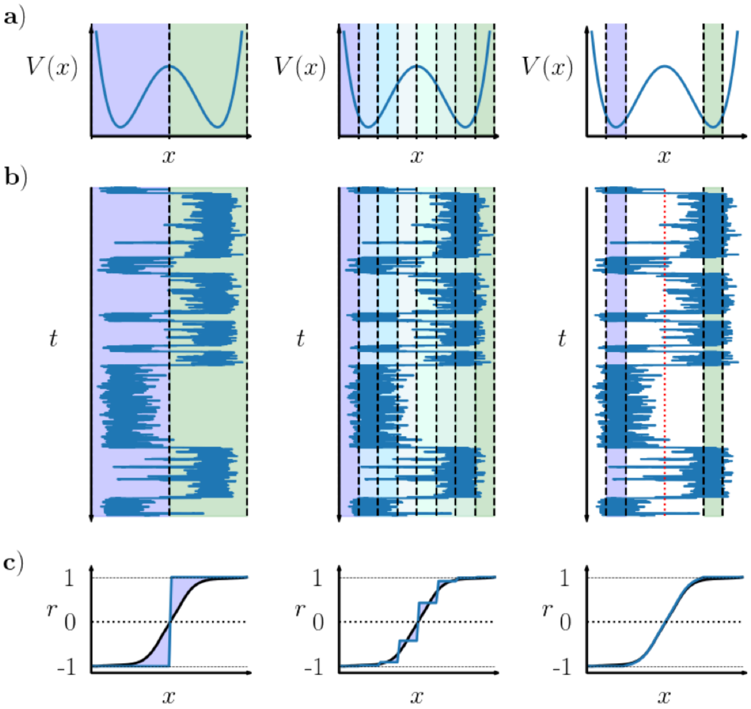


Figure 1.6: Depiction of a) a double-well potential and b) a time-dependent trajectory for a molecule moving in this potential; c) The dynamic process for the continuous dynamics is highlighted in black with an approximated process (blue) for different discretizations. The error that is induced by the discretization is highlighted by the filled area. The trajectory is projected onto two (left) and eight states (middle) covering the full sampled potential. In the right panel, two core sets were chosen. The position of the energy barrier of the potential is highlighted as a red, dashed line in the right panel.

During the estimation of the transition probability between two states, C_1 and C_2, crossings across the boundary between these states are usually observed, which are not observed in continuous dynamics (figure 1.6b) and are an artifact of the discretization. These "artificial" transitions incorporate memory, i.e. knowledge about the previous jump-process history, that harm the Markovianity. To minimize the discretization error, the discretization has to be chosen in such a way that the recrossing is reduced. This can, for example, be achieved by introducing core sets [12, 14, 107]. A core set describes a discrete state C_i, in which the trajectory stays for a long time. The core set is separated towards any other discrete state C_j by a transition region (as shown in figure 1.6 upper right). The transition region is not assigned to any state and modeled by committor or milestoning functions [108–111].

A definition of these core sets can be quite challenging since most data sets are high-dimensional. In addition, the core sets can feature arbitrary sizes and shapes. To utilize a core set discretization, an automatic approach that can detect core sets in a high dimensional data space is required. In the course of this work, an approach to solve this issue is introduced. The corresponding Markov Model based on this type of discretization, is referred to as core-set Markov State Models (cs-MSM) and is discussed in detail in section 2.2.2.2.

The reaction coordinates included in the construction of the model are another crucial point in finding a suitable set of discrete states. As it is often not feasible to discretize the full conformational space, owing to its high dimensionality, the conformational space is usually split into relevant and non-relevant coordinates [100]. Removing these non-relevant coordinates, however, introduces a projection error, which can cause a loss of information important for the analyzed kinetics [11, 112–114] as depicted in figure 1.7. Using dimensionality reduction techniques, such as principle component analysis (PCA) [115, 116] or time-lagged independent component analysis (TICA) [112, 117] to construct linear-optimized, independent reaction coordinates, can help to minimize this projection error. The theory behind these techniques is given in section 2.2.1.

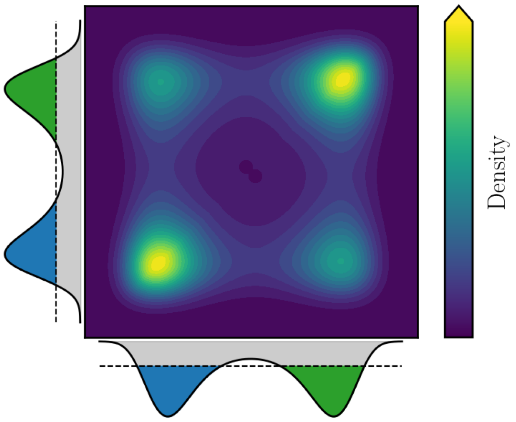


Figure 1.7: Depiction of a two-dimensional density projected onto the x- and y-axis, respectively. In both cases, the projection reduces all four peaks to two peaks. An extraction of these density peaks using a density threshold (e.g. by applying density-based clustering), is highlighted by a dashed line.

For MSMs constructed at a lag time that reduces the discretization error such that Markovianity is restored, an analysis of the eigenvectors and the corresponding eigenvalues of the transition probability matrix can be performed. Analyzing these quantities yields information about dynamic modes and their corresponding time scales. A detailed description of Markov Models and their analysis is given in section 2.2.2.

1.3.2 Clustering

Analyzing MD simulations involves the investigation of a large amount of data. To handle these data and to classify discrete states in the MD simulation, as for example necessary to construct MSMs, cluster algorithms can be applied. The challenge of the cluster algorithm is to group similar conformations in one cluster separated from other more diverse conformations [13, 118]. In order to do this, a set of descriptors or reaction coordinates that accounts for the diversity of the conformations has to be defined in a first step. Ideally, these discrete states characterize metastable conformations, i.e. minima in the free energy landscape of the simulated molecule [119, 120]. This, however, can get more difficult with increasing dimensionality of the system, since the distance between two conformations gets more and more equivalent [121].

In order to use cluster algorithms, every MD structure has to be interpreted as a data point in a high-dimensional space. Assuming that this space is constructed to separate different conformations well enough, the next challenge lies in the identification of these conformations. There is a huge variety of different cluster algorithms featuring, for example, partitioning [122–124], hierarchical [125, 126], fuzzy [127, 128] or density-based clustering [16, 129–133]. All algorithms use different criteria to decide whether two data points belong to the same cluster. For every data set, the type of clustering algorithm has to be chosen such that it fits the challenges of the analyzed data set. This includes, beside others, cluster sizes, shapes and number [134].

Since the goal of MD simulations is the identification of metastable conformations, i.e. conformations that frequently appear during the MD simulation, density-based cluster algorithms are a good choice [135–137]. In density-based clustering, clusters are defined as areas which cover a large number of data points. They have the advantage that the cluster number is not predefined and clusters of arbitrary shapes can be extracted. The outcome, however, always depends on the chosen density threshold that is used for the clustering [138].

An example for a density-based clustering outcome for a two-dimensional data set is shown in figure 1.8. This data set is quite challenging and features several properties also observed for MD simulation data. It includes clusters of different shapes and different sizes as well as noise. In addition, some clusters are intertwined and/or connected via a (sine-like) line.

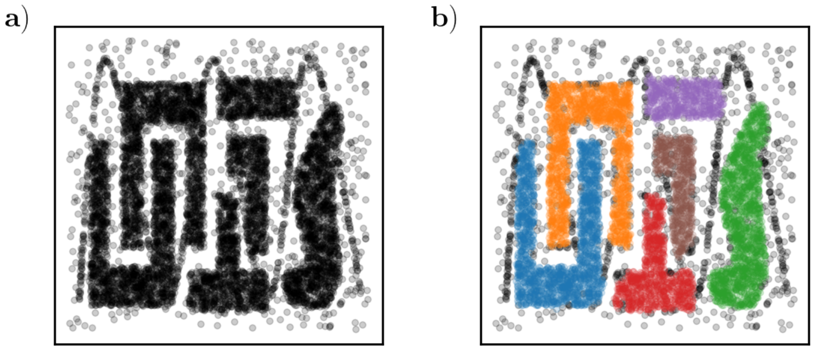


Figure 1.8: Example of a clustering outcome for a two-dimensional data set [139] with respect to a density criterion. a) Original data set and b) data set after clustering. Each cluster is highlighted with a separate color. Not assigned data points are shown in black.

Since density-based clustering algorithms use a density threshold to determine whether two data points belong to the same cluster, they can handle intertwined clusters of different shapes and noise. In addition, the outcome is not influenced by the line of data points connecting the clusters as the data point density of this line is below the threshold. The algorithms used in this thesis are described in section 2.2.3.

1.3.3 Theoretical Absorption Spectroscopy

For simulations of macromolecules containing chromophores, absorption spectra can be calculated. For these calculations, snapshots of the MD simulation can be taken. Using time-dependent density functional theory (TD-DFT), it is possible to calculate the electronic excitations of the chromophore from these snapshots [140]. Combining the calculation with a QM/MM optimization of the snapshot yields a geometry-optimization of the chromophore on the QM level, accounting for interactions and changes of the surroundings. The inclusion of the surroundings as a point charge field in the absorption spectra calculation further improves the excitation energies [141].

1.4 Molecular Docking

For modeling the binding modes of ligands towards biomolecules, often called target, molecular docking can be used. In molecular docking, a binding mode is predicted with respect to steric and energetic properties of the target molecule or rather its active site [142], as shown in figure 1.9.

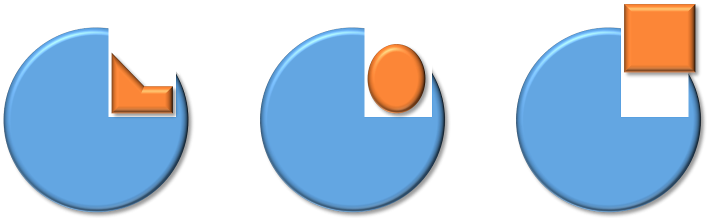


Figure 1.9: A simplified representation of a ligand (orange) binding to the active site of a target molecule (blue) with respect to steric properties. Left to right: optimal to sterically hindered ligand. A more detailed description including also the energetic properties, can be found in figure 1.4.

Docking can be used, amongst others, to examine potential binders out of a large set of ligands (virtual screening) [143] or to understand the important interactions that stabilize or destabilize bound ligands [144] as highlighted in figure 1.4. To do this, a scoring function is calculated and used to estimate the likelihood of the formation of a protein-ligand complex with respect to a certain ligand conformation. This scoring function accounts for van der Waals and electrostatic interactions as well as for entropy and desolvation effects [142, 145]. For the inclusion of dynamical effects into the docking procedure, a combination of molecular docking with MD simulations is often helpful. It can be used to obtain different starting conformations of the target [146, 147], to check for stability of the target-ligand complex analyzing favorable interactions [148], or to account for changes in the target upon docking of a ligand [149, 150]. The docking software and its procedure used for this work are discussed in section 2.4.

1.5 This Thesis

The work presented in this thesis, focuses on the theoretical analysis of the dynamical properties of proteins, on method development in the field of kinetic analyses and on binding mode prediction of biomolecule-ligand complexes. In section 2, the theoretical background of the methods used during this thesis is highlighted. Sections 3 to 5 account for the research done during the PhD studies and are set into context to current research questions. Every section is dedicated to a main focus:

- In section 3, the focus lies on the analysis of the water-soluble-chlorophyll-binding protein (WSCP), which shows a remarkably high stability towards environmental changes.

  By performing MD simulations, it is intended to identify the key components that grant this high stability, as most of the former analyses are based on a static view. Additionally, the influence of structural modifications such as the removal of chlorophylls or the formation of disulfide bridges, on the overall dynamics and stability is investigated. For the formation of disulfide bridges, no literature was found. Hence, it is examined on a theoretical level, how an introduction of disulfide bridges would affect former findings as this should increase the stability of the WSCP. The results are presented in section 3.1.

Based on these simulations, absorption spectra are calculated incorporating QM/MM optimization potentials, as presented in section 3.2. The spectra are calculated for different conformations extracted from the MD simulations. The influence of different treatments of these conformations on the calculated spectra such as different optimization potentials or the removal of the environment is compared. Furthermore, the influence of structural modifications on the absorption spectra, and the coupling between different chlorophylls are investigated.

- The main focus of section 4 is on method development with respect to cs-MSMs and the density-based Common-Nearest-Neighbor (CNN) clustering.

  For a long time no method for the identification of core sets was known. Therefore, it is investigated in section 4.1 whether a suitable definition of these core sets can be obtained by applying density-based clustering. For this purpose, different cluster algorithms are evaluated. A hierarchical clustering approach extracting clusters of different densities is introduced. Furthermore, the cs-MSMs based on the identified core sets are compared to conventional full-partitioning MSMs.

  In section 4.2, the most promising clustering algorithm, the CNN algorithm, is then benchmarked using a variety of different data sets. It is investigated how the CNN algorithm performs with respect to typical challenges of data sets. These challenges include, amongst others, clusters of different size, shape or density.

  In section 4.3, the combination of CNN clustering and cs-MSMs is applied to the cyclic peptide cyclosporine A and its derivative cyclosporine E simulated in water and chloroform, respectively. In this section, the advantages of a core-set discretization are discussed and compared to a full-partitioning discretization, highlighting disconnections in the data set. Additionally, it is investigated how specific reaction coordinates, included in the state definition, influence the quality of the MSM. In a last step, both molecules are compared using a joint discretization to characterize unique and shared conformations.

- Section 5 covers molecular modeling of biomolecule-ligand interactions.

  In section 5.1, the substrate scope of the enzyme tubuline-tyrosine ligase is interpreted utilizing flexible ligand docking. By combining this docking with MD-simulations, the flexibility and the stability of the protein-ligand complex are investigated. Furthermore, an insight into the important interactions and spatial properties for the protein-ligand binding is gained. Based on these findings, predictions with respect to other ligands can be made without former synthesis.

  Section 5.2 is attempted to find an explanation for the different inhibition efficiencies of DNA replication of quite-similar DNA intercalators. This is done by performing molecular modeling of different intercalator comformations and intercalation modes. Conformations that are only possible with a certain intercalator constitution are identified.

The thesis is concluded in section 6. In this section, the findings are summarized and the remaining research questions are highlighted. An outlook on further analyses that can be performed to deepen the knowledge and understanding of the investigated biomolecular system is emphasized. The framework of this thesis is highlighted in figure 1.10.

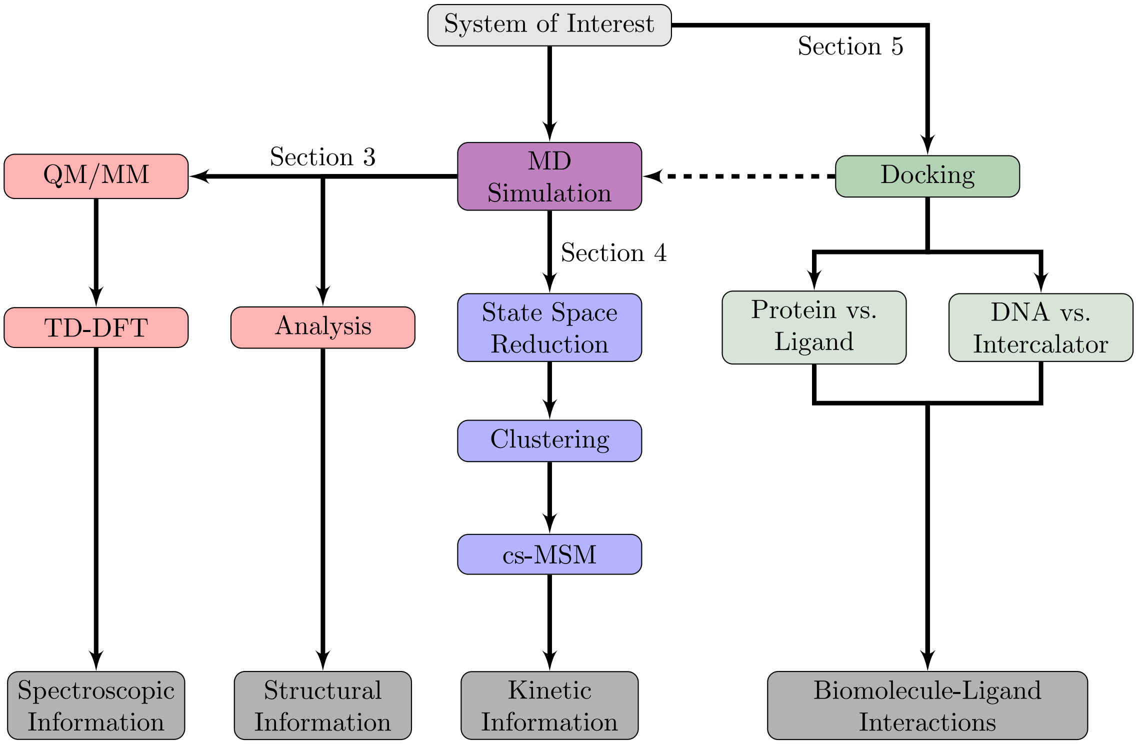


Figure 1.10: Framework of this thesis colored according to the corresponding section; MD simulations are involved in all three sections; The last row summarizes the information obtained in each corresponding section.

2 Methods and Theory

2.1 Simulations

2.1.1 Molecular Dynamics

The theory presented in this section was taken from Ref. [73, 151].

In molecular dynamics (MD) simulations, a system of atoms is propagated according to the laws of classical mechanics. The propagation is performed by integrating Newton's equations of motion:

$$\mathbf{F}_i = m_i \cdot \mathbf{a}_i \tag{1}$$

$\mathbf{F}_i$ denotes the force on a particle i with mass m_i and acceleration $\mathbf{a}_i$. Each atom in the simulation is treated as a mass point with a position $\mathbf{r_i} = x\hat{e}_x + y\hat{e}_y + z\hat{e}_z$ in a Cartesian coordinate system, with the unit vectors in direction j $\hat{e}_j$. The set of positions of all N_{atoms} atoms at a time t is given by $\mathbf{r}(t) = \{\mathbf{r}_i(t)\}_{i=1}^{N_{\text{atoms}}}$ and describes one configuration of the system. Associated to each coordinate $\mathbf{r}_i(t)$ at time t is a velocity $\mathbf{v}_i(t)$ with $\mathbf{v}_i = v_x\hat{e}_x + v_y\hat{e}_y + v_z\hat{e}_z$ and $\mathbf{v}(t) = \{\mathbf{v}_i(t)\}_{i=1}^{N_{\text{atoms}}}$. The time evolution of the configurations, and if needed of the velocities, is called trajectory $\mathbf{x}_t$. The connection of the position $\mathbf{r}_i$ to the velocity $\mathbf{v}_i$ and the acceleration $\mathbf{a}_i$ is defined by:

$$\frac{\mathrm{d}\mathbf{r}_i}{\mathrm{d}t} = \mathbf{v}_i(t) \tag{2}$$

$$\frac{\mathrm{d}^2\mathbf{r}_i}{\mathrm{d}t^2} = \frac{\mathrm{d}\mathbf{v}_i}{\mathrm{d}t} = \mathbf{a}_i(t) \tag{3}$$

To integrate Newton's equations of motion the continuous time t is described by a set of discrete time steps with a spacing of δt. For this case the propagated positions and velocities, if needed, at time $t + \delta t$ can be approximated as a Taylor series expansion by:

$$\mathbf{r}_i(t \pm \delta t) = \sum_{n=0}^{\infty} (\pm 1)^n \frac{\mathbf{r}_i^{(n)}(t)}{n!} \delta t^n \tag{4}$$

$$\mathbf{v}_i(t \pm \delta t) = \sum_{n=0}^{\infty} (\pm 1)^n \frac{\mathbf{v}_i^{(n)}(t)}{n!} \delta t^n \tag{5}$$

with $\mathbf{r}_i^{(n)}(t)$ denoting the nth derivative of $\mathbf{r}_i$ at the point t. According to equations 1 to 3 the new position of atom i is approximately obtained by taking the velocity of the atom as well as the force acting on the system into account in order to propagate it in time. Neglecting all higher order terms results in an error scaling with δt^3. By adding the forward propagated position $\mathbf{r}(t + \delta t)$ and the backward propagated position $\mathbf{r}(t - \delta t)$ and rearranging, equation 6 is obtained:

$$\mathbf{r}_i(t + \delta t) = 2\mathbf{r}_i(t) - \mathbf{r}_i(t - \delta t) + \frac{\mathbf{F}_i(t)}{m_i}\delta t^2 + \mathcal{O}(\delta t^4) \tag{6}$$

This algorithm is called Verlet-integrator [152] and has an error scaling with δt^4. In addition, the new position only depends on the acceleration or the force respectively and no longer on the velocities. Numerical problems may occur as a term scaling with δt^2 is added to two terms scaling with δt^0. A more robust way, compared to using the Verlet-integrator, is to propagate the system in time by applying the leap-frog integrator [153]. In the leap-frog integrator the position $\mathbf{r}_i(t)$ is propagated in time steps of $\delta t/2$. By adding up the forward and the backward propagation equation 7 is obtained:

$$\mathbf{r}_i(t + \delta t) = \mathbf{r}_i(t) + \mathbf{v}_i(t + \frac{\delta t}{2})\delta t + \mathcal{O}(\delta t^3) \tag{7}$$

The interesting property of this algorithm is that the velocities are estimated at time steps of $t \pm \delta t/2$, which causes the position and velocities to leap over each other, hence causing the name of the algorithm. The velocities are estimated via:

$$\mathbf{v}_i(t + \frac{\delta t}{2}) = \mathbf{v}_i(t - \frac{\delta t}{2}) + \frac{\mathbf{F}_i(t)}{m_i}\delta t + \mathcal{O}(\delta t^3) \tag{8}$$

giving an error scaling with δt^3. The algorithm has the advantage that the velocity is taken explicitly into account to estimate the new position. As already mentioned, position and velocity are not calculated at the same time step. This has the consequence that the velocity $\mathbf{v}(t)$ at time t, which is required to calculate the kinetic energy

$$E_{\mathrm{kin}}(t) = \frac{1}{2}m\mathbf{v}(t)^2, \tag{9}$$

has to be approximated by averaging over the velocities $\mathbf{v}(t + \delta t/2)$ and $\mathbf{v}(t - \delta t/2)$ by:

$$\mathbf{v}(t) = \frac{\mathbf{v}(t + \delta t/2) + \mathbf{v}(t - \delta t/2)}{\delta t} \tag{10}$$

The initial positions for the simulation are given by a reference structure such as a crystal structure obtained by X-ray crystallography. The initial velocities are drawn from a Maxwell-Boltzmann distribution. The probability $p(v_{x,i})$ for a certain velocity $v_{x,i}$ of atom i in direction x is obtained by

$$p(v_{x,i}) = \left(\frac{m_i}{2\pi k_{\mathrm{B}}T}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}\frac{m_i v_{x,i}^2}{k_{\mathrm{B}}T}\right) \ , \tag{11}$$

where T denotes the temperature of the simulated system and k_B denotes the Boltzmann constant.

The force acting on atom i depends on the atom's connectivity and surroundings. Taking the definition of the force

$$\mathbf{F}_i(t) = -\nabla_i V(\mathbf{r}(t)) \tag{12}$$

as the negative gradient ∇_i of a potential $V(\mathbf{r}(t))$ as well as the additivity of energies into account, the force can be described by a sum over all possible interactions that participate to the total force $\mathbf{F}_i$ acting on atom i. The potential $V(\mathbf{r}(t))$ is called force field and describes a Born-Oppenheimer potential energy surface on which the movement of atom i occurs. The potential can be interpreted containing bonded interactions $V(\mathbf{r})_\mathrm{bonded}$, which are based on the connectivity of the atoms as well as non-bonded interactions $V(\mathbf{r})_\mathrm{non-bonded}$ taking the surroundings into account:

$$V(\mathbf{r}(t)) = V(\mathbf{r}(t))_\mathrm{bonded} + V(\mathbf{r}(t))_\mathrm{non-bonded} \tag{13}$$

In this approximation the bonded part can be described by different contributions as:

$$V(\mathbf{r}(t))_\mathrm{bonded} = \sum_j^{N_\mathrm{bonds}} V(l_j(t)) + \sum_j^{N_\mathrm{angles}} V(\theta_j(t)) + \sum_j^{N_\mathrm{dihed.}} V(\omega_j(t)) + \sum_j^{N_\mathrm{imp.dihed.}} V(\zeta_j(t)) \tag{14}$$

A typical formulation for the different potentials is shown in figure 2.1. In addition to the formulation, a sketch of the various terms of the potential energy function is presented.

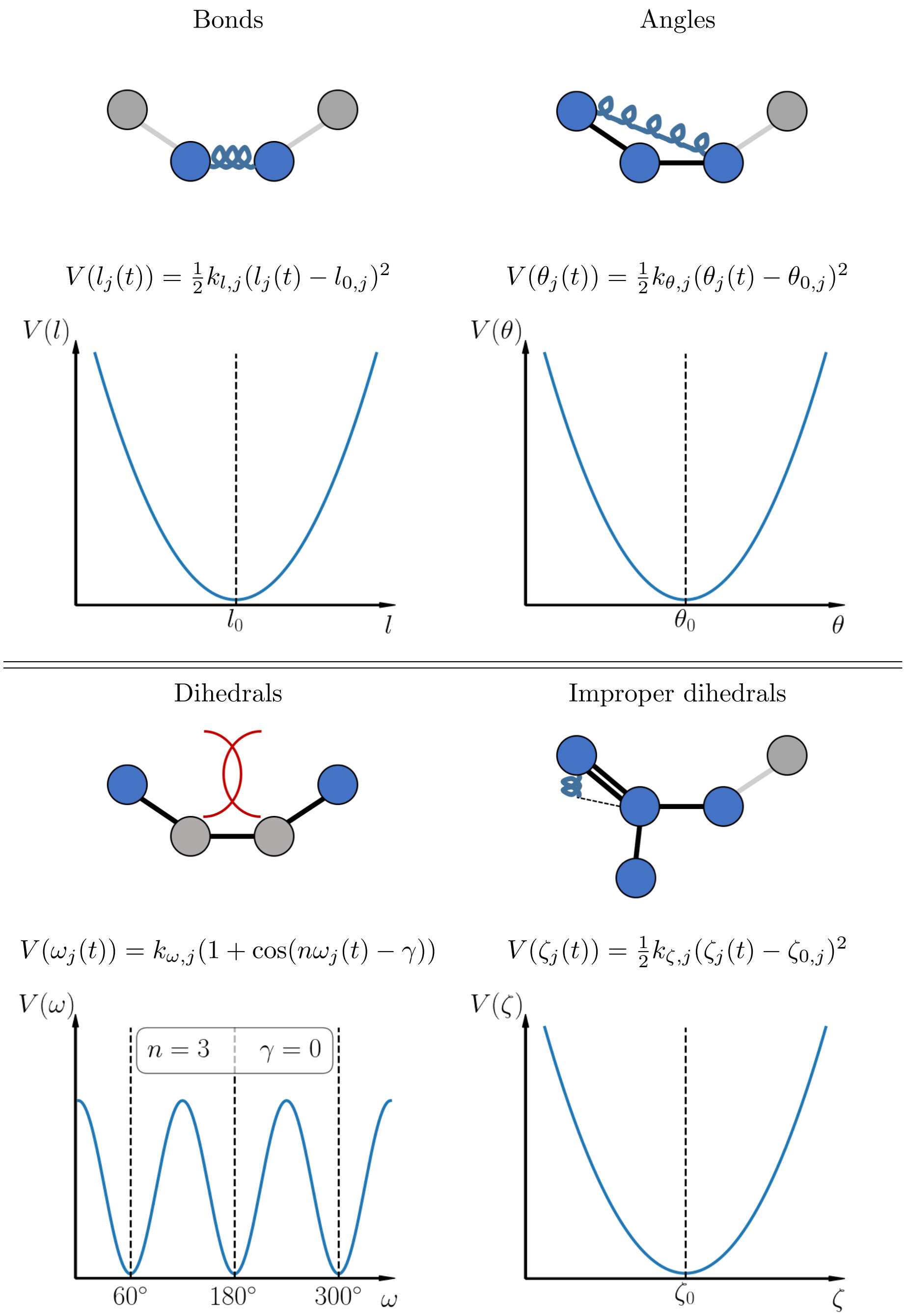

Bonds
$V(l_j(t)) = \frac{1}{2}k_{l,j}(l_j(t) - l_{0,j})^2$
$V(l)$
l_0
l
Angles
$V(\theta_j(t)) = \frac{1}{2}k_{\theta,j}(\theta_j(t) - \theta_{0,j})^2$
$V(\theta)$
θ_0
θ
Dihedrals
$V(\omega_j(t)) = k_{\omega,j}(1 + \cos(n\omega_j(t) - \gamma))$
$V(\omega)$
$n = 3$
$\gamma = 0$
60°
180°
300°
ω
Improper dihedrals
$V(\zeta_j(t)) = \frac{1}{2}k_{\zeta,j}(\zeta_j(t) - \zeta_{0,j})^2$
$V(\zeta)$
ζ_0
ζ

Figure 2.1: Depiction of the potentials included in the bonded part of the force field and their corresponding equations.

Typically, the bonds are modeled as a spring according to Hooke's law describing the bond vibration. A more accurate description can be obtained by using a Morse potential. However, as the deviation from the reference bond length $l_{0,j}$ is small in MD simulations, a harmonic approximation is often sufficient.

Each bond can be described by a force constant $k_{l,j}$ considering the bond strength and the bond length $l_j(t)$ at time t. The angle bending motion can also be approximated by a harmonic potential with a force constant $k_{\theta,j}$, a reference angle $\theta_{0,j}$ as well as the angle $\theta_j(t)$ at time t.

The last type of bonded interactions that is typically considered in the potential energy of bonded interactions are the dihedral angles. These angles describe the torsion of the molecule. There are two types of torsion motion: proper and improper torsions. Proper torsions describe the 1,4-interactions along a bond with respect to the angle $\omega_j(t)$ at time t. It is typically modeled by a cosine function connected to a force constant $k_{\omega,j}$. The number of minima is defined by the multiplicity n and can differ for different hybridizations, such as for two connected sp^3-hybridized atoms n=3, whereas for two connected sp^2-hybridized atoms n=2 is used. The position of the minima is defined by the phase shift γ. Improper dihedrals account for the out-of-plane bending motion of planar systems. This motion can be modeled by a harmonic potential with a force constant $k_{\zeta,j}$, a reference dihedral angle $\zeta_{0,j}$ and the dihedral angle $\zeta_j(t)$ at time t. Depending on the force field, other terms can be incorporated such as cross terms coupling two motions.

Alongside the bonded interactions also non-bonded interactions between two atoms i and j have to be taken into account:

$$V(\mathbf{r}(t))_{\text{non-bonded}} = \sum_{i}^{N_{\text{atoms}}} \sum_{j>i}^{N_{\text{atoms}}} \left(4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}(t)} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}(t)} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}(t)} \right) \tag{15}$$

with $r_{ij}(t) = |\mathbf{r}_i(t) - \mathbf{r}_j(t)|$. These interactions include van der Waals interactions, which can be modeled by a (6,12)-Lennard-Jones potential as well as a Coulomb potential accounting for charge-charge interactions. Both interactions depend on the distance r_{ij} between two atoms i and j. The Lennard-Jones potential consists of two terms and incorporates quantum effects into the model. The first term is a repulsive potential decaying with r^{-12} which incorporates the Pauli exclusion principle and is therefore called Pauli repulsion. The second term decays with r^{-6} and accounts for dispersive interactions. Dispersive interactions describe changes of the electron density generating instantaneous dipoles and therefore dipole-dipole interactions. The other two parameters in the (6,12)-Lennard-Jones potential ϵ_{ij} and σ_{ij} denote the depth of the potential with respect to $V_{LJ} = 0$ and the distance at which $V_{LJ} = 0$ holds. The connection between σ_{ij} and the reference distance (minimum) is $r_{0,ij} = 2^{1/6}\sigma_{ij}$.

To every atom a partial point charge q_i is assigned. The interactions between these point charges is modeled by a Coulomb potential decaying with r^{-1}. Therefore, the Coulomb potential represents long-range interactions while the Lennard-Jones potential models interactions at short-range. The parameter ϵ_0 denotes the vacuum permittivity. A sketch of the potentials is given in figure 2.2.

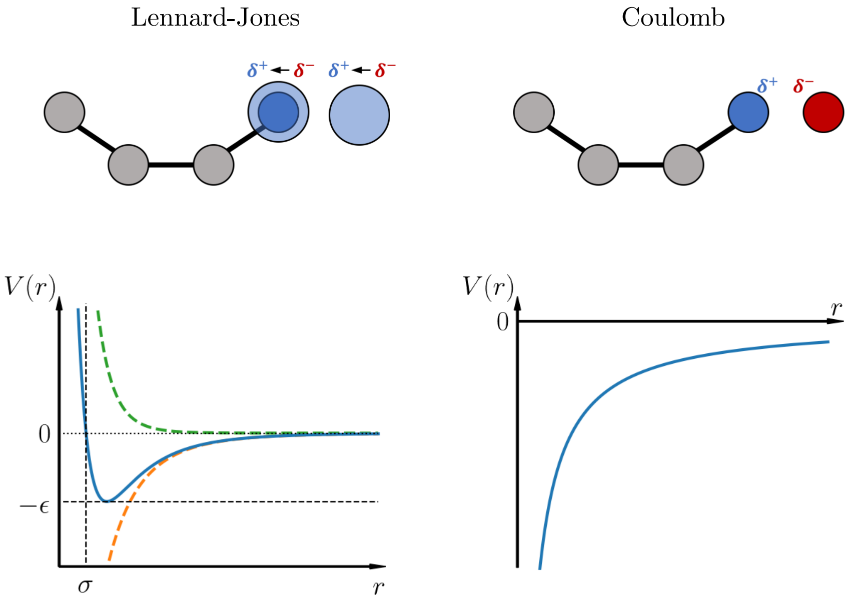


Figure 2.2: Depiction of the potentials included in the non-bonded part of the force field; In the Lennard-Jones potential plot the green curve represents the repulsive (r^{-12}) and the orange curve highlights the attractive (r^{-6}) interactions.

Since a simulation is typically run for a huge number of time steps, the choice of the time step is crucial. A too short chosen time step results in a decreased total simulation length, a too large chosen time step can cause numerical instabilities during the integration process. These numerical instabilities can result from atoms coming too close to each other resulting in a high energy, and therefore a high gradient, in the Lennard-Jones potential. As a consequence, a large force arises between these atoms, which can result in an unphysical behavior and a breakdown of the simulation. A typically used time step for an MD simulation of a biomolecular system is 0.5 fs, since the fastest motion, the X-H bond vibration, is in the range of 10 fs. Using a distance constraint for the X-H bond the time step can be increased to 2 fs. Several algorithms like SHAKE and LINCS [154, 155] were developed to achieve this by removing the vibration of the X-H bond.

To prevent the molecules from leaving the simulation box, periodic boundary conditions are introduced. Let's consider a cubic box with box length L. If a molecule leaves the box in x direction such that the new position will be $L + \delta x$ an equivalent molecule enters the box from the other side with a new position of δx. Thus, the system is treated as a unit cell with "infinite" exact copies in all directions. This fact is very important for the calculation of Coulomb interactions as those interactions occur on a long range decaying with r^{-1}. A depiction of the periodic boundary conditions is shown in figure 2.3.

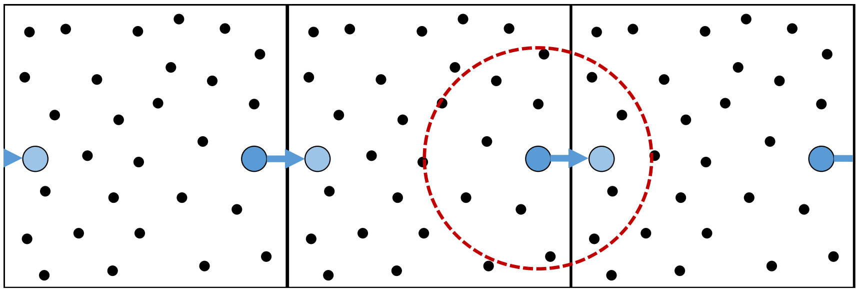


Figure 2.3: Periodic boundary conditions in two dimensions highlighting the movement of a particle leaving the box. The red circle indicates a cutoff e.g. for the calculation of non-bonded interactions.

So far, systems with constant energy which can not interact with their surroundings were discussed. Accordingly, the simulation setup accounts for an isolated system, also called micro-canonical or NVE ensemble, with a constant number of particles N, a constant volume V and a constant energy E. To account for more realistic systems, which can differ in energy, the simulation box can be coupled to a thermal reservoir. This results in a canonical or NVT ensemble with a constant temperature T representing a closed system. Numerically, this is achieved by introducing thermostats [156–158], which modify either the velocities at certain time steps or add additional degrees of freedom to the system. The probability that a system with energy $\epsilon_{\mathbf{r}}$ is realized in an NVT ensemble is given by the Boltzmann distribution:

$$p(\mathbf{r}) = \frac{\exp(-\epsilon_{\mathbf{r}}/k_{\mathrm{B}}T)}{\int \mathrm{d}\epsilon \exp(-\epsilon/k_{\mathrm{B}}T)} \tag{16}$$

Introducing non-rigid boundaries results in an isobaric-isothermal or NPT ensemble with a constant pressure P. This is achieved by introducing a barostat in addition to the thermostat [157, 159].

2.1.2 Quantum Mechanics/Molecular Mechanics

The theory described in this section is based on Ref. [3, 93–98, 160, 161].

Linking molecular dynamics (MD) simulations with quantum mechanical (QM) calculations combines the advantages of both methods. Using the MD level, one can benefit from the speed and hence acquire long simulation time scales, as well as incorporate the full complexity of the simulated system. Taking into account the QM level allows for bond-breaking and electronic rearrangement embedded in an electrostatic environment.

For practical applications the system is split into different layers: An inner layer treated on a QM level and an outer layer treated on a molecular mechanics (MM) level. If needed, more layers can be added on top to either freeze the rest of the system or to account for other interactions, like an implicit solvent model. The connection of both parts, the QM and the MM part, can either be performed using an additive or a subtractive scheme. In this work a two-layer system is used. The following equations thus will be written for a two-layer system. In an additive scheme, the potential of each layer is calculated on the respective level of theory. An additional term is added that accounts for the interaction between the two layers according to

$$V_{\mathrm{QM/MM}} = V_{\mathrm{QM}}(QM) + V_{\mathrm{MM}}(MM) + V_{\mathrm{QM-MM}}(QM + MM), \tag{17}$$

resulting in the full QM/MM potential $V_{\mathrm{QM/MM}}$. In general, $V_{\mathrm{K}}(x)$ denotes the potential energy of the layer K using the atoms associated with x. In a subtractive QM/MM scheme, the full system is calculated on an MM level. In a further step, the potential energy of the QM region computed on the MM level is subtracted and replaced by the corresponding potential calculated on QM level according to:

$$V_{\mathrm{QM/MM}} = V_{\mathrm{QM}}(QM) + V_{\mathrm{MM}}(QM + MM) - V_{\mathrm{MM}}(QM) \tag{18}$$

A subtractive scheme in its standard implementation lacks the explicit interaction between the MM charges with the QM region, which determines the spectroscopic properties inferred by the Coulombic protein interactions from the macromolecular environment. Thus, an additive QM/MM scheme will be used in this thesis.

The next important step is the embedding of the QM layer into the MM layer. The most basic approach is the treatment of the QM-MM interactions on an MM level (force field), called mechanical embedding. In this method the QM atoms are linked by force field terms to the MM level. Thus, no polarization of the electronic wave function on QM level is possible. The electrostatic embedding, where the MM atoms are treated as a point charge field (PCF) in the QM calculation, offers a more suitable way. This ensures that polarization effects can be treated as well. This is achieved by adding one-electron terms

$$\hat{h}_i^{\mathrm{QM/MM}} = \hat{h}_i^{\mathrm{QM}} - \sum_{j}^{N_{\mathrm{pc}}} \frac{e^2 Q_j}{4\pi\epsilon_0 |\mathbf{r}_i - \mathbf{R}_j|} \tag{19}$$

to the electronic Hamiltonian of the QM system, while the interactions with the nuclei are computed classically. In this equation $\hat{h}_i^{\mathrm{QM}}$ denotes the one-electron operator for electron i at position $\mathbf{r}_i$ interacting with the point charge j with partial charge Q_j and position $\mathbf{R}_j$ (accounting for kinetic, nuclear attraction). However, one has to adjust the charges properly to circumvent an over-polarization close to the boundary.

The most accurate way to embed the QM atoms into the MM region is the polarization embedding. Using polarizable force fields, it is possible to allow for a change in the MM charge field as well. In this work electrostatic embedding will be used.

All QM/MM calculations in this work are performed using the software *gmx2qmmm* developed by Dr. Jan P. Götze. For the MM calculations, a steepest descent algorithm is applied which propagates the MM-atoms with respect to

$$\mathbf{r}_i(t+\delta t) = \mathbf{r}_i(t) + \frac{\mathbf{F}_i(t)}{\max(|\mathbf{F}(\mathrm{t})|)} g(t) \tag{20}$$

until a certain energy threshold is reached. In this equation $\mathbf{r}_i(t)$ denotes the position of atom i at time t. $\mathbf{F}_i$ accounts for the force acting on atom i and $g(t)$ represents a maximal displacement. If the potential energy of the new state $\mathbf{r}_i(t+\delta t)$ is smaller than the potential energy of the old state $\mathbf{r}_i(t)$, the propagation step is accepted and the maximal displacement is modified according to $g(t+\delta t) = 1.2 \cdot g(t)$. For a higher potential energy the propagation step is rejected and has to be repeated with a new maximal displacement of $g(t) = 0.2 \cdot g(t)$. The QM-layer is treated at DFT level, which will be discussed in detail in section 2.3.1. For the electrostatic embedding link-atoms are added at the boundaries to satisfy the QM system. To circumvent overpolarization, a charge shift along the boundaries was used by including compensating dipoles along broken bonds.

2.2 Analysis on MD-Level

This section will focus on the analysis of the data generated by MD simulations with respect to dynamic properties and metastable states. Further analysis, like hydrogen-bond networks, secondary structure prediction (DSSP) [19] or root-mean square deviation (RMSD) analysis will not be discussed in this section.

2.2.1 State Space Reduction

This section is based on Ref. [112, 117, 162].

The outcome of an MD simulation is a trajectory $\mathbf{x}_t$ of length N_T containing the positions and the velocities for N system particles. This results in $6N$ coordinates per particle. Neglecting the velocities reduces the dimensionality to $3N$. By treating only the solute, and therefore removing the degrees of freedom of the solvent, the dimensionality can be reduced further. For realistic systems the remaining number of degrees of freedom is often still too large and needs further reduction. This can be achieved either manually by neglecting all degrees of freedoms which are considered "fast", or by methods like principal component analysis (PCA) or time-lagged independent component analysis (TICA).

In PCA and TICA a set of relevant input coordinates $\{z_j(\mathbf{x}_t)\}_{j=1}^{N_\mathrm{z}}$ with a number of N_z coordinates $z_j(\mathbf{x}_t)$ is extracted from the trajectory and transformed into mean free coordinates $\{\chi_j\}_{j=1}^{N_\mathrm{z}}$ by $\chi_j(\mathbf{x}_t) = z_j(\mathbf{x}_t) - \langle z_j(\mathbf{x}_t)\rangle$. For PCA, typically only the Cartesian coordinates are used. For TICA, these coordinates can be chosen arbitrarily like positions, angles or other data, but have to cover the slow dynamics of the analyzed system. The aim of both methods is to project these data onto a space of linearly uncorrelated coordinates. The projected coordinate $\zeta_i(\mathbf{x}_t)$ is given by:

$$\zeta_i(\mathbf{x}_t) = \sum_{j=1}^{N_\mathrm{z}} u_{ij}\chi_j(\mathbf{x}_t) \tag{21}$$

Using Ritz method [163] the weights $\mathbf{u}_j$ can be calculated by solving the eigenvalue problem

$$\tilde{\mathbf{C}}\mathbf{u_j} = \tilde{\lambda}_j\mathbf{S}\mathbf{u_j}, \tag{22}$$

where $\tilde{\mathbf{C}}$ denotes the covariance matrix of the basis $\{\chi_j\}_{j=1}^{N_\mathrm{z}}$ and $\tilde{\lambda}_j$ the j-th eigenvalue. In PCA $\tilde{\mathbf{C}}$ is given by the covariance matrix $\tilde{\mathbf{C}}(0)$ with elements $\tilde{c}_{ij}$ according to:

$$\tilde{c}_{ij} = \frac{1}{T-1}\sum_{t=1}^{T}\chi_i(\mathbf{x}_t)\chi_j(\mathbf{x}_t) \tag{23}$$

The matrix $\mathbf{S}$ is the identity matrix. The eigenvalues denote the autocovariance σ_j^2. Thus, equation 22 can be rewritten as:

$$\tilde{\mathbf{C}}(0)\mathbf{u}_j = \sigma_j^2\mathbf{u}_j \tag{24}$$

In TICA $\tilde{\mathbf{C}}$ denotes the time-lagged covariance matrix $\tilde{\mathbf{C}}(\tau)$. Their elements $\tilde{c}_{ij}$ are defined as:

$$\tilde{c}_{ij} = \frac{1}{T-k-1}\sum_{t=1}^{T-k}\chi_i(\mathbf{x}_t)\chi_j(\mathbf{x}_{t+\tau}) \tag{25}$$

$\tilde{\mathbf{C}}(\tau)$ accounts for the covariance after a lag time $\tau = k\Delta t$, where Δt is the time resolution of the trajectory. $\mathbf{S}$ denotes the overlap matrix and is defined by the covariance matrix $\tilde{\mathbf{C}}(0)$ (equation 23). For TICA equation 22 is given as:

$$\tilde{\mathbf{C}}(\tau)\mathbf{u_j} = \tilde{\lambda}_j(\tau)\tilde{\mathbf{C}}(0)\mathbf{u_j} \tag{26}$$

In PCA as well as in TICA the eigenvectors (weights) can be sorted with respect to the largest autocovariance σ_j^2 and time-lagged autocovariance $\tilde{\lambda}_j$, respectively. Choosing the N_r dominant eigenvectors a state space reduction with respect to the autocovariance is achieved and a projected trajectory $\hat{\mathbf{x}}_t$ is obtained by

$$\hat{\mathbf{x}}_t = \langle \mathbf{U}, \chi(\mathbf{x}_t) \rangle \tag{27}$$

with $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_{N_\mathrm{r}}]$.

2.2.2 Markov State Models

The theory described in this section focuses on Ref. [7–9, 11, 13, 14, 101, 164]. For the core-set Markov state models Ref. [10, 12, 102, 107–111] were additionally taken into account.

In a Markov State Model (MSM) analysis the trajectory $\mathbf{x}_t = [\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_T]$ of length N_T produced by MD simulation (or the projected trajectory $\hat{\mathbf{x}}_t$) is analyzed as a Markov chain. A Markov chain is a series of random experiments where every event depends only on the current state. Therefore, in Markov State Models the following properties are assumed to be valid:

- The trajectory is memory-less (Markovian) in state space Ω, i.e. the next sampled state only depends on the current state without any knowledge about the history. The probability $p(\mathbf{x}, \mathbf{y}; \tau)$ for this process is defined as:

  $$p(\mathbf{x}, \mathbf{y}; \tau)\mathrm{d}\mathbf{x} = \mathbb{P}\left[\mathbf{x}_{t+\tau} = \mathbf{y} | \mathbf{x}_t \in \mathbf{x}\mathrm{d}\mathbf{x}\right] \tag{28}$$

  with $\mathbf{x}, \mathbf{y} \in \Omega$ and $\tau \in \mathbb{R}_{0+}$. Therefore, a transition from state $\mathbf{x}\mathrm{d}\mathbf{x}$ to state $\mathbf{y}$ in time τ is equally probable at every time the system is in $\mathbf{x}\mathrm{d}\mathbf{x}$.

- The trajectory is ergodic, i.e. there exists no dynamically disconnected pair of states in Ω. Additionally, for an infinite simulation time, each state will be visited an infinite amount of times. If both criteria hold, the relative population of each state is represented by its equilibrium probability density μ. For a simulation at constant temperature T the corresponding Boltzmann distribution (Equation 16) will be sampled. This density is a unique invariant measure for the analyzed system.

- The trajectory is reversible, i.e. the simulation sampled equilibrium dynamics. These dynamics have a mean density flux between two states $\mathbf{x}$ and $\mathbf{y}$ of zero and fulfill the criterion of detailed balance:

  $$\mu(\mathbf{x})p(\mathbf{x}, \mathbf{y}; \tau) = \mu(\mathbf{y})p(\mathbf{y}, \mathbf{x}; \tau) \quad \forall\ \mathbf{x}, \mathbf{y} \tag{29}$$

In this section the description of the MSM will be done on a continuous state space Ω and is transferred to a discrete description in the subsections 2.2.2.1 and 2.2.2.2.

Within a continuous state space Ω one can define a propagator $\mathcal{P}(\tau)$ that propagates a density $p_t(\mathbf{y})$, which is not the stationary density $\mu(\mathbf{y})$, by a time step τ to a density $p_{t+\tau}(\mathbf{y})$. The density is given by an ensemble of molecular systems at time t.

$$p_{t+\tau}(\mathbf{y}) = \mathcal{P}(\tau)p_t(\mathbf{y}) = \int \mathrm{d}\mathbf{x}\ p(\mathbf{x},\mathbf{y};\tau)p_t(\mathbf{x}) \tag{30}$$

The propagation is done by taking the transition probability density $p(\mathbf{x},\mathbf{y};\tau)$ for all starting states $\mathbf{x}$ into account. Analogue to the propagator $\mathcal{P}(\tau)$ one can define the transfer operator $\mathcal{T}(\tau)$. Instead of densities, this operator transports functions u_t in time according to:

$$u_{t+\tau}(\mathbf{y}) = \mathcal{T}(\tau)u_t(\mathbf{y}) = \frac{1}{\mu(\mathbf{y})}\int \mathrm{d}\mathbf{x}\ p(\mathbf{x},\mathbf{y};\tau)\mu(\mathbf{x})u_t(\mathbf{x}) \tag{31}$$

The densities $p_t(\mathbf{x})$ and the functions $u_t(\mathbf{x})$ are connected via the stationary distribution $\mu(\mathbf{x})$ according to:

$$p_t(\mathbf{x}) = \mu(\mathbf{x})u_t(\mathbf{x}) \tag{32}$$

$$u_t(\mathbf{x}) = \mu(\mathbf{x})^{-1}p_t(\mathbf{x}) \tag{33}$$

By propagating $p_t(\mathbf{x})$ and $u_t(\mathbf{x})$ k times, respectively, the equations 30 and 31 can be reformulated to fulfill the Chapman-Kolmogorov property:

$$p_{t+k\tau}(\mathbf{x}) = \mathcal{P}(k\tau)p_t(\mathbf{x}) = \mathcal{P}(\tau)^k p_t(\mathbf{x}) \tag{34}$$

$$u_{t+k\tau}(\mathbf{x}) = \mathcal{T}(k\tau)u_t(\mathbf{x}) = \mathcal{T}(\tau)^k u_t(\mathbf{x}) \tag{35}$$

For $k \to \infty$, $p(\mathbf{x})$ converges to the stationary density $\mu(\mathbf{x})$ and $u(\mathbf{x})$ converges to a constant function $\mathbf{1}(\mathbf{x})$. The dynamics can be analyzed with respect to the eigenvalue spectrum. Both operators share the same eigenvalues λ_i with corresponding eigenfunctions $l_i(\mathbf{x})$ for $\mathcal{P}(\tau)$ and $r_i(\mathbf{x})$ for $\mathcal{T}(\tau)$:

$$\mathcal{P}(\tau)l_i(\mathbf{x}) = \lambda_i(\tau)l_i(\mathbf{x}) \tag{36}$$

$$\mathcal{T}(\tau)r_i(\mathbf{x}) = \lambda_i(\tau)r_i(\mathbf{x}) \tag{37}$$

Under the assumption of reversible dynamics the eigenfunctions are connected via:

$$l_i(\mathbf{x}) = \mu(\mathbf{x}) r_i(\mathbf{x}) \tag{38}$$

In addition, all eigenfunctions as well as the corresponding eigenvalues are real-valued. The eigenvalues are bound in the interval $]-1, 1]$, where the eigenvalue $\lambda_1 = 1$ is unique for an ergodic system according to the Perron-Frobenius theorem. The corresponding eigenfunctions l_1 and r_1 have only positive-valued entries and describe the equilibrium state. The sign structure of the other eigenfunctions with eigenvalues < 1 accounts for dynamic processes. Two eigenfunctions r_i and l_j are orthonormal with respect to the inner product $\langle r_i, l_j \rangle = \delta_{ij}$, where δ_{ij} denotes the Kronecker delta. This feature is a consequence of the self-adjointness of the operators with respect to a weighted inner product for a reversible Markov process:

$$\langle f | \mathcal{T}(\tau) g \rangle_{\mu^{-1}} = \langle g | \mathcal{T}(\tau) f \rangle_{\mu^{-1}} \tag{39}$$

with

$$\langle f | \mathcal{T}(\tau) g \rangle_{\mu^{-1}} = \int \mathrm{d}\mathbf{x} \; f(\mathbf{x}) \mu^{-1} \mathcal{T}(\tau) g(\mathbf{x}) \tag{40}$$

The eigenvalue spectrum can be split into the Perron-cluster with eigenvalues close to one $\{\lambda_1 = 1, \lambda_2 < 1, \ldots, \lambda_m \gg 0\}$, which are separated by a spectral gap to the other eigenvalues in the spectrum. Together with the Chapman-Kolmogorov property the transfer operator can be decomposed into a fast and a slow decaying part:

$$u_{t+k\tau} = \mathcal{T}(k\tau) u_t(\mathbf{x}) = \sum_{i=1} \lambda_i^k \langle u_t, r_i \rangle_\mu r_i(\mathbf{x}) = \sum_{i=1}^{m} \lambda_i^k \langle u_t, r_i \rangle_\mu r_i(\mathbf{x}) + \mathcal{T}_{fast}(k\tau) u_t(\mathbf{x}) \tag{41}$$

For $k \to \infty$ all processes corresponding to r_i, λ_i with $i > 1$ will decay exponentially towards 0. The higher the eigenvalue λ_i the slower the decay. Hence, the Perron-cluster forms m dominant processes decaying much slower than all other processes towards equilibrium. Due to this, we will only focus on the dominant processes. A consequence of the Markovian behavior is that the decay of each eigenvalue follows a single exponential decay with respect to τ / t_i. The parameter t_i denotes the implied timescale of the dynamic process and can be calculated by:

$$t_i = -\frac{\tau}{\ln |\lambda_i|} \tag{42}$$

Thus, t_i has to be constant for a valid MSM. An example for an MSM analysis is shown in figure 2.4.

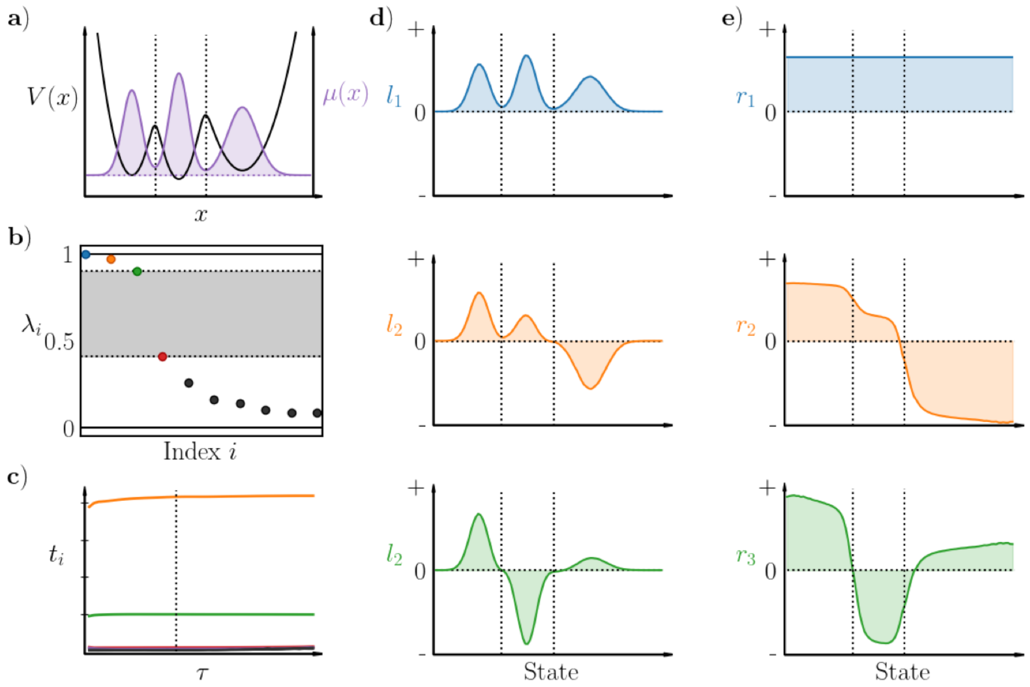


Figure 2.4: Sketch of an MSM analysis. In a) a triple-well potential with corresponding stationary density is shown. Constructing an MSM the eigenvalue spectrum (b) is obtained; The dominant and the non-dominant eigenvalues are separated by a spectral gap (grey). c) The corresponding implied timescales are highlighted. In d) the left and in e) the right eigenvectors are shown, colored with respect to the corresponding eigenvalue. The trajectory containing 10^6 time steps was generated using a Markov-Chain Monte-Carlo sampling algorithm ($\sigma = 1,\ \beta = 1$) [165].

For practical applications one shifts from a continuous description of the states towards a discrete one. Subsections 2.2.2.1 and 2.2.2.2 will discuss two ways to describe a MSM on a discrete space. A detailed description of how to determine the discrete states is presented in section 2.2.3.

2.2.2.1 Full-partitioning Markov State Models

Using the variational principle and the Ritz ansatz, the continuous eigenfunctions $r_i(\mathbf{x})$ of the Transfer operator $\mathcal{T}(\tau)$ can be approximated by a linear expansion in a finite basis $\{\chi_j\}_{j=1}^{N_\mathrm{C}}$ with unknown coefficients a_{ij}:

$$r_i(\mathbf{x}) \approx \sum_{j=1}^{N_\mathrm{C}} a_{ij}\chi_j(\mathbf{x}) \tag{43}$$

Each basis function describes one discrete state C_j of the system. Due to this, the complete system is projected onto N_C discrete states $\{C_j\}_{j=1}^{N_\mathrm{C}}$. In full-partitioning Markov State Models (fp-MSMs) the discrete states cover the full state space Ω and are not overlapping such that $\cup_{j=1}^{N_\mathrm{C}} C_j = \Omega$ and $C_i \cap C_j = \emptyset$.

A discussion of the definition of the states for an fp-MSM is described in section 2.2.3.2. To obtain an optimal set of coefficients the following eigenvalue problem can be solved:

$$\mathbf{T}(\tau)\mathbf{a}_i = \lambda_i(\tau)\mathbf{a}_i \tag{44}$$

Due to the variational approach the eigenvalue λ_i will always be smaller or equal to the corresponding eigenvalue of the continuous functions. Equation 44 is a discrete version of equation 37. The operator $\mathbf{T}(\tau)$ denotes the transition matrix, whose elements $T_{ij}(\tau)$ denote the probability to find the system in state C_i at time t and in state C_j at time $t+\tau$:

$$T_{ij}(\tau) = \mathbb{P}\left[\mathbf{x}_{t+\tau} \in C_j | \mathbf{x}_t \in C_i\right] \tag{45}$$

In this equation $\mathbf{x}_t$ denotes the analyzed trajectory as already described in section 2.2.2 or the projected trajectory $\hat{\mathbf{x}}_t$ as described in section 2.2.1. The transition matrix elements can be obtained by Galerkin discretization of $\mathcal{T}(\tau)$ with:

$$T_{ij} = \frac{\langle \chi_i, \mathcal{T}(\tau)\chi_j \rangle_\mu}{\langle \chi_i, \mathbb{1} \rangle_\mu} \tag{46}$$

For a full-partitioning discretization the continuous eigenfunctions can be expanded to a set of indicator or step functions. These indicator functions are equal to 1 if the system is in the corresponding discrete state or equal to 0 otherwise. In general they can be written as:

$$\chi_j(\mathbf{x}) = \begin{cases} 1, & \text{for } \mathbf{x} \in C_j \\ 0, & \text{for } \mathbf{x} \notin C_j \end{cases} \tag{47}$$

For practical applications the transition matrix elements can be estimated by a counting matrix $\mathbf{C}(\tau)$ using a sliding window approach. For this approximation τ is set as $\tau = k\Delta t$ which is a k-fold of the time step Δt separating the neighboring elements of the trajectory $\mathbf{x}_t$. The elements c_{ij} of the counting matrix can then be calculated by:

$$c_{ij}(\tau) = \sum_{t=0}^{T-k} \chi_i(\mathbf{x}_t)\chi_j(\mathbf{x}_{t+k}) \tag{48}$$

A set of trajectories can be treated as well by summing up the counting matrices for the single trajectories. As the simulation time is finite it is not possible to achieve perfectly reversible dynamics. Therefore, the reversibility can be introduced artificially by averaging over associated elements of the transition matrix:

$$c_{ij,\text{rev}}(\tau) = \frac{c_{ij} + c_{ji}}{2} \tag{49}$$

To ensure the validity of equations 45 and 46 the elements of the counting matrix are normalized by the row sum to obtain the transition matrix elements T_{ij}:

$$T_{ij}(\tau) = \frac{c_{ij,\mathrm{rev}}(\tau)}{\sum\limits_{j} c_{ij,\mathrm{rev}}(\tau)} \tag{50}$$

A direct calculation of $\mathbf{T}(\tau)$ from the non-reversible counting matrix would yield the maximum probability estimator of $\mathbf{C}(\tau)$. However, by enforcing reversibility the property gets lost. Algorithms have been developed to account for this, as described in Ref. [11], by finding the maximum probability estimator for the reversible case.

Projecting the continuous trajectory onto discrete states causes a loss of Markovianity as the exact information within a state C_j gets lost during the projection. Thus, to fulfill the Markov property several lag times have to be tested to find a range of lag times for which constant implied timescales t_i are achieved.

2.2.2.2 Core-set Markov State Models

Instead of using a full partitioning of the state space, a discretization into core sets can be chosen. Core sets C_i describe a subset of the state space Ω such that $\cup_{j=1}^{N_\mathrm{C}} C_j \subset \Omega$ and that there exists a space $\Omega \backslash \cup \{C_j\}_{j=1}^{N_\mathrm{C}}$ that is not assigned to any discrete state. As for the full-partitioning discretization the core sets are not overlapping $C_i \cap C_j = \emptyset$. A method to obtain the core set is discussed in section 2.2.3.1. The construction of the core-set MSM (cs-MSM) is in most steps similar to the full-partitioning case. As in equation 43 the continuous eigenfunctions can be expanded in a finite basis $\tilde{\chi}_i$. Instead of indicator functions for a core-set Markov State Model committor functions $q_i(\mathbf{x})$ are applied:

$$\tilde{\chi}_i(\mathbf{x}) = q_i(\mathbf{x}) = \begin{cases} 1, & \text{for } \mathbf{x} \in C_i \\ 0, & \text{for } \mathbf{x} \in C_j \ \forall j \neq i \\ q_i(\mathbf{x}) \in (0,1), & \text{for } \mathbf{x} \notin \cup_{j=1}^{N_\mathrm{C}} C_j \end{cases} \tag{51}$$

A committor function $q_i(\mathbf{x})$ can be interpreted as a probability to be in a certain discrete state. If the trajectory is in core set C_i the associated committor function $q_i(\mathbf{x})$ will be equal to 1 and all other functions $q_j(\mathbf{x})$ $\forall j \neq i$ will be equal to 0. If the trajectory is in no core set the commitor function $q_i(\mathbf{x})$ assigns a probability that the core set C_i will be hit before any other core set C_j $\forall j \neq i$. As a consequence the following equation has to be fulfilled:

$$\sum_{i=1}^{N_\mathrm{C}} q_i(\mathbf{x}) = 1 \tag{52}$$

Similar to the fp-MSM the expansion coefficients $\tilde{a}_{ij}$ for the expansion in the basis of the committor functions can be estimated solving the generalized eigenvalue problem

$$\mathbf{T}(\tau)\tilde{\mathbf{a}}_i = \lambda_i(\tau)\mathbf{M}\tilde{\mathbf{a}}_i \ , \tag{53}$$

where the operator $\mathbf{M}$ accounts for the overlap of the basis functions as the committor functions are not orthogonal anymore, compared to the indicator functions. This generalized eigenvalue problem is similar to equation 44, since for an orthogonal basis, as described in section 2.2.2.1, $\mathbf{M} = \mathbf{I}_{N_\mathrm{C}}$ is applied, where $\mathbf{I}_{N_\mathrm{C}}$ denotes the identity matrix with N_C rows and columns. In a non-orthogonal basis $\mathbf{M}$ can be approximated as:

$$M_{ij} = \frac{\langle \chi_i, \chi_j \rangle_\mu}{\langle \chi_i, \mathbb{1} \rangle_\mu} \tag{54}$$

Instead of committor functions, one can use milestoning functions for the Markov State Model construction as they are easier to calculate. One distinguishes between two different milestoning functions: The backward milestoning function $m_i^-(\mathbf{x})$

$$m_i^-(\mathbf{x}) = \begin{cases} 1, & \text{for } \mathbf{x} \in C_i \\ 1, & \text{for } \mathbf{x} \notin \cup_{j=1}^{N_\mathrm{C}} C_j \text{ and last came from } C_i \\ 0, & \text{else} \end{cases} \tag{55}$$

and the forward milestoning function $m_i^+(\mathbf{x})$

$$m_i^+(\mathbf{x}) = \begin{cases} 1, & \text{for } \mathbf{x} \in C_i \\ 1, & \text{for } \mathbf{x} \notin \cup_{j=1}^{N_\mathrm{C}} C_j \text{ and will go next to } C_i \\ 0, & \text{else} \end{cases} \quad . \tag{56}$$

In the case of the trajectory being in a core set C_i both milestoning functions will share the same value. If this is not true, the backward milestoning functions $m_i^-(\mathbf{x})$ is 1 if the last visited core set was C_i. Complementary, the forward milestoning function $m_i^+(\mathbf{x})$ is 1 if the next visited core set will be C_i.

Based on the milestoning functions the counting matrix can be described analogously to equations 48 to 50 for practical applications as:

$$c_{ij}(\tau) = \sum_{t=0}^{T-k} m^-(\mathbf{x}_t) m_j^+(\mathbf{x}_{t+k}) \tag{57}$$

Similar to the transition matrix, the overlap matrix can be estimated as:

$$m_{ij} = \sum_{t=0}^{T} m^{-}(\mathbf{x}_t) m_j^{+}(\mathbf{x}_t) \tag{58}$$

To obtain a reversible overlap matrix, an equivalent formalism as described in equation 49 is applied, followed by a normalization using the row sum according to:

$$M_{ij} = \frac{m_{ij,\mathrm{rev}}}{\sum\limits_{j} m_{ij,\mathrm{rev}}} \tag{59}$$

Equation 53 can be connected to equation 37 and 44 by defining an "effective" transition matrix $\mathbf{T}_Q(\tau) = \mathbf{T}(\tau)\mathbf{M}^{-1}$.

2.2.3 Clustering

A data set S can be described as a set of N data points $S = \{x_1, x_2, ..., x_N\}$. Each data point x_i denotes an M-dimensional vector in a space $\Omega \in \mathbb{R}^M$. The aim of cluster algorithms is to assign every data point $x_i \in S$ to a cluster C_j, with $C_j \subset S$. The assigned data point can either belong to a certain cluster ("hard" or "non-overlapping" clustering [122]) or to more than one cluster with a certain degree ("soft" or "overlapping" clustering [127]).

This thesis will focus on "non-overlapping" clustering. In "non-overlapping" clustering each data point is assigned to exactly one cluster such that $C_j \cap C_k = \emptyset$. The clusters can either cover the complete data set such that $\cup\{C_j\}_{j=1}^{N_\mathrm{C}} = S$ or only a subset such that $\cup\{C_j\}_{j=1}^{N_\mathrm{C}} \subset S$. For the latter case an additional subset $Z = S \backslash \cup \{C_j\}_{j=1}^{N_\mathrm{C}}$ is defined that includes all not assigned data points, later denoted as noise.

The main focus in this thesis will be on the density-based Common-Nearest-Neighbor (CNN) algorithm [16, 135, 166] as well as on the centroid-based k-Means++ algorithm [15, 122, 167].

2.2.3.1 Common-Nearest-Neighbor Algorithm

The CNN algorithm partitions the data set with respect to two parameters in a set of clusters $\{C_j\}_{j=1}^{N_\mathrm{C}}$ and a set of noise Z. The first parameter is a distance measure R describing the neighborhood of every data point x_i within the data set S, later denoted as R-neighborhood. The second parameter N denotes, in conjunction with R, a density threshold. In the CNN algorithm, two data points belong to the same cluster if they share at least N neighbors with respect to their R-neighborhood (figure 2.5). In addition, both data points themselves have to be neighbors with respect to R. The CNN algorithm is a cluster growing algorithm [168] and can be described by the following scheme:

1. **Parameter choice**: Define a parameter set $\{R, N\}$

2. **Cluster initialization**:

   i Initiate a new cluster $C_j = \{\}$

   ii Choose an arbitrary data point x_i from the set of unclustered data points U, add it to C_j and remove it from U. Store x_i to a set of newly added data points L.

3. **Cluster expansion**:

   i Select a data point x_{new} from L

   ii Add all data points from the set of unclustered data points U for which the density criterion with respect to x_{new} is fulfilled to C_j and L and remove them from the set U.

   iii Remove x_{new} from L

   iv Repeat steps i to iii until no new data point is assigned ($L = \{\}$).

4. **Clustering**: Repeat steps 2 to 3 until the set of unclustered data points is empty ($U = \{\}$)

5. **Termination**: Assign all clusters with only 1 assigned data point to Z

A depiction of this scheme is given in figure 2.5. The CNN algorithm is fully deterministic. Therefore, the selection of x_{new} can be done arbitrarily. However, to speed up the algorithm the data point in U with the highest number of neighbors is chosen to initiate a new cluster (Step 2). An additional speed up is achieved by the criterion that two data points have to be neighbors of each other. Hence, all data points that can be added in Step 3 have to be located within R of x_{new}.

If two data points x_i and x_j are closer than a distance R and fulfill the density criterion with respect to each other, they are density-reachable. However, as the algorithm is a cluster growing algorithm also two data points x_i and x_k that are not density-reachable from each other can belong to the same cluster. This property is called density-connectivity [130]. Two data points x_i and x_k are density-connected if there is a chain of density-reachable data points connecting x_i and x_k (figure 2.5, lower right). For example, if x_j is density-reachable from x_i and x_k then x_i and x_k are density-connected via x_j.

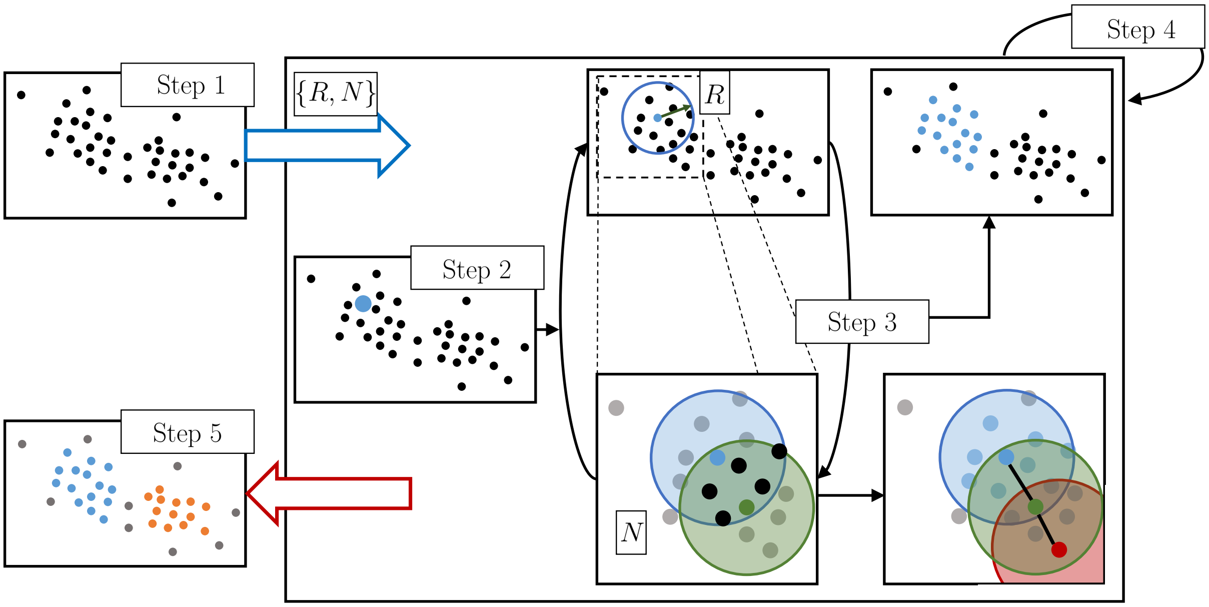


Figure 2.5: Depiction of the clustering process according to the presented scheme: The data set is clustered with respect to a parameter set $\{R, N\}$ (Step 1); A cluster is initiated (Step 2) and expanded with respect to the cluster criterion (Step 3); Using an iterative procedure the complete data set is assigned to clusters (Step 4); After a termination step the final outcome is obtained (Step 5); In the lower right the principle of density-connectivity is presented: Although the blue and the red data point are not density-reachable with respect to $\{R, N\}$, they are density-connected with respect to the green data point.

Former studies [135] showed that the algorithm can yield a high number of small clusters containing only a few data points. To compensate these “artifacts” a third parameter M is added. M denotes the minimal cluster size and is typically set to 0.1 % of the data set size. If $M > 2$ all clusters that are smaller than M will be added to Z and removed from the set of clusters. This modifies steps 1 and 5 as follows:

1. **Parameter choice**: Define a parameter set $\{R, N, M\}$

5. **Termination**: Assign all clusters with less than M assigned data points to Z

A drawback of density-based-cluster algorithms is the extraction of clusters with a huge difference in data point density between these two clusters. In the CNN algorithm this can be compensated by applying a hierarchical clustering scheme as introduced in Ref. [135]. In a hierarchical approach the data set is clustered with a “low” density threshold $\{R_s, N_s\}$ in a first step. A “low” density threshold refers to either a large cutoff R_s or a small number of shared neighbors N_s. In the second step all clusters that are not refined satisfactorily are clustered again using a “higher” density threshold $\{R_h, N_h\}$, i.e. either decreasing R_h or increasing N_h with respect to $\{R_s, N_s\}$. This procedure can be repeated until the clustering is satisfactory. With this approach clusters of different data point density can be separated, which would not be possible using a single parameter set as illustrated in figure 2.6. A discussion on how to estimate a good parameter set as well as on how to estimate the quality of the clustering can be found in Ref. [16, 135, 166].

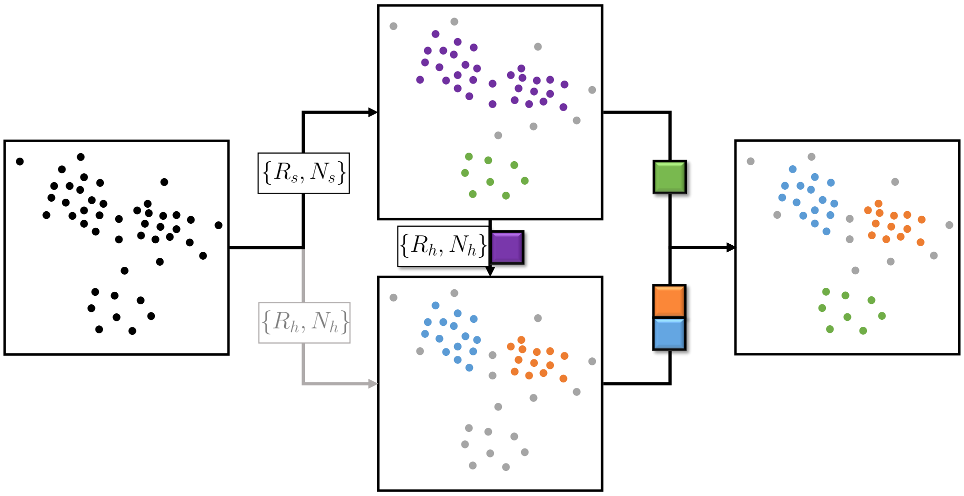


Figure 2.6: Depiction of a hierarchical clustering approach using the density thresholds $\{R_s, N_s\}$ and $\{R_h, N_h\}$; Both parameter sets themselves are not feasible to extract all three clusters. Using an hierarchical approach by applying first $\{R_s, N_s\}$, extracting the green cluster and reclustering the violet cluster with $\{R_h, N_h\}$ yields the expected clustering; Extracted/Transferred clusters are highlighted by a colored box.

2.2.3.2 k-Means++ Algorithm

The k-Means algorithm is a centroid-based algorithm partitioning the data-set into Voronoi-cells (figure 2.7). The algorithm needs only the number of clusters n as an input parameter. Based on this input a set of n centroids $C = \{c_j\}_{j=1}^{n}$ is randomly initiated and optimized such that a function ϕ is minimized:

$$\phi = \sum_{i=1}^{N} \min_{c_j \in C} ||x_i - c_j||^2 \tag{60}$$

The k-Means algorithm can be described by the following scheme:

1. **Parameter choice**: Define a number of clusters n
2. **Cluster initialization**: Choose n centroids randomly
3. **Cluster optimization**:

    i Assign all data points to their closest centroids

    ii For each cluster compute the center of mass as the new centroid iii Repeat steps i to ii until the change of the centroid drops below a predefined threshold

In general the cluster centroids are drawn from a uniform distribution. The random choice of the initial clusters can bias the outcome [169]. Consequently, different techniques were developed to optimize step 2. One method was proposed in Ref. [167] named the k-Means++ algorithm. In the k-Means++ algorithm step 2 is modified in the following way:

2. **Cluster initialization**:

   i Choose the first centroid from a uniform distribution

   ii Choose a further centroid with a probability of $p(x) = D(x)^2 / \sum_{x \in S} D(x)^2$ (with $D(x)$ being the distance to the closest centroid)

   iii Repeat step ii until n centroids are chosen

The addition of these steps speeds up the k-Means algorithm and yields a better partitioning of the data set compared to a random selection of the initial centroids. The drawbacks of the algorithm are, on the one hand, the non-deterministic behavior (dependency on the initial centroids) and, on the other hand, the characterization of different shapes due to the use of the distance to the nearest centroid as a clustering criterion.

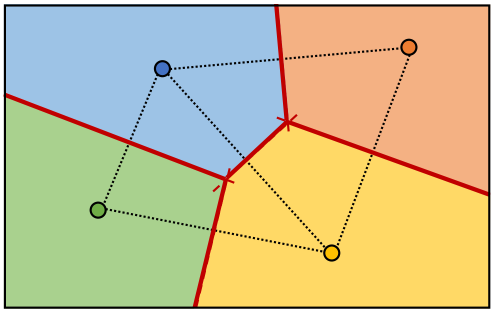


Figure 2.7: Depiction of a Voronoi partitioning. The circles represent the centroids of the cluster, the solid red lines represent the border; connected centroids (dashed lines) are bisected midway.

2.3 Analysis on QM-Level

The theory explaining density functional theory is based on Ref. [170–172]. For time-dependent density functional theory it is additionally referred to Ref. [173–176].

2.3.1 Density Functional Theory

All equations in this chapter are given in atomic units ($\hbar = 1$ a.u, $m_e = 1$ a.u., $e = 1$ a.u., $\frac{1}{4\pi\epsilon_0} = 1$ a.u.). As MD simulations based on classical mechanics are limited with respect to bond breaking and are strongly biased by the used force field, one can increase the level of theory to get an insight into the electronic structure. Assuming a system without time dependency, we can obtain information about the electronic structure by solving the time-independent Schrödinger equation

$$\hat{H}\Psi(\mathbf{r},\mathbf{r}_2,\ldots,\mathbf{r}_{N_e},\mathbf{R}) = E\Psi(\mathbf{r},\mathbf{r}_2,\ldots,\mathbf{r}_{N_e},\mathbf{R}), \tag{61}$$

where $\hat{H}$ denotes the Hamiltonian of the system with eigenfunctions $\Psi(\mathbf{r},\mathbf{r}_2,\ldots,\mathbf{r}_{N_e})$, also called wave functions and eigenvalues E, accounting for an energy. The vector $\mathbf{r_i}$ denotes the position of electron i of a total N_e electrons and $\mathbf{R}$ denotes the position of all nuclei. By applying the Born-Oppenheimer approximation, which assumes the position of the nuclei as fixed, the Hamiltonian can be split into a nuclear $\hat{H}_\mathrm{n}$ and an electronic Hamiltonian $\hat{H}_\mathrm{el}$. The electronic Hamiltonian consists of three terms according to

$$\hat{H}_\mathrm{el} = \hat{T} + \hat{V}_\mathrm{ee} + \hat{V}_\mathrm{ext}\ , \tag{62}$$

where $\hat{T}$ represents the kinetic energy operator of the electrons:

$$\hat{T} = -\frac{1}{2}\sum_{j=1}^{N_e}\nabla_j^2 \tag{63}$$

∇_j^2 is the Laplace operator with respect to j defined as $\nabla_j^2 = \frac{\partial^2}{\partial x_j^2} + \frac{\partial^2}{\partial y_j^2} + \frac{\partial^2}{\partial z_j^2}$ in Cartesian coordinates. The term $\hat{V}_{ee}$ accounts for the electron-electron repulsion and is defined as:

$$\hat{V}_\mathrm{ee} = \frac{1}{2}\sum_{i\neq j}^{N_e}\frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} = \sum_{i=1}^{N_e}\sum_{j>i}^{N_e}\frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} \tag{64}$$

In order to avoid double counting, the factor $\frac{1}{2}$ is applied. The last term $\hat{V}_\mathrm{ext}$ accounts for a one-body potential

$$\hat{V}_\mathrm{ext} = \sum_{j=1}^{N_e} v_\mathrm{ext}(\mathbf{r}_j) \tag{65}$$

and describes in most cases the interaction of the electrons with the nuclei. Additional terms may arise in case, e.g., of the application of an external potential such as an electric field. In density functional theory (DFT), one uses the electron density $\rho(\mathbf{r})$ instead of a wave function representation. The electron density $\rho(\mathbf{r})$ addresses the probability of detecting an electron in the volume $\mathrm{d}^3\mathbf{r}$ around $\mathbf{r}$ according to:

$$\rho(\mathbf{r}) = N_e \int \mathrm{d}^3\mathbf{r}_2 \ldots \int \mathrm{d}^3\mathbf{r}_{N_e} |\Psi\left(\mathbf{r},\mathbf{r}_2,\ldots,\mathbf{r}_{N_e};\mathbf{R}\right)|^2 = \sum_{j=1}^{N_e}|\phi_j(\mathbf{r})|^2 \tag{66}$$

with normalization:

$$\int \mathrm{d}^3\mathbf{r}\ \rho(\mathbf{r}) = N_e \tag{67}$$

A common way to obtain the electronic kinetic energy within DFT is the use of Kohn-Sham functions ϕ_i. ϕ_i is a single electron wave function of a system of fictitious non-interacting electrons. The Kohn-Sham functions are used to obtain the total density according to equation 66 and can be obtained similar to equations 61-65 by solving the Kohn-Sham equation:

$$\left(-\frac{1}{2}\nabla^2 + v_{\mathrm{eff}}(\mathbf{r})\right)\phi_i = \epsilon_i\phi_i \tag{68}$$

In this equation, $v_{\mathrm{eff}}(\mathbf{r})$ denotes the effective potential

$$v_{\mathrm{eff}}(\mathbf{r}) = v_{\mathrm{ext}}(\mathbf{r}) + v_{\mathrm{H}}(\mathbf{r}) + v_{\mathrm{xc}}(\mathbf{r}) \tag{69}$$

accounting for an external potential v_{ext}, a Hartree potential v_{H} defined as

$$v_{\mathrm{H}} = \int \mathrm{d}^3\mathbf{r}'\,\frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} \tag{70}$$

and an exchange-correlation potential v_{xc}

$$v_{\mathrm{xc}} = \frac{\delta E_{\mathrm{xc}}}{\delta\rho(\mathbf{r})} \tag{71}$$

with the exchange-correlation energy $E_{\mathrm{xc}} = E_{\mathrm{x}} + E_{\mathrm{c}}$. The exchange-correlation energy is necessary since the approximation of non-interacting electrons causes an error in the total energy. It can be broken down into an exchange energy E_{x} and a correlation energy E_{c}. Both energies compensate missing effects like exchange and correlation of electrons.

Different functionals can be used to define the exchange-correlation energy. The most basic functional is the local-density approximation (LDA) which assumes the electrons to be distributed in a uniform electron gas. For this simple approximation, $E_{\mathrm{xc}}^{\mathrm{LDA}}$ can be written as

$$E_{\mathrm{xc}}^{\mathrm{LDA}}[\rho] = \int \mathrm{d}^3\mathbf{r}\ \rho(\mathbf{r})\epsilon_{\mathrm{xc}}(\rho(\mathbf{r})) \tag{72}$$

with $\epsilon_{\mathrm{xc}}(\rho(\mathbf{r}))$ denoting the exchange-correlation energy per particle. The inclusion of electron spins α and β, with $\rho = \rho_\alpha + \rho_\beta$, allows for spin-polarization and one obtains:

$$E_{\mathrm{xc}}^{\mathrm{LSDA}}\left[\rho_\alpha, \rho_\beta\right] = \int \mathrm{d}^3\mathbf{r}\ \rho(\mathbf{r})\epsilon_{\mathrm{xc}}(\rho_\alpha(\mathbf{r}), \rho_\beta(\mathbf{r})) \tag{73}$$

To arrive at a more accurate model for systems where the assumption of a uniform electron gas does not hold, one can apply the generalized gradient approximation (GGA) instead of LDA. In GGA one adds the gradient of the electron density $\nabla\rho(\mathbf{r})$ to take the non-homogeneous distribution of the electrons into account. In general, the functional for GGA can be written as:

$$E_{\mathrm{xc}}^{\mathrm{GGA}}\left[\rho_\alpha, \rho_\beta\right] = \int \mathrm{d}^3\mathbf{r}\ f(\rho_\alpha(\mathbf{r}), \rho_\beta(\mathbf{r}), \nabla\rho_\alpha(\mathbf{r}), \nabla\rho_\beta(\mathbf{r})) \tag{74}$$

The exchange energy dominates in most cases. Thus, using hybrid functionals that contain the exact exchange energy E_x and are obtained from a wave function based ansatz, one can further modify the exchange-correlation energy via

$$E_{\mathrm{xc}}^{\mathrm{Hyb}}\left[\rho_\alpha, \rho_\beta\right] = a\left(E_{\mathrm{x}} - E_{\mathrm{x}}^{\mathrm{GGA}}\left[\rho_\alpha, \rho_\beta\right]\right) + E_{\mathrm{xc}}^{\mathrm{GGA}}\left[\rho_\alpha, \rho_\beta\right] \tag{75}$$

with a scaling factor a.

2.3.2 Linear Response Time-dependent Density Functional Theory

Spectroscopy is intrinsically a time-dependent process involving an external electromagnetic field and the corresponding response of the material. To include time-dependent properties into the DFT calculations, one has to solve the time-dependent Kohn-Sham equation

$$i\frac{\partial\phi_i(\mathbf{r}, t)}{\partial t} = \left[-\frac{1}{2}\nabla^2 + v_{\mathrm{eff}}(\mathbf{r}, t)\right]\phi_i(\mathbf{r}, t) \tag{76}$$

with a time-dependent effective potential $v_{\mathrm{eff}}(\mathbf{r}, t)$

$$v_{\mathrm{eff}}(\mathbf{r}, t) = v_{\mathrm{ext}}(\mathbf{r}, t) + v_{\mathrm{H}}(\mathbf{r}, t) + v_{\mathrm{xc}}(\mathbf{r}, t) \tag{77}$$

and time-dependent Kohn-Sham functions $\phi_i(\mathbf{r}, t)$. Similar to equation 66 one can formulate a time-dependent electron density $\rho(\mathbf{r}, t)$ according to:

$$\rho(\mathbf{r}, t) = \sum_{j=1}^{N_e} |\phi_i(\mathbf{r}, t)|^2 \tag{78}$$

For calculations of small perturbations of the external potential

$$v_{\text{ext}}(\mathbf{r},t) = v_{\text{ext}}(\mathbf{r}) + \delta v_{\text{ext}}(\mathbf{r},t) \tag{79}$$

with $\delta v_{\text{ext}}(\mathbf{r},t) \ll v_{\text{ext}}(\mathbf{r})$, like that obtained by standard spectroscopy, linear response theory can be applied. For small perturbations and the assumption that the initial state is the ground state (GS), one can write the changes in the densities with respect to the ground state as

$$\rho(\mathbf{r},t) = \rho_{\text{GS}}(\mathbf{r},t) + \delta\rho(\mathbf{r},t) \tag{80}$$

with incremental, time-dependent changes $\delta\rho(\mathbf{r},t)$ and $\delta v_{\text{ext}}(\mathbf{r},t)$. Defining the susceptibility of the ground state towards small changes $\chi\left[\rho_{\text{GS}}\right](\mathbf{r},\mathbf{r}',t-t')$ as

$$\chi\left[\rho_{\text{GS}}\right](\mathbf{r},\mathbf{r}',t-t') = \left.\frac{\delta\rho(\mathbf{r},t)}{\delta v_{\text{ext}}(\mathbf{r}',t')}\right|_{v_{\text{ext},0}} \tag{81}$$

yields a way to describe the changes in the electron density according to:

$$\delta\rho(\mathbf{r},t) = \int \mathrm{d}t' \int \mathrm{d}^3\mathbf{r}' \chi\left[\rho_{\text{GS}}\right](\mathbf{r},\mathbf{r}',t-t')\delta v_{\text{ext}}(\mathbf{r}',t') \tag{82}$$

Linking this linear response approach to the Kohn-Sham formalism with respect to the susceptibility of a Kohn-Sham system

$$\chi_{\text{KS}}\left[\rho_{\text{GS}}\right](\mathbf{r},\mathbf{r}',t-t') = \left.\frac{\delta\rho(\mathbf{r},t)}{\delta v_{\text{eff}}(\mathbf{r}',t')}\right|_{v_{\text{eff}}\left[\rho_{\text{GS}}\right]} \tag{83}$$

equation 82 can be reformulated according to

$$\delta\rho(\mathbf{r},t) = \int \mathrm{d}t' \int \mathrm{d}^3\mathbf{r}' \chi_{\text{KS}}\left[\rho_{\text{GS}}\right](\mathbf{r},\mathbf{r}',t-t')\delta v_{\text{eff}}(\mathbf{r}',t') \tag{84}$$

since both equations have to yield the same outcome. The change in the effective potential is defined as:

$$\delta v_{\text{eff}}(\mathbf{r},t) = \delta v_{\text{ext}}(\mathbf{r},t) + \int \mathrm{d}^3\mathbf{r}' \frac{\delta\rho(\mathbf{r}',t)}{|\mathbf{r}-\mathbf{r}'|} + \int \mathrm{d}t' \int \mathrm{d}^3\mathbf{r}' f_{\text{xc}}\left[\rho_{\text{GS}}\right](\mathbf{r},\mathbf{r}',t-t')\delta\rho(\mathbf{r}',t') \tag{85}$$

with

$$f_{\text{xc}}\left[\rho_{\text{GS}}\right](\mathbf{r},\mathbf{r}',t-t') = \left.\frac{\delta v_{\text{xc}}(\mathbf{r},t)}{\delta\rho(\mathbf{r}',t')}\right|_{\rho=\rho_{\text{GS}}} \tag{86}$$

The functional $f_{\mathrm{xc}}\left[\rho_{\mathrm{GS}}\right](\mathbf{r},\mathbf{r}',t-t')$ denotes the exchange-correlation kernel, which is a functional of the ground state density.

In the course of this thesis, TD-DFT is used to calculate UV/Vis absorption spectra. Every peak in these spectra accounts for the energy difference between two solutions of the time-dependent Kohn-Sham equation (equation 76). For the excitations, the same geometry of the ground and the excited state is assumed. Every excitation from the ground to an excited state incorporates an oscillator strength. The oscillator strength accounts for the absorption probability and thus for the intensity of the peak in the spectrum.

2.4 Docking

The theory of the docking and of the used docking software were taken from Ref. [142, 177–179].

In molecular docking an optimal binding mode between a small molecule (ligand) and a target is predicted. This prediction can either be done on a rigid level, where only a "rigid" rotation of both compounds is allowed, following the lock-and-key model or on a flexible level enabling conformational changes of the ligand as well as (partially) of the target. The latter case would refer to an induced-fit or conformational-selection model. Different binding modes are ranked with respect to a scoring function, which is based on the interaction between ligand and target. Other features like desolvation, entropy or accessible surface area of the solvent can also be included into the scoring function. There is a wide variety of software and scoring functions, which is too complex to be discussed in the course of this thesis. The following section will thus focus on the theory behind the applied software *AutoDock4* [179], which allows a flexible docking of both, ligand and target.

During the docking, different binding modes are evaluated with respect to a semi-empirical scoring function estimating the free energy of binding ΔG_{bind} according to:

$$\Delta G_{\mathrm{bind}} = (V^{\mathrm{L-L}}_{\mathrm{bound}} - V^{\mathrm{L-L}}_{\mathrm{unbound}}) + (V^{\mathrm{P-P}}_{\mathrm{bound}} - V^{\mathrm{P-P}}_{\mathrm{unbound}}) + (V^{\mathrm{P-L}}_{\mathrm{bound}} - V^{\mathrm{P-L}}_{\mathrm{unbound}} + \Delta S_{\mathrm{conf}}) \tag{87}$$

The first two parentheses of the scoring function take the energy difference of the bound and unbound ligand L as well as the protein P into account. The third parenthesis accounts for the interactions between the protein and the ligand including an entropic term that accounts for the conformational entropy loss, which is caused by the binding procedure. The unbound state for the ligand can be treated as either an extended conformation, the bound conformation or an optimized conformation. For the protein a reference structure, like a crystal structure, can be used, i.e. in the case of a rigid protein the bound and the unbound energy are equal and the term vanishes. For the protein-ligand interactions it is assumed that at a long distance, the interactions will decay towards zero.

The potential energy V for pairwise interactions is given by the following equation:

$$\begin{aligned} V = W_{\text{vdw}} \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + W_{\text{Hbond}} \sum_{i,j} E(\theta) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) \\ + W_{\text{elec}} \sum_{i,j} \frac{q_i q_j}{\epsilon(r_{ij}) r_{ij}} + W_{\text{solv}} \sum_{i,j} (S_i V_j + S_j V_i) \exp\left(-\frac{r_{ij}^2}{2\sigma^2} \right) \end{aligned} \tag{88}$$

All terms depend on the pair distance r_{ij} and a weight W that accounts for non-bonded interactions and effects including van der Waals interactions (vdw), hydrogen bonding (Hbond), electrostatic interactions (elec) and desolvation (solv). $E(\theta)$ is a function depending on the angle of the hydrogen bonding, $A_{ij}, B_{ij}, C_{ij}, D_{ij}$ are constants with respect to the potential. For the electrostatic interactions partial charges q_i are taken into account. In the desolvation potential S_i accounts for a solvation parameter, V_i for the spherical atomic volume, and σ for a distance weight. The conformational entropy loss due to binding is calculated based on the number of rotatable bonds $N_{torsions}$:

$$\Delta S_{\text{conf}} = W_{\text{conf}} N_{\text{torsions}} \tag{89}$$

For a fast computation of the scoring function a grid-based approach is applied. In this approach, the interesting part of the target is described by a fine grid. For each grid-point an energy for a probe with a certain feature, like hydrogen donor/hydrogen acceptor capability or atom type, is calculated. This is possible due to the additivity of the energy. This method enables the creation of some kind of dictionary for every grid point and a fast evaluation of different binding modes.

The generation of different docking poses for the ligand is performed by a Lamarckian genetic algorithm (LGA). In a genetic algorithm a random population of starting poses is generated, which is optimized with respect to a fitness score. In every generation the poses with the highest fitness are selected, modified and evaluated again. This procedure is repeated until a threshold criterion is met. In the modification step two different modifications are performed. In a first step some conformational features of random selected poses are mixed (crossover mutation). In a second step random mutations are introduced by adding a real number drawn from a Cauchy distribution to the conformational features. In LGA, for each generation a local search of a subset of the population is performed allowing to scan the local conformational space and increasing the fitness score for this subset.

3 Analysis of the Properties of the Water-Soluble Chlorophyll-binding Protein (WSCP)

Molecular dynamics (MD) simulations are a powerful tool for the investigation of dynamical properties of biomolecules. MD simulations can be applied, among other things, to predict the stability and to highlight flexible parts of the investigated system, or to sample the conformational space the biomolecule can adapt. Including small structural changes in the biomolecule, MD can additionally be used to observe changes in the dynamics induced by these constitutional modifications. By adding further analyses such as hydrogen bond network analysis, force distribution analysis or clustering, a deep insight into the dynamical behavior of the biomolecule can be achieved.

The tetrameric water-soluble chlorophyll binding protein (WSCP) is a quite remarkable protein. Although WSCP has not been observed to be participating in any photosynthetic process it contains up to four chlorophyll (Chl) molecules [180–182]. It further shows a high stability towards extreme temperature and extreme pH [182–184] as well as a higher photostability of the Chls compared to free Chls [185–188]. While its biological function is still unknown, there are several functions proposed. As the apo-WSCP is located in the endoplasmic reticulum [189, 190], it was assumed to be part of the pathogen recognition [187]. Another proposed function is the role as a scavenger for free Chls during chloroplast degradation [189].

The research presented in this work focuses on the finding of explanations for the high stability of WSCP and the bound Chls. The work is split into two parts. The first topic, described in section 3.1, focuses on the stability of WSCP using MD simulations. Within this context, a possible function of WSCP is discussed. Section 3.2 deals with the examination of the spectral properties of the bound Chls.

3.1 On the Stability of the Water-Soluble Chlorophyll-binding Protein (WSCP) Studied by Molecular Dynamics Simulations

In this section, the stability of WSCP is discussed based on MD simulations. WSCP consists of four non-covalently bound subunits, each capable of binding one Chl molecule. The overall structure forms a dimer of dimer structure where the porphyrine-rings of the Chls within one dimer are closer to each other than the porphyrine-rings of the other dimer. It was found that WSCP is stable if at least two Chls in one dimer are bound [182].

In this study, the stability of WSCP is estimated using structural properties, root-mean-squared deviation (RMSD), force distribution (FDA) and hydrogen bonding analysis. Additionally, the dynamical changes by removing Chl molecules are investigated. During the structural analysis of WSCP, two cysteines per protein subunit were observed to form a disulfide bridge within the subunit. In the MD simulation, these cysteines showed no strong conformational changes. Since the cysteines are located directly at the interface between two dimers, a reconnection of these cysteines between different subunits was tested in order to determine the effects on the dynamical properties. This was done under the assumption that the stability of the tetrameric structure is improved. The research was submitted to *J. Phys. Chem. B*.

On the stability of the water-soluble chlorophyll-binding protein (WSCP) studied by molecular dynamics simulations

Oliver Lemke and Jan P. Götze*

Freie Universität Berlin, Department of Chemistry and Biochemistry, Arnimallee 22, 14195 Berlin

E-mail: jgoetze@zedat.fu-berlin.de

Abstract

The water-soluble chlorophyll-binding Protein (WSCP) is assumed to be not part of the photosyntetic process. Applying molecular dynamics (MD) simulations we aimed to obtain insight into the exceptional stability of WSCP. We analyzed dynamical features such as the hydrogen bond network, flexibility and force distributions. The WSCP structure contains two cysteines at the interfaces of every protein chain, which are in close contact to the cysteines of the other dimer. We tested if a connection of these cysteines between different protein chains influences the dynamical behavior to investigate any influences on the thermal stability. We find that the hydrogen bond network is very stable regardless of the presence or absence of the hypothetical disulfide bridges and/or the chlorophyll units. Furthermore, it is found that the phytyl chains of the chlorophyll units are extremely flexible, much more than what is seen in crystal structures. Nonetheless, they seem to protect a photochemically active site of the chlorophylls over the complete simulation time. Finally, we also find that a cavity in the chlorophyll-surrounding sheath exists, which may allow access for individual small molecules to the core of WSCP.

1 Introduction

Molecular dynamics (MD) simulations have become a valuable tool in the last decades to study the dynamical properties of biological systems.[1–5] With the help of these simulations combined with other analysis techniques it is possible to make statements regarding conformational dynamics,[6–10] allostery,[11] protein misfolding[12] and many other relevant biological features. The investigated systems are quite diverse and can reach from small[13–15] to larger biomolecules[3,11,16] or even complexes consisting of multiple biomolecules and/or ligands.[17–22]

Using MD simulations the water-soluble chlorophyll-binding protein (WSCP) is analyzed in the course of this article. WSCP differs from other chlorophyll carriers such as reaction centres or antenna systems as it is located in the cytoplasm. It is not inserted into the thylakoid membrane and fully water-soluble.[23] WSCP was not detected to be involved in any photosynthetic processes.[24] Its function, however, is still unknown and there are several suggestions towards it, like the function as a chlorophyll scavenger,[25] as a signaler for pathogen attack[26] or as an agent involved in programmed cell death.[27,28]

The protein itself consists of a tetrameric structure, showing a pseudo-tetrahedral conformation. Every protein subunit contains one chlorophyll (Chl), which was detected by X-Ray crystal structure in 2006.[29] It was reported that WSCP can bind chlorophyll a (Chla), chlorophyll b (Chlb) as well as other chlorophyll derivatives. The ratio of selectivity is depending on the primary sequence of the WSCP.[30] Due to the overall 222 symmetry, WSCP can be described as a dimer of dimers, where the porphyrin rings within a dimer are in close contact.[29] Former studies[31] showed that if only one dimer is binding 2 Chls, the tetrameric WSCP complex is still stable towards dissociation. WSCP features additionally the capability for uptake of further Chls from thylakoid membranes[23,32] to saturate the dimer lacking Chls. Lacking all Chls, no stability was observed. It was proposed that the phytyl chains of the Chls stabilize the WSCP complex as they are forming a hydrophobic cavity in the protein.[29,33] Other studies showed that the phytyl chains are not needed for the stability at ambient, but at higher temperatures.[26]

The WSCP is formed under stressful conditions[34–37] and therefore was assumed to have a protective function.[38] It shows a high thermal[39,40] as well as a high pH stability.[31] Furthermore, the bound Chls possess a high resistance towards photobleaching although the Chls within a dimer form an excitonic couple[41] and no photoprotective carotenoids are present.[23,32] It is known that excited Chls in a triplet state can transfer their excitation energy towards oxygen, which leads to the formation of singlet oxygen or other reactive oxygen species (ROS).[42] With respect to the singlet oxygen production Schmidt *et al.*[23] reported that Chl bound to WSCP shows a reduced singlet oxygen formation, which might explain the higher stability. Based on the crystal structure and this observation it was assumed that the protein forms a diffusion barrier towards the interaction of oxygen and excited Chls.[38] In contrast, Agostini *et al.*[26] recorded that the singlet oxygen formation is equal as for free Chl and the phytyl chains of the Chls act as a shield for a photochemically active sites of the protein and therefore are the key for their photostability.

Most of the suggestions regarding the stability and properties of WSCP are based on crystal structures and not on the dynamical behavior, which could strongly influence these properties. To strengthen or refute the found characteristics the dynamical behavior has to be investigated as well. To our knowledge, this has not been done up to now and is an open field of research. In this work, we therefore focus on the analysis of the dynamical behavior of WSCP during molecular dynamics (MD) simulations at room temperature starting from the crystal structure.[29] Additionally, we aim to identify the important interactions and capture their dynamical behavior that can increase the stability of the WSCP complex. By removing Chls from the simulation setup, changes in the properties are investigated, which might be of importance to understand the uptake of Chls.

During the course of this study we discovered that on the interfaces between the dimers (later referred to as interfaces $\mathrm{I_{CA,CD}}$ and $\mathrm{I_{CB,CC}}$) 2 cysteines per subunit (C45 and C92) are present (Figure 1). In the crystal structure,[29] which was used as a starting structure for the simulation, the cysteines form a disulfide bridge within the subunit. Due to the close proximity of the cysteines it might be conceivable that a reconnection of these cysteines between different subunits is possible. Further, we found that at these interfaces hydrogen bonds are formed that are in close proximity to the cysteines. At this point the question arises if these disulfide bridges are needed to get the hydrogen bonds into contact. On the contrary, it might be that the hydrogen bonds allow the cysteines of the subunits to get close enough for a covalent connection between the subunits. An interesting fact is that only 2 cysteines are present in every subunit. Of these 2 cysteines per subunit, both are located at potentially dimer-connecting interface positions. Combining this observation with the hydrogen bonding of the neighboring amino acids, strengthens the hypothesis that a crosslinking of the cysteines between the dimers is possible. Furthermore, it might be a key for the high stability of WSCP. Utilizing MD simulations we investigate how a (hypothetical) crosslinking of these cysteines influences the properties examined in this article. The information gained by this analysis can on the one hand be used to check if this crosslinking distorts the quarternary structure of the WSCP and on the other hand be used to identify characteristics that are only present if the cysteines are crosslinked.

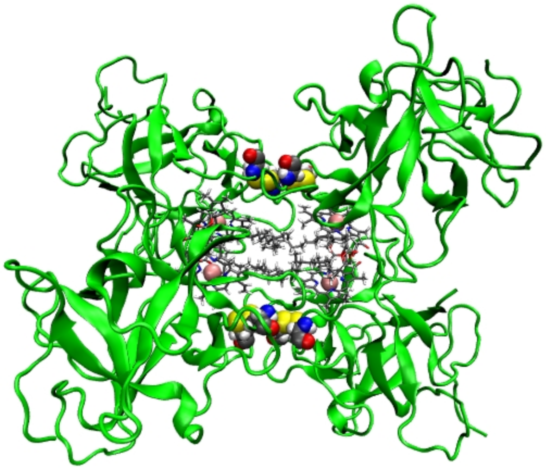


Figure 1: Tetrameric structure of WSCP with highlighted cysteines C45 and C92 at the interfaces $I_{CA,CD}$ and $I_{CB,CC}$.

This publication aims to characterize the dynamical properties of WSCP and investigate their changes when either Chls are removed to a certain amount or the tertiary structure is modified by crosslinking specific cysteine residues. In Section 2 the setup for the simulations and analyses are highlighted. Based on these simulations, several analyses with respect to dynamic behavior, flexibility and energetics were carried out (Section 3) and set into context to properties observed in experiments (Section 4). In Section 5 the observations are summarized.

2 Methods

MD Simulations

MD simulations were performed using the GROMACS 2016.1 simulation package.[43] All calculations were run in an NPT ensemble at T=300 K using leap-frog integration,[44] the CHARMM36 force field[45,46] and the TIP3P water model.[47] As a starting structure the PDB entry 2DRE[29] was used (protonated: N-terminus, Lys, Arg; deprotonated: C-terminus, Asp, Glu). Missing amino acids were added manually assuming 4 equal subunits followed by an energy minimization using the steepest decent algorithm (*emtol* = 100 kJ mol^{-1} nm^{-1}). The parameters for Chla were obtained from Ref. 48–52. After solvation and neutralization, adding 40 sodium ions, the system was energy minimized using a steepest decent algorithm, followed by an NVT and an NPT equilibration of 100 ps each. The temperature was kept constant using a v-rescale thermostat[53] (*ref_ t* = 300 K, *tau_ t* = 0.01 ps). The WSCP complex as well as the rest of the system (solvent and ions) were coupled to separate thermostats. For the NPT calculations a Parrinello-Rahman barostat[54] (*ref_ p* = 1.0 bar, *tau_ p* = 2 ps) was added as well. Bonds between hydrogen and heavy-atoms were constrained using the LINCS algorithm[55] (*lincs_ iter* = 1, *lincs_ order* = 4). Long range interactions were calculated using a Particle-Mesh-Ewald summation[56] (*pme_ order* = 4, *fourierspacing* = 0.16) and cut-offs (*rlist* = 1.2 nm, *rvdw* = 1.2 nm, *rcoulomb* = 1.2 nm) with a force-switch (*rvdw* = 1.0 nm). After equilibration, each investigated system was simulated with a time step of 2 fs, saving system's coordinates every 10 ps. The described setup was used for all simulations if not stated differently. For all simulations of complexes, which differ from the original 4 Chls with 4 protein chain, modifications such as removal of Chls or reconnection of bonds were performed manually on an equilibrated structure of the holoprotein and solvated again. All simulations with no crosslinking of the cysteines were started from the same equilibrated structure. For the simulations with a crosslinking of the disulfide bridges, a snapshot taken at 150 ns of the simulation containing 4 Chls was used as a starting structure.

In total 5 different systems were simulated containing 4, 2 or 0 Chls. In the case of 2 Chls, the 2 Chls of one dimer were kept, whereas the Chls of the other dimer were removed. For the formation of disulfide bridges, which will be discussed in Section 3, calculations for 4 and 2 Chls were performed. The 8 symmetric disulfide bridges were formed manually (using GROMACS trjconv -ss flag) and energy minimized to obtain a starting structure for the simulation. The simulation length, modifications as well as the later used labels are summarized in Table 1.

Table 1: Labeling of the different simulation setups with respect to the number of Chls, structural modifications and the total simulation length; †For the system Nat4_Pulled 3 simulations with 300 ns each were performed.

Label	#Chl	Structure modification	Duration/ns
Nat4	4	–	450
Nat2	2	–	300
Nat0	0	–	300
Sulf4	4	Formation of disulfide bridges between the dimers	300
Sulf2	2		200
Nat4_P_all	4	Subunits pulled apart by 1 nm	900†
Nat4_P_protein	4	Protein chains pulled apart by 1 nm	300
Nat_Chl	4	Removal of the protein chains	100
Nat_Chl_P	4	Removal of the protein starting from Nat4_P_all	100

To study the dynamical behavior of the monomers as well as the tetramer formation further simulations were performed using the same simulation setup as described above. In these simulations, the subunits were pulled apart along the axes of a tetrahedron by 1 nm each for further simulations. For this, the coordinates of the protein chains and the Chls were modified manually to obtain starting structures. A detailed description for the pulling is given in the supporting information. Different simulations were performed where either complete subunits or only the protein chains were pulled apart. Two additional simulations were performed that contained no apoprotein and thus only Chls; one with the Chls in their original conformation and another with the Chls pulled apart as described above saving the coordinates every 1 ps. All further information are reported in Table 1.

Analysis

RMSD calculations were done using the GROMACS 2016.1 simulation package with a reference structure taken at $t = 1$ ns from each simulations. RMSDs were averaged over 1ns and plotted with their corresponding minimal and maximal value every nanosecond. Force distribution analysis was performed for the simulation Nat2 using the *gmx_fda* package.[57] For the analysis the non-bonded forces (Coulomb, Lennard-Jones (LJ)) were calculated and summed for different groups. These groups were defined as subunits of the protein (CA, CB, CC, CD) with coordinated Chls for the dimer formed by CA and CB. Non-binding energies were calculated using the GROMACS 2016.1 simulation package for the same classes. Additionally, the energy between the Chl of subunit CA and CC as well as their counterparts were taken into account. The Coulomb and LJ energy values were added up and averaged over the complete trajectory.

Hydrogen bond analysis was performed using the GROMACS 2016.1 simulation package with an angle cutoff of 30° and a distance cutoff of 0.35 nm. All hydrogen bonds within a subunit were neglected and only hydrogen bonds between different subunits or the protein and Chl were taken into account. In addition a overall population of 10 % was set as a minimal threshold to characterize the significant hydrogen bonds. In the analysis hydrogen bonds between equivalent parts of the protein were averaged with the counter part of the other dimer, assuming a symmetric structure.

3 Results

3.1 Benchmark and Dynamical behavior

In a first step the dynamical properties of the simulated systems were compared with static reference values (theoretical and experimental). Beside the systems containing 4, 2 or 0

Chls, 2 further simulations are included in the analysis with a modification of the tertiary structure. As already described in section 1, we tested how a (hypothetical) reconnection of the cysteines, covalently linking the dimers, influences the dynamical properties.

The investigated properties are reported in Table 2. These features include the angle of the Mg ion with respect to the porphyrin ring, the angle between the porphyrin rings (denoted as plane) within the dimer or between different dimers as well as the distance between the Mg ions. Horigome *et al.*[29] reported that the coordination of the Mg ion is essential for the binding towards the protein. Therefore, the distance between the coordinating P36 backbone oxygen and the Mg ion was tracked. Palm *et al.*[30] proposed that the phytyl chains block a photochemically active site of the Chl. Due to this, the shortest distance between the photochemically active methine-20 and the phytyl chains is reported as well; numbering of Chl is depicted in Figure S1.

Table 2: Dynamic chlorophyll properties (mean value ± standard deviation) of selected systems. Plane-Plane and Mg-Mg properties between chlorophylls of the same (s) or the opposite (o) dimer are reported. The angle between the planes was calculated using the plane's perpendicular. For the distance between the methine-20 atom of the prophyrin system, which was reported to be a photochemically active site of the Chl,[30] the shortest distance between the methine-20 carbon and any atom of the phytyl chains of all Chls is reported.

Label		Nat4	Disulf4	Nat2	Disulf2	Ref.
Angle (Mg - Plane)/°		5.8 ± 2.4	4.8 ± 2.4	6.3 ± 1.9	4.2 ± 2.3	11.7[29]
Angle (Plane - Plane)/°	s	34.7 ± 5.4	38.0 ± 7.0	30.3 ± 4.5	38.3 ± 4.8	30[58]
	o	98.1 ± 5.6	98.8 ± 7.3	–	–	–
Distance (Mg - Mg)/Å	s	10.3 ± 0.6	10.5 ± 0.6	10.1 ± 0.4	10.3 ± 0.6	10[26]
	o	20.2 ± 1.0	19.4 ± 1.3	–	–	20[26]
Distance (Mg - P36)/Å		2.3 ± 0.3	2.5 ± 0.4	2.2 ± 0.1	2.5 ± 0.4	2.1[29]
Distance (Phythyl - C20)/Å		4.3 ± 0.6	4.2 ± 0.6	4.2 ± 0.4	4.3 ± 0.7	4.0[29]

Analyzing the dynamical properties it can be observed that the Mg ion is permanently coordinated by P36, with a slightly increased average distance compared to the reference values. In addition, it can be seen that a phytyl chain is always in close contact to the methine-20 carbon and shows only a small fluctuation. The distance between the Mg ions matches the reference values as well. From Table 2 three notable observations can be made. Firstly, the angle of the Mg ion with respect to the porphyrin plane does not match the reference value. This is an artifact of the force field as the atoms are not treated on an electronic level, confirmed by preliminary QM/MM calculations (data not shown, optimization of a snapshot taken at 1 ns: 8.5 → 9.9 °). Due to the tight packing and the presence of the protein, acting as a cage for the Chls, only the magnesium position seems to be affected by this artifact. To confirm this, the distance of the magnesium to the center of the porphyrine plane was calculated as well for Nat4 with 0.24 ± 0.04 Å(Reference:[29] 0.49 Å). The difference of 0.2–0.3 Å is counteracted by the enlarged distance between Mg and P36, which can be considered a fortuitous error cancellation.

Secondly, removing 2 Chls (Nat2) reduces the fluctuation of the angle between the prophyrine rings within a dimer. This may be explained as the binding cavity containing only 2 Chls is not as packed compared to the presence of 4 Chls. As a consequence, the porphyrin rings may more likely form favorable interactions that stabilize the conformation. This can also be observed for the P36-Mg distance. Lastly, we can see that introducing disulfide bridges (Disulf4 and Disulf2) has no strong influence on the observed properties with exception to the angle between different porphyrins. The explanation for this can be found in a small conformational change of the protein by reconnecting the cysteines. However, as for all other properties the changes in the protein conformation are surprisingly small.

For the coordination of the phytyl chains towards the methine-20 carbon of the Chls, we also examined to which Chl the phytyl chain belonged. In roughly 22 % of all cases the phytyl chains belonged to the same molecule, in roughly 74 % to the same dimer. For Nat4 in roughly 6 % of all frames a coordination of a phytyl chain of the other dimer was observed. In some cases the phytyl chain of one Chl was able to coordinate towards both methine-20 within the dimer. This was possible due to the methyl groups of the phytyl chain and can also be observed in the crystal structure of a WSCP mutant.[30] This indicates a strong variance of the phytyl chain structure over time.

All investigated systems remained in their tetrameric form during the complete simulation. For Nat0, this is not in agreement with the experimental observations. The reason for this can be either found in an over-stabilization of the protein interactions induced by the force field or an dissociation time scale beyond >300 ns. In the latter case it might be as well a consequence of the chosen starting structure. As the starting structure was taken from the simulation of the system containing 4 Chls, all hydrogen bonds, hydrophobic and ionic contacts were already formed stabilizing the molecule. Starting from 4 monomers it might be very improbable to form these kind of interactions spontaneously. Different simulations with separated monomers either with bound or unbound Chl were performed, see Section 3.5.

3.2 Effect of Chl presence and Cys-Cys linking

We continued to investigate the effect of Chls and/or disulfide bridges on the structure. The first investigated observable is the root-mean-squared-deviation (RMSD). The mean RMSD for every nanosecond (100 frames) and the fluctuation (minimal and maximal RMSD per nanosecond) was analyzed (Figure 2). In the former section we observed that the tetramer was stable over the full 300 ns that were simulated. Hence, all changes in the RMSD are caused by conformational changes and not by dissociation. For Nat4 a constant RMSD with only a small fluctuation is observed, which indicates the high stability of the holoprotein. As the starting structures of all other simulations were taken from Nat4, they can be compared directly.

Removing 2 Chls (Nat2) increases the RMSD after roughly 40 ns. This observed time, however, is only based on one simulation. As no reversibility is observed during the simu-

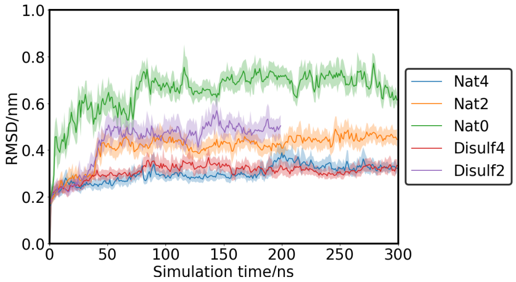


Figure 2: Mean RMSD over 1 ns for all investigated simulations. The error bars depict the minimal and maximal RMSD value within an interval of 1 ns.

lation this time is not significant and can only be seen as an indicator. The increase in the RMSD is caused by a movement of the monomers not containing any Chl and might be a requirement for the uptake of 2 further Chls. Connecting the disulfide bridges (Disulf4 and Disulf2) between different dimers yields a comparable behavior with respect to Nat4 and Nat2. Reconnecting the cysteines does not affect the flexibility.

After removal of the last 2 Chls (Nat0) we observed that this fluctuation gets larger and faster. After a few ns, an RMSD increase compared to Nat2 and Nat4 is observed, which shows a much stronger fluctuation within 1 ns compared to Nat4. In addition, a larger overall fluctuation over the complete simulation is observed compared to the other systems.

3.3 Hydrogen Bond Network

In a next step the hydrogen bond network was analyzed to find an explanation for the behavior of the RMSD and to investigate the conformational changes of the systems Nat2 and Disulf2. For the analysis only hydrogen bonds between either different subunits or between protein and Chl were analyzed. Using a threshold of at least 10 % occurrence over the whole simulation, the high number of hydrogen bonds can be reduced to the most important interactions (Figure 3).

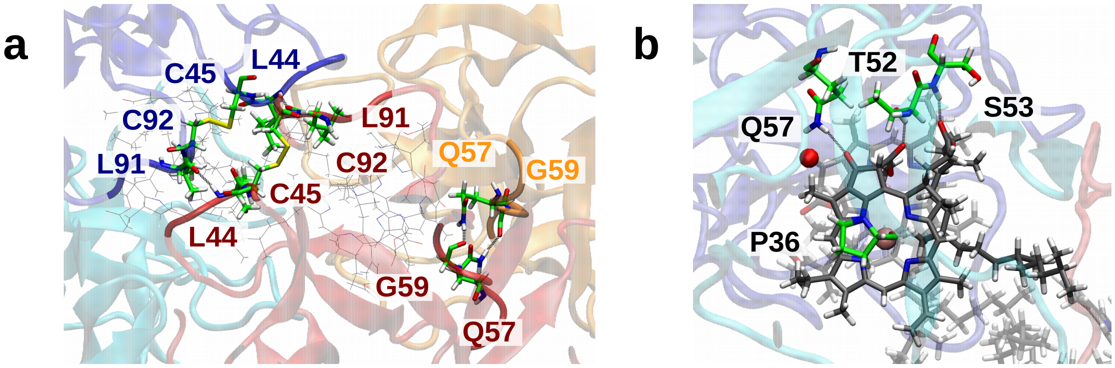


Figure 3: Dominant hydrogen bond donors and acceptors; (a) between the protein chains CA (blue), CB (cyan), CC (orange) and CD (red), highlighting the interfaces $I_{CC,CD}$ and the interfaces $I_{CA,CD}$. The amino acid labels are colored according to their parent protein chain. Note, although only 2 interfaces are presented a symmetric behavior is observed for the interfaces $I_{CA,CB}$ and $I_{CB,CC}$; (b) between the protein and Chl highlighting the coordinating amino acids. Q57 can either coordinate towards the same-chain Chl (as shown) or towards the other Chl within the same dimer (alternative coordination site, oxygen, depicted as a red sphere).

For the interactions between the 2 subunits within a dimer, symmetric hydrogen bonding between Q57 and G59 was observed with a population of 40-60 %. Q57 seems to play an important role for the overall stability as it can also coordinate towards both Chls within the dimer. Horigome *et al.*[29] already described the bonding of Q57 towards the Chl, however, only to the Chl of the other subunit. In the MD simulation a more complex behavior of Q57 could be observed (Figure 4).

Further hydrogen bonding between the protein and Chl was observed by the amino acids T52 and S53 showing a high population over the complete trajectory. Remarkably, S53 can coordinate with both, the backbone as well as the side chain, towards the porphyrin ring of Chl.

For the interaction between the dimers at the interfaces $I_{CA,CC}$ and $I_{CB,CD}$, no hydrogen bonding matching the threshold of 10 % occurrence was observed. These interfaces are formed by hydrophobic residues (L41, L153 and W154). In addition, the area between these

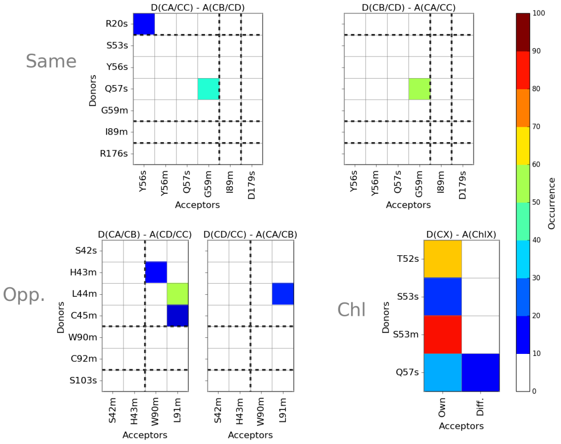


Figure 4: Hydrogen bonds of simulation Nat4 between donor 'D' and acceptor 'A' with an occurrence larger than 10 % within the same dimer, between different dimers and between the protein and Chl; Coordination via main chain 'm' or side chain 's'. Labelling D(X/Y)-A(X'/Y') reads as follows: donor X towards acceptor X' and donor Y towards acceptor Y'.

subunits is by a factor of three smaller compared to the interfaces $I_{CA,CD}$ and $I_{CB,CC}$.[29] At the latter interfaces, hydrogen bonding between L44 towards L91 is present with an occurrence of 50-60 %. Both amino acids are directly neighboring a cysteine (C45 or C92). This close proximity at the interface strengthens the hypothesis for a covalent linking of the dimers by reconnecting the cysteines.

Forming all possible inter-dimer cysteine bridges (CA-CD and CB-CC) results only in a small change for the overall dominant hydrogen bond network (Figure S2). In this situation, hydrogen bonds between the newly linked interfaces ($I_{CA,CD}$ and $I_{CB,CC}$) are strengthened. The hydrogen bond between Q57 and G59 becomes weaker, however, a stronger coordination of Q57 towards both Chls is observed. In addition, a high stability between the monomers is achieved by the interactions of the Chls, which will be discussed in detail in the next section.

Removing 2 Chls (Nat2, Disulf2) in one dimer weakens the interaction at the interface $I_{CC,CD}$ (Figures S3 and S4). As a consequence, a small opening at the interface $I_{CC,CD}$ is possible as no stabilizing Chls are present anymore. This results in a conformational change as observed in the RMSD. The interfaces $\mathrm{I}_{\mathrm{CA,CD}}$ and $\mathrm{I}_{\mathrm{CB,CC}}$ show only small changes in the population. Hence, a reconnection of the cysteines at these interfaces does not affect the conformational flexibility as already observed in the RMSDs. Removing the last 2 Chls (Nat0) results in a change of the overall hydrogen bond network (Figure S5). Although some hydrogen bonds stay in contact, a change over time as observed in the RMSD is detected which also affects the hydrogen bond network. Further discussion is omitted as the stability of the starting structure is likely artificial due to the pre-formed hydrogen bond network.

3.4 Non-bonded Forces and Energies

To get a better understanding of the protein-protein and protein-Chl interactions, non-bonded energies (electrostatics and van-der-Waals interactions) as well as the force distribution during the simulation (Figures 5 and S6) were analyzed. The simulated system Nat2 consists of one dimer containing Chls and one dimer lacking Chls, and was thus chosen as the reference system. Here, the interaction within a dimer as well as the interaction between the two differently constituted dimers can be analyzed. As reference points, the systems Nat4 and Nat0 were analyzed, too, but will only partly be included in the discussion. The force distribution analysis as well as the non-bonded energies for all three analyzed systems are reported in the supporting information (Figures S8 and S9 as well as Table S1).

The force distributions and non-bonded energies between the two dimers and subunits of different dimers were examined in a first step. For the interactions between the dimers, a Gaussian distribution with a mean close to 0 nN was observed pointing towards equilibrium dynamics. This was also observed for the systems Nat0 and Nat4 with a small shift to repulsive forces for Nat4, most probably induced by the tight packing of the Chls. A remarkable behavior can be found analyzing the force distribution of Nat0. For every combination of subunits different distributions are obtained pointing towards a distortion of the tetrameric

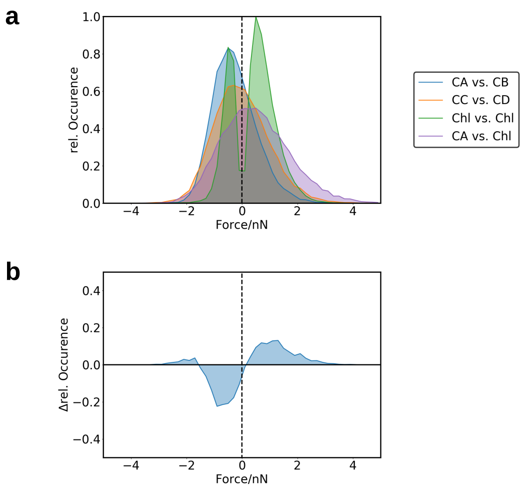


Figure 5: Force distribution analysis between different parts of the protein (Nat2), where the subunits are denoted as CA to CD, with CA and CB belonging to the Chl-containing and CC and CD to the Chl-deficient dimer; In (a) only the discussed interactions are shown. A full depiction can be found in the SI (Figure S6); in (b) the difference between the dimer CC+CD (orange in (a)) and the Chl-containing dimer CA+CB (blue in (a)) is depicted.

structure. This is in agreement with changes in the hydrogen bond network and the increased RMSD. Analyzing the non-bonded energies an attractive energy of -113.2 ± 12.0 kJ/mol is obtained for the system Nat2 for the interaction of the protein chains of different dimers, which is comparable to both Nat0 and Nat4.

For the interactions of the Chls towards the protein chains of the Chl-free dimer, we observed only a small amount of attractive non-bonded energies (-20.5 ± 5.5 kJ/mol), most likely due to the phytyl chains interacting with the hydrophobic interfaces ($I_{CA,CC}$ and $I_{CB,CD}$). This is in agreement to the values obtained for Nat4. For the system Nat4 the interaction between Chls located in different dimers was examined as well. In the force distribution analysis only weak forces are observed with a mean slightly shifted to attractive forces. This weak interaction can also be observed analyzing the non-bonded energies between the Chls. For the interactions between the Chls of different dimers a non-bonded energy of -14.5 ± 2.8 kJ/mol is observed, which is weaker than all other analyzed energies.

In a next step the interactions within the dimers were analyzed. For the protein chains within the Chl-free dimer (Figure 5a, orange), more attractive than repulsive forces are observed over time. The amount and strength of these attractive forces increases even more for the Chl-containing dimer (Figure 5a+b, blue) resulting in a higher stability compared to the dimer not containing any Chl. Hence, the close packing of the Chl affects the interactions and conformations within a dimer and thus generates stronger attractive forces. This can also be found in the case of Nat4 where a comparable force distribution is observed.

Comparing the non-bonded energies, a slightly stronger interaction between the protein chains belonging to the Chl-free dimer with -236.7 ± 12.0 kJ/mol is observed, compared to -174.8 ± 32.8 kJ/mol between the protein chains of the Chl-containing dimer. However, it has to be kept in mind that the Chl-containing dimer is further stabilized by the interactions between the Chls (-136.3 ± 9.8 kJ/mol) and the interaction between the Chl and the other protein chain of the Chl-containing dimer (-30.3 ± 9.5 kJ/mol) i.e. Chl of CA interacts with CB and Chl of CB interacts with CA. For Nat4 an energy of -224.5 ± 37.0 kJ/mol is observed, which is larger as for Nat2. Remarkably, for the system Nat0 an even stronger non-bonded energy of -311.0 ± 68.0 kJ/mol was observed. This increase is caused by a rearrangement of the protein chains optimizing the non-bonded interactions as observed in the force distribution analysis and the change of the hydrogen bond network. This is possible as no Chls are present that can prevent this rearrangement. However, it may very well be that such an arrangement would never come to be *in vivo*, as the protein chains would likely not come together without Chls bound in the first place.

The force distribution between the Chls within a dimer (Figure 5a, green) shows a behavior differing from the other cases as it does not display a Gaussian-like shape. For the interaction of the Chls, a dual-peak distribution is obtained with a drop in density at 0 nN. This is observed for both systems, Nat2 and Nat4. Hence, a breathing motion due to compact packing of the Chls in the center of the WSCP can be assumed. Analyzing the time series of the forces (SI, Figure S7), a fast change between attractive and repulsive forces can be observed. This strong breathing motion can be seen as an indicator for the stability caused by the close packaging of the Chl. If the distance between the Chls gets too large attractive forces cause the Chls to move together again.

In a last step, the forces within a subunit are investigated. For the forces within a subunit, a weakly repulsive mean is observed between the protein and the Chl (Figure 5, violet). The protein-Chl non-bonded energy is, however, with -355.6 ± 15.7 kJ/mol the highest among all analyzed non-bonded energies causing the protein and Chl to stick close together. This behavior is also observed for analyzing the stability of this contact using MD-simulations, which will be discussed in the next section of in this article.

3.5 Formation of Tetrameric Structure

To investigate the formation of the tetrameric structure, the behavior of the mostly hydrophobic Chls without the protein was examined. In a first simulation the protein was removed and the Chl tetramer was simulated in explicit water (Nat_Chl). In a second simulation the Chls were separated along the tetrahedron axis by 1 nm in the beginning (Nat_Chl_P) as described in section 2.

Nat_Chl showed no dissociation during the simulation time, likely due to the mostly hydrophobic character of the Chls. Using Common-Nearest-Neighbor-Clustering[7,59–61] based on the internal coordinates of the Mg ions towards the center of mass (see SI), it was possible to characterize different conformations. The most prominent conformation was a weakly distorted tetrahedron-like conformation with the Mg ions on the vertices (Figure 6). The phytyl-chains are clustered inside the tetrahedron. Thus, a micelle-like structure can be assumed. The distortion arises from the phytyl chains not fitting perfectly inside the cavity. This structure is different from the one in the WSCP complex; it minimizes the solvent accessible surface of the molecule. The Mg ions are pointing to the outside. Hence, it is possible for other molecules to coordinate. In addition, other hydrogen bond acceptor are present on the porphyrin ring that can coordinate towards the solvent.

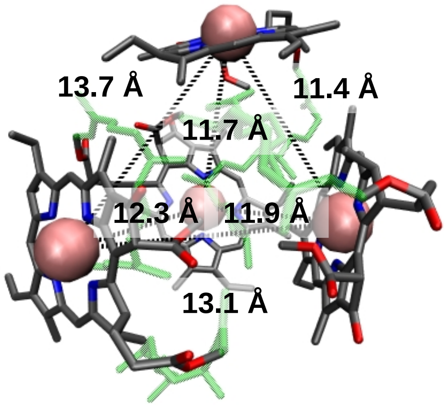


Figure 6: Most dominant conformation of isolated Chls in explicit water; the Mg ions as well as their distorted tetrahedral arrangement are highlighted.

In Nat_Chl_P, the phytyl chains may likely act as a scanner for other phytyl chains, due to their strongly hydrophobic character. The first dimer was formed within 1 ns and the tetramer after 3 ns. Afterwards, as observed in the former simulation no separation of the Chls was observed.

When repeating this kind of analysis while keeping the protein present, a notable behavior is observed. If the protein chains are pulled apart and the Chls kept in place (Nat4_P_protein), all protein chains coordinate towards the Chls after 150 ns. However, the native coordination of P36 towards the Mg ion is not observed. In contrast, when separating the subunits including Chls (Nat4_P_all), the Chl coordination pattern is not affected. In Nat4_P_all, only a dimer formation can be observed (SI, Figure S10) where the phytyl chains again can act as a scanner towards other hydrophobic molecules. This dimer lacks at one side an open spot towards the Chls and needs to be shielded from the solvent. In the Nat4_P_all simulation, a formation of trimeric structures is also found, which is not stabilized over a long period. Based on these observations, we can assume that in a first step a dimer has to be formed containing Chls. In a next step either another dimer of the same constitution has to associate, forming a homo-tetrameric structure or two Chl-deficient subunits attach, shielding the hydrophobic parts of the dimerized Chls. However, both described scenarios were not observed within the simulated amount of time and it is therefore still unclear how the tetrameric structure is formed.

4 Discussion

Disulfide bridges connecting different subunits

The MD simulations showed that the cysteines C45 and C92 are extremely well suited to be covalently linked towards the subunit of another dimer, which should increase the thermal stability of the WSCP further as already observed for other proteins.[62] In addition it was already reported that the apo-protein of the WSCP is located in the endoplasmic reticulum[25,63] where disulfide bond formation occurs[64] catalyzed by the protein disulfide isomerases.[65,66] We could also show that forming the disulfide bridges induces no large conformational changes. In addition, we highlighted that the structure shows no loss in flexibility despite the overall increase in stability, which may allow for unmitigated chance of Chl uptake even if a Chl dimer is already bound.

Palm *et al.*[67] showed that the *Lepidium virginicum* WSCP favors binding of Chlb over Chla. They proposed that a hydrogen bond between the amide of L91 (directly neighbouring C92) and a Chlb-specific formyl-group at the Chlb carbon C7' (see Figure S1), is one key factor to favor Chlb over Chla. Connecting the cysteines causes a small conformational change of the protein. However, the contact to the amide of L91 is not lost. Although we did not simulate Chlb, we expect only minute differences as Chlb and Chla are identical except for the aforementioned formyl-group. The C7' of the Chl shows roughly a distance of 3 - 4 Å to the amide of L91. Hence, it is still in the range for hydrogen bonding. In addition, a water molecule is observed close to the L91 amide, serve as a bridge between Chl and L91.

Other possible interactions that were not analyzed

Another attractive interaction of Chls with their respective subunits are π-stacking interactions by W90 and W154. This interaction is based on induced dipole moments, which are not properly accounted for using the regular CHARMM force field. As such, we will omit the corresponding discussion here.

Singlet Oxygen Production

Agostini *et al.* proposed recently[26] that WSCP shows a $^1O_2$ production comparable to free Chls and that the phytyl chains are involved in the photoprotection of the Chls. During the MD-simulations it could be observed that it was possible for a water molecule to diffuse into the hydrophobic cavity of the WSCP. The diffusion occurred via a pore (L41, S42, I89, L91, W154) that was already described by Horigome *et al.*[29]

Up to now it was considered that this pore was blocked by the Chls in the cavity. However, due to the dynamic movement and the breathing motion of the Chls it was possible for water to move into the pocket (Figure 7). The water molecule remained in the pocket during the course of our simulations. This is either an artifact of the force field as the water molecule was permanently coordinating the Mg ion of one Chl, or an indicator for the rare event of the diffusion into the hydrophobic cavity of the WSCP. Although the diffusion of water itself can also be an artifact of the force field, it can not be excluded that other small molecules like oxygen can diffuse into the cavity. For molecular oxygen, the likelihood to diffuse into the pocket is probably higher than that of water, as the electrostatic interaction with the solvent is much weaker.

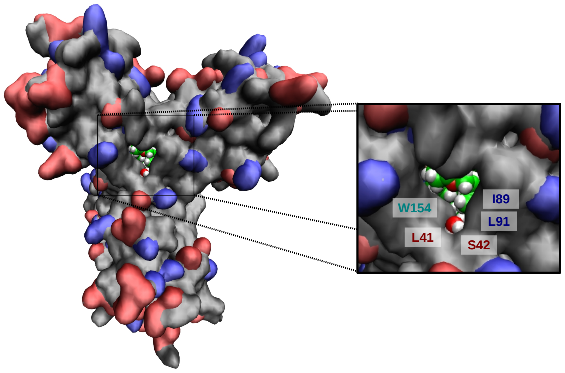


Figure 7: Surface plot of the WSCP highlighting the pore for which a diffusion of a water molecule towards the inside of the cavity was observed. Red areas denote acidic, blue areas denote basic residues, The Chls are shown as green spheres. Amino acid labels colored according to their parent protein chain, see Figure 3.

Protein Charge

The protein-Chl complex shows a charge of -10 per subunit at neutral pH, which was assumed in the simulation, resulting in a total charge of -40. Most of the charged residues are located on the surface of the complex and are most likely involved in the solubility[68] of the molecule. Apparently, only the C-terminal amino acids R176, E177 and D179 are involved in stabilizing the dimer at the interface of the subunits. A charge-charge interaction on the interfaces towards other dimers is not observed in the simulation.

Another reason for the high number of charged residues might be the β-barrel-like tertiary structure arising from the supersecondary structure, stabilizing the subunits. However, the quite high negative total charge of the protein is remarkable and might, beside the increased solubility, be important for the function of the WSCP. Taking into account the highly negative surface of membranes one possibility might be a protection of the cell nucleus and therefore of the DNA with respect to oxidative damage by singlet oxygen, due to repulsive interactions. However, this assumption seems to contradict the extraction of Chls from thylakoid membranes as described by Satoh *et al.*[32]

Considering that most of the charged residues are on the surface of the tetrameric structure, which is not stable without Chl, we have to investigate the monomeric structure. Here, only one positively charged residue (R51) is present in a 1 nm distance to the potential Mg binding site. In a larger radius of 1.5 nm, we find a total of 2 positively and 2 negatively charged residues. Hence, we find no evidence which contradicts the possibility of WSCP extracting Chls from membranes, as the lack of negative charges towards the Chl-binding site would indeed allow for a weak WSCP/membrane interaction. Upon formation of the tetramer, the less charged site becomes buried inside to complex. Yet, it was shown that WSCP may be stable with only 2 Chls present, while retaining the ability for uptake of more Chl molecules.[31] Due to the unknown binding mode towards the membrane and due to the high computational costs associated with corresponding simulations, however, we did not pursue this line of thought further.

Quarternary Structure Formation

Analyzing the formation of the quarternary structure, we could observe that the Chls are needed at least in a dimeric form for the stabilization of the protein Chl complex. In this dimeric form further protein chains are required to shield the open hydrophobic site. However, we were unable to simulate a complete regeneration of the native coordination pattern or complex formation. For determining the mechanism of the tetramer formation, the uptake of the Chls may be the key. A potential mechanism could involve pre-formation of Chl aggregates, such as the presented tetramer (Figure 6), either in solution or in the thylakoid membranes. These aggregates might then be bound by WSCP. Another possibility would be that a small amount of micelle-like Chl agglomerates can leave the membrane and WSCP acts as a scavenger. As WSCP is formed under stress conditions,[34–37] this Chl scavenging function might either preserve the Chl compounds or protect the cell from harmful photoproducts.

Another open question is how the Chls are released from the protein, as it is highly stable towards pH and temperature changes.[31,39,40] One possibility might be the presence of another protein acting as a extraction agent of the Chls. Another possibility would be an injection of the Chls back into the thylakoid membranes. However, as long as the function of WSCP is unknown this questions is hard to answer. As a tentative prediction, if our scavenger role is indeed correct, older or stressed cells may accumulate WSCP-bound Chl as they seek to protect themselves from the remains of damaged chloroplasts.

Stability of the Chl-deficient system

Studying the stability of the protein lacking Chls, we observed a stable protein. The reason for this might be the starting structure (equilibrated crystal structure with removed Chls) as all attractive interactions are already in place. Further, the lack of Chls leads to a reduced steric hindrance, leading to a small rearrangement of the subunits. As a result, the interactions are optimized and the initial structure becomes more stable. An indicator for this process is found by analysis of the non-bonded protein-protein interactions of Chl-deficient dimers. In addition a distortion of the tetrameric structure could be observed in the force distribution analysis. Due to these strong non-bonding energies, long dissociation times can be expected, which were not achieved within the simulated amount of time. Without the phytyl-chain interactions of Chl units bound to their apoprotein subunits, the tetrameric structure might not be formed at all. Combined with the simulations of separated Chl units we can assume that the Chls are needed to bring these interactions into the right contact as every subunit has to match the perfect binding position with respect to the other subunits. Additionally, it shows a larger fluctuation compared to the WSCP containing 4 Chls, which is an indicator for a reduced stability.

5 Conclusion

Using MD simulations of WSCP complexes with different numbers of Chls as well as tertiary structure modification, it was possible to point out the main contributors stabilizing the WSCP-Chl complex. The binding of the Chl towards the Mg-coordinating subunit was the strongest non-bonded energy contribution observed in the analysis. This binding is supported by the coordination of P36 towards the Mg ion of the Chl as well as hydrogen bonding by the amino acids T52, S53 and Q57. The phytyl chains of the Chls in the protein are intertwined forming a hydrophobic cavity at the center of the protein. In addition, they are always in close contact to the photochemically active site of the Chls, which strengthens a photoprotective role. Analyzing the force distribution for the Chl interaction a strong breathing motion was observed, which is assumed to come into play because of the close packing of the hydrophobic cavity.

At the protein interfaces within a dimer ($I_{CA,CB}$ and $I_{CC,CD}$), charge-charge interactions between R176, E177 and D179 were observed. However, as at least one pair of Chl has to be present for a stable complex, this interaction between the interfaces seems either not strong enough to avoid dissociation or it is not formed if no Chls are present as assumed by pulling simulations. In addition, it was observed that the presence of Chls results in more and stronger attractive forces between the protein chains within a dimer, stabilizing the complex further. Another stabilizing factor comes into play with amino acid Q57, as it shows the capability to coordinate between the protein chains and additionally towards both Chls within the dimer. If a subunit does not contain Chls, a higher flexibility was observed in the simulation pointing towards the possibility of uptake of 2 Chls.

At the interfaces between different dimers either hydrophobic interactions formed by L41, L153, W154 and the phytyl chains ($I_{CA,CC}$ and $I_{CB,CD}$) or strong hydrogen bonding ($I_{CA,CD}$ and $I_{CB,CC}$) were observed. For the latter interfaces, bidirectional hydrogen bonding between L44 and L91 backbones were observed, which are directly neighboring C45 and C92. We showed that it is hypothetical possible to reconnect the cysteines between different protein chains resulting in no strong effect on the overall dynamical properties. However, they can be expected to greatly increase WSCP stability via means not captured by our simulations, as the crosslinking leads to two covalent disulfide bonds between the subunits which stabilize the complex already by themselves. As such, they do not necessarily need to have an impact on the hydrogen bonding or the interaction energies analyzed in this study.

For all analyzed systems, simulation times of 200–450 ns were achieved. In this time it was possible to characterize the dynamic behvaior of different contributors to the stability of WSCP. For upcoming studies the conformational space of the WSCP can be investigated further. For this, much longer simulation times are needed. Due to the high stability and size of the WSCP-complex also other techniques than classical all-atomic MD simulations should be considered such as coarse-graining or enhanced sampling techniques. The latter one can additionally be applied to generate starting structures for classical MD-simulations to obtain an insight into the confromational space of the WSCP. Another interesting field of research are the dynamical changes in the optical properties of WSCP analyzing the photochemical protection in more detail on a quantum mechanical level. This research is part of an upcoming publication.

Acknowledgement

OL acknowledges support by the DFG collaborative research center CRC 765. JPG is currently funded by the Deutsche Forschungsgemeinschaft, project no. 393271229. We appreciate support and resources provided by Prof. Dr. Bettina Keller, as well as Dr. habil. Heiko Lokstein for reading and commenting on the manuscript. Further, we thank the North-German Supercomputing Alliance (HLRN) for providing computational resources that were used for the MD simulations.

Supporting Information Available

In the supporting information additional figures are included that are discussed in this publication. In addition, a short discussion on the CNN-clustering and the pulling is done.

References

(1) Swope, W. C.; Pitera, J. W.; Suits, F. Describing Protein Folding Kinetics by Molecular Dynamics Simulations. *J. Phys. Chem. B* **2004**, *108*, 6571–6581.

(2) Buch, I.; Giorgino, T.; Fabritiis, G. D. Complete Reconstruction of an Enzyme-Inhibitor Binding Process by Molecular Dynamics Simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 10184–10189.

(3) Karplus, M.; McCammon, J. A. Molecular Dynamics Simulations of Biomolecules. *Nat. Struct. Biol.* **2002**, *9*, 646–652.

(4) Adcock, S. A.; McCammon, J. A. Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins. *Chem. Rev.* **2006**, *106*, 1589–1615.

(5) Maragakis, P.; Lindorff-Larsen, K.; Eastwood, M. P.; Dror, R. O.; Klepeis, J. L.; Arkin, I. T.; Jensen, M. Ø.; Xu, H.; Trbovic, N.; Friesner, R. A. et al. Microsecond Molecular Dynamics Simulation Shows Effect of Slow Loop Dynamics on Backbone Amide Order Parameters of Proteins. *J. Phys. Chem. B* **2008**, *112*, 6155–6158.

(6) Vitalini, F.; Mey, A. S. J. S.; Noé, F.; Keller, B. G. Dynamic properties of force fields. *J. Chem. Phys.* **2015**, *142*, 084101.

(7) Lemke, O.; Keller, B. G. Density-Based Cluster Algorithms for the Identification of Core Sets. *J. Chem. Phys.* **2016**, *145*, 164104.

(8) Witek, J.; Keller, B. G.; Blatter, M.; Meissner, A.; Wagner, T.; Riniker, S. Kinetic Models of Cyclosporin A in Polar and Apolar Environments Reveal Multiple Congruent Conformational States. *J. Chem. Inf. Model.* **2016**, *56*, 1547–1562.

(9) Witek, J.; Mühlbauer, M.; Keller, B. G.; Blatter, M.; Meissner, A.; Wagner, T.; Riniker, S. Interconversion Rates between Conformational States as Rationale for the Membrane Permeability of Cyclosporines. *ChemPhysChem* **2017**, *18*, 3309–3314.

(10) Pinamonti, G.; Paul, F.; Noé, F.; Rodriguez, A.; Bussi, G. The Mechanism of RNA Base Fraying: Molecular Dynamics Simulations Analyzed with Core-Set Markov State Models. *J. Chem. Phys.* **2019**, *150*, 154123.

(11) Hanske, J.; Aleksić, S.; Ballaschk, M.; Jurk, M.; Shanina, E.; Beerbaum, M.; Schmieder, P.; Keller, B. G.; Rademacher, C. Intradomain Allosteric Network Modulates Calcium Affinity of the C-Type Lectin Receptor Langerin. *J. Am. Chem. Soc.* **2016**, *138*, 12176–12186.

(12) Sirur, A.; Sancho, D. D.; Best, R. B. Markov State Models of Protein Misfolding. *J. Chem. Phys.* **2016**, *144*, 075101.

(13) Kiran, P.; Bhatia, S.; Lauster, D.; Aleksić, S.; Fleck, C.; Peric, N.; Maison, W.; Liese, S.; Keller, B. G.; Herrmann, A. et al. Exploring Rigid and Flexible Core Trivalent Sialosides for Influenza Virus Inhibition. *Chem. Europ. J.* **2018**, *24*, 19373–19385.

(14) Witek, J.; Wang, S.; Schroeder, B.; Lingwood, R.; Dounas, A.; Roth, H.-J.; Fouché, M.; Blatter, M.; Lemke, O.; Keller, B. et al. Rationalization of the Membrane Permeability Differences in a Series of Analogue Cyclic Decapeptides. *J. Chem. Inf. Model.* **2018**, *59*, 294–308.

(15) Yao, G.; Joswig, J.-O.; Keller, B. G.; Süssmuth, R. D. Total Synthesis of the Death Cap Toxin Phalloidin: Atropoisomer Selectivity Explained by Molecular-Dynamics Simulations. *Chem. Europ. J.* **2019**, *25*, 8030–8034.

(16) Dror, R. O.; Dirks, R. M.; Grossman, J. P.; Xu, H.; Shaw, D. E. Biomolecular Simulation: A Computational Microscope for Molecular Biology. *Annu. Rev. Biophys.* **2012**, *41*, 429–452.

(17) Götze, J. P.; Greco, C.; Mitrić, R.; Bonačić-Koutecký, V.; Saalfrank, P. BLUF Hydrogen Network Dynamics and UV/Vis Spectra: A Combined Molecular Dynamics and Quantum Chemical Study. *J. Comput. Chem.* **2012**, *33*, 2233–2242.

(18) Plattner, N.; Doerr, S.; Fabritiis, G. D.; Noé, F. Complete Protein–Protein Association Kinetics in Atomic Detail Revealed by Molecular Dynamics Simulations and Markov Modelling. *Nat. Chem.* **2017**, *9*, 1005–1011.

(19) Schumacher, D.; Lemke, O.; Helma, J.; Gerszonowicz, L.; Waller, V.; Stoschek, T.; Durkin, P. M.; Budisa, N.; Leonhardt, H.; Keller, B. G. et al. Broad Substrate Tolerance of Tubulin Tyrosine Ligase Enables One-Step Site-Specific Enzymatic Protein Labeling. *Chem. Sci.* **2017**, *8*, 3471–3478.

(20) Perilla, J. R.; Hadden, J. A.; Goh, B. C.; Mayne, C. G.; Schulten, K. All-Atom Molecular Dynamics of Virus Capsids as Drug Targets. *J. Phys. Chem. Lett.* **2016**, *7*, 1836–1844.

(21) Perilla, J. R.; Goh, B. C.; Cassidy, C. K.; Liu, B.; Bernardi, R. C.; Rudack, T.; Yu, H.; Wu, Z.; Schulten, K. Molecular Dynamics Simulations of Large Macromolecular Complexes. *Curr. Opin. Struct. Biol.* **2015**, *31*, 64–74.

(22) Zhao, G.; Perilla, J. R.; Yufenyuy, E. L.; Meng, X.; Chen, B.; Ning, J.; Ahn, J.; Gronenborn, A. M.; Schulten, K.; Aiken, C. et al. Mature HIV-1 Capsid Structure by Cryo-Electron Microscopy and All-Atom Molecular Dynamics. *Nature* **2013**, *497*, 643–646.

(23) Schmidt, K.; Fufezan, C.; Krieger-Liszkay, A.; Satoh, H.; Paulsen, H. Recombinant Water-Soluble Chlorophyll Protein from Brassica oleracea Var. Botrys Binds Various Chlorophyll Derivatives. *Biochemistry* **2003**, *42*, 7427–7433.

(24) Satoh, H.; Uchida, A.; Nakayama, K.; Okada, M. Water-Soluble Chlorophyll Protein in Brassicaceae Plants Is a Stress-Induced Chlorophyll-Binding Protein. *Plant Cell Physiol.* **2001**, *42*, 906–911.

(25) Takahashi, S.; Yanai, H.; Nakamaru, Y.; Uchida, A.; Nakayama, K.; Satoh, H. Molecular Cloning, Characterization and Analysis of the Intracellular Localization of a Water-

Soluble Chl-Binding Protein from Brussels Sprouts (Brassica oleracea var. gemmifera). *Plant Cell Physiol.* **2012**, *53*, 879–891.

(26) Agostini, A.; Palm, D. M.; Schmitt, F.-J.; Albertini, M.; Valentin, M. D.; Paulsen, H.; Carbonera, D. An Unusual Role for the Phytyl Chains in the Photoprotection of the Chlorophylls Bound to Water-Soluble Chlorophyll-Binding Proteins. *Sci. Rep.* **2017**, *7*.

(27) Boex-Fontvieille, E.; Rustgi, S.; Reinbothe, S.; Reinbothe, C. A Kunitz-Type Protease Inhibitor Regulates Programmed Cell Death During Flower Development in Arabidopsis thaliana. *J. Exp. Bot.* **2015**, *66*, 6119–6135.

(28) Boex-Fontvieille, E.; Rustgi, S.; von Wettstein, D.; Reinbothe, S.; Reinbothe, C. Water-Soluble Chlorophyll Protein Is Involved in Herbivore Resistance Activation During Greening of Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, 7303–7308.

(29) Horigome, D.; Satoh, H.; Itoh, N.; Mitsunaga, K.; Oonishi, I.; Nakagawa, A.; Uchida, A. Structural Mechanism and Photoprotective Function of Water-soluble Chlorophyll-binding Protein. *J. Biol. Chem.* **2006**, *282*, 6525–6531.

(30) Palm, D. M.; Agostini, A.; Pohland, A.-C.; Werwie, M.; Jaenicke, E.; Paulsen, H. Stability of Water-Soluble Chlorophyll Protein (WSCP) Depends on Phytyl Conformation. *ACS Omega* **2019**, *4*, 7971–7979.

(31) Palm, D. M.; Agostini, A.; Tenzer, S.; Gloeckle, B. M.; Werwie, M.; Carbonera, D.; Paulsen, H. Water-Soluble Chlorophyll Protein (WSCP) Stably Binds Two or Four Chlorophylls. *Biochemistry* **2017**, *56*, 1726–1736.

(32) Satoh, H.; Nakayama, K.; Okada, M. Molecular Cloning and Functional Expression of a Water-soluble Chlorophyll Protein, a Putative Carrier of Chlorophyll Molecules in Cauliflower. *J. Biol. Chem.* **1998**, *273*, 30568–30575.

(33) Bednarczyk, D.; Dym, O.; Prabahar, V.; Peleg, Y.; Pike, D. H.; Noy, D. Fine Tuning of Chlorophyll Spectra by Protein-Induced Ring Deformation. *Angew. Chem. Int. Ed.* **2016**, *55*, 6901–6905.

(34) Downing, W. L.; Mauxion, F.; Fauvarque, M.-O.; Reviron, M.-P.; Vienne, D. D.; Vartanian, N.; Giraudat, J. A Brassica napus Transcript Encoding a Protein Related to the Künitz Protease Inhibitor Family Accumulates upon Water Stress in Leaves, not in Seeds. *Plant J.* **1992**, *2*, 685–693.

(35) Reviron, M.-P.; Vartanian, N.; Sallantin, M.; Huet, J.-C.; Pernollet, J.-C.; de Vienne, D. Characterization of a Novel Protein Induced by Progressive or Rapid Drought and Salinity in Brassica napus Leaves. *Plant Physiol.* **1992**, *100*, 1486–1493.

(36) Annamalai, P.; Yanagihara, S. Identification and Characterization of a Heat-Stress Induced Gene in Cabbage Encodes a Kunitz Type Protease Inhibitor. *J. Plant Physiol.* **1999**, *155*, 226–233.

(37) Nishio, N.; Satoh, H. A Water-Soluble Chlorophyll Protein in Cauliflower May Be Identical to BnD22, a Drought-Induced, 22-Kilodalton Protein in Rapeseed. *Plant Physiol.* **1997**, *115*, 841–846.

(38) Renger, G.; Pieper, J.; Theiss, C.; Trostmann, I.; Paulsen, H.; Renger, T.; Eichler, H. J.; Schmitt, F.-J. Water Soluble Chlorophyll Binding Protein of Higher Plants: A Most Suitable Model System for Basic Analyses of Pigment–Pigment and Pigment–Protein Interactions in Chlorophyll Protein Complexes. *J. Plant Physiol.* **2011**, *168*, 1462–1472.

(39) Bektas, I.; Fellenberg, C.; Paulsen, H. Water-Soluble Chlorophyll Protein (WSCP) of Arabidopsis Is Expressed in the Gynoecium and Developing Silique. *Planta* **2012**, *236*, 251–259.

(40) Kamimura, Y.; Mori, T.; Yamasaki, T.; Katoh, S. Isolation, Properties and a Possible

Function of a Water-Soluble Chlorophyll a/b-Protein from Brussels Sprouts. *Plant Cell Physiol.* **1997**, *38*, 133–138.

(41) Theiss, C.; Trostmann, I.; Andree, S.; Schmitt, F. J.; Renger, T.; Eichler, H. J.; Paulsen, H.; Renger, G. Pigment-Pigment and Pigment-Protein Interactions in Recombinant Water-Soluble Chlorophyll Proteins (WSCP) from Cauliflower. *J. Phys. Chem. B* **2007**, *111*, 13325–13335.

(42) Damaraju, S.; Schlede, S.; Eckhardt, U.; Lokstein, H.; Grimm, B. Functions of the Water Soluble Chlorophyll-Binding Protein in Plants. *J. Plant Physiol.* **2011**, *168*, 1444–1451.

(43) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: Fast, Flexible, and Free. *J. Comput. Chem.* **2005**, *26*, 1701–1718.

(44) Hockney, R. W.; Goel, S. P.; Eastwood, J. W. Quiet High-Resolution Computer Models of a Plasma. *J. Comput. Phys.* **1974**, *14*, 148–158.

(45) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain $\chi1$ and $\chi2$ Dihedral Angles. *J. Chem. Theory Comput.* **2012**, *8*, 3257–3273.

(46) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat. Methods* **2016**, *14*, 71–73.

(47) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926–935.

(48) Guerra, F.; Adam, S.; Bondar, A.-N. Revised Force-Field Parameters for Chlorophyll-a, Pheophytin-a and Plastoquinone-9. *J. Mol. graph. Model.* **2015**, *58*, 30–39.

(49) Kuczera, K.; Kuriyan, J.; Karplus, M. Temperature Dependence of the Structure and Dynamics of Myoglobin. *J. Mol. Biol.* **1990**, *213*, 351–373.

(50) Foloppe, N.; Ferrand, M.; Breton, J.; Smith, J. Structural Model of the Photosynthetic Reaction Center of Rhodobacter Capsulatus. *Proteins: Struct. Funct.* **1995**, *22*, 226–244.

(51) Damjanovic, A.; Kosztin, I.; Kleinekathöfer, U.; Schulten, K. Excitons in a Photosynthetic Light-Harvesting System: a Combined Molecular Dynamics, Quantum Chemistry, and Polaron Model Study. *Phys. Rev.* **2002**, *65*, 031919.

(52) Foloppe, N.; Brenton, J.; Smith, J. C. *Potential Energy Function for Photosynthetic Reaction Centre Chromophores: Energy Minimisations of a Crystalline Bacteriophytin A Analog*; 1992; in J. Brenton, A. Vermeglio (Eds.), The Photosynthetic Bacterial Reaction Center II, Plenum Press, New York.

(53) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling Through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.

(54) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, *52*, 7182–7190.

(55) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**, *18*, 1463–1472.

(56) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N·log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089.

(57) Costescu, B. I.; Gräter, F. Time-resolved Force Distribution Analysis. *BMC Biophys.* **2013**, *6*, 5.

(58) Renger, T.; Trostmann, I.; Theiss, C.; Madjet, M. E.; Richter, M.; Paulsen, H.; Eichler, H. J.; Knorr, A.; Renger, G. Refinement of a Structural Model of a Pigment-Protein Complex by Accurate Optical Line Shape Theory and Experiments. *J. Phys. Chem. B* **2007**, *111*, 10487–10501.

(59) Keller, B.; Daura, X.; van W. F. Gunsteren, Comparing Geometric and Kinetic Cluster Algorithms for Molecular Simulation Data. *J. Chem. Phys.* **2010**, *132*, 074110.

(60) Lemke, O.; Keller, B. G. Common Nearest Neighbor Clustering—A Benchmark. *Algorithms* **2018**, *11*, 19.

(61) Lemke, O.; Keller, B. G. CNNClustering. `https://github.com/BDGSoftware/CNNClustering`, 2017.

(62) Bashirova, A.; Pramanik, S.; Volkov, P.; Rozhkova, A.; Nemashkalov, V.; Zorov, I.; Gusakov, A.; Sinitsyn, A.; Schwaneberg, U.; Davari, M. Disulfide Bond Engineering of an Endoglucanase from Penicillium verruculosum to Improve Its Thermostability. *Int. J. Mol. Sci.* **2019**, *20*, 1602.

(63) Takahashi, S.; Aizawa, K.; Nakayama, K.; Satoh, H. Water-Soluble Chlorophyll-Binding Proteins from Arabidopsis thaliana and Raphanus sativus Target the Endoplasmic Reticulum body. *BMC Res. Notes* **2015**, *8*, 365.

(64) Braakman, I. Folding of Influenza Hemagglutinin in the Endoplasmic Reticulum. *J. Cell Biol.* **1991**, *114*, 401–411.

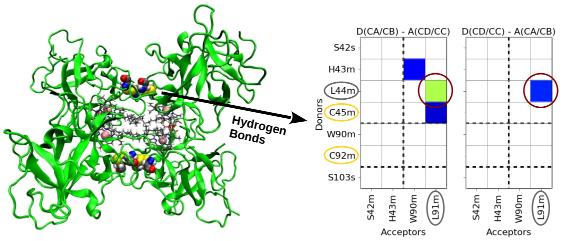
(65) Frand, A. R.; Kaiser, C. A. Ero1p Oxidizes Protein Disulfide Isomerase in a Pathway for Disulfide Bond Formation in the Endoplasmic Reticulum. *Mol. Cell* **1999**, *4*, 469–477.

(66) Freedman, R. B.; Hirst, T. R.; Tuite, M. F. Protein Disulphide Isomerase: Building Bridges in Protein Folding. *Trends Biochem. Sci.* **1994**, *19*, 331–336.

(67) Palm, D. M.; Agostini, A.; Averesch, V.; Girr, P.; Werwie, M.; Takahashi, S.; Satoh, H.; Jaenicke, E.; Paulsen, H. Chlorophyll a/b Binding-Specificity in Water-Soluble Chlorophyll Protein. *Nat. Plants* **2018**, *4*, 920–929.

(68) Kramer, R. M.; Shende, V. R.; Motl, N.; Pace, C. N.; Scholtz, J. M. Toward a Molecular Understanding of Protein Solubility: Increased Negative Surface Charge Correlates with Increased Solubility. *Biophys. J.* **2012**, *102*, 1907–1915.

6 TOC Graphic



Supporting Information

Oliver Lemke and Jan P. Götze*

Freie Universität Berlin, Department of Chemistry and Biochemistry, Arnimallee 22, 14195 Berlin

E-mail: jgoetze@zedat.fu-berlin.de

1 Supporting Figures

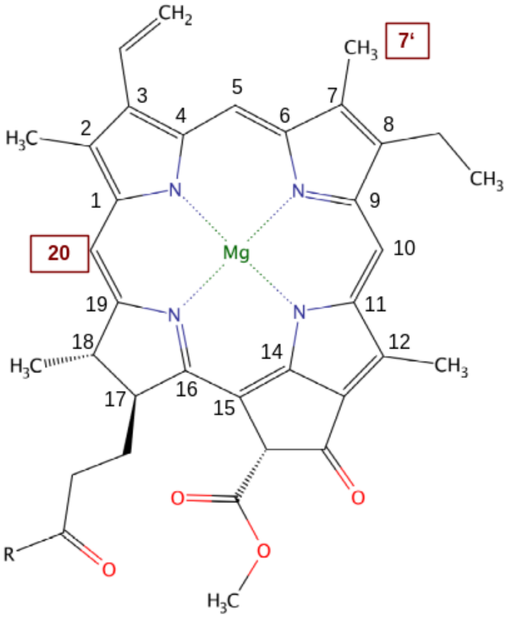


Figure S1: Numbering of the porphyrin ring of chlorophyll a. The positions mentioned in the paper are highlighted.

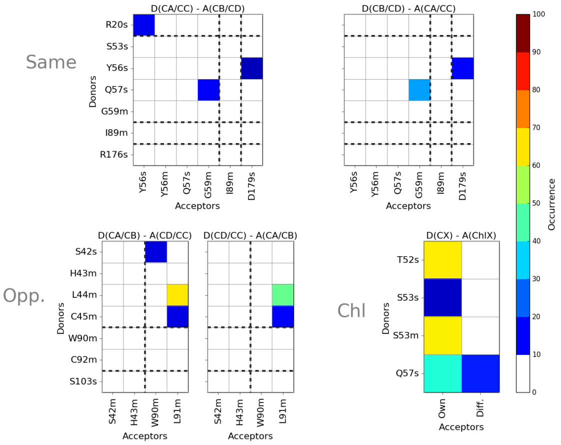


Figure S2: Hydrogen bonds of simulation Disulf4 between donor 'D' and acceptor 'A' with an occurrence larger than 10 % within the same dimer, between different dimers and between the protein and Chl; Coordination via main chain 'm' or side chain 's'. Labelling D(X/Y)-A(X'/Y') reads as follows: donor X towards acceptor X' and donor Y towards acceptor Y'.

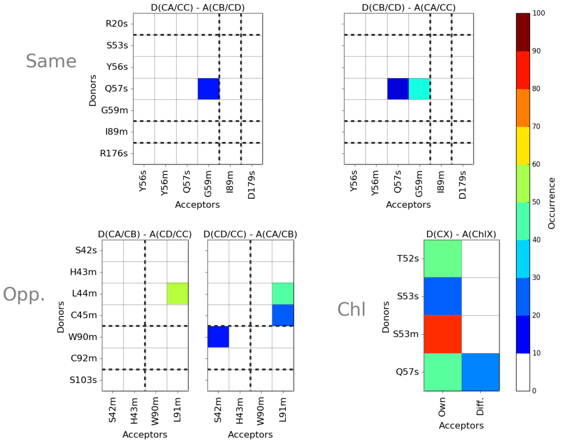


Figure S3: Hydrogen bonds of simulation Nat2 between donor 'D' and acceptor 'A' with an occurrence larger than 10 % within the same dimer, between different dimers and between the protein and Chl; Coordination via main chain 'm' or side chain 's'. Labelling D(X/Y)-A(X'/Y') reads as follows: donor X towards acceptor X' and donor Y towards acceptor Y'.

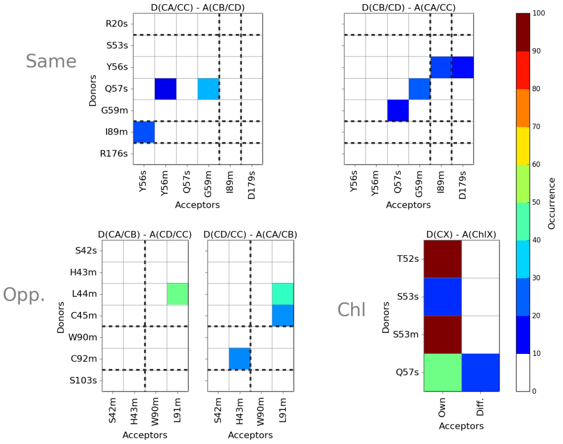


Figure S4: Hydrogen bonds of simulation Disulf2 between donor 'D' and acceptor 'A' with an occurrence larger than 10 % within the same dimer, between different dimers and between the protein and Chl; Coordination via main chain 'm' or side chain 's'. Labelling D(X/Y)-A(X'/Y') reads as follows: donor X towards acceptor X' and donor Y towards acceptor Y'.

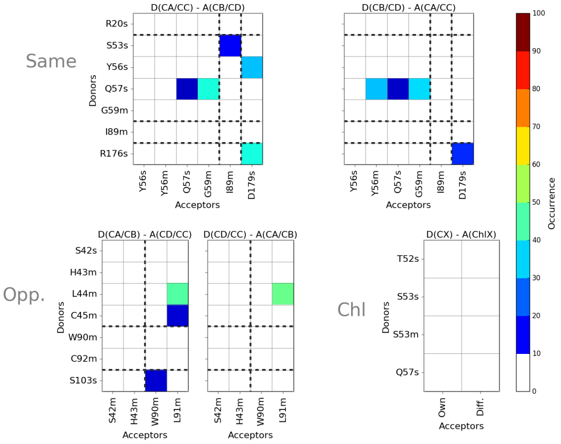


Figure S5: Hydrogen bonds of simulation Nat0 between donor 'D' and acceptor 'A' with an occurrence larger than 10 % within the same dimer, between different dimers and between the protein and Chl; Coordination via main chain 'm' or side chain 's'. Labelling D(X/Y)-A(X'/Y') reads as follows: donor X towards acceptor X' and donor Y towards acceptor Y'.

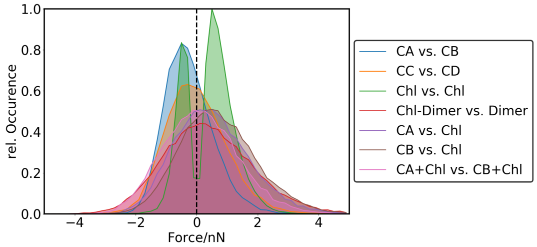


Figure S6: Force distribution analysis between different parts of the protein (Nat2), where the sub units are denoted as CA to CD, with CA and CB belonging to the Chl-containing dimer and CC and CD to the Chl-deficient dimer; All analyzed distributions are highlighted.

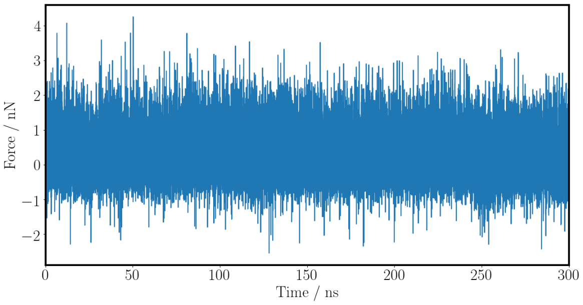


Figure S7: Fluctuation of the forces between the Chl during the simulation Nat2.

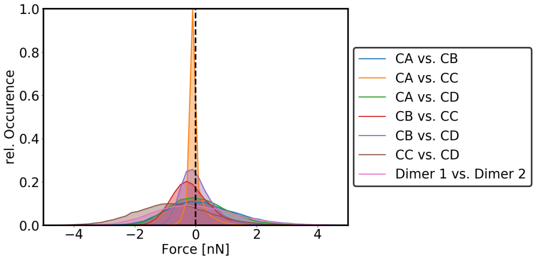


Figure S8: Force distribution analysis between different parts of the protein (Nat0), where the sub units are denoted as CA to CD; All analyzed distributions are highlighted.

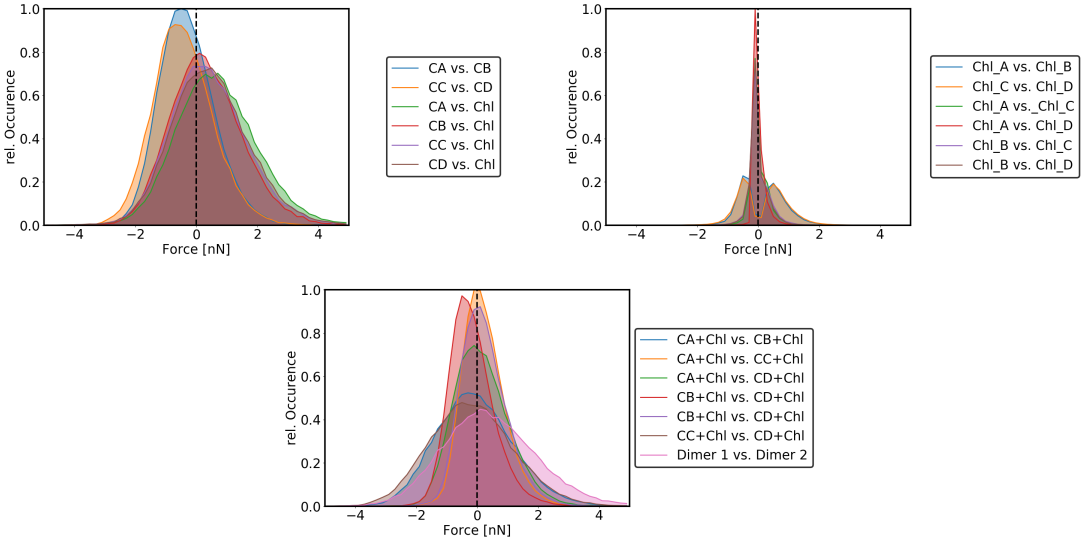


Figure S9: Force distribution analysis between different parts of the protein (Nat4), where the sub units are denoted as CA to CD. The Chls are denoted as Chl (belonging to the same subunit) or Chl_X (belonging to subunit X); All analyzed distributions are highlighted.

Table S1: Non-bonded energies between different parts for the systems Nat4, Nat2, Nat0; (Chl) indicates dimers containing Chl; Additional labels: 's.d.': same and 'o.d.': opposite dimer and 's.su.' 'same' subunit; † Interactions between Chl of subunit CA with protein chain CC and between Chl of subunit CB with protein chain CD. The interaction with the other subunit (CD/CC) is neglectable. All energy values are given in kJ/mol.

		Nat4	Nat2	Nat0
Protein - Protein	s.d. (Chl)	-224.5 ± 37.0	-174.8 ± 32.8	–
	s.d.	–	-236.7 ± 12.0	-311.0 ± 68.0
	o.d.	-91.9 ± 15.0	-113.2 ± 12.0	-100.8 ± 10.5
Protein - Chl	s.su.	-334.5 ± 11.4	-355.6 ± 15.7	–
	s.d.	-38.6 ± 3.5	-30.3 ± 9.5	–
	o.d.†	-24.1 ± 3.3	-20.5 ± 5.5	–
Protein - Chl	s.d.	-122.8 ± 8.7	-136.3 ± 9.8	–
	o.d.	-14.5 ± 2.8	–	–

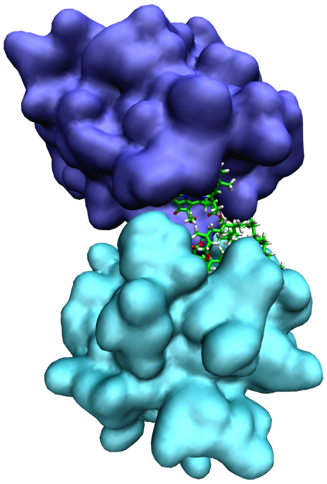


Figure S10: Dimer formed during the simulation Nat4_P_all. Every subunit is highlighted by a separate color.

2 Clustering

To analyze the conformations of chlorophyll a in solution a Common-Nearest-Neighbor clustering was performed[1–3] using the software provided at *GitHub*.[4] For further explanation with respect to the input parameter, the algorithm and the outcome we refer to Ref. 3.

As input coordinates the positions of the magnesium ions were chosen. To remove the rotational degrees of freedom all positions were transformed into internal coordinates (distance r, polar angle θ and azimuthal angle ϕ) with respect to the center of mass, located at (0,0,0). For each frame, the magnesium ion closest to the center of mass was rotate onto the z axis. The next closest magnesium ion was rotated along the z-axis and settled in the x-z-plane. This selection of the two closest magnesium ions as reference points ensures the exchangeability of the chlorophylls. To account for periodicity the sine and cosine of the angles were taken for further analysis. By normalization all data were bound between [0,1].

For the clustering the data set was reduced by a taking every tenth frame resulting in a data set size of 10,001 data points. Between all data points the euclidean distance between the normalized coordinates was calculated and used as a distance measure. In a first clustering step (parameter set: R=0.15, N=5, M=10) 28 clusters were isolated with 26 % of the data set declared as noise. The 2 most populated clusters were refined (parameter set: R=0.15, N=10, M=10) in an hierarchical approach according to Ref. 2 yielding 23 new clusters, resulting in a total of 49 clusters. The noise was increased to a total of 40 %. The most dominant clusters, representing slightly distorted tetrahedron structures, are reported in Figure S11.

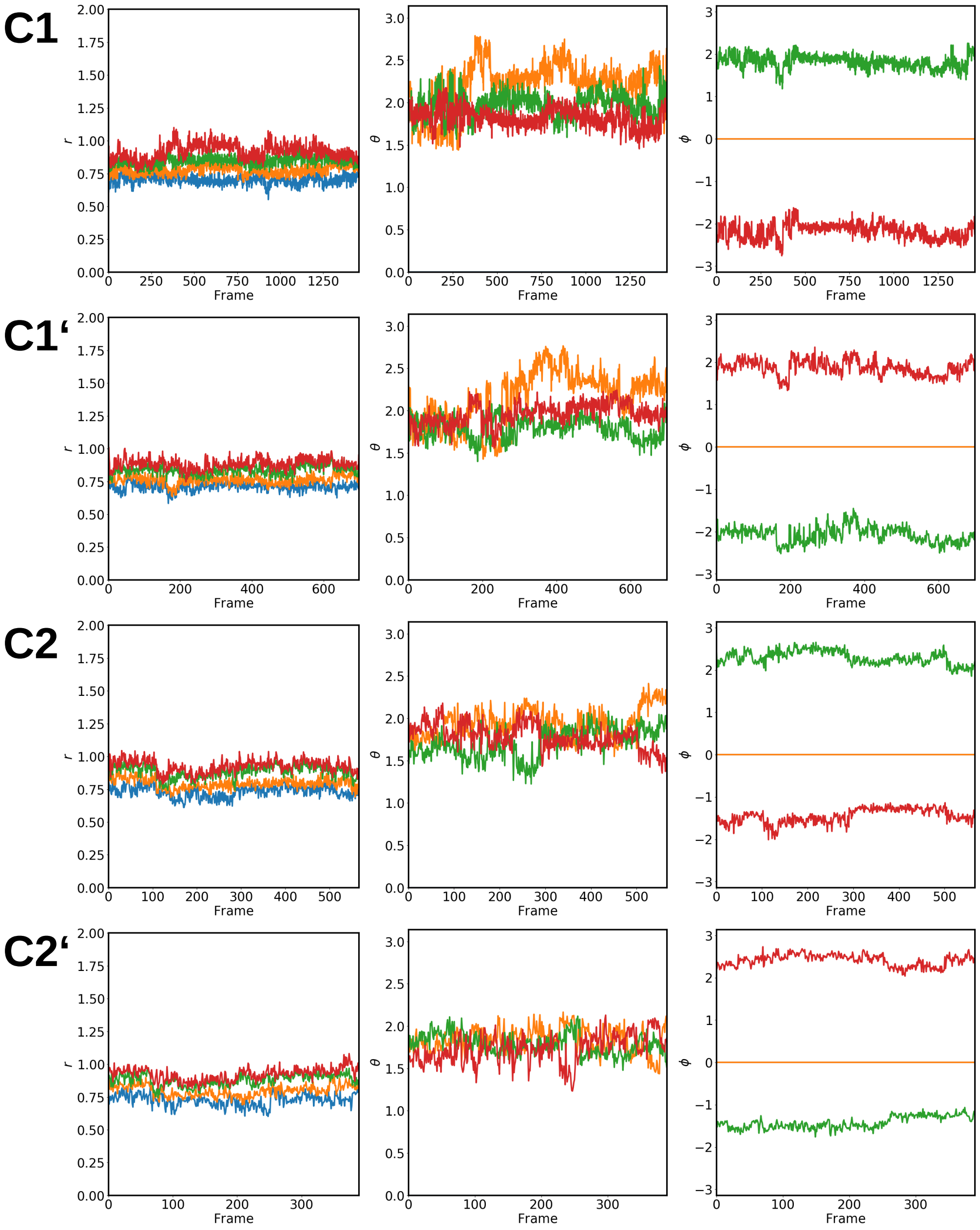


Figure S11: Plot of the internal coordinates (r, θ, ϕ) for the conformations of the 4 largest clusters. The coloring is according to the distance between the magnesium ion and the center of mass from closest to largest: blue, orange, green, red. The ’ indicates that the cluster is a mirror image with respect to the exchange of the red and green labeled magnesium ions; Note that as the closest magnesium ion was rotated on the z-axis (blue) ϕ and θ are always 0, for the magnesium ion settled in the x-z-plane (orange) ϕ is always 0.

3 Pulling of the WSCP subunits

For the pulling of the WSCP a conformation extracted at 1 ns of the simulation Nat4 was used. The system was translated to set the center of mass to (0,0,0). The system was rotated such that the average point on the vector connecting Ile124-C_α of subunits CA and CB was located onto the z-axis to a coordinate (0,0,z). A second rotation along the z-axis was performed to locate the Ile124-C_α of subunit CA in the xz-plane onto coordinate (x,0,z'). The coordinates of subunit CA and its chlorophyll were modified by adding (c,0,c) with $c = \frac{1}{\sqrt{2}}$ to their coordinates. For the other subunits, modifications of: CB (-c,0,c), CC (0,c,-c) and CD (0,-c,-c), were applied. For Nat4_P_all, these modfications were done for all atoms, for Nat_Chl_P only for the Chls with the protein chains being removed. For Nat4_P_protein, only the protein coordinates were modified and the Chl coordinates kept in place. A starting structure where all coordinates were modified is shown in figure S12

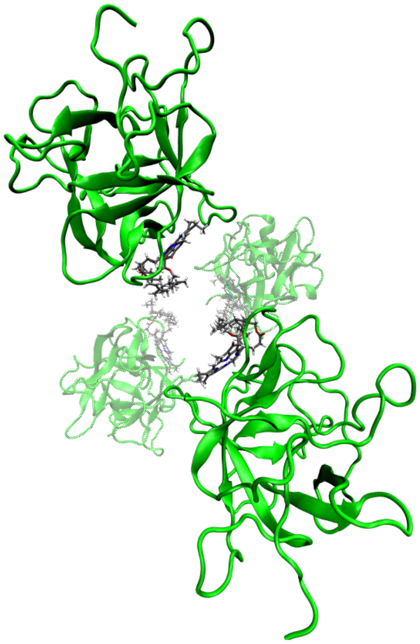


Figure S12: Pulled starting structure for the simulation Nat4_P_all. The dimer CA-CB is shown in the foreground, the dimer CC-CD in the background.

References

(1) Keller, B.; Daura, X.; van W. F. Gunsteren, Comparing Geometric and Kinetic Cluster Algorithms for Molecular Simulation Data. *J. Chem. Phys.* **2010**, *132*, 074110.

(2) Lemke, O.; Keller, B. G. Density-Based Cluster Algorithms for the Identification of Core Sets. *J. Chem. Phys.* **2016**, *145*, 164104.

(3) Lemke, O.; Keller, B. G. Common Nearest Neighbor Clustering—A Benchmark. *Algorithms* **2018**, *11*, 19.

(4) Lemke, O.; Keller, B. G. CNNClustering. `https://github.com/BDGSoftware/CNNClustering`, 2017.

3.2 Analyzing the Spectral Properties of WSCP Comparing Different Calculation Setups

3.2.1 Introduction

Besides the high stability of the WSCP holoprotein, a high resistance of the chlorophyll (Chl) with respect to photobleaching was reported [185]. This is quite remarkable since no carotenoids (Crts) are present that could protect the Chl from oxidative damage. As it is known that Chl can transfer energy towards oxygen generating highly reactive singlet oxygen, Crts are assumed to act as Chl triplet quenchers [191, 192], oxygen scavenger [193] and harvesters of high energetic light [141, 194].

Former studies [185] proposed that the resistance of WSCP towards photobleaching is achieved by a lower singlet oxygen production. The reason for this is assumed to be a diffusion barrier formed by the protein preventing oxygen to diffuse into a cavity within the WSCP. Other mechanisms such as intersystem crossing reducing the lifetime of the excited state, or Chl triplet quenching, were ruled out [195]. In the MD simulations presented in the previous section, however, we discovered that this diffusion barrier might not be that strong, since diffusion of water into the cavity of the WSCP was detected. This is in agreement with recent studies reporting a singlet oxygen production comparable to that of free Chls. In this research [188], the phytyl chains are assumed to act as protection against oxidative damage.

Since the electronic excitations of the Chls are important for the photostability, the spectral properties of Chl bound to WSCP are investigated in this section. Chl absorbs in two regimes of the UV/Vis spectrum: A low energy regime around 650 nm (red/yellow) containing two states Q_y and Q_x, and a high energy regime about 440 nm (blue) containing several states, the so-called Soret-states [141, 196, 197]. The spectrum features a “green gap” between these two regions resulting in the green color of Chl. It was found that the two Chls within a dimer are strongly excitonic coupled. The excitation energy transfer between the other Chls in the WSCP was expected to occur via a Förster-type energy transfer [195, 198].

For the research described in the previous section 3.1, a large amount of simulation data for WSCP containing different numbers of Chls or disulfide bridges within or between subunits were produced. Within the simulations, different conformations of the WSCP were sampled, which can be used to calculate absorption spectra of the Chls using time-dependent density function theory (TD-DFT). This has the advantage that an average absorption spectrum for an ensemble of different WSCP conformations can be obtained. In the following section, it is investigated how this absorption spectrum varies by comparing different simulation setups. Additionally, it is examined how different optimization potentials such as

QM/MM influence the absorption spectra as well as the coupling of the Chls within the WSCP. The research focuses only on chlorophyll a. Thus, the abbreviation Chl refers to chlorophyll a in this section.

3.2.2 Methods

3.2.2.1 Computational Details

Snapshots from MD simulations performed with the GROMACS simulation package 2016.1 [199], were taken as a basis for the calculations. For all calculations, the same MD simulations as discussed in section 3.1 were used. Optimization of the snapshots at an MM level were performed using the steepest descent algorithm implemented in the GROMACS simulation package 2016.1. For optimizations at a QM/MM level, the software *gmx2qmmm* developed by Dr. Jan Götze was applied. For this purpose, the system was separated in 3 layers: The inner layer consisting of the porphyrine ring of the Chl and the coordinating P36 residue of the protein (QM), the central layer containing all residues and solvent molecules within a 1.2 nm range of the magnesium ion of the optimized Chl (MM) and the outer layer including the rest of the WSCP and all solvent molecules at a distance larger than 1.2 nm and smaller than 4 nm with respect to the magnesium ion (frozen). Solvent molecules at the interface between two layers were always assigned to the respective inner layer. The MM-layer was optimized using the steepest descent algorithm. For the QM-layer, DFT calculations using the CAM-B3LYP functional [200] and the 6-31G* basis set [201–204] were performed. The QM system is electrostatically embedded in the MM system and utilizes link atoms to saturate the QM system, and charge shifts to counteract overpolarization. The QM/MM optimizations were truncated after 20 optimization steps to save computational costs as no significant changes in geometries were observed thereafter.

Absorption spectra were calculated with TD-DFT (CAM-B3LYP/6-31G*) using Gaussian 16 [205–212]. 10 excited states per Chl were computed in the TD-DFT calculation. The TD-DFT system was identical to the QM-layer, as mentioned above. The environment consisting of the rest of the protein and all water molecules within a 4 nm range of the magnesium ion of the Chl was treated as a point charge field, unless not stated differently. For every investigated system, several snapshots from the MD simulations were extracted and optimized as summarized in table 3.1, taking into account different conformations. Since every snapshot contains up to four Chls, several absorption spectra could be calculated from one snapshot. This allows for further comparison such as the relative orientations of transition dipole moments within one snapshot.

Table 3.1: Systems for the TD-DFT calculations based on snapshots extracted from MD simulations; N_{Chl} denotes the number of Chls in the MD simulation. N_{snap} snapshots were taken every Δt ns. Thus, a total number of $N_{Chl} \cdot N_{snap}$ TD-DFT calculations were performed; † The TD-DFT calculation includes both Chl and the coordinating P36 within the dimer reducing the total number of calculations to $N_{Chl} \cdot N_{snap}/2$; * Changes in the MD simulation setup.

N_{Chl}	Optimization potential	Δt/ns	N_{snap}	Total	Notes
4	–	1.0	25	100	–
4	MM	1.0	25	100	–
4	QM/MM	1.0	25	100	–
4	QM/MM	3.0	8	32	in vacuo
4	MM	1.0	25	50	Two Chls on QM level†
2	QM/MM	1.5	17	34	Two Chl in MD simulation*
4	QM/MM	2.0	12	48	Disulfide bridges*

The first five calculations mentioned in table 3.1 were based on the MD simulation starting from the crystal structure (see simulation Nat4, section 3.1), varying the optimization potential or the environment for the TD-DFT calculation. In these calculations the absorption spectrum of one Chl was calculated. Additionally, absorption spectra for both Chls within one dimer were calculated using an MM optimization of the snapshots. Snapshots from MD simulations containing only two Chl (see simulation Nat2, section 3.1) or protein units linked via disulfide bridges (see simulation Disulf4, section 3.1) were examined as well.

For the depicted absorption spectra, the calculated excitations were grouped with a bin size of 10 nm by summing up the oscillator strength and normalizing by the total number of calculations. Therefore, the absorption spectra can be seen as an average spectrum over all snapshots with an artificial broadening of 10 nm. Further broadening accounts for conformational changes. To obtain a smooth spectrum, a spline function with a resolution of 1 nm was calculated. As the order of the molecular orbitals (MOs) and the excitations can change for every snapshot, we developed a mapping procedure, which is described in section 3.2.2.2.

3.2.2.2 Mapping Procedure

For the mapping of the excitations to a reference set, a two-step procedure was used: In the first step, the order of the MOs is adjusted. The comparison of the excitations is done in a second step. For the mapping of the MOs, the full snapshot is rotated by superposing the QM system with respect to a reference structure. In this reference structure, the Chl of interest is set such that the magnesium ion denotes the origin, the coordinating oxygen of P36 denotes the z-axis and the nitrogen of the ring carrying the phytyl chain is located in the x-z-plane. For the rotated system, the 855 MO coefficients of the 10 highest occupied (HOMO) and the 10 lowest unoccupied MOs (LUMO) are compared to the MO coefficients of a reference set by calculating the euclidean distance. For every MO of the reference set, the MO with the closest distance is mapped using a maximal distance cutoff of 0.8. An error function for the MOs is utilized, which accounts for all orbitals that could not be mapped.

In the second step, the rearranged MOs are used to map the excitations with respect to a reference set. As a reference set, the excitations of the same calculation as for the MO-mapping were used. This is necessary because multiple transitions are involved to a certain degree. The transitions are arranged in a 10×10 matrix where each row represents a LUMO and each column a HOMO. The matrix elements contain the squared contribution of the corresponding transition to the overall excitation. The element-wise euclidean distance is calculated to compare two matrices. The closest distance (cutoff 0.3) is set as a mapping criterion. As for the MOs, an error function for the excitations is included, accounting for all excitations that could not be mapped to the reference set. The reference set is chosen in a way that only a small error function arises, and the spectrum can be reproduced by different reference sets. A depiction of the mapping is shown in figure 3.1.

In the case of two Chls in the TD-DFT calculation, the mapping algorithm is modified to detect equivalent MOs of both Chls. These MOs are labeled equally. As a consequence, it is possible to detect coupled excitations. For this analysis, the excitations per snapshot are separated into a higher and a lower energy excitation.

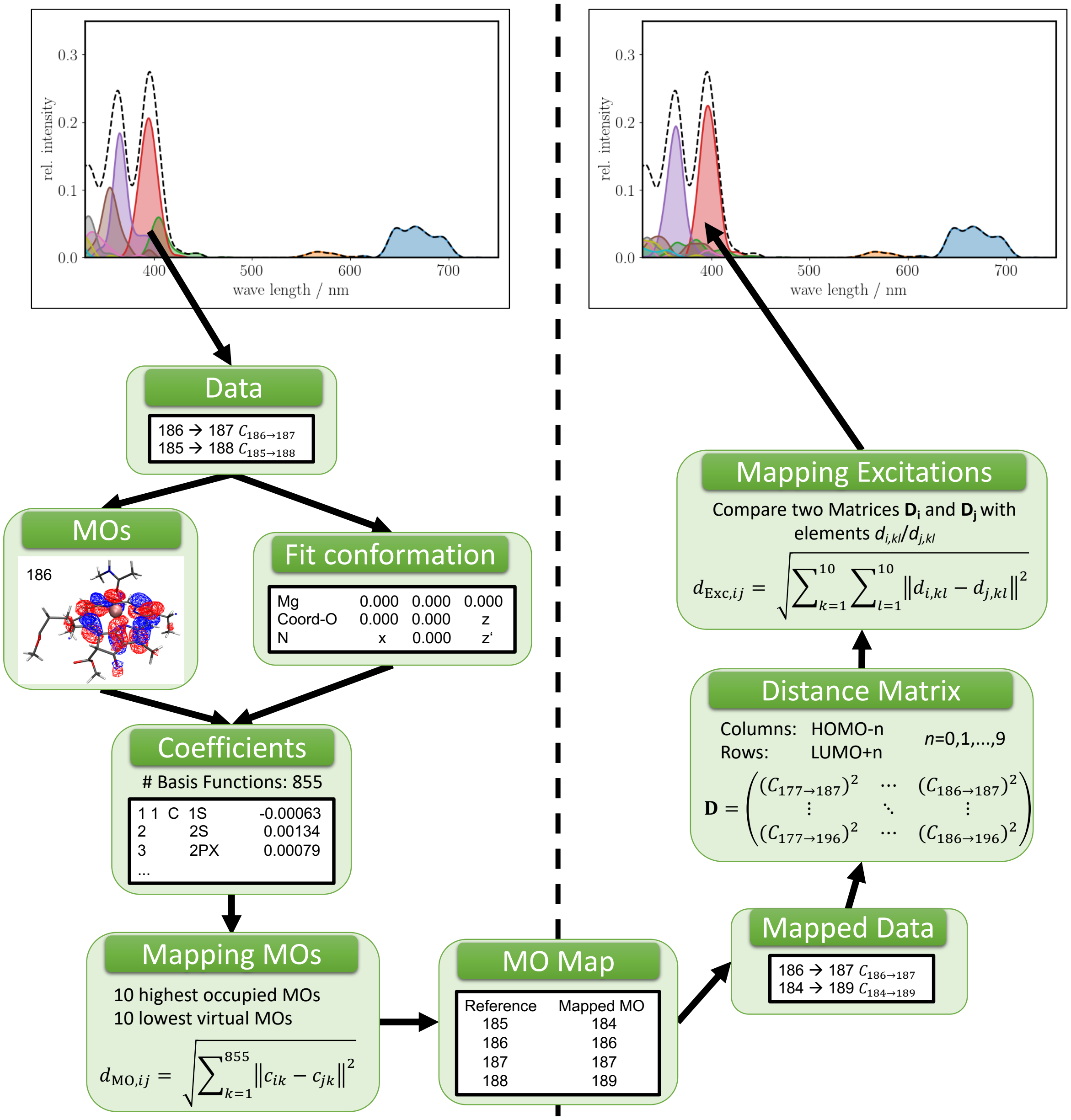

0.3
0.2
0.1
0.0
rel. intensity
400
500
600
700
wave length / nm
Data
186 → 187 $C_{186\to187}$
185 → 188 $C_{185\to188}$
MOs
186
Fit conformation
Mg 0.000 0.000 0.000
Coord-O 0.000 0.000 z
N x 0.000 z'
Coefficients
# Basis Functions: 855
1 1 C 1S -0.00063
2 2S 0.00134
3 2PX 0.00079
...
Mapping MOs
10 highest occupied MOs
10 lowest virtual MOs
$d_{\mathrm{MO},ij} = \sqrt{\sum_{k=1}^{855} \|c_{ik} - c_{jk}\|^2}$
MO Map
Reference Mapped MO
185 184
186 186
187 187
188 189
Mapped Data
186 → 187 $C_{186\to187}$
184 → 189 $C_{184\to189}$
Distance Matrix
Columns: HOMO-n
Rows: LUMO+n
n=0,1,…,9
$\mathbf{D} = \begin{pmatrix} (C_{177\to187})^2 & \cdots & (C_{186\to187})^2 \\ \vdots & \ddots & \vdots \\ (C_{177\to196})^2 & \cdots & (C_{186\to196})^2 \end{pmatrix}$
Mapping Excitations
Compare two Matrices $\mathbf{D}_i$ and $\mathbf{D}_j$ with elements $d_{i,kl}/d_{j,kl}$
$d_{\mathrm{Exc},ij} = \sqrt{\sum_{k=1}^{10} \sum_{l=1}^{10} \|d_{i,kl} - d_{j,kl}\|^2}$

Figure 3.1: Mapping procedure starting from the raw data. In a first step (left), the MOs are mapped. In a second step (right), the excitations are relabeled and mapped using the mapped MOs, resulting in a spectrum where every peak accounts for a specific excitation character (right) rather than for the energetic ordering of the excitations in every snapshot (left).

3.2.3 Results and Discussion

3.2.3.1 Comparison of Different Optimization Potentials

In a first step, it is examined how different optimization potentials influence the absorption spectrum of Chl. The MD snapshots were either considered directly without further optimization (w/o) or optimized by either an MM potential or a QM/MM potential. For the QM/MM potential, it was investigated how many optimization steps are necessary until the excitation energy and the transition dipole moments converged. Thus, for one snapshot an optimization was performed until the truncation criterion of the QM/MM scheme was met. The two investigated properties, excitation energy and transition dipole moment, are shown in figure 3.2 as a function of the optimization progress.

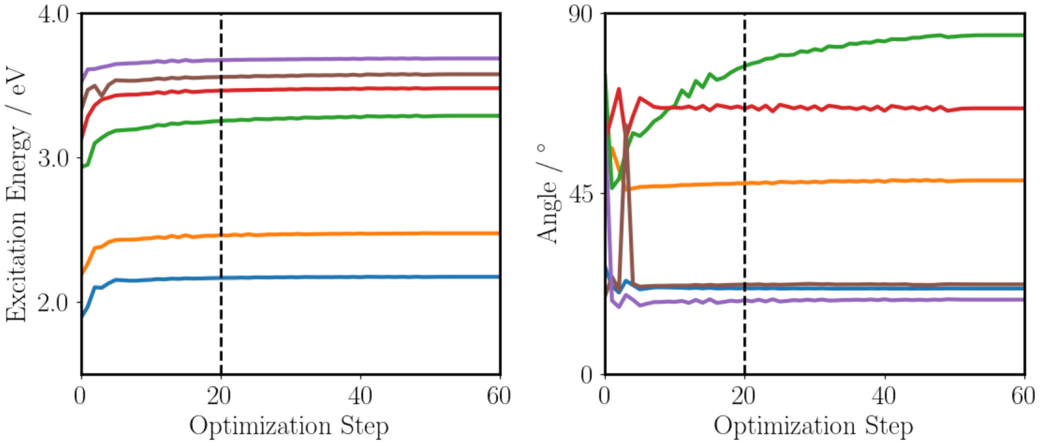


Figure 3.2: Convergence of the excitation energy (left) and of the angle between the transition dipole moment and the z-axis (right) depending on the progress of the QM/MM optimization. Only data for the first six excitations are shown.

Since all investigated excitation energies and transition dipole moment orientations (except for green) were converged after 20 optimization steps, the QM/MM optimization was truncated after 20 steps to save computational resources. The third excited state (green) only shows a low oscillator strength and will therefore not be considered in the further analysis. The spectra obtained for different optimization methods are shown in figure 3.3, and a qualitative scheme of the four dominant transitions is depicted in figure 3.4.

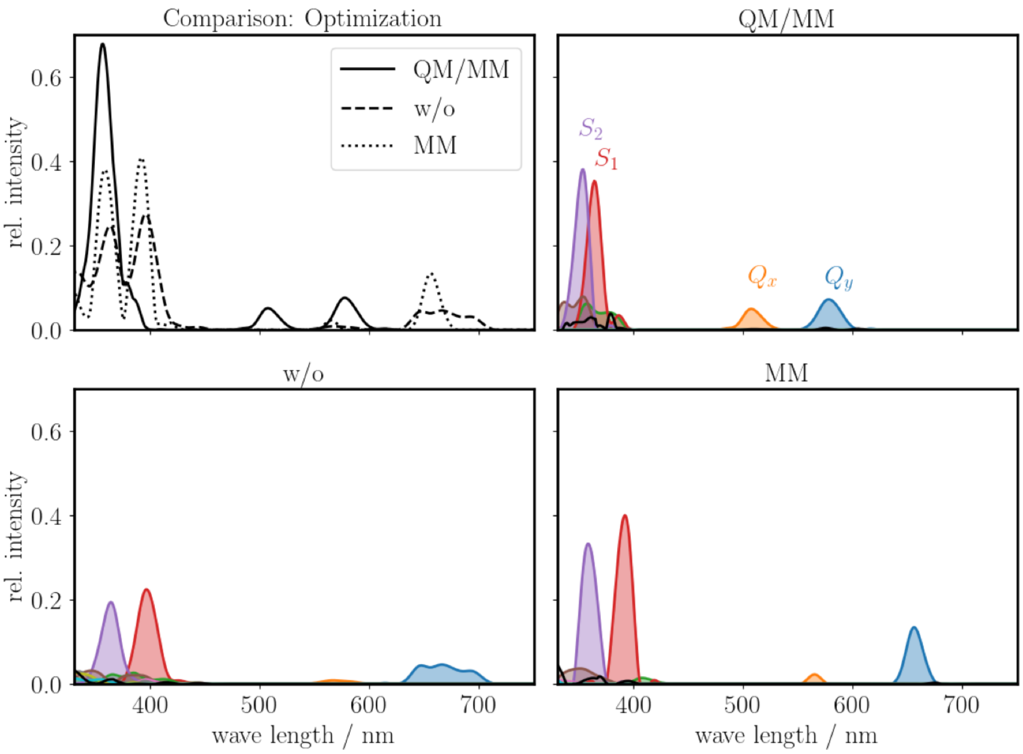


Figure 3.3: Comparison of different optimization potentials: Vertical absorption spectrum of 1 Chl averaged over several snapshots. The overall as well as the detailed absorption spectrum are shown; For every detailed spectrum, the black curve denotes an error function for those excitations which could not be mapped.

Comparing the overall spectra, similar peak positions with only a small blue-shift are obtained for the non-optimized and the MM-optimized systems. This is assumed to be caused by a movement of the WSCP conformation around the minimum during the MD simulation. The steepest descent algorithm moves the extracted conformations further towards the minimum. As a consequence, sharper peaks for the spectrum are obtained, compared to those using no optimization potential. For the calculations using no optimization, a broadening of the peaks can be seen, which is caused by structural fluctuations. The obtained absorption spectrum thus incorporates different conformations of the WSCP which are all located around one minimum and get lost using an optimization potential.

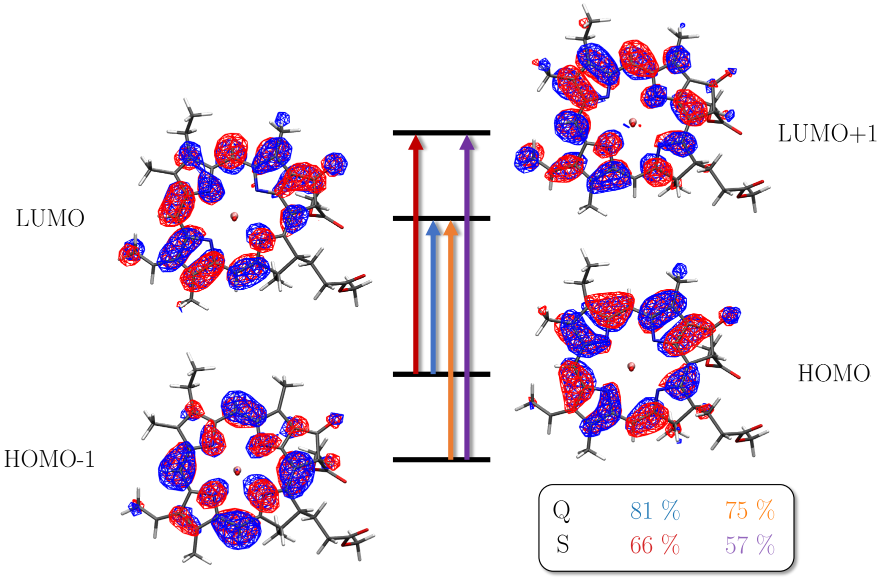


Figure 3.4: MOs involved in the four dominant excitations Q_y, Q_x, S_1 and S_2. The highlighted excitations account for the MO pair with the highest contribution. The contribution is given and colored according to the excitation and to figure 3.3 (QM/MM).

Treating the snapshots at a QM/MM level causes a strong blue shift. This is caused by an adaptation of the potential energy landscape, which is not allowed with the applied MM-method. Since the conformational change follows the ground state gradient, the optimization results in an increase of the energy gap between the ground and the excited state. Vertical excitations do not account for any excited state relaxation, which will likely result in higher excitation energies when optimizing the ground state. The largest conformational change in the WSCP ground state using a QM/MM optimization, occurred in the out-of-plane movement of the magnesium ion. In the previous section, we showed that the position of the magnesium ion above the ring is not well supported by the used force field. An optimization at QM level thus corrects this error by moving the angle between the magnesium ion and the ring closer to the experimentally observed value.

In addition to the absorption spectra of the single Chls, the coupling between the states of the Chls within a snapshot was also examined. For this examination, the collinearity c according to Ref. [141] was calculated. The collinearity is defined by the angle between relative transition dipole moments $\theta(\vec{\mu}_{0n}\vec{\mu}_{0m})$ as

$$c = \frac{|\theta(\vec{\mu}_{0n}\vec{\mu}_{0m}) - \frac{\pi}{2}|}{\frac{\pi}{2}} \tag{90}$$

and can yield values between 0 and 1. A value of 0 denotes an angle of 90°, whereas a value of 1 denotes an angle of 0° or 180°. Note, the sign of the transition dipole vector is arbitrary since they are taken from DFT calculations. The variable $\vec{\mu}_{0n}$ denotes the transition dipole moment between the ground state and the excited state n. If the collinearity is close to 1, a stronger coupling of the two states is present. In contrast, a collinearity close to 0 prevents this coupling [213]. For both peaks of the Q-band (Q_y and Q_x) and the two dominant peaks of the Soret-band (S_1 and S_2), the collinearity is given in table 3.2. In a first analysis, only the coupling between the same excited states of different Chls was investigated. As every snapshot contains four Chls bound to subunits CA-CD, two values for c within a dimer (CA-CB and CC-CD) and four values for c between different dimers can be calculated. The latter four can be split in two times two values for c taking into account the dimer of dimer structure. Note, besides the collinearity, the strength of this state coupling depends on the length of the transition dipole moment as well as on the distance between the two Chls.

Table 3.2: Comparison of different optimization potentials: Collinearity c between the transition dipoles of the same excited states and relative intensity I of the dominant peaks in the absorption spectrum of Chl for Chls in the same (s) or in the opposite (o) dimer. The opposite dimer is split in Chl coupling between CA-CC and CB-CD (o_1) as well as between CA-CD and CB-CC (o_2); The relative intensity denotes the, with respect to Q_y, normalized sum of the oscillator strength for the corresponding transition.

		w/o		MM		QM/MM	
		c	rel. I	c	rel. I	c	rel. I
Q_y	s	0.62 ± 0.04	1.00	0.62 ± 0.04	1.00	0.62 ± 0.03	1.00
	o_1	0.79 ± 0.07		0.77 ± 0.05		0.70 ± 0.06	
	o_2	0.67 ± 0.08		0.69 ± 0.04		0.77 ± 0.07	
Q_x	s	0.58 ± 0.21	0.11	0.63 ± 0.08	0.12	0.20 ± 0.07	0.62
	o_1	0.26 ± 0.17		0.27 ± 0.10		0.51 ± 0.05	
	o_2	0.17 ± 0.13		0.16 ± 0.08		0.05 ± 0.03	
S_1	s	0.77 ± 0.15	2.07	0.76 ± 0.07	2.63	0.68 ± 0.16	3.28
	o_1	0.24 ± 0.15		0.25 ± 0.11		0.27 ± 0.08	
	o_2	0.22 ± 0.14		0.20 ± 0.09		0.20 ± 0.07	
S_2	s	0.53 ± 0.14	1.69	0.59 ± 0.04	2.23	0.56 ± 0.09	3.22
	o_1	0.63 ± 0.19		0.71 ± 0.07		0.70 ± 0.11	
	o_2	0.58 ± 0.21		0.69 ± 0.05		0.63 ± 0.16	

Analyzing the collinearity, a value for $c > 0.5$ for most of the cases of coupling within the same dimer is observed. In the case of opposing dimers, a strong dependency on the excited state is obtained. For the Q_y transition, for instance, a strong collinearity around 0.7 is observed. For coupling within S_1, values for c around 0.2 are found. The optimization shows no big influence on the average value of c with the exception of the Q_x peak. On the one hand, c drops from 0.6 to 0.2 within the same dimer, i.e. moving from a strong to a weak coupling. On the other hand, the subunits of the opposing dimer are no longer equivalent. This is due to the fact that in one case c decreases close to 0 denoting a perpendicular arrangement, whereas for the other Chl pair it increases to 0.5. In addition to the change in the collinearity, an increased intensity for Q_x is observed in the case of a QM/MM optimization, which is accompanied by a drop of the intensity in Q_y. By comparing the fluctuations, an increased fluctuation is observed in the case of using no further optimization techniques. As for the spectral broadening, this is also assumed to be caused by a fluctuation of the Chl conformation.

Up to this point, only the coupling of the same excited states of different Chls has been investigated. However, a coupling between different excited states can be possible, if the energy difference is not too large. To determine whether this occurs, the collinearity for the coupling $Q_x \rightarrow Q_y$ and $S_2 \rightarrow S_1$ for the snapshots optimized at a QM/MM level was examined. Additionally, the coupling from S_2 to S_1 via an intermediate excited state S' (brown in figure 3.3) was investigated to determine its influence on the energy transfer. The values for c are reported in table 3.3.

Table 3.3: Collinearity c between the transition dipoles of different excited states based on the snapshots that were optimized at a QM/MM level for Chls in the same (s) or in the opposite (o) dimer. The opposite dimer is split in Chl coupling between CA-CC and CB-CD (o_1) as well as between CA-CD and CB-CC (o_2).

	$Q_x \rightarrow Q_y$	$S_2 \rightarrow S_1$	$S_2 \rightarrow$ S'	S' $\rightarrow S_1$
s	0.35 ± 0.07	0.15 ± 0.11	0.45 ± 0.13	0.27 ± 0.21
o_1	0.56 ± 0.06	0.34 ± 0.12	0.59 ± 0.19	0.35 ± 0.13
o_2	0.23 ± 0.06	0.12 ± 0.12	0.55 ± 0.21	0.15 ± 0.12

For the coupling between the Q-peaks, Q_x and Q_y, a better collinearity is observed compared to the coupling within the Q_x itself, as reported in table 3.2. A differentiation between the two Chls of different dimers can be observed for this coupling as well. The coupling between the two dominant Soret-states S_1 and S_2 shows a low collinearity within a dimer, however, this improves for Chls of different dimers. For the coupling towards an intermediate state S', an even better collinearity is observed, which can then again couple to S_1.

Within the same Soret-state, a comparable collinearity was observed for the coupling between Chls of the different dimers (o_1 and o_2). Thus, the Chl of the opposing dimer could be assumed to be equivalent. For the collinearity between different Soret-states, however, a larger difference can be observed. This difference emphasizes that the Chls of the opposing dimer have to be treated separately, not only within the Q-states but also within the Soret-states.

3.2.3.2 Comparison of Different Modifications

In a next step, it is investigated how different structural modifications influence the spectra. For this, the MD simulation with introduced disulfide bridges between the dimers and the MD simulation containing only 2 Chl molecules were analyzed. Since the QM/MM optimization showed a remarkable effect on the peak position and the collinearity within the Q_x-state, further investigations were conducted at a QM/MM-optimized level. The spectra for the investigated system are shown in figure 3.5. As a reference system the QM/MM-optimized spectrum presented in figure 3.3 is chosen.

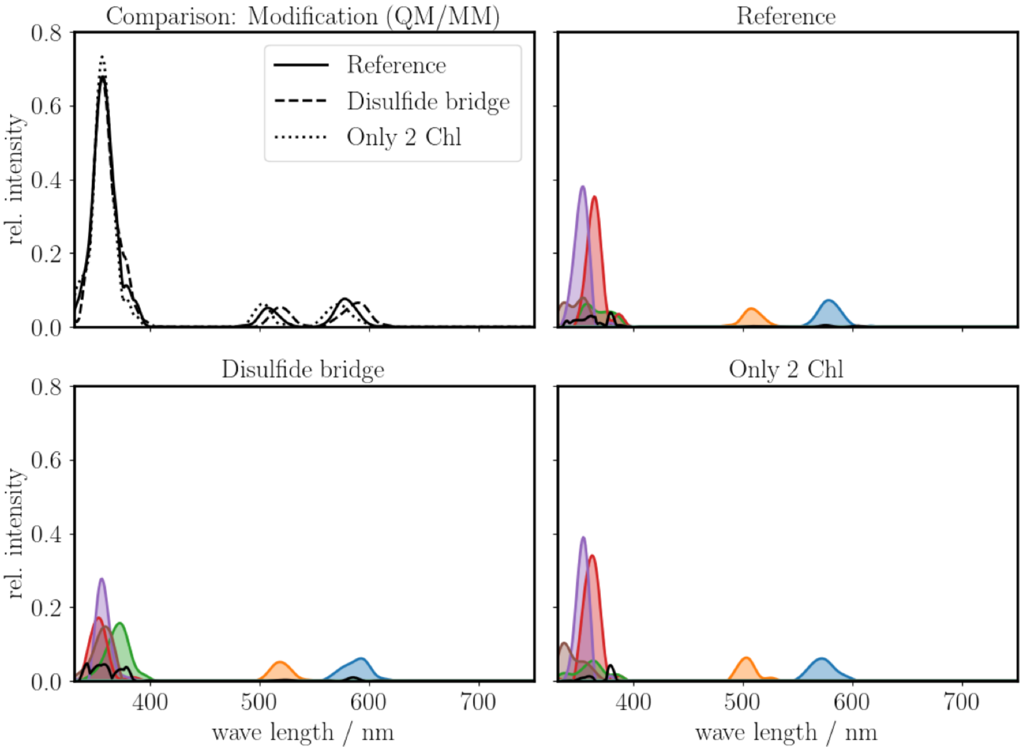


Figure 3.5: Comparison of different modifications: Vertical absorption spectrum of 1 Chl averaged over several snapshots. The overall as well as the detailed absorption spectrum are shown; For every detailed spectrum, the black curve denotes an error function for those excitations which could not be mapped.

The analysis of this figure shows that introducing disulfide bridges has only a minute influence on the overall spectrum. Only a small red shift of the Q-band and small changes in the Soret-band are observed. In the Soret-band, however, four dominant states are detectable now. The reason for this is most likely the change in the angle between the porphyrine planes, as reported in section 3.1. Removing 2 Chl causes a small blue shift of the Q-band. The collinearity is reported in table 3.4.

Table 3.4: Comparison of different modifications: Collinearity c between the transition dipoles of the same excited states and relative intensity I of the dominant peaks in the absorption spectrum of Chl for Chls in the same (s) or in the opposite (o) dimer. The opposite dimer is split in Chl coupling between CA-CC and CB-CD (o_1) as well as between CA-CD and CB-CC (o_2). Since in the case of the system with disulfide bridges between the dimers four dominant peaks in the Soret-band are present, they are labeled S_i and S_i'; The relative intensity denotes the, with respect to Q_y, normalized sum of the oscillator strength for the corresponding transition.

		Disulfides		Only 2 Chl		Reference (QM/MM)	
		c	rel. I	c	rel. I	c	rel. I
Q_y	s	0.56 ± 0.04	1.00	0.67 ± 0.04	1.00	0.62 ± 0.03	1.00
	o_1	0.64 ± 0.08		–		0.70 ± 0.06	
	o_2	0.75 ± 0.07		–		0.77 ± 0.07	
Q_x	s	0.20 ± 0.10	0.65	0.12 ± 0.06	0.72	0.20 ± 0.07	0.62
	o_1	0.56 ± 0.07		–		0.51 ± 0.05	
	o_2	0.07 ± 0.04		–		0.05 ± 0.03	
S_1	s	0.80 ± 0.14	2.20	0.63 ± 0.17	3.57	0.62 ± 0.27	0.98
	o_1	0.25 ± 0.07		–		0.23 ± 0.12	
	o_2	0.22 ± 0.07		–		0.20 ± 0.14	
S_1'	s	0.36 ± 0.23	1.81	–		0.68 ± 0.16	3.28
	o_1	0.35 ± 0.17		–		0.27 ± 0.08	
	o_2	0.28 ± 0.23		–		0.20 ± 0.07	
S_2	s	0.42 ± 0.20	2.25	0.63 ± 0.08	3.02	0.42 ± 0.15	1.21
	o_1	0.44 ± 0.25		–		0.51 ± 0.21	
	o_2	0.41 ± 0.21		–		0.47 ± 0.22	
S_2'	s	0.33 ± 0.13	1.81	–		0.56 ± 0.09	3.22
	o_1	0.52 ± 0.11		–		0.70 ± 0.11	
	o_2	0.27 ± 0.29		–		0.63 ± 0.16	

For a better comparison, the collinearity c for the small peaks in the reference system is also calculated. The system containing disulfide bridges that are linking the dimers shows different values for c in the Q_x-peaks for o_1 and o_2. This is in agreement with the behavior observed in the reference system. Although the Soret-band does not change in the overall spectrum if disulfide bridges are formed, the transition dipole moment and the intensities of the single excitations are affected. The changes are caused by a slightly different conformational orientation of the Chl compared to the reference system. Thus, it is difficult to compare the collinearity as the contribution of the MOs to the transitions in the Soret-band changes as well. Fitting the excitations based on the contribution of the MOs that are involved in the excitation process would link “Reference (QM/MM)” S_1 to “Disulfides” S_1’ and “Reference (QM/MM)” S_1’ to “Disulfides” S_1. For the states S_2 and S_2’ the same relationship is observed. Removing 2 Chls causes a small change in *c*. The trend, however, remains the same compared to the reference system, and the relative intensities are in agreement. It can thus be concluded that the two dimers may interact with each other, but the structural effects on the Chl excitations are local.

3.2.3.3 Comparison of Different Environments

In a next step, it is compared how the environment in the TD-DFT calculation affects the absorption spectrum. For this, the absorption spectra for the chromophore were computed *in vacuo*, neglecting the point charge field (PCF). The spectra are presented in figure 3.6.

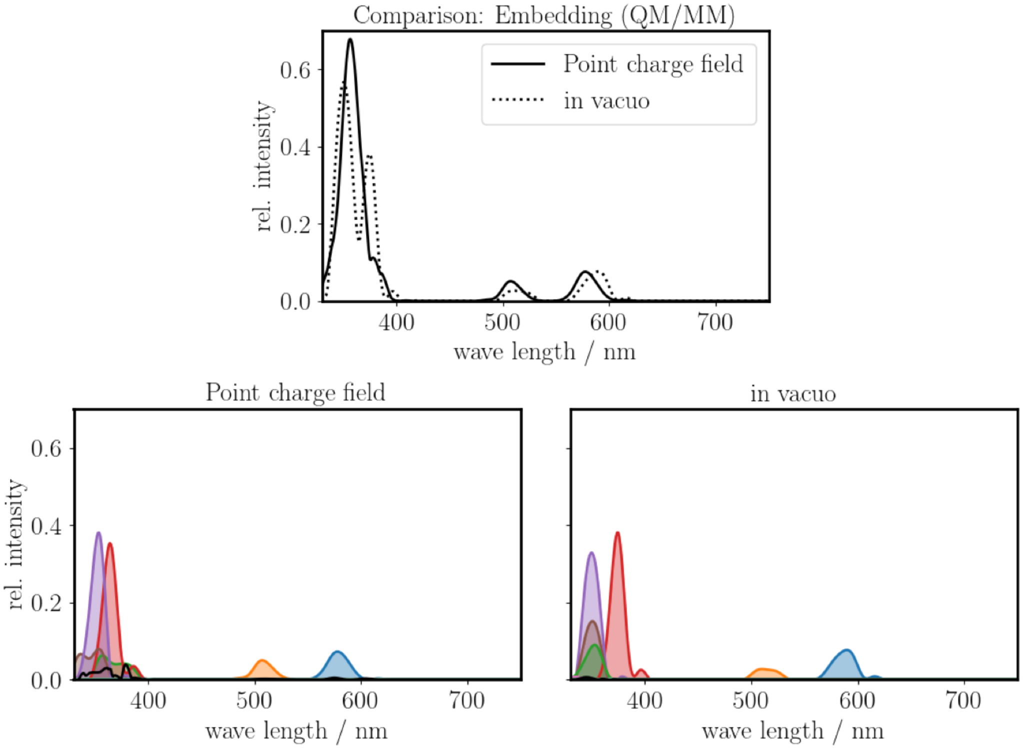


Figure 3.6: Comparison of different environments during the TD-DFT calculation: Vertical absorption spectrum of 1 Chl averaged over several snapshots. The overall as well as the detailed absorption spectrum are shown; For every detailed spectrum, the black curve denotes an error function for those excitations which could not be mapped.

Comparing the spectra, a small red shift of the two Q-peaks and of one dominant Soret-peak is observed. Like for the systems using either no optimization or an MM optimization, two dominant Soret-peaks are present in the overall spectrum now. These separate peaks merge into one using the QM/MM optimization and a PCF environment. Comparing the collinearity as summarized in table 3.5, only minor influence of the environment can be observed. The only significant difference can be found in the collinearity of the transition dipole moments of the S_1-peak within a dimer. Comparing the relative intensities (see table 3.5), a weaker Q_x-peak is observed for the system treated *in vacuo* during the TD-DFT calculation. This indicates that the function of the protein is more likely to be found in the structural arrangement of the Chls than in a direct modification of the spectroscopic properties.

Table 3.5: Comparison of different environments during the TD-DFT calculation: Collinearity c between the transition dipoles of the same excited states and relative intensity I of the dominant peaks in the absorption spectrum of Chl for Chls in the same (s) or in the opposite (o) dimer. The opposite dimer is split in Chl coupling between CA-CC and CB-CD (o_1) as well as between CA-CD and CB-CC (o_2); The relative intensity denotes the, with respect to Q_y, normalized sum of the oscillator strength for the corresponding transition.

		in vacuo		Point charge field	
		c	rel. I	c	rel. I
Q_y	s	0.63 ± 0.04	1.00	0.62 ± 0.03	1.00
	o_1	0.75 ± 0.03		0.70 ± 0.06	
	o_2	0.73 ± 0.05		0.77 ± 0.07	
Q_x	s	0.22 ± 0.04	0.36	0.20 ± 0.07	0.62
	o_1	0.50 ± 0.04		0.51 ± 0.05	
	o_2	0.05 ± 0.03		0.05 ± 0.03	
S_1	s	0.89 ± 0.06	2.95	0.68 ± 0.16	3.28
	o_1	0.21 ± 0.04		0.27 ± 0.08	
	o_2	0.20 ± 0.04		0.20 ± 0.07	
S_2	s	0.53 ± 0.13	2.77	0.56 ± 0.09	3.22
	o_1	0.64 ± 0.17		0.70 ± 0.11	
	o_2	0.60 ± 0.12		0.63 ± 0.16	

In the previous analysis, it was observed that removing the PCF has only little influence on the collinearity for most of the transition dipole moments. Since this only accounts for environmental effects, it is of interest how the transition dipole moments behave with respect to the orientation of the chromophore. In the MD simulation we observed that the angle between the Chls can change slightly over time. Therefore, it is of interest whether these conformational changes influence the angle between the transition dipole moments.

To get an idea of the fluctuation of the transition dipole moments with respect to conformational changes of the chromophores, the angle between the transition dipole moments and the perpendicular of the porphyrine plane is analyzed. For this analysis, 12 perpendicular lines with respect to the position of three out of the four nitrogens of the porphyrine plane are calculated and averaged. In a next step, the angle θ between the transition dipole moments of the states Q_x, Q_y, S_1 and S_2 and the averaged perpendicular is calculated.

For all four states, angles close to 90° were detected. This means that all transition dipole moments are located within or close to the ring-plane. Thus, changes in the porphyrine orientation between two Chl affect the dipole moment orientation to each other. For the Q-peaks, standard deviations around 1°, and for the dominant S-peaks, deviations around 2° were observed. A comparable behavior was observed by calculating this angle, neglecting the PCF (Q_y: 88.7 ± 1.0 °, Q_x: 88.4 ± 1.0 °, S_1: 86.9 ± 1.7 °, S_2: 86.9 ± 3.2 °) with the only difference being a larger fluctuation of the S_2-state.

Apart from the conformational changes, a fluctuation of the transition dipole moment is also possible. To test this, the dihedral angle ϕ between the transition dipole moments of the same excited state was calculated. The magnesium-magnesium distance was chosen as a connection between the transition dipole vectors. The calculated properties are summarized in table 3.6.

Table 3.6: Analysis of the fluctuation between two transition dipole moments of the same excited state. θ denotes the angle between the planes' averaged perpendicular and the transition dipole moment, ϕ denotes the dihedral angle between two transition dipole moments and c denotes the collinearity; The analysis was based on the snapshots that were optimized at a QM/MM level with the QM part being embedded in a PCF during the absorption spectrum calculation.

		θ/°	ϕ/°	c
Q_y	s	88.4 ± 1.1	13.6 ± 7.3	0.62 ± 0.03
	o_1		43.7 ± 9.5	0.70 ± 0.06
	o_2		40.6 ± 15.4	0.77 ± 0.07
Q_x	s	88.2 ± 1.2	80.7 ± 6.8	0.20 ± 0.07
	o_1		44.7 ± 4.1	0.51 ± 0.05
	o_2		29.5 ± 8.4	0.05 ± 0.03
S_1	s	86.4 ± 2.0	29.7 ± 15.3	0.68 ± 0.16
	o_1		62.7 ± 15.5	0.27 ± 0.08
	o_2		57.5 ± 7.4	0.20 ± 0.07
S_2	s	87.6 ± 1.6	12.9 ± 11.2	0.56 ± 0.09
	o_1		35.2 ± 15.8	0.70 ± 0.11
	o_2		54.4 ± 20.6	0.63 ± 0.16

The analysis of the dihedral angle shows that a strong fluctuation of the dihedral angle can be observed. This fluctuation is stronger for the Soret-states than for the Q-states. This is in agreement with the corresponding fluctuation of the collinearity. In summary, changes in the conformational orientation of the rings with respect to each other as well as fluctuations of the dihedral angle between the transition dipoles can influence the coupling strength between two Chls.

3.2.3.4 Comparison of Different Number of Chlorophylls

Since the two Chls within a dimer are strongly excitonically coupled, it is subsequently investigated how the inclusion of both Chls in the TD-DFT calculation affects the absorption spectrum. In addition, it can directly be checked whether coupling of the degenerated excited states of the Chls exist. Since an optimization at the QM/MM-level for two Chls in the active part is computationally expensive to be carried out for multiple snapshots, the focus lies on a discussion based on the MM-optimization. The mapping of the excitations gets more complicated for treating more than one chromophore since degenerated states have to be considered now. A discussion of the mapping for this system can be found following the analysis of the absorption spectrum (figure 3.7).

In figure 3.7, nearly no change in the overall absorption spectrum is observed when two Chls are included in the TD-DFT calculation. The major difference can be found in the Q_y-peak which is split into two peaks with a relative intensity ratio of 0.26. Remarkably, the oscillator strength for the state at lower energy/higher wave length is smaller compared to the state at higher energy. This behavior is not typical. In strongly coupled chromophores of light harvesting complexes, for example, the opposite behavior is observed [195, 214]. Taking into account the detailed spectrum, a coupling of the excited states can be detected for all excitations featuring a high oscillator strength. In those cases, the same behavior as for the Q_y-state is observed with a dominating intensity for the excited state at higher energy. The ratios are: Q_x (orange) 0.61, S_1 (red) 0.42 and S_2 (violet) 0.46. Note, the wave length difference between the peaks does not scale linearly with the energy difference. Analyzing the mean energy difference induced by the coupling results in: Q_y 57 meV and Q_x 21 meV as well as S_1 66 meV and S_2 99 meV. The weak coupling in the Q_x-band is in agreement with a small collinearity as observed for a single Chl optimized at a QM/MM level. This could not be achieved by only using an MM optimization. In future work, it should be investigated how this coupling and the ratios change using a QM/MM optimization for both Chls simultaneously.

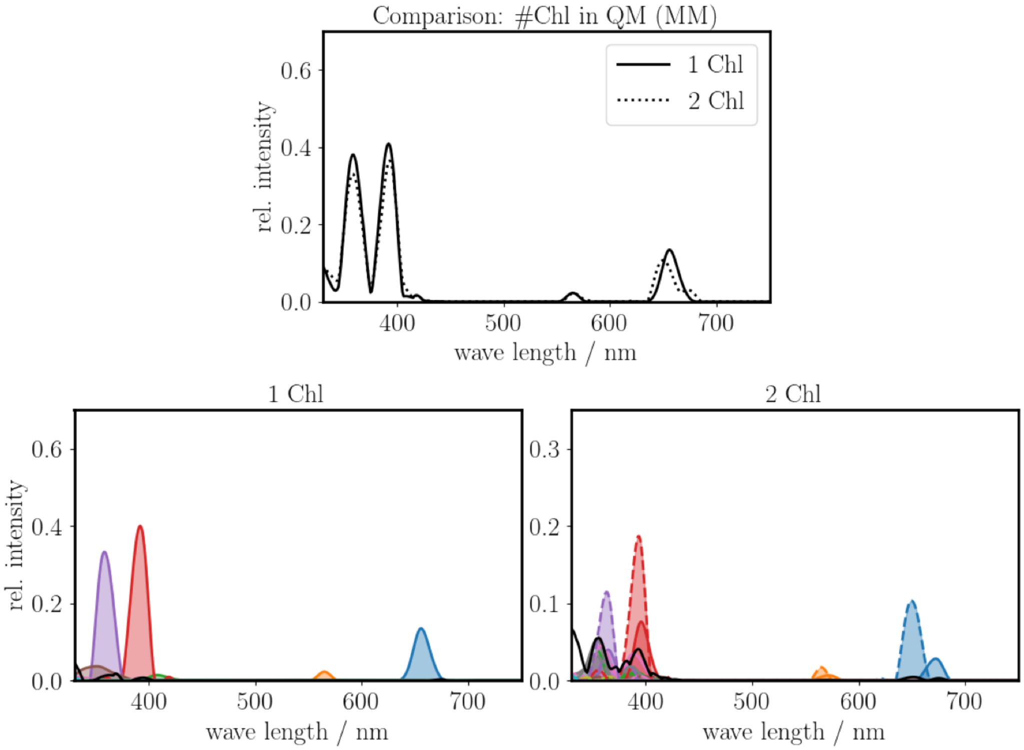


Figure 3.7: Comparison of different numbers of Chls in the TD-DFT calculation: Vertical absorption spectrum of 1 Chl or 2 Chls averaged over several snapshots. The overall as well as the detailed absorption spectrum are shown; Note, due to the degeneracy and the coupling of the excited states including two Chls, the y-axis of the detailed spectrum is scaled down by a factor of 2. Equivalent excitations are highlighted in the same color. The coupled excitation with lower energy is drawn with a solid line, the other with a dashed line; For every detailed spectrum the black curve denotes an error function for those excitations which could not be mapped.

The increase in dimensionality might be a problem for the mapping of the MOs for multiple Chls. For 1 Chl, 855 expansion coefficients are compared by calculating the euclidean distance. This number increases for 2 Chls to 1710. Due to the curse of dimensionality, the distance becomes less significant when doubling the number of dimensions. Another problem arises since degenerated MOs have to be taken into account now. These MOs are no longer independent of each other, since they can show coupling. Furthermore, MOs do not have to be localized at one chromophore but can be found on both.

4 Kinetic Analysis of MD Simulations Based on Core Sets

To extract kinetic information from MD simulations, Markov state models (MSMs) are applied frequently [13, 136, 137, 215, 216]. In these models, the transition probability between different discrete states is estimated. The quality of the MSM, however, depends strongly on the definition of these states. If every simulation frame has to be assigned to a discrete state, the optimal boundary between two states has to be set on top of the energy barrier between these states. However, as MD simulations are performed in a high-dimensional potential energy surface, the exact location of these boundaries is hard to estimate. To counteract this problem, a high number of states is usually necessary [9, 102].

Since the interest usually is not on the exact location of the boundaries, but rather on the metastable states of the investigated molecule, a core-set definition of the discrete states can be applied [12, 14, 107]. A core set denotes an area in which the simulation stays for a long time until a fast transition towards another core set is observed. This definition thus describes metastable conformations of the analyzed system. A general definition of these states, however, was not known.

As presented in section 4.1, the capability of density-based clustering for the definition of these core sets was therefore examined by comparing different density-based cluster algorithms. The extracted core sets were then used for the construction of a core-set MSM (cs-MSM) [12, 14, 107–111] analyzing the dynamics of the simulated systems. Among these, the most promising algorithm, the Common-Nearest-Neighbor (CNN) algorithm, was benchmarked with respect to a variety of different challenges. The results of this study are reported in section 4.2. In the final investigation, presented in section 4.3, the developed workflow (using CNN clustering) was applied to the cyclosporines A and E, simulated in water and chloroform, respectively. In this work the focus lies on a comparison between the two molecules in the two solvents. The importance of certain reaction coordinates for the construction of an MSM is discussed.

4.1 Density-based Cluster Algorithms for the Identification of Core Sets

Although the core-set approach was already known for several years, it was not clear how to find a suitable definition of the core sets. Since core sets account for metastable states and therefore for frequently sampled minima in the potential energy surface, we proposed that a suitable definition of core sets should be obtained using density-based cluster algorithms. For this clustering, it is assumed that every frame can be represented as a data point in this high dimensional data space. The density-based cluster algorithm is then capable of extracting areas, which show a drop in data point density towards other areas, with each area representing a core set.

In the presented work, it is shown that an extraction of these core sets can be done by density-based clustering. In this context, three different algorithms, the Jarvis-Patrick- [129], the DBSCAN- [130] and the Common-Nearest-Neighbor algorithm [16], were compared as every algorithm uses another criterion to determine whether two data points belong to the same cluster. The investigated systems were a two-dimensional potential, the alanine dipeptide as well as a 14-residue β-hairpin peptide, ordered by increasing complexity. For the biomolecular systems, the difference between two conformations was described by an RMSD of the backbone's atom positions. Since the sample frequency and therefore the data point density of a metastable conformation highly depends on its energy, a hierarchical clustering approach was developed that can be used to extract clusters with varying density.

The presented research was published in: Lemke, O.; Keller, B. G. "Density-based cluster algorithms for the identification of core sets", *J. Chem. Phys.* **2016**, *145*, 164104; doi: 10.1063/1.4965440.

4.2 Common-Nearest-Neighbor Clustering – a Benchmark

In section 4.1 it is shown that it is possible to extract core sets from MD simulations applying density-based clustering and to use them to construct suitable cs-MSMs. Thus, the performance of the most-promising algorithm, the Common-Nearest-Neighbor (CNN) algorithm, was evaluated on a series of benchmark data sets. These data sets were designed to exhibit features which are known to cause wrong clustering results in commonly used cluster algorithms. For this investigation, the performance with respect to different properties of data sets was examined. These properties included dimensionality, cluster size, cluster number, cluster shape, overlapping clusters, clusters with different data point density and noise. Lastly, the performance with increasing data set size was investigated. The clustering results were compared to the k-Means++ algorithm [122, 167] and the DBSCAN algorithm [130]. The implemented cluster algorithm is available, see Ref. [217].

The presented research was published in: Lemke, O.; Keller, B. G. "Common Nearest Neighbor Clustering – A Benchmark", *Algorithms* **2018**, *11*, 19; doi: 10.3390/a11020019.

4.3 Comparison of Kinetic Models of the Cyclosporines A and E

4.3.1 Introduction

We set up a workflow for the construction of cs-MSMs based on density-based clustering, as shown in section 4.1. Since the CNN algorithm showed a remarkable performance towards typical challenges related to data sets, as shown in section 4.2, it was included in this workflow. With this workflow at hand, conformational differences between the cyclic peptides cyclosporines A and E are examined in a next step.

Cyclic peptides have gained high interest recently [218–221]. However, they often suffer in bioavailability because of their complexity, size and solubility [222]. Cyclosporine A (CsA), a cyclic undecamer with seven methylated backbone amides, is an exception. It is applied as an immunosupressive drug for kidney and liver transplants [223, 224], since it can bind to the cytosol protein cyclophilin [225] and it consequently diminishes the T-cells' function [226]. The high bioavailability is achieved by passive diffusion through the membrane [44, 45]. It is assumed that CsA can change between "open" and "closed" conformations and can thus adapt to the polarity of its environment [36, 37, 44, 45]. Furthermore, it is assumed that "congruent" conformations exist, which are soluble in both the cytosol and the membrane [36, 37]. The reason for the diffusion is thus located in the dynamics of the molecule. By removing one backbone methylation, the derivative cyclosporine E (CsE) is obtained, whose membrane permeability is one order of magnitude lower than that of CsA. Sticking to the assumptions concerning the membrane diffusion, this worse membrane permeability is caused by changes in the molecule's kinetics. Utilizing MD simulations of both molecules, in water and chloroform respectively, as well as full-partitioning MSMs (fp-MSMs), these changes were discussed in Ref. [36, 37].

In Ref. [36, 37], several distinct conformations, including "congruent" conformations, could be observed. For CsE, one "congruent" conformation that was observed for CsA was missing. In addition, higher conformational conversion times for CsE were observed compared to CsA which are assumed to be one reason for the worse membrane permeability of CsE [36, 37]. Within these studies, a relatively fast *cis*-*trans*-isomerization of the 9-MeLeu-10-MeLeu peptide bond compared to other conformational changes was observed for some cyclosporine-solvent combinations. This seems to be very unlikely since a *cis*-*trans*-isomerization along a partial double bond is linked to a transition above a high energy barrier and would therefore be suggested to occur on slower timescales than other conformational changes. An acceleration of this configurational change might come into play due to the ring structure, since it can tense the molecule.

The work presented in this section is an extension of these former studies in collaboration with Dr. Jagna Witek and Prof. Dr. Sereina Riniker (ETH Zurich, Switzerland) who provided the simulation trajectories.

The work can therefore be set into relation to these former studies as the same trajectories were analyzed. Applying the developed workflow, cs-MSMs were constructed for CsA and CsE in water and chloroform respectively. For the construction of the cs-MSMs, different sets of input coordinates were tested. Both molecules show the ability of a configurational isomerization of the peptide bond between MeLeu-9 and MeLeu-10, changing between a *cis*- and a *trans*-configuration. The effect of including or neglecting this information in the cs-MSM construction was investigated, since the dihedral angle of the peptide bond ω is usually assumed to be fixed.

Since the clustering of the MD simulation data is hard to realize in a high dimensional state space due to the curse of dimensionality [101], i.e. the higher the dimensionality is, the worse it gets to find a good discretization, the dimensionality has to be reduced. To achieve this dimensionality reduction, the time-lagged independent component analysis (TICA) [112, 117] was included in the workflow. In this analysis all slow degrees of freedom are provided as input data and the linearly optimized reaction coordinates that show the highest time-lagged variance are extracted as an outcome. The big advantage is that further information can be included, which is extractable from the full state space. This information can include, for example, hydrogen bonding, atom-atom distances or other structural information. Recently, a combination of TICA and MSMs was applied to obtain kinetic information of different biomolecules [43, 137, 227]. Besides the construction of cs-MSMs, joint spaces were set up utilizing TICA [112, 117] for the direct comparison of either both molecules in one solvent or of one molecule in two different solvents. Using this kind of analysis, it is possible to characterize unique conformations as well as “congruent” states.

The cs-MSMs are compared to fp-MSMs, which are constructed with the use of the commonly applied k-Means algorithm and its variation k-Means++[11, 15, 122, 167]. Full-partitioning discretizations often suffer from a recrossing problem by “badly” assigned borders [12] and thus the quality of the MSM is often diminished. To check this quality, for example, the implied timescales obtained by the MSM can be examined with respect to convergence, which is sometimes only achieved on a logarithmic scale [9, 36, 37, 228, 229].

4.3.2 Methods

4.3.2.1 Simulation Details

In total, 4 setups were simulated: CsE and CsA in either water or chloroform. The simulation data for CsA were taken from Ref. [36], and for CsE from Ref. [37]. The simulations were performed using the GROMOS package [230, 231]. As a force field, GROMOS 54A7 was applied [232]. The simulations were carried out in an NPT ensemble at a temperature $T = 300$ K using periodic boundary conditions and a time step of 2 fs. Solute coordinates were written to file every 5 ps. For the solvents, the SPC water model [233] and the GROMOS chloroform model [234] were used. The system was coupled to a temperature bath with a coupling constant of 0.1 ps [157]. The pressure was set to 1 atm and kept constant using a pressure bath with a coupling constant of 0.5 ps. Non-bonded interactions were calculated using a twin-range cutoff scheme (0.8 and 1.4 nm). Bond lengths were constrained using the SHAKE algorithm [154].

To generate starting structures for the production simulations, enhanced sampling techniques were applied. For detailed information, it is referred to Ref. [36] and Ref. [37]. Starting from 100 different starting conformations (196 for CsE in water), each structure was simulated for 100 ns resulting in a total simulation length of 10 μs (19.6 μs for CsE in water) per setup.

4.3.2.2 Reaction Coordinates and State Space Reduction

For the analysis, the focus was laid on three main reaction coordinates: The backbone dihedral angles ϕ and ψ, hydrogen bonds and the dihedral angle of the 9-10-peptide bond $\omega_{9,10}$. Backbone dihedral angles and hydrogen bonds were isolated using the GROMOS software package [230, 231]. For the hydrogen bonds, a cutoff distance of 2.5 Å and an angle cutoff of 135° were applied. The dihedral angle of the 9-10-peptide bond was extracted using the *Python* package *MDTraj* [18]. For the state space reduction, the dihedral angles were shifted to the interval $]0°, 360°]$ and normalized to the interval $]0, 1]$. For the 9-10-peptide bond, its absolute values were used and also normalized to the interval $]0, 1]$. The hydrogen bonds were translated into binary information. Each hydrogen bond is represented as a single array with "1" if the hydrogen bond is present, and "0" if not.

The input data were processed by time-lagged independent component analysis (TICA) using the *pyEMMA* package [15]. The lag time was fixed to 5 ns as for this lag time a convergence of the implied timescales was observed. For the research presented in this section, four different TICA-spaces were constructed for each system: Space 1 included only the backbone dihedral angles without the dihedral angle of the 9-10-peptide bond. For Spaces 2 and 3, the hydrogen bonds and the dihedral angle of the 9-10-peptide bond were added. For Space 3, all replicas which contain at least one conformation with a 9-10-peptide bond in *cis*-configuration were rejected. The results for Spaces 1 to 3 are presented in section 4.3.3.1. For Space 4, a combined TICA-space of two systems (only those replicas where the 9-10-peptide bond is in *trans*-configuration the whole simulation time), sharing either the solvent or the molecule, was constructed (section 4.3.3.3). To simplify the differentiation of the investigated reduced state spaces, the spaces are summarized according to table 4.1.

Table 4.1: Names and included information of the different setups for the reduced state space construction.

Name	Information included in reduced space construction	Section
Space 1	backbone dihedral angles ϕ, ψ	4.3.3.1
Space 2	Space 1 + hydrogen bonds + 9-10-peptide bond $\omega_{9,10}$	
Space 3	Space 2 - trajectories containing *cis*-states	
Space 4	joint space of two Space 3	4.3.3.3

For the Spaces 1–3, the first 5 time-lagged independent components (TICs) were used for further analysis. For Space 4, 7 TICs were used. Using more TICs was tested as well, but showed no significant changes. In addition, with increasing dimensionality of the reduced space some connectivity issues can come into play, which will be explained in detail in section 4.3.3.1.

4.3.2.3 Core-set Markov State Models

The trajectories projected on the reduced state spaces were clustered using the CNN algorithm [217]. The data set for the clustering was reduced by extracting every 200th frame for Spaces 1–3 (section 4.3.3.1). For Space 4 (section 4.3.3.3), where two systems were combined into one space, the stride was reduced to 150 for CsE in chloroform and increased to 350 for CsE in water to guarantee equally sized parts within the resulting data set.

As a distance metric for the clustering, the euclidean distance was applied. For each setup, a screen over 80 different parameter sets was performed (10 R values and 8 N values). This is possible as the calculation of the distance matrix, which is the most expensive step, is only done once. All values of R were set to be located before the first maximum of the distance distribution of the data set [16]. N was increased from 2 to 30. The parameter M was fixed to 10 for all data sets smaller or equal than 10,000 data points. For CsE in water as well as for the combined spaces, M was increased to 20. The outcome of every clustering step was ranked empirically according to three properties: the number of clusters N_{C}, the percentage of the largest cluster p_{l} and the percentage of data points assigned to noise p_{n} according to

$$\sqrt{N_{\mathrm{C}}} \cdot \exp\left(-\frac{(p_{\mathrm{l}}-\mu_{\mathrm{l}})^2}{2\sigma_{\mathrm{l}}^2}\right) \cdot \exp\left(-\frac{(p_{\mathrm{n}}-\mu_{\mathrm{n}})^2}{2\sigma_{\mathrm{n}}^2}\right) \tag{91}$$

with $\mu_{\mathrm{l}} = 0.25$, $\mu_{\mathrm{n}} = 0.25$, $\sigma_{\mathrm{n}} = 0.2$ and

$$\sigma_{\mathrm{l}} = \begin{cases} 0.1; & \text{for } p_{\mathrm{l}} < \mu_{\mathrm{l}} \\ 0.2; & \text{for } p_{\mathrm{l}} \geq \mu_{\mathrm{l}} \end{cases} .$$

For the chosen parameter sets, all data points were projected back onto the clusters with respect to the cluster parameters. A data point is assigned to a cluster, if it has at least N neighbors in the reduced data set within a cutoff distance R. The drawback of this approach is that data points of the reduced data set can be rejected from the cluster since some neighbors in the reduced data set are declared as noise. However, this happens only to the border points of the cluster and can therefore be neglected. The parameter sets chosen to construct the cs-MSMs are summarized in table 4.2.

Table 4.2: Parameters and outcome of the CNN clustering.

System	Solvent	Space	cis?	Data set size	Parameter			Outcome	
					R	N	M	#Cluster	Noise / %
CsE	$CHCl_3$	1	yes	10,000	0.20	5	10	10	34
		2	yes	10,000	0.25	10	10	13	16
		3	no	6,700	0.15	3	10	11	32
	H_2O	1	yes	19,600	0.20	2	20	6	15
		2	yes	19,600	0.20	3	20	6	11
		3	no	15,800	0.20	5	20	7	32
CsA	$CHCl_3$	1	yes	10,000	0.35	3	10	12	9
		2	yes	10,000	0.50	3	10	13	1
		3	no	9,600	0.35	5	10	6	7
	H_2O	1	yes	10,000	0.25	5	10	16	26
		2	yes	10,000	0.25	5	10	17	26
		3	no	9,300	0.25	5	10	14	28
CsE	Both	4	no	18,142	0.25	2	20	19	25
CsA		4	no	18,900	0.25	5	20	10	54
Both	$CHCl_3$	4	no	18,578	0.25	3	20	21	22
	H_2O	4	no	18,464	0.30	3	20	17	28

Using the resulting discrete trajectories, cs-MSMs were constructed according to Ref. [12, 135]. For the construction of the cs-MSMs, the detailed balance criterion was applied by counting each transition from C_i to C_j also as a transition from C_j to C_i [11]. The analysis was focused on the largest connected set rejecting all disconnected states. For CsE in chloroform, one cluster for the cs-MSM was removed manually, as it was only connected by one transition leading to an implied timescale much larger than the total simulation length of 10 μs and therefore to a statistical error. In cs-MSMs, every trajectory is truncated at the position of the first and the last assigned frame. As the trajectories are relatively short (20,000 frames) and not all frames are assigned to a discrete state, a minimal length of the truncated trajectories was set to $10 \cdot \Delta t$. Δt denotes the time step for the construction of the MSM. All trajectories which were smaller were rejected.

For all systems, the cs-MSM was constructed for lag times τ in the range of 0.5 ns to a maximum of 50 ns (if not limited by the shortest discrete state trajectory). Δt was set to 0.5 ns. For every lag time τ, the corresponding implied timescale (equation 42) was calculated and plotted in dependency to the lag time. Since a valid MSM has to be independent of the lag time, the region in which the implied timescales are converged was investigated further. The chosen lag times and further parameters are summarized in table 4.3.

Table 4.3: Parameters for the construction of the fp-MSM and the cs-MSM with the maximal evaluated lag time τ_{Max} and the lag time used for the model construction τ_{Mkv}. For Space 4: The first row denotes the MSM parameters for $CHCl_3$ or for CsE, respectively.

System	Solvent	Space	cis?	cs-MSM		fp-MSM	
				τ_{Max} / ns	τ_{Mkv} / ns	τ_{Max} / ns	τ_{Mkv} / ns
CsE	$CHCl_3$	1	yes	46.0	15.0	50.0	15.0
		2	yes	21.0	5.0	50.0	30.0
		3	no	50.0	20.0	50.0	30.0
	H_2O	1	yes	34.0	9.0	50.0	15.0
		2	yes	50.0	10.0	50.0	20.0
		3	no	36.0	10.0	50.0	20.0
CsA	$CHCl_3$	1	yes	50.0	20.0	50.0	15.0
		2	yes	50.0	20.0	50.0	30.0
		3	no	35.5	20.0	50.0	20.0
	H_2O	1	yes	21.5	20.0	50.0	15.0
		2	yes	29.5	20.0	50.0	30.0
		3	no	29.5	15.0	50.0	30.0
CsE	Both	4	no	50.0	20.0	–	–
				36.0	5.0	–	–
CsA		4	no	13.0	10.0	–	–
				15.0	10.0	–	–
Both	$CHCl_3$	4	no	50.0	20.0	–	–
				24.0	20.0	–	–
	H_2O	4	no	38.0	10.0	–	–
				36.0	20.0	–	–

4.3.2.4 Full-partitioning Markov State Models

Fp-MSMs were constructed as reference systems. For every setup, the reduced state space was divided into 500 clusters using the k-Means++ algorithm that is implemented in the *pyEMMA* package [15]. The MSM was constructed like the cs-MSM to provide an easier comparison. The eigenvalue spectrum was calculated for 50 ns with a time step of 0.5 ns. As for the cs-MSM, the lag time interval, in which the implied timescales were converged, was used for further analysis. The parameters for the fp-MSM are summarized in table 4.3.

For an easy interpretation of the outcome of the MSM (and the cs-MSM), the discrete states were lumped into metastable states using the PCCA+ algorithm [235] that is implemented in *EMMA* [105]. The number of metastable states was chosen manually by the number of dominant implied timescales.

4.3.3 Results

The importance of I) the input data used in the reduced space construction (section 4.3.3.1), and of II) the discretization method (section 4.3.3.2) is investigated with respect to the reproduction of the dynamics of CsA and CsE observed in the MD simulations. For I), the main focus is set on the explicit treatment of the dihedral angle formed by the 9-10-peptide bond ($\omega_{9,10}$), as these dihedral angles are usually neglected in the analysis of MD data. For an easier understanding, conformations with *cis*-configuration are called *cis*-states and conformations with *trans*-configuration are called *trans*-states analogously. For II), full-partitioning and core-set discretizations within these different spaces are compared. For every model, the dynamics of CsA and CsE in water as well as in chloroform are investigated. A full-partitioning discretization is obtained using the k-Means++ algorithm. For a core-set discretization, the CNN algorithm is used instead. By using a joint reduced space, two systems are compared in a last step (section 4.3.3.3).

As a reference for an accurate model, four criteria are defined:

1. The implied timescales of the constructed MSM are converged (on a linear timescale), indicating a sufficient assignment of the discrete states.
2. The isolated *cis*-states are “pure” and do not contain any conformation in *trans*-state.
3. The lumping of discrete states to metastable states using PCCA+ does not merge clusters in *cis*-states to *trans*-states assuring a slow *cis*-*trans*-isomerization.
4. The kinetics for the *cis*-*trans*-isomerization are comparable to those observed in the MD simulations.

For the last point, the maximal and minimal absolute dihedral angle of the 9-10-peptide bond was analyzed for every replica (figure 4.1). This analysis showed that for CsE and CsA in chloroform only in one replica a transition is observed, switching between a *cis*- and a *trans*-state. For CsE in water, in no replica a complete *cis*-*trans* transition (180° flip) is observed. CsA in water shows two transitions between *cis*- and *trans*-configurations. In replica 14 the transition occurs at the end of the simulation containing only a few *cis*-states. The coloring of the trajectories containing *cis*-states is according to the dynamical connectivity. Except for CsA in water, the dynamically connected *cis*-states (highlighted in red) are for all systems "open" ones, the dynamically disconnected conformations are "closed" ones (highlighted in yellow). For CsA in water, the opposite behavior is observed. In summary, for CsE in water, no transition between a *cis*- and a *trans*-state is expected. For the other three systems, the transition should be in the μs regime or close to it.

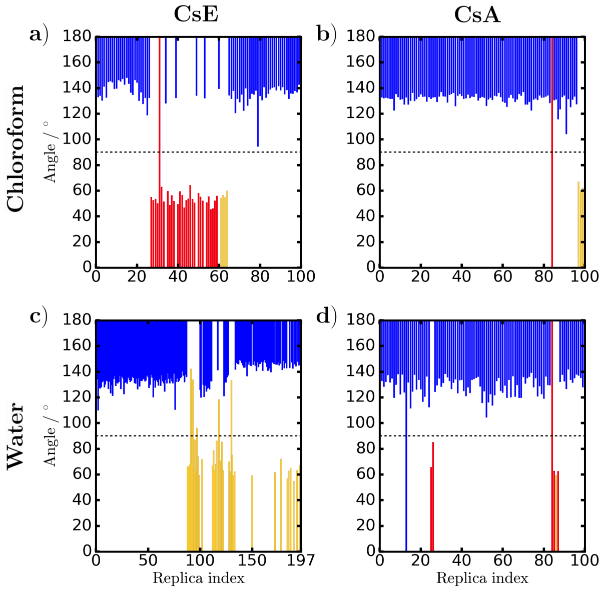


Figure 4.1: Minimal and maximal dihedral angle of the 9-10-peptide bond for every replica for CsE (a) in chloroform and (c) in water as well as CsA (b) in chloroform and (d) in water; Categorization of the replicas (according to figures 4.4 and 4.5): Replicas that stay only in *trans*-state and are not connected to any isolated *cis*-state (blue), that contain connected *cis*-states (red), that contain isolated *cis*-states (figure 4.6) (yellow).

4.3.3.1 Choice of Input Coordinates Using a Core-set Discretization

Dihedral angles ϕ and ψ without a treatment of the 9-10-peptide bond explicitly

To investigate the influence of the input data for the reduced state space construction on the outcome of the MSM analysis, a reduced space was constructed on the basis of the backbone angles ϕ and ψ (Space 1) first. The ω dihedral angle along the peptide bond is usually assumed to be fixed by the partial double bond character and therefore often neglected. Hence, the dihedral angle of the 9-10-peptide bond, for which an impact on the dynamics was already reported, was not included in the reduced space construction.

The first analysis focused on CsE analyzing the dynamics in chloroform (figure 4.2a) and water (figure 4.2b) using a core-set discretization. For both systems, it is possible to isolate core sets, which dominantly contain *cis*-states without treating the 9-10 peptide-bond in the state space construction explicitly. This conformation is in both solvents a "closed" conformation. In addition, in chloroform an "open" conformation is achieved. The overall purity of these *cis*-states is 99.9% in chloroform and 99.9% in water.

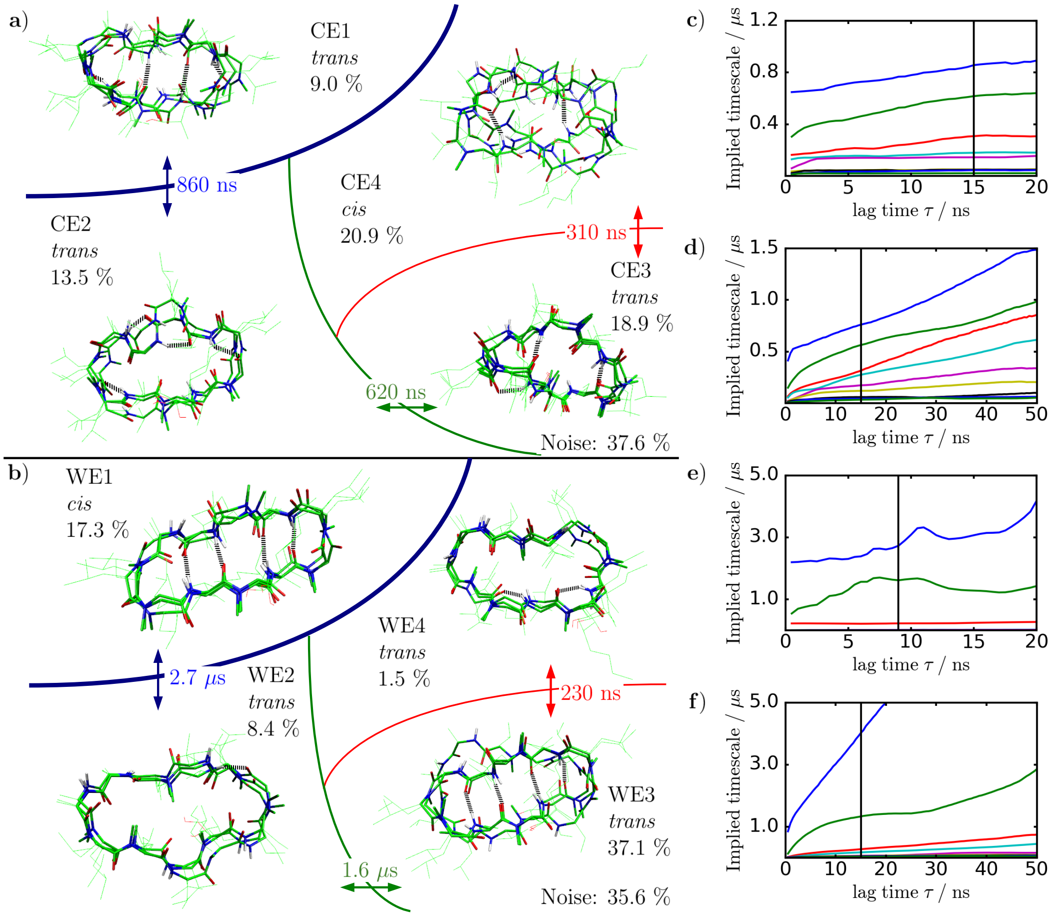


Figure 4.2: Dynamics of CsE (a) in chloroform and (b) in water (Space 1); The configuration of the 9-10-peptide bond is highlighted; The populations are calculated with respect to the clusters (they do not sum up to 100 %). Implied timescales in dependence of the lag time τ: CsE in chloroform for (c) cs-MSM and (d) fp-MSM; CsE in water for (e) cs-MSM and (f) fp-MSM. The fp-MSMs were constructed using the k-Means++ algorithm; The vertical line represents the lag time τ that was used for the analysis.

To examine whether this purity is sufficient enough to reproduce the dynamics that were observed in the MD simulation, the quality of the cs-MSMs based on these core sets is investigated in a next step. Therefore, the convergence of the implied timescales is analyzed first.

The well converged implied timescales of the cs-MSMs (figure 4.2c+e), which only show a small drift on a linear scale, indicate a sufficient assignment of the discrete states. However, the *cis*-states for CsE in chloroform ("open" and "closed") are merged into the same metastable state using the PCCA+ algorithm indicating a fast dynamical exchange between both conformations. This does not match the behavior observed in the MD simulation, as described above. In water, the PCCA+ cluster that contains *cis*-states is formed by only one core set, which strengthens the assumption that the assignment suffices the expectations for this system. In both solvents, a *cis*-*trans*-isomerization of the 9-10-peptide bond can be detected, which was also observed by Witek *et al.* [37]. However, in chloroform this *cis*-*trans*-isomerization is on a relatively fast timescale around 300 ns, which again does not match the findings described above, although a purity of 99.9% was achieved. In water, where no *cis*-*trans*-isomerization should occur according to the former analysis, a *cis*-*trans*-isomerization around 3 μs is observed. The other dynamic modes represent opening and closing movements of *trans*-states. Although the implied timescales indicate a good discretization, the behavior observed in figure 4.1a+c could not be reproduced by a cs-MSM constructed on a reduced space in which the *cis*-*trans*-isomerization of the 9-10-peptide bond was neglected. Hence, the model for CsE has to be improved.

The full analysis (figure 4.3) was repeated for the dynamics of CsA in Space 1 using a core-set discretization. In water, a "closed" *cis*-state was isolated with a purity of 97.8 % using the core-set discretization. In chloroform, an "open" (purity of 96.6 %) as well as a "closed" *cis*-state (purity of 63.0 %) were observed. The purity of the former two core sets is in the same range as for CsE. Therefore, it is examined again how this impurity affects the MSM analysis.

As for CsE, well converged implied timescales are observed on a linear scale. However, due to the poor purity of the "closed" *cis*-state in chloroform, it is not directly involved in the slow dynamics and is merged in the PCCA+ assignment into *trans*-states. The "open" *cis*-state forms an own PCCA+ cluster. A timescale close to the μs regime is obtained matching the expected dynamics. In water, the *cis*-*trans*-isomerization is again on a fast timescale, which is not reproducing the behavior described in the reference. Nevertheless, the *cis*-states form an own PCCA+ cluster as in the case of CsE in water, suggesting a sufficient discrete state definition.

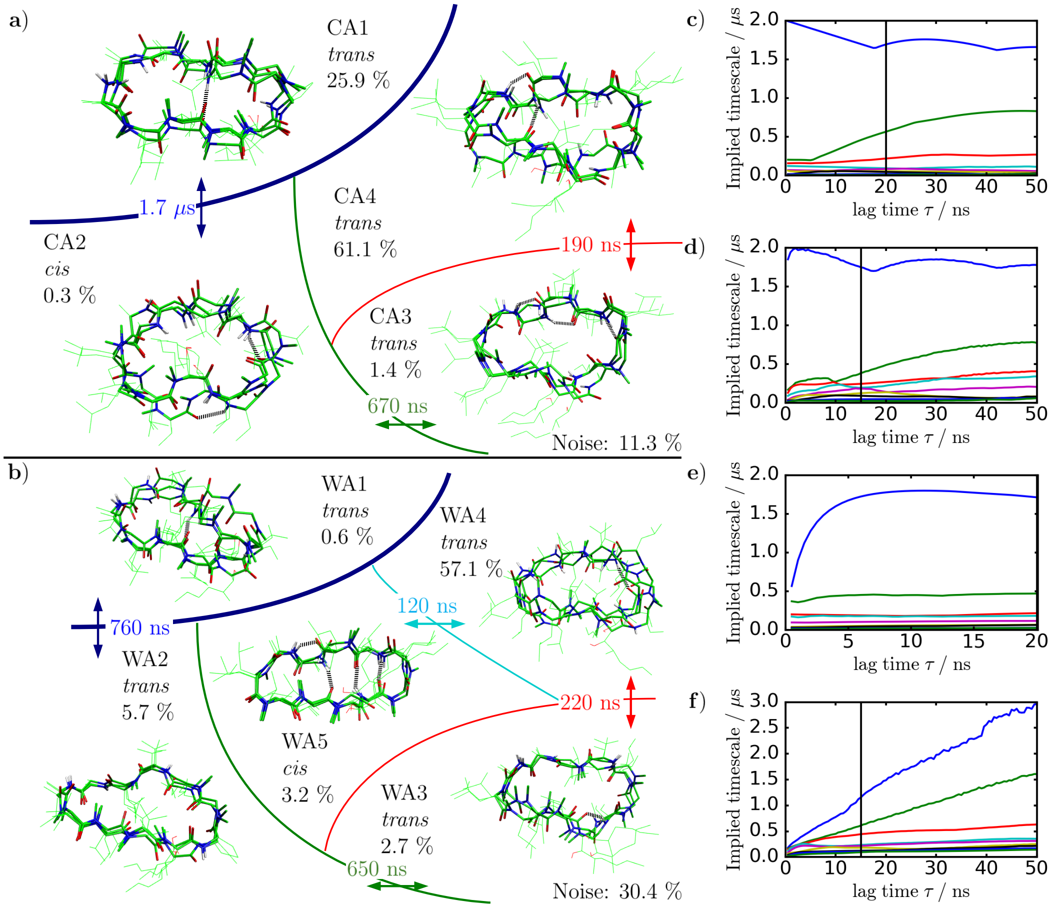


Figure 4.3: Dynamics of CsA (a) in chloroform and (b) in water (Space 1); The configuration of the 9-10-peptide bond is highlighted; The populations are calculated with respect to the clusters (they do not sum up to 100 %). Implied timescales in dependence of the lag time τ: CsA in chloroform for (c) cs-MSM and (d) fp-MSM; CsA in water for (e) cs-MSM and (f) fp-MSM. The fp-MSMs were constructed using the k-Means++ algorithm; The vertical line represents the lag time τ that was used for the analysis.

For all investigated systems in Space 1, it was not possible to reproduce the dynamics of the MD data using a core-set discretization. To evaluate the error source, another discretization method is tested first. Therefore, the core-set discretization is compared to a full-partitioning discretization. For the full-partitioning discretization, it is again possible to isolate metastable states that dominantly contain *cis*-states. The purity of these states is slightly lower than for the core-set discretization, with 98.1 % for CsE in chloroform and 99.7 % for CsE in water. For CsA in chloroform, a purity of 67.9 % and for CsA in water, a purity of 91.0 % was achieved. The purities were averaged over all metastable states mainly containing *cis*-states. For the implied timescales of the fp-MSM (figure 4.2d+f), also a comparable behavior is observed. Thus, the discretization method is not the error source for the wrong dynamical behavior that is obtained by the MSMs. Hence, the error source is likely to be found in the reduced space construction and therefore in the used input coordinates. To test this, more specific coordinates have to be included in the reduced space construction such as the dihedral angle of the 9-10-peptide bond or the hydrogen bond network. A detailed discussion and further comparison between core-set and full-partitioning discretization as well as their influence on the quality of the MSMs will be covered in section 4.3.3.2.

Treating the 9-10-peptide bond explicitly

To examine whether the reduced space can be improved including more specific input coordinates, the absolute value of the dihedral angle of the 9-10-peptide bond as well as hydrogen bonds were added into the reduced space construction (Space 2). As for Space 1, it is possible to extract *cis*-states using a core-set discretization. However, now a purity of 100 % is achieved for all 4 investigated systems. All *trans*-states show a purity of 100 % as well.

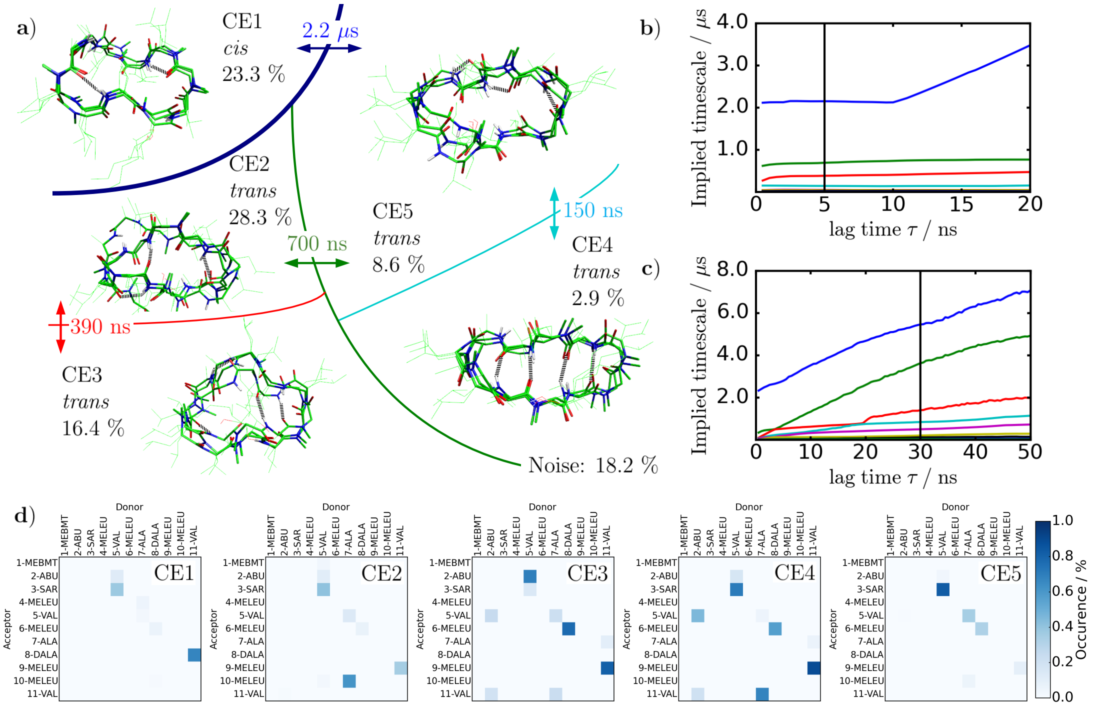


Figure 4.4: Dynamics of CsE (a) in chloroform (Space 2); The configuration of the 9-10-peptide bond is highlighted; The populations are calculated with respect to the clusters (they do not sum up to 100 %); (b-c) Implied timescales in dependence of the lag time τ using different cluster algorithms: (b) CNN and (c) k-Means++; (d) Hydrogen bond population.

To evaluate the effect of the 100 % purity of the isolated *cis*-states, the reproduction of the reference data by a cs-MSM is examined in the next step. As for Space 1, the dynamics of CsE in chloroform and water are analyzed first. The 100 % purity leads to a big improvement of the quality of the model. On the one hand, a better convergence to higher implied timescales in comparison to figure 4.2c+e can be observed in both solvents. On the other hand, a better reproduction of the dynamics as described above is obtained.

For CsE in chloroform, the *cis*-*trans*-isomerization is on the slowest implied timescale around 2 μs (figure 4.4a+b) now, which is much higher compared to the cs-MSM in Space 1. Also the PCCA+ clusters show a 100 % purity (i.e. core sets containing *cis*-states are neither merged into *trans*-states nor into other *cis*-states showing a different conformation). At this point, it has to be mentioned that the dynamically connected *cis*-conformation (CE1) is an "open" one (present in the red highlighted replicas in figure 4.1a). The "closed" conformation was isolated by the CNN algorithm as well, but it is not found to be dynamically connected in the cs-MSM, which again matches the MD data. For the *trans*-states, "open" as well as "closed" conformations are found with dynamic modes occurring on timescales of hundreds of ns.

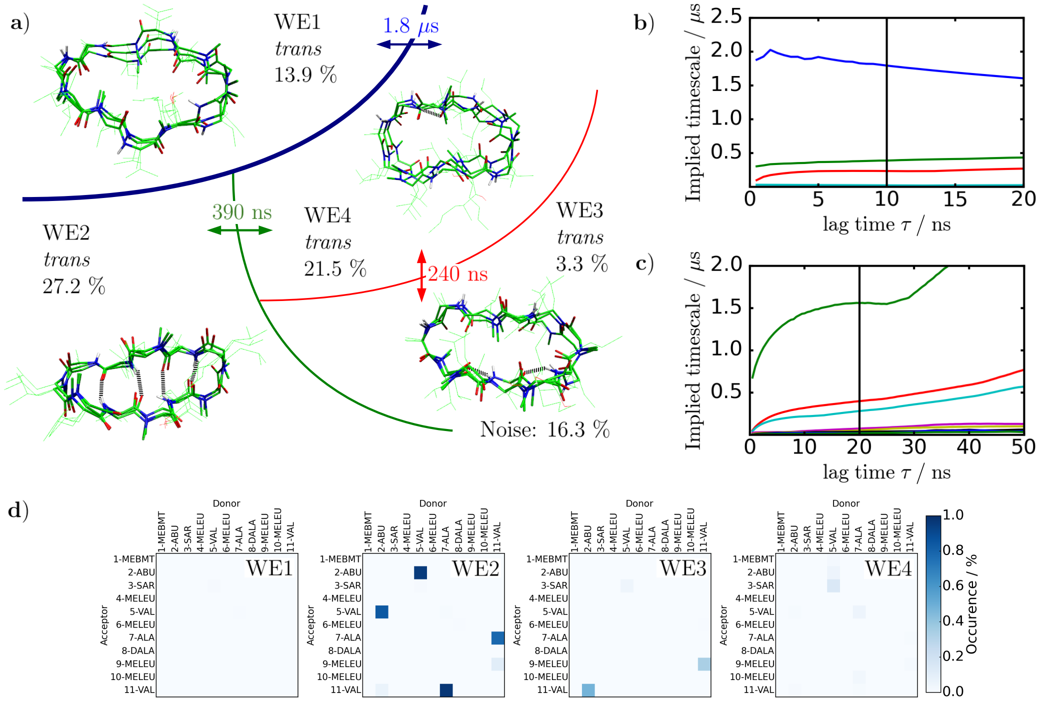


Figure 4.5: Dynamics of CsE (a) in water (Space 2); The configuration of the 9-10-peptide bond is highlighted; The populations are calculated with respect to the clusters (they do not sum up to 100 %); (b-c) Implied timescales in dependence of the lag time τ using different cluster algorithms: (b) CNN and (c) k-Means++; (d) Hydrogen bond population.

For CsE in water, a comparable behavior of the improved purity as in chloroform is observed (figure 4.5). The *cis*-*trans*-isomerization is not present anymore, matching the observations of figure 4.1c. The "closed" *cis*-state was isolated by the CNN algorithm and is, as in chloroform, dynamically not connected to other core sets of the cs-MSM. In addition, a highly populated, "closed" *trans*-state is observed showing a higher purity than observed in Space 1. This shows that including more specific coordinates into the reduced space construction does not only improve the *cis*-*trans*-isomerization but also other dynamic modes.

Similar analyses for CsA in chloroform and water were performed (supporting figures: figure 4.12 and figure 4.13), but will not be discussed in detail, since the same behavior as for CsE in Space 2 is observed. A direct comparison of CsE and CsA is made in section 4.3.3.3.

For all setups in Space 2 (CsA/CsE in chloroform/water), one dynamically disconnected *cis*-state is observed (figure 4.6) using a core-set discretization. The corresponding replicas are highlighted in yellow in figure 4.1. For 3 out of 4 setups, this conformation is a "closed" one. Hence, a high stability of this conformation can be assumed. This is in agreement with NMR measurements, where a stable, "closed" *cis*-state was characterized for CsA and CsE in chloroform [36, 37], as well as with the behavior directly observed within the MD data. Only for CsA in water, the "closed" *cis*-conformation shows a transition towards the *trans*-states in the model. Note, for the "closed" *cis*-state one additional hydrogen bond between Val-11-NH and DAla-8-CO can be observed for CsE, which is not present in CsA due to the backbone methylation at Val-11. Based on this dynamical disconnection an insufficient sampling can be assumed. Hence, for a statistical treatment the sampling has to be improved.

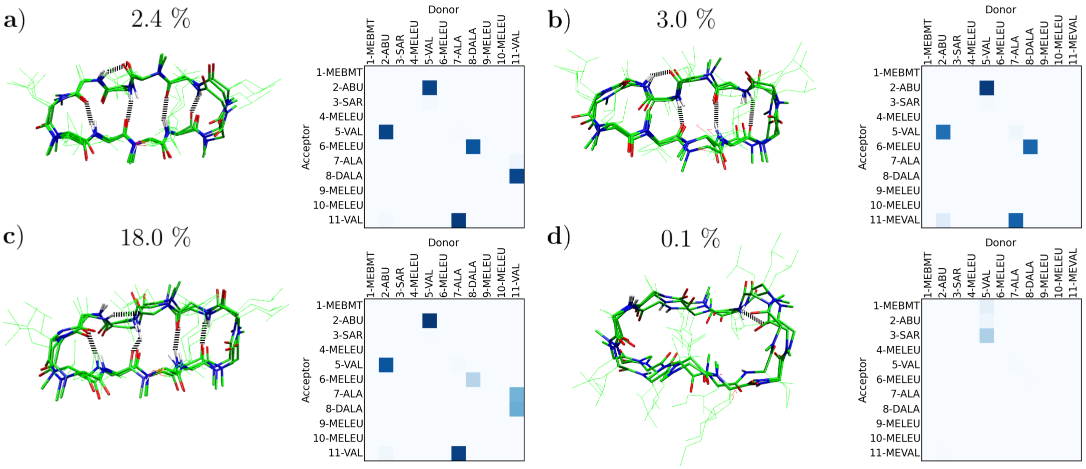


Figure 4.6: Superimposed disconnected *cis*-states (structure and hydrogen bond population) for CsE (a) in chloroform and (c) in water and CsA (b) in chloroform and (d) in water; The population with respect to the CNN clustering is highlighted.

Excluding trajectories containing *cis*-states

Since a significant statistical analysis of the *cis*-*trans*-isomerization is not feasible with the current data, only the changes in the cs-MSMs exclusively containing *trans*-states are investigated in the next step. Therefore, it is tested whether an exclusion of replicas containing *cis*-states changes the kinetics of the *trans*-states. The results are summarized in figure 4.7 and in the supporting figures 4.14 to 4.16.

For all setups in space 3, comparable kinetics as for space 2 are observed with small differences in the population and the timescales. Both can be explained due to missing replicas containing *cis*-states. Only for CsA in water, a small drop of the slowest implied timescale is observed. However, the underlying dynamics are not affected. The most constant system is CsE in water as the MD data did not show any *cis*-*trans*-isomerization (figure 4.7). A detailed discussion of this behavior is presented in section 4.3.3.2. Based on these findings the *cis*-states were excluded for further analysis (section 4.3.3.3).

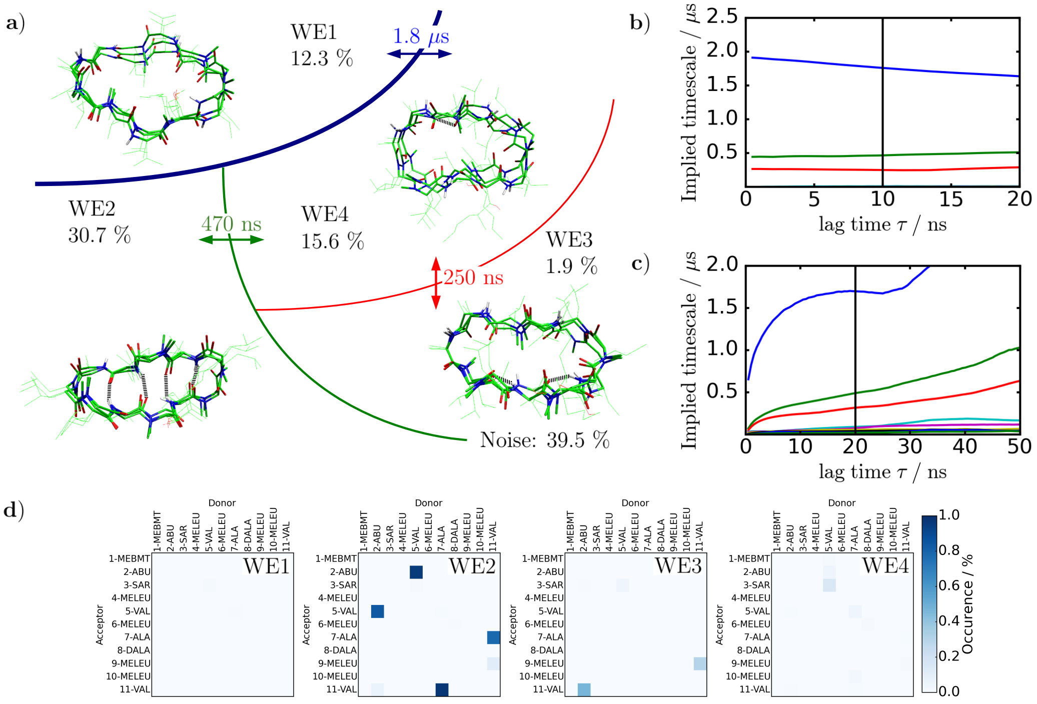


Figure 4.7: Dynamics of CsE (a) in water (Space 3); The configuration of the 9-10-peptide bond is highlighted; The populations are calculated with respect to the clusters (they do not sum up to 100 %); (b-c) Implied timescales in dependence of the lag time τ using different cluster algorithms: (b) CNN and (c) k-Means++; (d) Hydrogen bond population.

Summarizing section 4.3.3.1, it could be observed that the information of the *cis*-*trans*-isomerization of the 9-10-peptide bond has to be treated explicitly in the construction of the reduced state space. Within these reduced state spaces, it is possible to obtain a 100 % separation of *cis*- and *trans*-states as well as a differentiation between different conformations within both classes using a core-set discretization. Due to this good separation, it could also be observed that neglecting the *cis*-states for further analysis no or only small changes for the description of the dynamics of the *trans*-states occur. In a next step, the impact and differences between a core-set and a full-partitioning discretization in the Spaces 1–3 are examined.

4.3.3.2 Comparing Full-partitioning and Core-set Discretization

In the former section, it was shown that using specific input coordinates for the reduced state space construction, it is possible to obtain accurate models matching the behavior observed in the MD data if a core-set discretization is used. In this section, it is examined whether this outcome is also possible with a full-partitioning discretization using the k-Means++ algorithm.

The first investigated space is Space 1, using only the backbone dihedral angles ϕ and ψ. As described in section 4.3.3.1, *cis*-states with a purity comparable to the core-set discretization are obtained. The purity for CsE is 98.1 % in chloroform and 99.7 % in water. For CsA in chloroform, a purity of 67.9 % and in water a purity of 91.0 % was observed. Although the purity of the *cis*-states is comparable, the convergence of the implied timescales is not (figure 4.2d+f for CsE and figure 4.3d+f for CsA). For 3 out of the 4 systems, the implied timescales for the fp-MSM are not converging on a linear scale at all. For the core-set discretization, this was not the case. The only investigated system for which comparable results as for a core-set discretization and well converging implied timescales can be observed is CsA in chloroform. Hence, the cs-MSM outperforms the fp-MSM in Space 1.

For Space 2, *cis*-states with 100 % purity have been observed for all investigated systems using a core-set discretization. For the full-partitioning discretization (figures 4.17 and 4.18), this is also achieved for 2 out of 4 systems: CsE in chloroform and CsA in water. For CsE in water, a *cis*-state with a purity of 99.9 % is obtained. For CsA in chloroform, the PCCA+ algorithm lumped *cis*-states together with *trans*-states resulting in a purity of 70.7 % for this metastable state. The remaining metastable *trans*-states, however, are pure in the case of CsA in chloroform. The same is true for CsE in water. This is not the case for the other systems as for CsA in water two metastable *trans*-states contain *cis*-states to a small amount. For CsE in chloroform, a comparable behavior is observed for one metastable *trans*-state. This analysis again showed that including the 9-10-peptide bond explicitly into the reduced space construction, more accurate reduced spaces with respect to the quality of the MSMs are obtained.

However, for all investigated systems, except for CsA in chloroform, the cs-MSM yielded more accurate models with respect to the reference data than the full-partitioning MSM. As already observed for Space 1, the implied timescales of the fp-MSM are worse or not converging compared to the cs-MSM. Although the slowest implied timescales tend to converge to higher values compared to the cs-MSMs, no convergence of these implied timescales is observed after half of the simulation time of a single replica (50 ns). For the cs-MSM, this is not the case and all implied timescales converged after a few nanoseconds. The reason behind this behavior is recrossing and is discussed in detail in section 4.3.4. Although the purity of the *cis*-states is 100 % in most cases, the “closed” and “open” *cis*-states are not separated well enough in the full-partitioning discretization, since mixing of *cis*- and *trans*-states is observed to a small amount. Hence, recrossing between conformations, which are dynamically not connected in the MD data, is present. As shown in section 4.3.3.1, this issue can be solved using the cs-MSM as core sets do not need any definition of the exact border between two states.

Using the core-set discretization, one dynamically disconnected *cis*-state in either an “open” or a “closed” conformation was observed for each system. For the full-partitioning discretization, this is only achieved for CsA in chloroform. For this system, the same “closed” conformation as for the core-set discretization is found to be not connected in the fp-MSM, which might explain the similar quality of the fp-MSM and the cs-MSM for this system. This is possible, because of well separated minima in the reduced state space of CsA in chloroform for Space 2. To estimate this property, the distance distribution of all data points in this space can be analyzed. If the minima are well separated, sharp peaks with only a poorly populated transition region will be present (figure 4.8b). Hence, the accuracy of the fp-MSM strongly depends on the constitution of the reduced space. If the reduced state space consists of well separated minima, the fp-MSM yields a comparable reproduction of the reference data as the cs-MSMs with respect to the slow dynamic modes and dynamical connectivity.

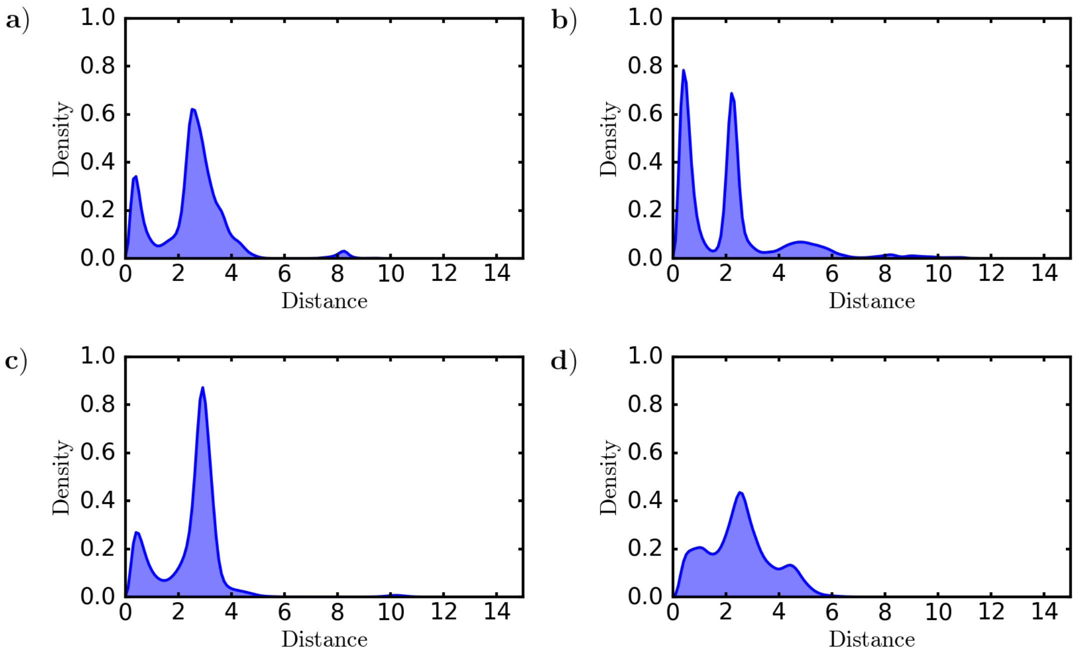


Figure 4.8: Distance distribution in the TICA-space (Space 2) for CsE (a) in chloroform and (c) in water and CsA (b) in chloroform and (d) in water.

If the *cis*-states are excluded in the reduced space construction (Space 3), the quality of the fp-MSMs can be improved. In most cases, except for CsA in chloroform, the slow not-converging implied timescales vanish (figures 4.14c to 4.16c). This observation strengthens the assumption, that they were caused by recrossing of the *cis*-states due to a full-partitioning of the conformational space. However, although some not-converging implied timescales, caused by the badly connected *cis*-states, are missing, the overall convergence of the fp-MSM is worse than for the cs-MSM. This is again caused by recrossing between the remaining *trans*-states, which is not present in a core-set discretization.

4.3.3.3 Comparison of Two Different Models in a Joint State Space

To compare the kinetics of two systems that share either the same molecule or the same solvent, a joint reduced space (Space 4) can be constructed. The condition for this construction is that both compared systems can be described by the same set of input coordinates. Since only the solvent or a backbone methylation is varied in the case of CsA and CsE, it can be used as a strong tool to characterize overlaps within the conformational spaces of these systems. In section 4.3.3.1, it could be shown that the exclusion of *cis*-states does not affect the dynamics between the *trans*-states. Thus, only *trans*-states are taken into account for this analysis.

Performing a clustering on this joint space results in three kinds of clusters: Clusters that are only present for system A, clusters that are only present for system B and clusters that are shared by both systems. Since the interest lies in the dynamics and therefore the dominant metastable states, cs-MSMs for both systems were constructed separately. After lumping the discrete states using PCCA+, the outcome is analyzed with respect to shared and unique conformations. For the analysis, all possible combinations are investigated. The implied timescale tests for the cs-MSMs of each system are depicted in figure 4.19 for spaces sharing the same solvent, and figure 4.20 for spaces sharing the same molecule. These implied timescale tests are comparable with those of Space 3. A detailed discussion can be found in section 4.3.4.

In figure 4.9 the conformations and in figure 4.10 their hydrogen bond pattern are presented. All systems share one “semi-closed” conformation. This conformation is characterized by a partially closing of the C-terminal part of the molecule that is indicated by an hydrogen bond between Ala-7-NH and MeLeu-10-CO (highlighted in blue in figure 4.10). Although this hydrogen bond is not always present, this closing can be observed analyzing the distance between Ala-7-NH and MeLeu-10-CO in all setups. The conformation shows a population of 30 to 60 % and is quite diffuse. In addition to this shared conformation between all four systems, also other shared conformations either present in one solvent or for one molecule can be observed. One state is only present in 3 of the 4 setups and yields another “congruent” conformation for CsA. This state is an "open" conformation showing a relatively large distance between Ala-7-NH and MeLeu-10-CO.

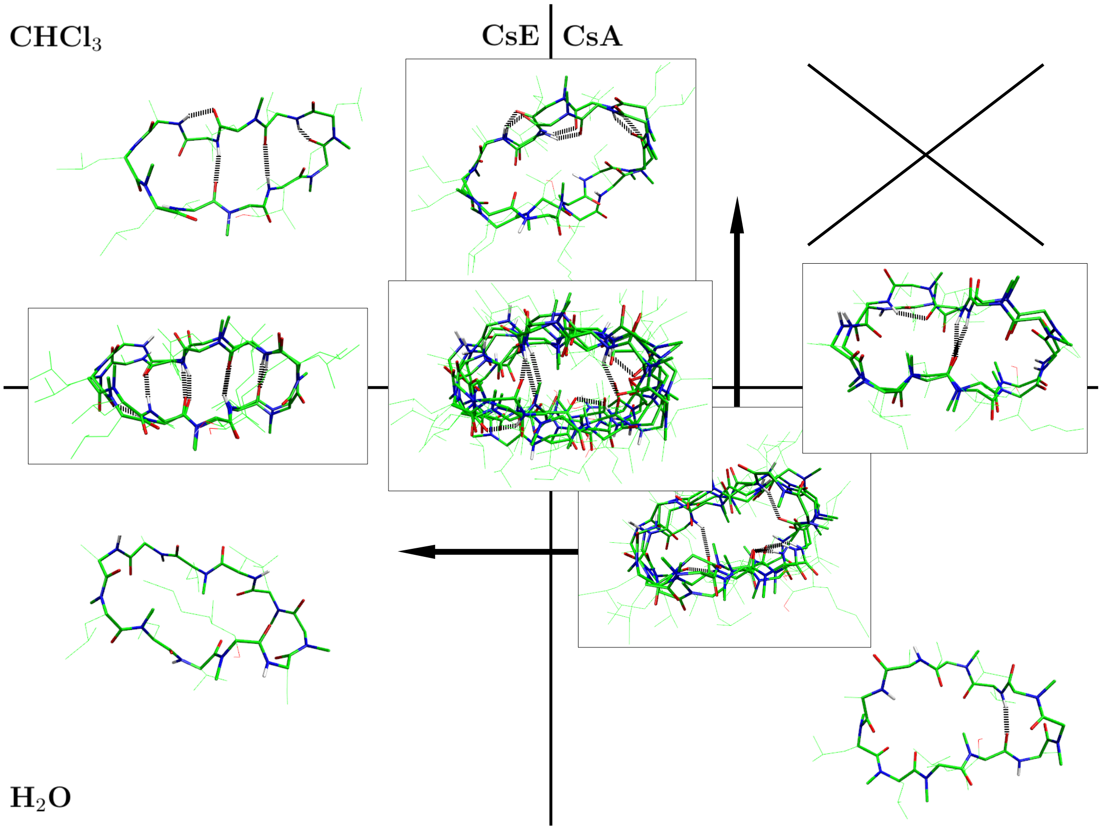


Figure 4.9: Analysis of the joint TICA-spaces; All framed states are shared by at least two systems.

Another "congruent" state is only present in CsA with a population of 20 to 30 %. This state was also observed as a "congruent" state in an earlier analysis [36]. It denotes a "semi-closed" conformation with a hydrogen bond between Val-5-NH and MeLeu-10-CO (highlighted in magenta in figure 4.10). Another shared state could be found for CsE and CsA in chloroform, which is not present in water. This state is an "open" one showing a strongly populated hydrogen bond between DAla-8-NH and MeLeu-6-CO (highlighted in red in figure 4.10).

The same hydrogen bond is also strongly present in the unique, "closed" state of CsE in chloroform, but is not observed for CsE in water. In addition, a strongly populated hydrogen bond is present with Val-11-NH as a donor, which can not be formed in CsA due to the methylation at this position. In chloroform, the hydrogen bond is dominantly formed towards MeLeu-9-CO (highlighted in cyan in figure 4.10), and in water dominantly towards Ala-7-CO (highlighted in orange in figure 4.10). Due to this behavior and the hydrogen bond between DAla-8-NH and MeLeu-6-CO, which might cause this change in hydrogen bonding of Val-11-NH, this “closed” conformation is unique for CsE in chloroform. Nevertheless, a shared, “closed” conformation for CsE is observed as well. Note, that for both “closed” conformations a hydrogen bond between Ala-7-NH and Val-11-CO is observed (highlighted in green in figure 4.10), which might be disfavored in CsA due to steric hindrance.

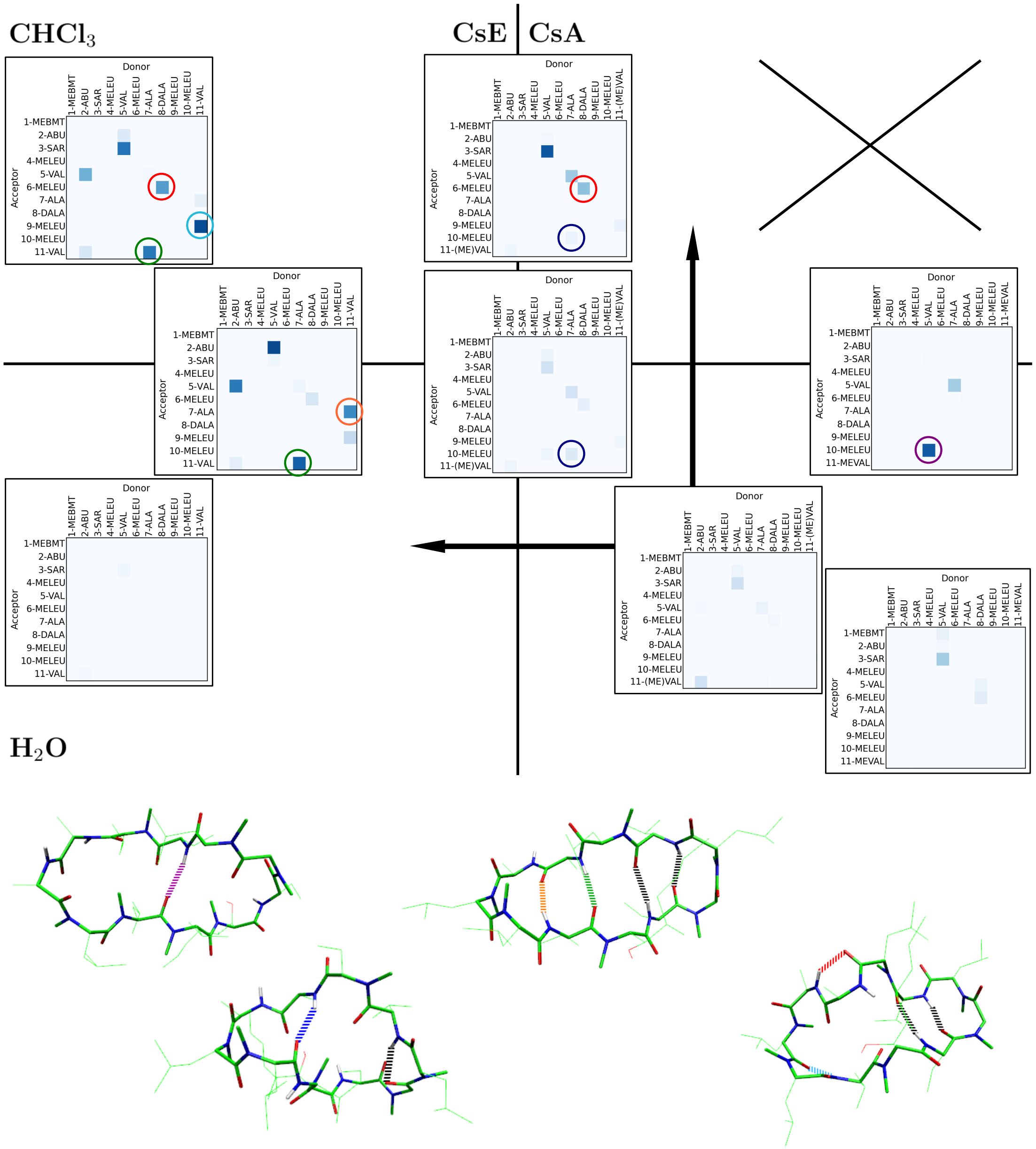


Figure 4.10: Hydrogen bonds of the conformations extracted using a joint discretization; Characteristic hydrogen bonds are highlighted.

Based on this analysis, two “congruent” states for CsE and three for CsA could be observed in total. At this point, it has to be remembered that only *trans*-states were taken into account and that there might be “congruent” *cis*-states as well. Especially, the “closed” *cis*-state, which was present in the simulation data of all four systems, might be a “congruent” state as well.

4.3.4 Discussion

4.3.4.1 Results

In the previous section, it was shown that the choice of the input data is of great importance. Leaving out important reaction coordinates results in a projection error. This error can reduce the quality of the model, since it can cause a loss of important information like the *cis*-*trans*-isomerization. Although it was possible in State 1 to extract different *cis*-states, they were either on a too fast timescale with respect to the *cis*-*trans*-isomerization or merged together using the PCCA+ algorithm. Therefore, they did not match the reference data described in section 4.3.3.

An explicit treatment of the important reaction coordinate resulted in an increased quality of the model, with higher valued and faster converging implied timescales. For some systems, a relatively fast statistical instability was observed for timescales in the μs regime. This might be caused on the one hand by insufficient sampling (total simulation time 10 - 20 μs) or on the other hand by the short length of the single replica (100 ns per replica).

For all systems, it could be observed that the *cis*-states are either not or only poorly connected. Hence, a quantitative interpretation of the implied timescales is not possible as transitions were not sampled well enough. Therefore, for further analysis, trajectories containing *cis*-states were removed. However, it has to be kept in mind that they are important for the dynamics. To circumvent this problem in future, enhanced sampling techniques can be used. However, using these techniques, the dynamic information gets lost. To counteract this, reweighting methods such as introduced in Ref. [84] for MSMs have to be used, which can be a future extension of this work.

Nonetheless, it was shown that if only the *trans*-states are analyzed using a joint discretization, a deeper inside into the dynamics of CsE and CsA can be obtained. In the joint discretization, one “semi-closed” conformation was observed which is present in all setups and can therefore work as a “congruent” state for the diffusion through the membrane. In addition, one has to keep in mind that, although *cis*-states were excluded in the joint model construction, the analysis on Space 2 revealed a “closed” *cis*-state. This *cis*-state featured a comparable hydrogen bond pattern in all systems and could therefore work as a “congruent” state as well. For CsE, an additional hydrogen bond between Val-11-NH and DAla-8-CO, which is not possible for CsA due to the methylation of the backbone amide at Val-11, was present.

For CsA, two additional "congruent" states were observed. These two states (one of them populated with more than 20 % in both solvents, also reported in Ref. [36]) might be one reason for the faster diffusion of CsA since for CsE only one additional "congruent" state could be observed. Another reason for the enhanced membrane permeability of CsA compared to CsE could be revealed analyzing the unique conformation of CsE in chloroform. In this simulation, a "closed" *trans*-state was observed, which was not shared beyond the solvents. Hydrogen bond network analysis showed that it differs from other "closed" *trans*-states by several hydrogen bonds, which can be assumed to be solvent-dependent. Taking into account the presence of a "closed" *cis*-state shared between all four systems, the unique, "closed" *trans*-state might act as an antagonist. For CsA, comparable "closed" conformations are present as well, but only to a small amount (0.6 % for CsA in water and 0.3 % in chloroform). They could be isolated applying a hierarchical cluster approach as introduced in section 4.1, but they were merged to the "congruent" states by PCCA+. Hence, they show a fast change towards other conformations. The reason for this can again be found in the hydrogen bond network. Besides the additional hydrogen bonds in CsE that are possible due to the missing backbone methylation and can stabilize the "closed" conformation, a (nearly) unique hydrogen bond between Val-7-NH and Val-11-CO is observed for CsE, further stabilizing the "closed" *trans*-states.

A third reason for the enhanced membrane permeability of CsA compared to CsE can be observed by analyzing the implied timescales for the formation of the shared "semi-closed" conformation as summarized in table 4.4 and proposed in Ref. [37]. The shared conformation shows a 2.5-time faster formation for CsA than for CsE. This observation strengthens the experimental outcome.

Table 4.4: Implied timescales (Space 3) for the formation of the shared "semi-closed" *trans*-state; For the implied timescales, all replicas containing *cis*-states were rejected due to insufficient sampling.

System	Solvent	Implied timescale (Shared)
CsE	$CHCl_3$	520 ns
	H_2O	250 ns
CsA	$CHCl_3$	200 ns
	H_2O	100 ns

To link the clusters isolated on the basis of a joint discretization (Space 4) to the clusters isolated on a non-joint space (Space 3), one can compare the frames assigned to each cluster, as exemplarily depicted in figure 4.11.

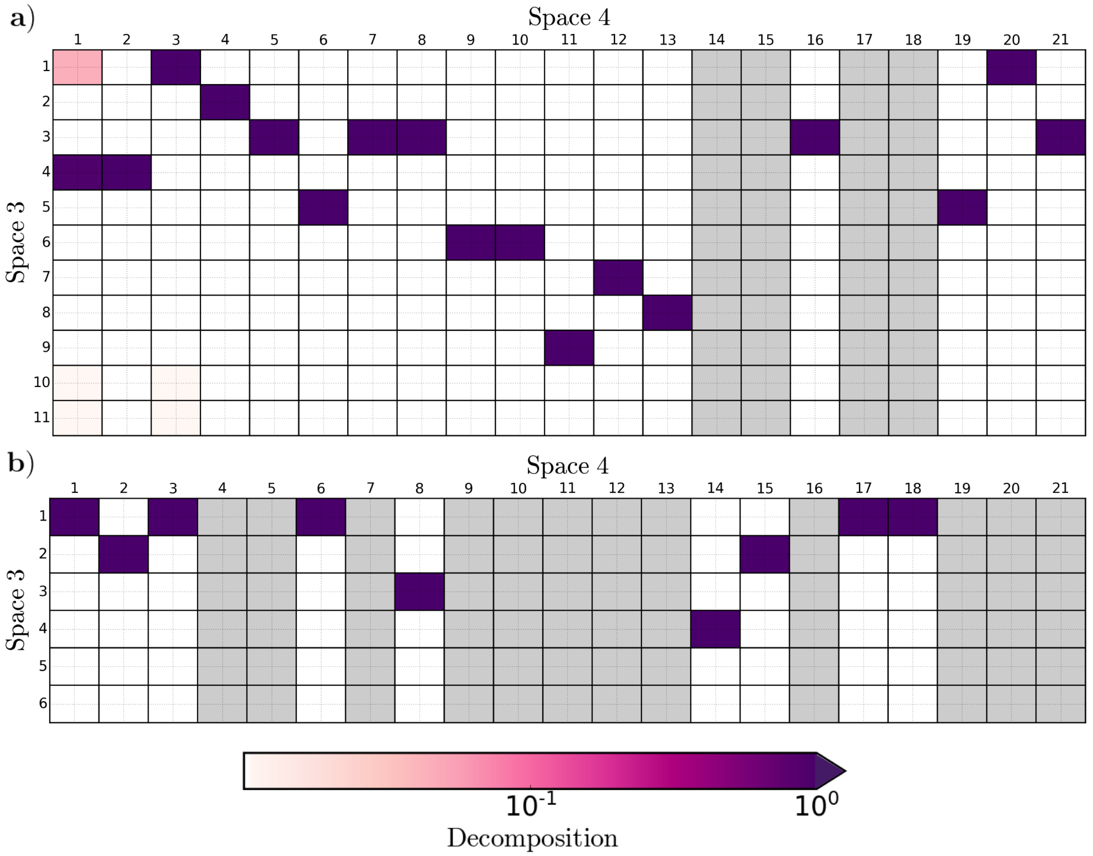


Figure 4.11: Analysis of the cluster decomposition (Space 4): for (a) CsE and (b) CsA in chloroform; As a reference the clusters of Space 3 are used; each column is normalized to 1. Clusters that are not present in one system are highlighted by grey columns.

Applying this kind of analysis results not only in the direct translation of the isolated clusters to earlier isolated structures but also yields three further advantages. First, unique clusters can be observed by one glance as they are not present in the other model (highlighted in figure 4.11). Second, one can observe whether some clusters were lost in the new space. This is, for example, the case in figure 4.11b, where the small clusters 5 and 6 are missing for CsA. Analyzing these clusters, however, showed their minor importance as they are merged to a much larger cluster using PCCA+ and therefore validate the joint discretization. Third, it can be checked whether isolated clusters are a mix of two or more reference clusters. If this is the case, the clustering may not be sufficient enough and has to be repeated or modified including a hierarchical approach. Alternatively, the PCCA+ assignment of the reference set can be checked. If, as in the case presented in figure 4.11a, the not separated clusters are merged together in the PCCA+ analysis of the reference set, the clustering can be seen as sufficient enough.

4.3.4.2 Comparison to Former Models

As stated in the introduction, the same data as in Ref. [36, 37] were used. Thus, a direct comparison of the cs-MSMs presented in this section and the MSMs constructed in Ref. [36, 37] is possible. In contrast to the use of an euclidean distance in a reduced state space, the assignment of discrete states was based on a clustering on the basis of the backbone root-mean-square derivation (RMSD) in these MSMs.

Based on the hydrogen bond network, an assignment of the metastable states to the reference model was possible. For CsE [37] in chloroform, a *cis*-*trans*-isomerization and an interconversion between the "closed" and "open" *cis*-states on timescales of several hundreds ns were reported. In the cs-MSM presented in this section, the latter transition was not observed at all. For the *cis*-*trans*-isomerization, a timescale in the μs regime was achieved.

For CsE in water, no complete *cis*-*trans*-isomerization was observed in the MD-simulation data at all, as discussed in section 4.3.3. In the MSM reported in Ref. [37], however, this isomerization, which was not observed in the cs-MSM, is on a timescale of approximately 200 ns. As this transition was characterized between a "closed" *cis*-state and a "closed" *trans*-state, it can be assumed that it is caused by impurity of the clusters due to the used distance measure. The backbone RMSD might not be well suited to differentiate between a "closed" *cis*-state and a "closed" *trans*-state, since they consist of a comparable atom position pattern. The formation of the "closed" *trans*-state was characterized on a timescale of sereval hundred ns and is comparable to the timescale that was obtained by the cs-MSM.

A comparable behavior for the *cis*-*trans*-isomerization was reported for CsA [36] in chloroform. For the formation of the "closed" *cis*-state an interconversion timescale of approximately 100 ns was found. In the cs-MSM, this interconversion could not be detected. Furthermore, the formation of the "semi-closed" *trans*-state (CA1 in figure 4.12) is one order of magnitude smaller than in the cs-MSM. An indicator that this behavior is caused by the used distance measure can be found, since also in the case of the fp-MSM a comparable timescale like for the cs-MSM was present.

For CsA in water, the formation of the "closed" *cis*-state was again reported to be on a relatively fast timescale (smaller than 100 ns). It is thus one order of magnitude below the value obtained in the cs-MSM. The formation of the "semi-closed" *trans*-state was reported in the same range. In the cs-MSM, this transition could not be detected at all. In the fp-MSM, it was found in the μs-regime. However, since the implied timescale was not converged, this connection can be interpreted as an artifact of the full-partitioning discretization.

For all systems, comparable populations of the metastable states in the fp-MSM and in the MSMs described in Ref. [36, 37] were found with only small deviations. Thus, the improvement of the dynamic behavior is caused by the inclusion of a state space reduction method into the MSM construction. As already discussed in the former section, a further improvement is obtained when a core-set discretization instead of a full-partitioning discretization is used.

4.3.4.3 Model Construction

For the clustering, a screen over different parameter sets was applied to obtain a good cs-MSM. This screen is relatively robust as small differences in the parameters do not affect the dominant kinetics in most cases. However, applying such a scheme has three advantages. First, some clusters can get lost if a too small cutoff parameter is applied as shown in section 4.2. Testing different combinations, however, can circumvent this error. Second, the clusters of different parameter sets can be compared. Using this comparison (analogously to figure 4.11), it is easy to analyze which clusters can be split or which clusters might get lost using another parameter set. With this information, a hierarchical clustering scheme can be applied as introduced in section 4.1 to optimize the clustering and isolate clusters of different data point density. For the systems analyzed in this section, this was not necessary. Third, an evaluation function can be used to determine which parameter set is most promising with respect to cluster number, cluster size, and percentage of data points assigned to noise.

For the construction of a joint discretization, it is harder to find a suitable set of parameters as it has to match with both systems. However, using a parameter screen, this step can be simplified and it was possible to identify at least one suitable parameter set for every combination. The implied timescales of the models constructed in Space 4 (supporting figures 4.19 and 4.20) are comparable to the implied timescales constructed in Space 3. Nonetheless, it can become harder to stabilize transitions, which are badly sampled. For CsE in chloroform (same solvent), one conformation with badly sampled transitions had to be removed manually. However, due to the analysis described in figure 4.11, it was possible to include this conformation in the analysis, such that no information got lost. The same holds for CsA in water, where the “congruent” state was dynamically not connected, but could be identified analyzing the cluster decomposition.

4.3.4.4 Density-based Clustering

The k-Means++ algorithm partitions the data set into N_C Voronoi-cells, where each data point is closer to the centroid of its associated cluster than to the centroid of any other cluster. For this algorithm, the number of clusters N_C has to be defined as an input parameter for the clustering. To obtain a sufficient enough clustering, in most cases N_C has to be set to values with at least three digits. The biggest problem with the k-Means algorithm is the non-deterministic behavior as the initial cluster centroids determine the final outcome. There are several approaches to optimize this algorithm [167, 236, 237]. However, due to its objective function it fails extracting convex or non-spherical clusters [16].

The CNN algorithm is a deterministic clustering algorithm that can isolate non-spherical clusters, as shown in Ref. [16] and sections 4.1 and 4.2. It uses a pairwise density estimation to determine whether two frames belong to the same cluster or not. The pairwise density estimation requires a pairwise distance calculation which is computationally expensive. However, this calculation has to be only done once. The clustering step itself can cluster 10,000 data points within seconds. In section 4.1, an approach using these characteristics was presented. Instead of the full trajectory only a sub part of the sampled conformational space was clustered, typically in the range of 10,000 to 20,000 frames (extracted equidistantly). This approach yields two big advantages: On the one hand, hundreds of parameter sets can be evaluated within minutes. On the other hand, these tests can be run in parallel on multiple CPUs. To assign the complete trajectory to the clusters, the complete trajectory is mapped onto the extracted clusters applying the cluster parameters in a second step. The final outcome assigns all data points either to a cluster or as noise. Hence, they can be treated as core sets. Although the outcome is only an approximation of the clustering of the complete data set, we could show that the approximation is sufficient enough to build cs-MSMs on it, which outperform conventional fp-MSMs.

4.3.4.5 Full-partitioning MSMs versus Core-set MSMs

In section 4.3.3.2, it was shown that cs-MSMs based on the CNN algorithm outperformed fp-MSMs based on the k-Means++ algorithm. If the core sets are defined well, this behavior can be generalized, since cs-MSMs can minimize the discretization error [238]. The reason for this is that only the metastable states have to be defined and no knowledge about the energy barrier is needed. In comparison, in fp-MSMs the position of the energy barriers has to be guessed exactly to minimize this error. The convergence behavior of the implied timescales can be used as an indicator for this error. The larger the error gets, the slower the implied timescales converge. This behavior can, for example, be seen by comparing the cs- and fp-MSMs, as done in section 4.3.3.2. For the fp-MSMs, a worse convergence compared to a cs-MSM or no convergence at all were observed for the slow implied timescales. The reason for this bad convergence are badly connected states as for all systems except for CsA in chloroform all 500 states were connected. Badly assigned borders between the clusters in a full-partitioning discretization generate recrossing and therefore “pseudo-transitions”. As these “pseudo-transitions” are not sampled in the MD simulation and are an artifact of the discretization, they can destabilize the MSM as observed in figure 4.4 and 4.5 as well as figure 4.13. As shown in section 4.3.3.1, this also holds true, if the reduced space is not capable to represent the correct dynamics due to projection errors. In Space 1 for example, impure core sets were obtained for CsE as well as for CsA. These isolated core sets dominantly featured *cis*-states but also contained a small number of *trans*-states. This mixing introduced recrossing, resulting in the count of "pseudo-transitions" also in the case of the core sets.

4.3.5 Supporting Figures

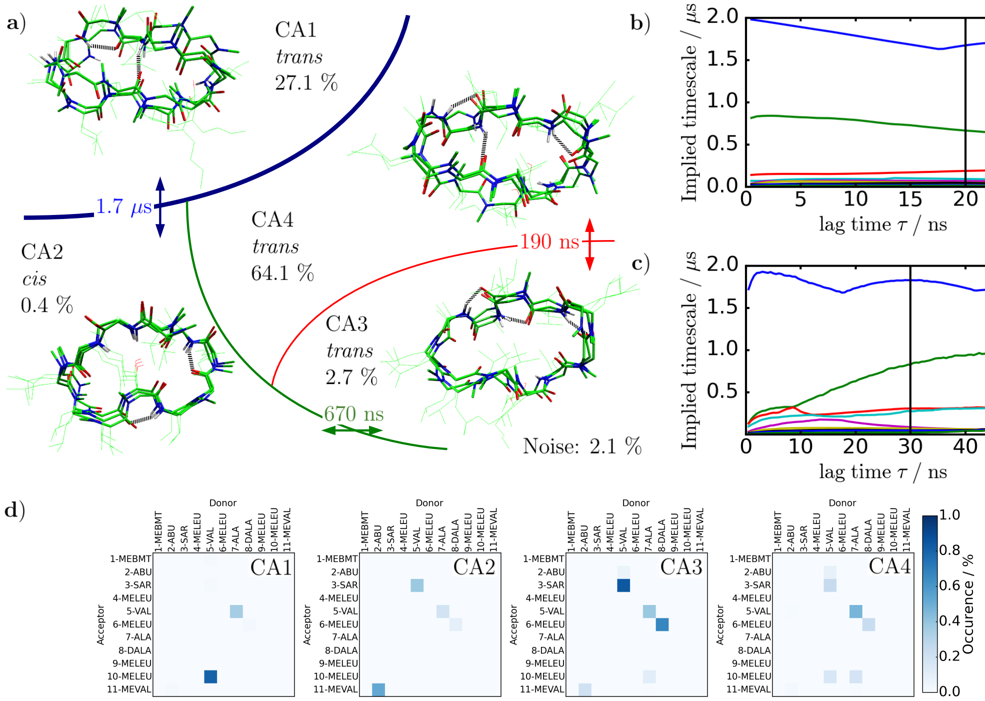


Figure 4.12: Dynamics of CsA (a) in chloroform (Space 2); The configuration of the 9-10-peptide bond is highlighted; The populations are calculated with respect to the clusters (they do not sum up to 100 %); (b-c) Implied timescales in dependence of the lag time τ using different cluster algorithms: (b) CNN and (c) k-Means++; (d) Hydrogen bond population.

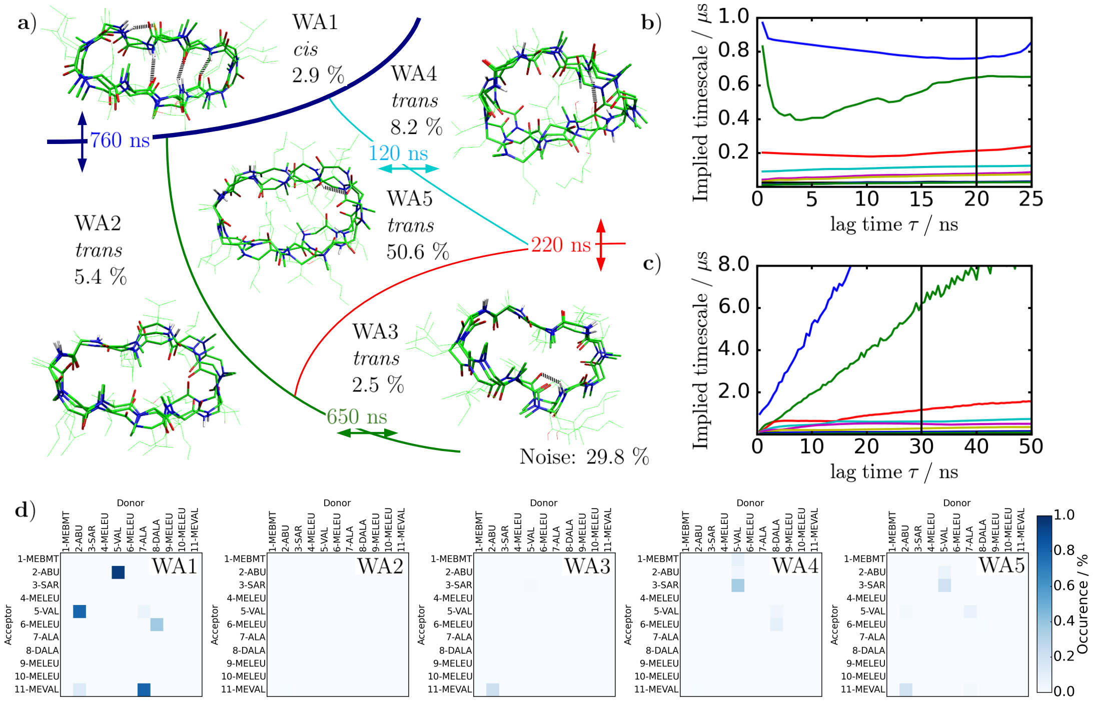


Figure 4.13: Dynamics of CsA (a) in water (Space 2); The configuration of the 9-10-peptide bond is highlighted; The populations are calculated with respect to the clusters (they do not sum up to 100 %); (b-c) Implied timescales in dependence of the lag time τ using different cluster algorithms: (b) CNN and (c) k-Means++; (d) Hydrogen bond population.

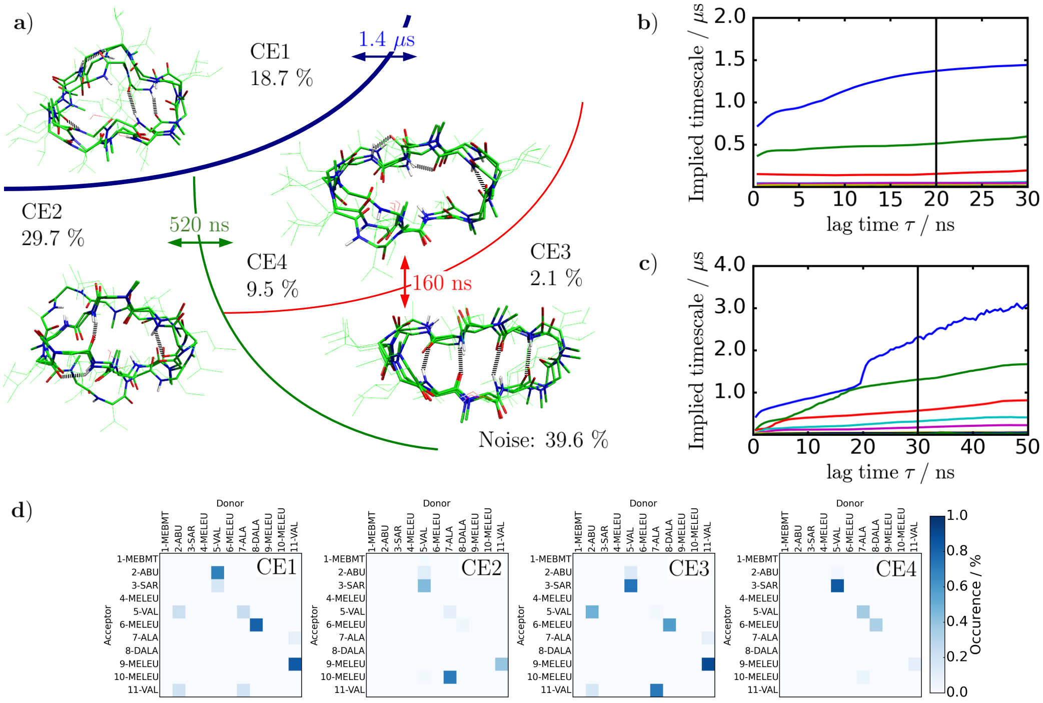


Figure 4.14: Dynamics of CsE (a) in chloroform (Space 3); The configuration of the 9-10-peptide bond is highlighted; The populations are calculated with respect to the clusters (they do not sum up to 100 %); (b-c) Implied timescales in dependence of the lag time τ using different cluster algorithms: (b) CNN and (c) k-Means++; (d) Hydrogen bond population.

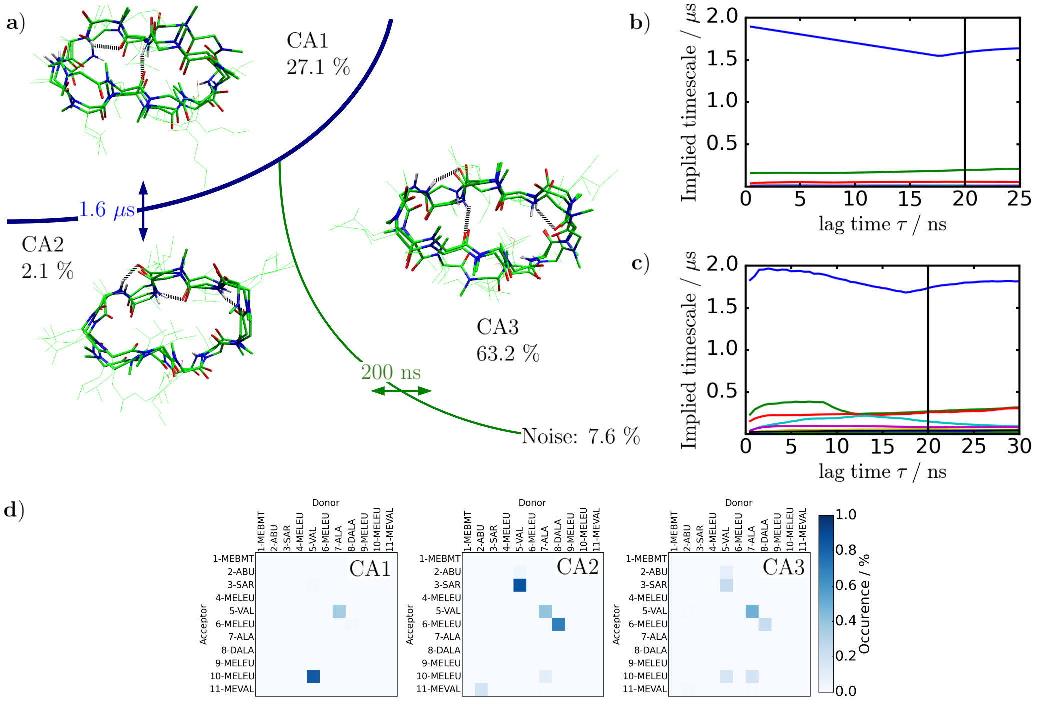


Figure 4.15: Dynamics of CsA (a) in chloroform (Space 3); The configuration of the 9-10-peptide bond is highlighted; The populations are calculated with respect to the clusters (they do not sum up to 100 %); (b-c) Implied timescales in dependence of the lag time τ using different cluster algorithms: (b) CNN and (c) k-Means++; (d) Hydrogen bond population.

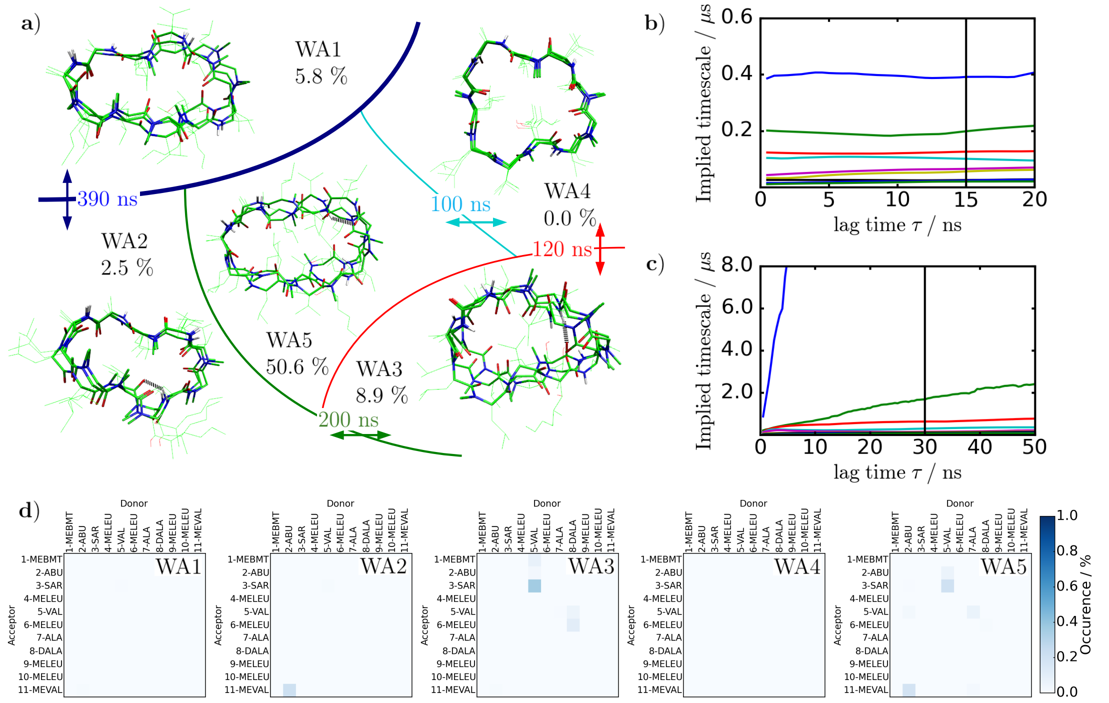


Figure 4.16: Dynamics of CsA (a) in water (Space 3); The configuration of the 9-10-peptide bond is highlighted; The populations are calculated with respect to the clusters (they do not sum up to 100 %); (b-c) Implied timescales in dependence of the lag time τ using different cluster algorithms: (b) CNN and (c) k-Means++; (d) Hydrogen bond population.

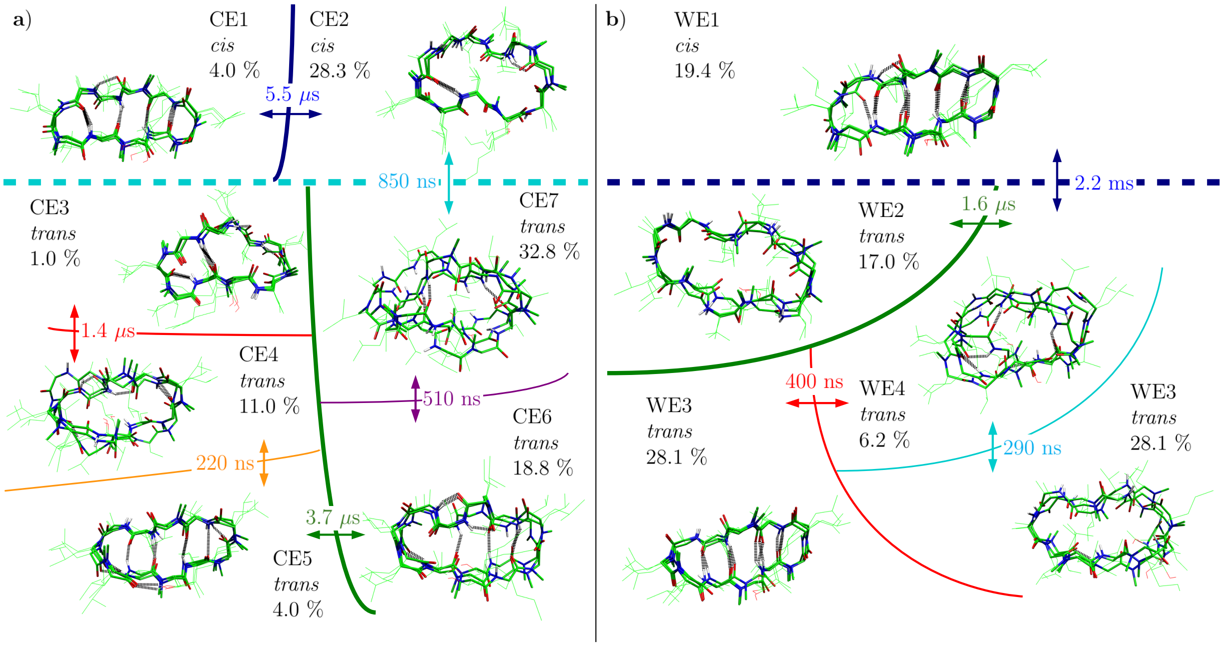


Figure 4.17: Dynamics of CsE (a) in chloroform and (b) in water (Space 2) using the k-Means++ algorithm; The configuration of the 9-10-peptide bond is highlighted.

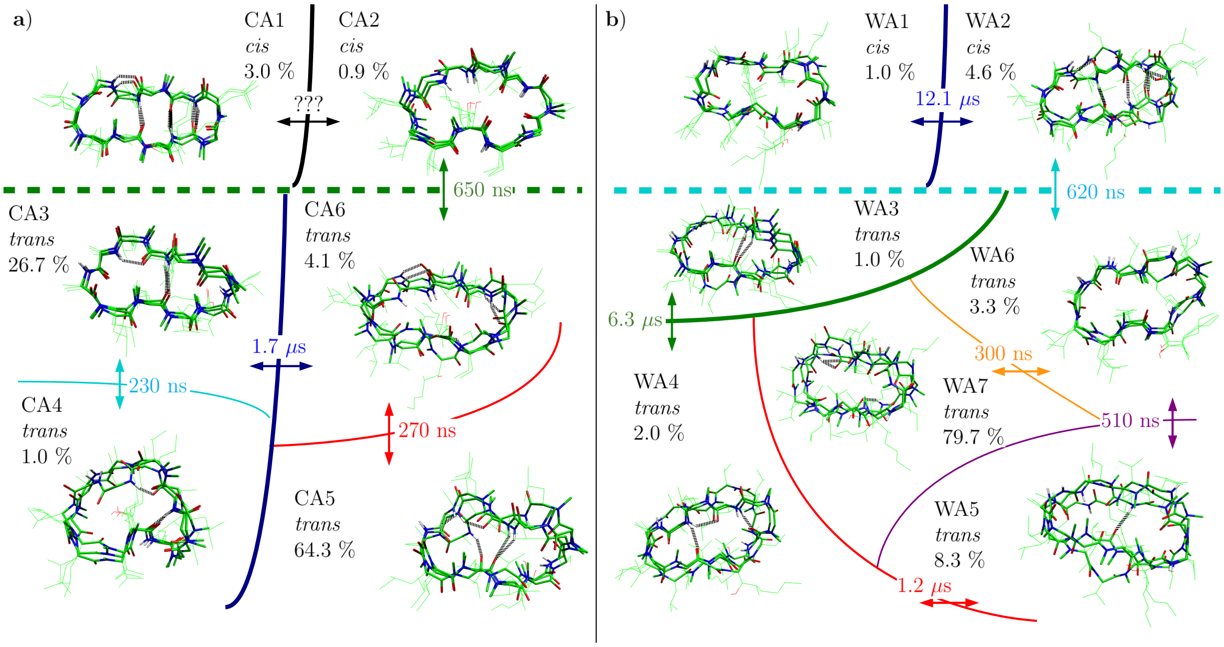


Figure 4.18: Dynamics of CsA (a) in chloroform and (b) in water (Space 2) using the k-Means++ algorithm; The configuration of the 9-10-peptide bond is highlighted.

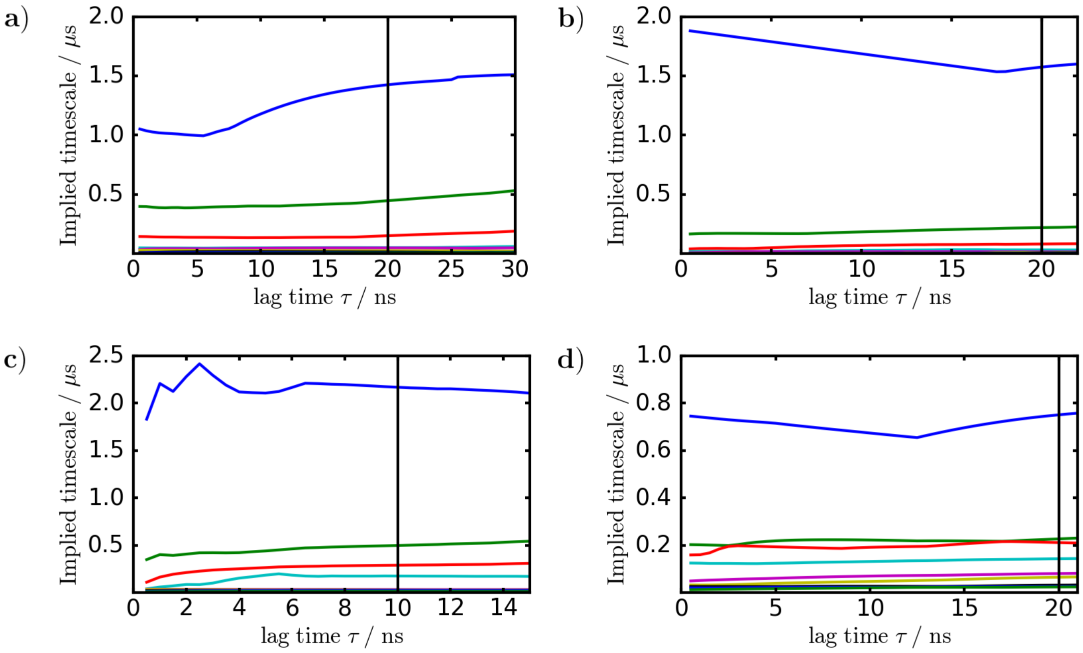


Figure 4.19: Implied timescales for the joint space (Space 4) for CsE (a) in chloroform and (c) in water and CsA (b) in chloroform and (d) in water sharing the same solvent.

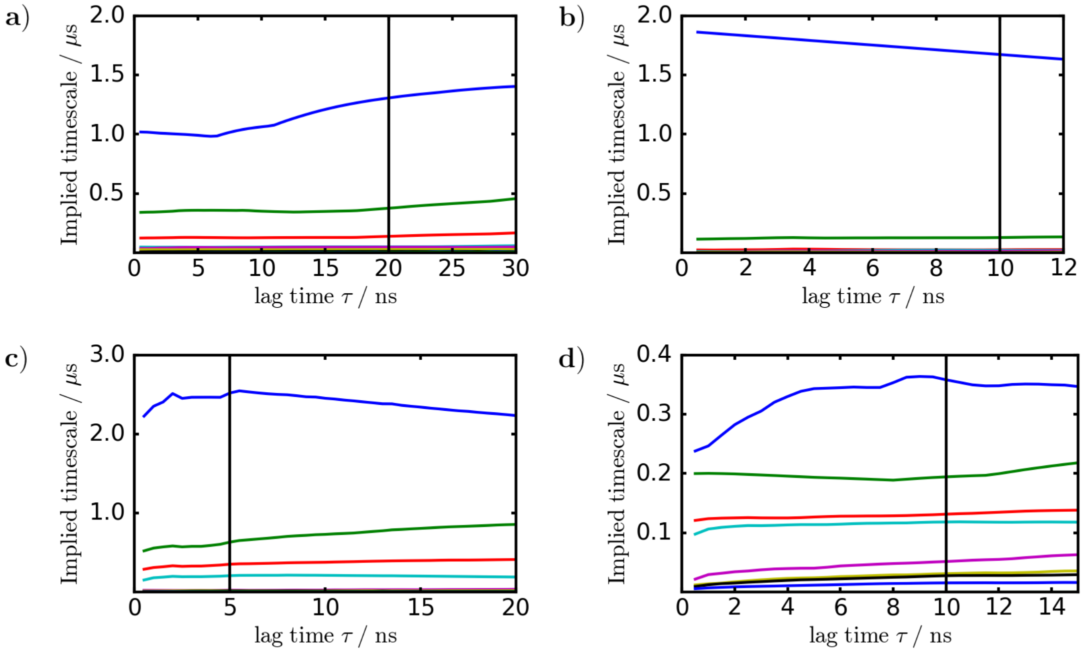


Figure 4.20: Implied timescales for the joint space (Space 4) for CsE (a) in chloroform and (c) in water and CsA (b) in chloroform and (d) in water sharing the same molecule.

5 Molecular Modeling of the Binding Modes of Ligands towards Biomolecules

Since for the function of biomolecules further molecules might be necessary, the cooperation between two or more molecules is of importance. This cooperation can include interactions between two biomolecules such as protein-protein [239] or protein-DNA [240], or interactions between a biomolecule and one or multiple ligands. The latter case includes, for example, enzyme-substrate interactions, which are important for the enzyme's catalytic function as well as specificity [31], or the intercalation of planar molecules into DNA [64]. The prediction of these binding modes and the characterization of important interactions between these molecules thus form another interesting field of research.

The work presented in section 5.1, is focused on modeling the interactions between an enzyme and small organic molecules that can function as a substrate for the enzyme. In section 5.2, different intercalation modes of DNA intercalators that can inhibit DNA replication are discussed.

5.1 Broad Substrate Tolerance of Tubulin Tyrosine Ligase Enables One-step Site-Specific Enzymatic Protein Labeling

Enzymes form a large class of proteins, catalyzing chemical reactions in the cell. They are usually highly specific with respect to the substrates that can bind to their active site. This specificity is caused by a particular arrangement of the amino acids pointing into the active site. The amino acids can serve as hydrogen bond donors or acceptors, can contribute charges or π-stacking interactions, or can create a hydrophobic environment. Due to this high specificity of the active site, the substrate scope of an enzyme is usually quite narrow. Therefore, using enzymes in chemical catalysis is limited to a small set of substrates, without an additional protein engineering of the active site [241].

The enzyme tubulin tyrosine ligase (TTL) shows a remarkable substrate scope towards unnatural amino acids. In cells, TTL ligates a tyrosine-moiety to the C-terminus of a detyrosinated α-tubulin using ATP as a coenzyme [242, 243]. Former studies [31], however, showed that any protein that incorporates a "Tub-tag", a specific amino acid sequence, at its C-Terminus, can be used as a target for ligation of a tyrosine. It was shown that TTL is not only limited to ligate tyrosine since it is also able to add tyrosine-derivatives carrying functional groups. These functional groups can later be used in bioorthogonal reactions to enable further modifications.

The presented study focuses on an expansion of this substrate scope by adding molecules that are larger and sterically more demanding than tyrosine such as a coumarine derivative. The work was a collaboration with the chemical biology group of Prof. Dr. Christian Hackenberger (FMP Berlin, Germany), the biochemical group of Prof. Dr. Heinrich Leonhardt (LMU Munich, Germany) and the biocatalytical group of Prof. Dr. Nediljko Budisa (University of Manitoba, Canada). During this work, molecular docking analysis was utilized to explain the broad substrate scope of TTL. Analyzing the observed docking poses of the ligand, the important interactions present in the active site such as π-stacking and hydrogen bonding, were characterized. Combining molecular docking with MD simulations yielded additional information concerning the stability and the strength of specific interactions.

The presented research was published in: Schumacher, D.; Lemke, O.; Helma, J.; Gerszonowicz, L.; Waller, V.; Stoschek, T.; Durkin, P. M.; Budisa, N.; Leonhardt, H.; Keller, B. G.; Hackenberger, C. P. R. “Broad substrate tolerance of tubulin tyrosine ligase enables one-step site-specific enzymatic protein labeling”, *Chem. Sci.* **2017**, *8*, 3471–3478; doi: 10.1039/c7sc00574a.

5.2 Multiply Intercalator-Substituted Cu(II) Cyclen Complexes as DNA Condensers and DNA/RNA Synthesis Inhibitors

DNA is the information storage within the cells. Errors in the DNA induced by damage or mutations can thus have a big impact on the cell's function. If the damage occurs in replicating cells and is not repaired correctly or not inhibited by cell death, it can cause diseases like cancer [244, 245]. To inhibit the replication of these cells, DNA intercalators can be used. DNA intercalators are small planar molecules that can bind to the DNA by intercalating between two base pairs. The induction of this binding can inhibit DNA replication [65, 66] or, if the intercalator carries a metal ion, cleave the DNA [67–69]. They can also be used as DNA condensing agent [246].

The presented work was a collaboration with the bioinorganic group of Prof. Dr. Nora Kulak (FU Berlin, Germany), the biophysical group of Prof. Dr. Viktor Brabec (Czech Academy of Science, Czech Republic) and the organometallic group of Prof. Dr. Ingo Ott (TU Braunschweig, Germany). It was focused on the ability of anthraquinone-based (AQ-based) DNA intercalators for DNA condensation and DNA replication inhibition. Cu(II)-carrying cyclenes modified with one to three AQ units, were examined with respect to these properties. Different binding modes of these molecules were tested to explain the different activities in DNA condensation and DNA replication inhibition. The studies were based on the generation of different conformations of the intercalator, which were investigated with respect to their DNA intercalation ability. We proposed a *bis*-intercalation binding mode that matches the experimental findings.

The presented research was published in: Hormann, J.; Malina, J.; Lemke, O.; Hülsey, M. J.; Wedepohl, S.; Potthoff, J.; Schmidt, C.; Ott, I.; Keller, B. G.; Brabec, V.; Kulak, N. "Multiply Intercalator-Substituted Cu(II) Cyclen Complexes as DNA Condensers and DNA/RNA Synthesis Inhibitors", *Inorg. Chem.* **2018**, *57*, 5004–5012; doi: 10.1021/acs.inorgchem.8b00027.

6 Conclusion and Outlook

In the presented work, different theoretical methods were applied to obtain information about structural and/or dynamical properties of biomolecules. The investigated biomolecules ranged from small linear or cyclic peptides to larger proteins or DNA. In addition protein/DNA-ligand interactions were examined. The work was split into three main topics:

- In section 3, the dynamical and spectroscopic properties of the water-soluble chlorophyll-binding protein (WSCP) were discussed with respect to its remarkable stability. Using MD simulations, an ensemble of different WSCP conformations was sampled. Including simulations with a varied number of chlorophylls, it was possible to point out the important interactions that stabilize the molecule. In addition to this, it was examined how structural modifications by reconnecting cysteine bridges affect the important interactions, since these cysteine bridges should increase the stability. As no experimental literature was found that discusses the possibility of disulfide-bridge formation, theoretical studies were used to investigate how this structural modification affects the molecule's properties. Subsequently, experiments are necessary to determine whether the found properties can be validated.

  Applying QM/MM and TD-DFT calculations, the spectral properties of the bound chlorophylls (Chls) were examined. Within the course of this analysis, it was shown that the setup for the absorption spectrum calculation is very important, since several factors can influence the position and strength of the excitation. Calculating the collinearity, it was possible to investigate the coupling between the different Chls bound to the WSCP. In this study, it was shown that an optimization at a QM/MM level can strongly influence the absorption spectrum. Optimizing multiple Chls, however, can be computationally expensive, but needs to be done in future. In addition, we developed a mapping scheme comparing different TD-DFT calculations. For a single Chl, this scheme worked well. For more Chls, however, problems due to high dimensionality and degeneracy were detected. Thus, it is of importance to optimize the mapping with respect to these properties in future.

- In section 4, a method to estimate core sets for the construction of core-set Markov state models (cs-MSM) was discussed, which was an unsolved problem for a long time. It was shown that by applying density-based clustering a reliable core-set discretization is obtained since metastable states of the analyzed system can be extracted. Additionally, it was highlighted how cs-MSMs outperform conventional MSMs owing to the fact that the discretization error can be reduced. Recently, several studies were published applying density-based clustering for the definition of core sets [43, 137]. Also related approaches, like dynamical coring, were introduced using density-based clustering [136].

  In this study, the Common-Nearest-Neighbor (CNN) algorithm [16, 135, 166] (available at Ref. [217])

was the most promising algorithm. It is capable of extracting clusters of different size, shape or number, and can also handle different densities by using the introduced hierarchical clustering scheme. The outcome, however, is strongly biased on the used parameter set. We proposed two schemes to find a suitable set of parameters. The next step would be the development of an algorithm that can find an optimal parameter set, since this is strongly data-set dependent. In a further step, these algorithms could be enhanced to determine multiple parameter sets enabling hierarchical clustering. Also the introduction of weights for every data point could enhance the algorithm [247]. Recently, the CNN algorithm was used in a pharmacophore-prediction software [248].

Applying the setup for the construction of cs-MSMs combined with a TICA state space reduction, yielded early converging implied timescales for the cyclosporines A and E. We showed that the information included in the TICA-space construction bias the analysis as leaving out slow reaction coordinates such as the *cis*-*trans*-isomerization of a peptide bond results in a projection error and introduces recrossing. However, as the reaction coordinates are not independent of each other it was still possible to extract conformations mainly in the *cis*-configuration. Including the *cis*-*trans*-isomerization explicitly in the TICA-space construction improved the results significantly. In addition, we highlighted a method to compare both molecules directly by using a joint space describing their properties. During this analysis, it was shown that by using a core-set discretization, disconnections within the sampled conformational space can be detected if the analyzed data consist of multiple simulation trajectories. In some cases this may not be possible, if all data points are assigned to a cluster (full-partitioning). To remove the disconnection either longer simulation times or enhanced sampling techniques such as metadynamics [81–83], are required. Longer simulation times might yield a connection of all data. The slow interconversion times, however, may not be significant, if these interconversions are not sampled several times. Enhanced sampling techniques bias the kinetics. To remove this bias, reweighting methods such as Girsanov reweighting [84, 85] have to be applied.

- In section 5, the important energetic and steric properties for the conversion of non-natural amino acids by the tubulin tyrosine ligase (TTL) were determined. On the basis of this knowledge, it is possible to predict the binding efficiency of other potential ligation candidates. The exception of this are ligands carrying a moiety connected via a long linker. Due to the open binding site, it is hard to predict how the linker would behave. For smaller molecules, however, the knowledge was applied to decide whether the synthesis of an interesting ligand is promising or not.

  Applying theoretical intercalation studies with respect to DNA, it was possible to predict a *bis*-intercalating binding mode, explaining the experimental findings. For this analysis, the structural and energetic properties of both the DNA and the intercalator, have to be taken into account. As this requires complex experiments, a theoretical analysis can be done in a first step. Based on this analysis, the properties can be discussed and it can be decided whether experiments regarding this question are worth of being carried out afterwards.

References

[1] Dror, R. O.; Dirks, R. M.; Grossman, J. P.; Xu, H.; Shaw, D. E. "Biomolecular Simulation: A Computational Microscope for Molecular Biology". *Annu. Rev. Biophys.* **2012**, *41*, 429–452.

[2] Karplus, M.; McCammon, J. A. "Molecular dynamics simulations of biomolecules". *Nature Struct. Biol.* **2002**, *9*, 646–652.

[3] Senn, H. M.; Thiel, W. "QM/MM Methods for Biomolecular Systems". *Angew. Chem. Int. Ed.* **2009**, *48*, 1198–1229.

[4] Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. "Protein folding kinetics and thermodynamics from atomistic simulation". *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 17845–17850.

[5] Florián, J.; Goodman, M. F.; Warshel, A. "Computer simulations of protein functions: Searching for the molecular origin of the replication fidelity of DNA polymerases". *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6819–6824.

[6] Maragakis, P.; Lindorff-Larsen, K.; Eastwood, M. P.; Dror, R. O.; Klepeis, J. L.; Arkin, I. T.; Jensen, M. Ø.; Xu, H.; Trbovic, N.; Friesner, R. A.; Palmer, A. G.; Shaw, D. E. "Microsecond Molecular Dynamics Simulation Shows Effect of Slow Loop Dynamics on Backbone Amide Order Parameters of Proteins". *J. Phys. Chem. B* **2008**, *112*, 6155–6158.

[7] Swope, W. C.; Pitera, J. W.; Suits, F. "Describing Protein Folding Kinetics by Molecular Dynamics Simulations". *J. Phys. Chem. B* **2004**, *108*, 6571–6581.

[8] Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C. "Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics". *J. Chem. Phys.* **2007**, *126*, 155101.

[9] Noé, F.; Horenko, I.; Schütte, C.; Smith, J. C. "Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states". *J. Chem. Phys.* **2007**, *126*, 155102.

[10] Buchete, N.-V.; Hummer, G. "Coarse Master Equations for Peptide Folding Dynamics". *J. Phys. Chem. B* **2008**, *112*, 6057–6069.

[11] Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. "Markov models of molecular kinetics: Generation and validation". *J. Chem. Phys.* **2011**, *134*, 174105.

[12] Sarich, M.; Banisch, R.; Hartmann, C.; Schütte, C. "Markov State Models for Rare Events in Molecular Dynamics". *Entropy* **2014**, *16*, 258–286.

[13] Husic, B. E.; Pande, V. S. “Markov State Models: From an Art to a Science”. *J. Am. Chem. Soc.* **2018**, *140*, 2386–2396.

[14] Schütte, C.; Noé, F.; Lu, J.; Sarich, M.; Vanden-Eijnden, E. “Markov state models based on milestoning”. *J. Chem. Phys.* **2011**, *134*, 204105.

[15] Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noé, F. “PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models”. *J. Chem. Theory Comput.* **2015**, *11*, 5525–5542.

[16] Keller, B.; Daura, X.; van Gunsteren, W. F. “Comparing geometric and kinetic cluster algorithms for molecular simulation data”. *J. Chem. Phys.* **2010**, *132*, 074110.

[17] Jain, A.; Stock, G. “Identifying Metastable States of Folding Proteins”. *J. Chem. Theory Comput.* **2012**, *8*, 3810–3819.

[18] McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. “MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories”. *Biophys. J.* **2015**, *109*, 1528–1532.

[19] Kabsch, W.; Sander, C. “Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features”. *Biopolymers* **1983**, *22*, 2577–2637.

[20] Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. “MDAnalysis: A toolkit for the analysis of molecular dynamics simulations”. *J. Comput. Chem.* **2011**, *32*, 2319–2327.

[21] Shoichet, B. K.; Kuntz, I. D.; Bodian, D. L. “Molecular docking using shape descriptors”. *J. Comput. Chem.* **1992**, *13*, 380–397.

[22] Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. “Improved protein-ligand docking using GOLD”. *Proteins* **2003**, *52*, 609–623.

[23] Brooijmans, N.; Kuntz, I. D. “Molecular Recognition and Docking Algorithms”. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 335–373.

[24] Gaigeot, M.-P. “Theoretical spectroscopy of floppy peptides at room temperature. A DFTMD perspective: gas and aqueous phase”. *Phys. Chem. Chem. Phys.* **2010**, *12*, 3336–3359.

[25] Neugebauer, J. “Subsystem-Based Theoretical Spectroscopy of Biomolecules and Biomolecular Assemblies”. *ChemPhysChem* **2009**, *10*, 3148–3173.

[26] Eisenberg, D.; Jucker, M. “The Amyloid State of Proteins in Human Diseases”. *Cell* **2012**, *148*, 1188–1203.

[27] Selkoe, D. J. "Folding proteins in fatal ways". *Nature* **2003**, *426*, 900–904.

[28] Uversky, V. N.; Oldfield, C. J.; Dunker, A. K. "Intrinsically Disordered Proteins in Human Diseases: Introducing the D^2 Concept". *Annu. Rev. Biophys.* **2008**, *37*, 215–246.

[29] Furuhashi, M.; Hotamisligil, G. S. "Fatty acid-binding proteins: role in metabolic diseases and potential as drug targets". *Nat. Rev. Drug Discov.* **2008**, *7*, 489–503.

[30] Zaks, A.; Klibanov, A. M. "Enzyme-catalyzed processes in organic solvents". *Proc. Natl. Acad. Sci. U.S.A.* **1985**, *82*, 3192–3196.

[31] Schumacher, D.; Helma, J.; Mann, F. A.; Pichler, G.; Natale, F.; Krause, E.; Cardoso, M. C.; Hackenberger, C. P. R.; Leonhardt, H. "Versatile and Efficient Site-Specific Protein Functionalization by Tubulin Tyrosine Ligase". *Angew. Chem. Int. Ed.* **2015**, *54*, 13787–13791.

[32] Meunier, B.; de Visser, S. P.; Shaik, S. "Mechanism of Oxidation Reactions Catalyzed by Cytochrome P450 Enzymes". *Chem. Rev.* **2004**, *104*, 3947–3980.

[33] Takeichi, M. "Functional correlation between cell adhesive properties and some cell surface proteins". *J. Cell Biol.* **1977**, *75*, 464–474.

[34] Somerville, C.; Bauer, S.; Brininstool, G.; Facette, M.; Hamann, T.; Milne, J.; Osborne, E.; Paredez, A.; Persson, S.; Raab, T.; Vorwerk, S.; Youngs, H. "Toward a Systems Approach to Understanding Plant Cell Walls". *Science* **2004**, *306*, 2206–2211.

[35] Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V. "Stereochemistry of polypeptide chain configurations". *J. Mol. Biol.* **1963**, *7*, 95–99.

[36] Witek, J.; Keller, B. G.; Blatter, M.; Meissner, A.; Wagner, T.; Riniker, S. "Kinetic Models of Cyclosporin A in Polar and Apolar Environments Reveal Multiple Congruent Conformational States". *J. Chem. Inf. Model.* **2016**, *56*, 1547–1562.

[37] Witek, J.; Mühlbauer, M.; Keller, B. G.; Blatter, M.; Meissner, A.; Wagner, T.; Riniker, S. "Interconversion Rates between Conformational States as Rationale for the Membrane Permeability of Cyclosporines". *ChemPhysChem* **2017**, *18*, 3309–3314.

[38] Yao, G.; Joswig, J.-O.; Keller, B. G.; Süssmuth, R. D. "Total Synthesis of the Death Cap Toxin Phalloidin: Atropoisomer Selectivity Explained by Molecular-Dynamics Simulations". *Chem. Europ. J.* **2019**, *25*, 8030–8034.

[39] Altstein, M.; Ben-Aziz, O.; Daniel, S.; Schefler, I.; Zeltser, I.; Gilon, C. "Backbone Cyclic Peptide Antagonists, Derived from the Insect Pheromone Biosynthesis Activating Neuropeptide, Inhibit Sex Pheromone Biosynthesis in Moths". *J. Biol. Chem.* **1999**, *274*, 17573–17579.

[40] Hayouka, Z.; Levin, A.; Hurevich, M.; Shalev, D. E.; Loyter, A.; Gilon, C.; Friedler, A. "A comparative study of backbone versus side chain peptide cyclization: Application for HIV-1 integrase inhibitors". *Bioorg. Med. Chem.* **2012**, *20*, 3317–3322.

[41] White, C. J.; Yudin, A. K. "Contemporary strategies for peptide macrocyclization". *Nat. Chem.* **2011**, *3*, 509–524.

[42] Kessler, H. "Conformation and Biological Activity of Cyclic Peptides". *Angew. Chem. Int. Ed.* **1982**, *21*, 512–523.

[43] Witek, J.; Wang, S.; Schroeder, B.; Lingwood, R.; Dounas, A.; Roth, H.-J.; Fouché, M.; Blatter, M.; Lemke, O.; Keller, B.; Riniker, S. "Rationalization of the Membrane Permeability Differences in a Series of Analogue Cyclic Decapeptides". *J. Chem. Inf. Model.* **2019**, *59*, 294–308.

[44] Rezai, T.; Yu, B.; Millhauser, G. L.; Jacobson, M. P.; Lokey, R. S. "Testing the Conformational Hypothesis of Passive Membrane Permeability Using Synthetic Cyclic Peptide Diastereomers". *J. Am. Chem. Soc.* **2006**, *128*, 2510–2511.

[45] Rezai, T.; Bock, J. E.; Zhou, M. V.; Kalyanaraman, C.; Lokey, R. S.; Jacobson, M. P. "Conformational Flexibility, Internal Hydrogen Bonding, and Passive Membrane Permeability: Successful in Silico Prediction of the Relative Permeabilities of Cyclic Peptides". *J. Am. Chem. Soc.* **2006**, *128*, 14073–14080.

[46] Buckton, L. K.; McAlpine, S. R. "Improving the Cell Permeability of Polar Cyclic Peptides by Replacing Residues with Alkylated Amino Acids, Asparagines, and D-Amino Acids". *Org. Lett.* **2018**, *20*, 506–509.

[47] Joo, S. H. "Cyclic Peptides as Therapeutic Agents and Biochemical Tools". *Biomol. Ther. (Seoul)* **2012**, *20*, 19–26.

[48] Kahraman, A.; Thornton, J. M. *"Methods to Characterize the Structure of Enzyme Binding Sites"*; 2008; pp 189–221, in Schwede, T.; Peitsch, M. *Computational Structural Biology*, World Scientific Publishing.

[49] Koshland, D. E. "The Key–Lock Theory and the Induced Fit Theory". *Angew. Chem. Int. Ed. Engl.* **1995**, *33*, 2375–2378.

[50] Koshland, D. E. "Application of a Theory of Enzyme Specificity to Protein Synthesis". *Proc. Natl. Acad. Sci. U.S.A.* **1958**, *44*, 98–104.

[51] Monod, J.; Wyman, J.; Changeux, J.-P. "On the nature of allosteric transitions: A plausible model". *J. Mol. Biol.* **1965**, *12*, 88–118.

[52] Tsai, C.-J.; Ma, B.; Nussinov, R. "Folding and binding cascades: Shifts in energy landscapes". *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9970–9972.

[53] Liu, H.; Jin, L.; Koh, S. B. S.; Atanasov, I.; Schein, S.; Wu, L.; Zhou, Z. H. "Atomic Structure of Human Adenovirus by Cryo-EM Reveals Interactions Among Protein Networks". *Science* **2010**, *329*, 1038–1043.

[54] Topf, M.; Lasker, K.; Webb, B.; Wolfson, H.; Chiu, W.; Sali, A. "Protein Structure Fitting and Refinement Guided by Cryo-EM Density". *Structure* **2008**, *16*, 295–307.

[55] Drenth, J. *Principles of protein X-Ray Crystallography*; Springer Science and Business Media, New York, 2007.

[56] Herrmann, T.; Güntert, P.; Wüthrich, K. "Protein NMR Structure Determination with Automated NOE Assignment Using the New Software CANDID and the Torsion Angle Dynamics Algorithm DYANA". *J. Mol. Biol.* **2002**, *319*, 209–227.

[57] Bax, A.; Grzesiek, S. "Methodological advances in protein NMR". *Acc. Chem. Res.* **1993**, *26*, 131–138.

[58] Yong, W.; Lomakin, A.; Kirkitadze, M. D.; Teplow, D. B.; Chen, S.-H.; Benedek, G. B. "Structure determination of micelle-like intermediates in amyloid β-protein fibril assembly by using small angle neutron scattering". *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 150–154.

[59] Zhang, F.; Skoda, M. W. A.; Jacobs, R. M. J.; Martin, R. A.; Martin, C. M.; Schreiber, F. "Protein Interactions Studied by SAXS: Effect of Ionic Strength and Protein Concentration for BSA in Aqueous Solutions". *J. Phys. Chem. B* **2007**, *111*, 251–259.

[60] Hura, G. L.; Menon, A. L.; Hammel, M.; Rambo, R. P.; Poole II, F. L.; Tsutakawa, S. E.; Jenney Jr, F. E.; Classen, S.; Frankel, K. A.; Hopkins, R. C.; Yang, S.; Scott, J. W.; Dillard, B. D.; Adams, M. W. W.; Tainer, J. A. "Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS)". *Nat. Methods* **2009**, *6*, 606–612.

[61] Hays, F. A.; Teegarden, A.; Jones, Z. J.; Harms, M.; Raup, D.; Watson, J.; Cavaliere, E.; Ho, P. S. "How sequence defines structure: A crystallographic map of DNA structure and conformation". *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 7157–7162.

[62] Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P. *Molecular Biology of the Cell*; Garland Science, 2008.

[63] Horton, H. R.; Moran, L. A.; Scimgeour, K. G.; Perry, M. D.; Rawn, J. D. *Principles of Biochemistry*; Pearson Education, Inc, 2006.

[64] Tse, W. C.; Boger, D. L. “A Fluorescent Intercalator Displacement Assay for Establishing DNA Binding Selectivity and Affinity”. *Acc. Chem. Res.* **2004**, *37*, 61–69.

[65] Woo, S. H.; Sun, N.-J.; Cassady, J. M.; Snapka, R. M. “Topoisomerase II inhibition by aporphine alkaloids”. *Biochem. Pharmacol.* **1999**, *57*, 1141–1145.

[66] Birg, F.; Praseuth, D.; Zerial, A.; Thuong, N. T.; Doan, T. L.; Hélène, C. “Inhibition of Simian Virus 40 DNA replication in CV-1 cells by an oligodeoxynucleotide covalently linked to an intercalating agent”. *Nucleic Acids Res.* **1990**, *18*, 2901–2908.

[67] Sugiura, Y.; Shiraki, T.; Konishi, M.; Oki, T. “DNA intercalation and cleavage of an antitumor antibiotic dynemicin that contains anthracycline and enediyne cores”. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 3831–3835.

[68] Ly, D.; Kan, Y.; Armitage, B.; Schuster, G. B. “Cleavage of DNA by Irradiation of Substituted Anthraquinones: Intercalation Promotes Electron Transfer and Efficient Reaction at GG Steps”. *J. Am. Chem. Soc.* **1996**, *118*, 8747–8748.

[69] Uma Maheswari, P.; Palaniandavar, M. “DNA binding and cleavage properties of certain tetrammine ruthenium(II) complexes of modified 1,10-phenanthrolines – effect of hydrogen-bonding on DNA-binding affinity”. *J. Inorg. Biochem.* **2004**, *98*, 219–230.

[70] Brana, M. F.; Cacho, M.; Gradillas, A.; de Pascual-Teresa, B.; Ramos, A. “Intercalators as Anticancer Drugs”. *Curr. Pharm. Des.* **2001**, *7*, 1745–1780.

[71] Ralph, R. K.; Marshall, B.; Darkin, S. “Anti-cancer drugs which intercalate into DNA: How do they act?”. *Trends Biochem. Sci.* **1983**, *8*, 212–214.

[72] Halgren, T. A.; Damm, W. “Polarizable force fields”. *Curr. Opin. Struct. Biol.* **2001**, *11*, 236–242.

[73] Leach, A. R. *Molecular Modelling*; Addison Wesley Longman, Essex, 2001.

[74] Shaw, D. E.; Dror, R. O.; Salmon, J. K.; Grossman, J. P.; Mackenzie, K. M.; Bank, J. A.; Young, C.; Deneroff, M. M.; Batson, B.; Bowers, K. J.; Chow, E.; Eastwood, M. P.; Ierardi, D. J.; Klepeis, J. L.; Kuskin, J. S.; Larson, R. H.; Lindorff-Larsen, K.; Maragakis, P.; Moraes, M. A.; Piana, S.; Shan, Y.; Towles, B. “Millisecond-scale molecular dynamics simulations on Anton”. In Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis - SC '09, ACM Press, New York. 2009.

[75] Sosnick, T. R.; Mayne, L.; Hiller, R.; Englander, S. W. “The barriers in protein folding”. *Nat. Struct. Biol.* **1994**, *1*, 149–156.

[76] Sugita, Y.; Okamoto, Y. "Replica-exchange molecular dynamics method for protein folding". *Chem. Phys. Lett.* **1999**, *314*, 141–151.

[77] Affentranger, R.; Tavernelli, I.; di Iorio, E. E. "A Novel Hamiltonian Replica Exchange MD Protocol to Enhance Protein Conformational Space Sampling". *J. Chem. Theory Comput.* **2006**, *2*, 217–228.

[78] Andrec, M.; Felts, A. K.; Gallicchio, E.; Levy, R. M. "Protein folding pathways from replica exchange simulations and a kinetic network model". *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6801–6806.

[79] Torrie, G. M.; Valleau, J. P. "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling". *J. Comput. Phys.* **1977**, *23*, 187–199.

[80] Kästner, J. "Umbrella sampling". *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 932–942.

[81] Barducci, A.; Bussi, G.; Parrinello, M. "Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method". *Phys. Rev. Lett.* **2008**, *100*, 020603.

[82] Barducci, A.; Bonomi, M.; Parrinello, M. "Metadynamics". *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 826–843.

[83] Laio, A.; Gervasio, F. L. "Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science". *Rep. Prog. Phys* **2008**, *71*, 126601.

[84] Donati, L.; Hartmann, C.; Keller, B. G. "Girsanov reweighting for path ensembles and Markov state models". *J. Chem. Phys.* **2017**, *146*, 244112.

[85] Donati, L.; Keller, B. G. "Girsanov reweighting for metadynamics simulations". *J. Chem. Phys.* **2018**, *149*, 072335.

[86] Miao, Y.; Sinko, W.; Pierce, L.; Bucher, D.; Walker, R. C.; McCammon, J. A. "Improved Reweighting of Accelerated Molecular Dynamics Simulations for Free Energy Calculation". *J. Chem. Theory Comput.* **2014**, *10*, 2677–2689.

[87] Zhang, J.; Zhang, H.; Wu, T.; Wang, Q.; van der Spoel, D. "Comparison of Implicit and Explicit Solvent Models for the Calculation of Solvation Free Energy in Organic Solvents". *J. Chem. Theory Comput.* **2017**, *13*, 1034–1043.

[88] Rudd, R. E.; Broughton, J. Q. "Coarse-grained molecular dynamics and the atomic limit of finite elements". *Phys. Rev. B* **1998**, *58*, R5893–R5896.

[89] Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S.-J. "The MARTINI Coarse-Grained Force Field: Extension to Proteins". *J. Chem. Theory Comput.* **2008**, *4*, 819–834.

[90] Stansfeld, P. J.; Sansom, M. S. P. "From Coarse Grained to Atomistic: A Serial Multiscale Approach to Membrane Protein Simulations". *J. Chem. Theory Comput.* **2011**, *7*, 1157–1166.

[91] Vitalini, F.; Mey, A. S. J. S.; Noé, F.; Keller, B. G. "Dynamic properties of force fields". *J. Chem. Phys.* **2015**, *142*, 084101.

[92] Banks, J. L.; Kaminski, G. A.; Zhou, R.; Mainz, D. T.; Berne, B. J.; Friesner, R. A. "Parametrizing a polarizable force field from ab initio data. I. The fluctuating point charge model". *J. Chem. Phys.* **1999**, *110*, 741–754.

[93] van der Kamp, M. W.; Mulholland, A. J. "Combined Quantum Mechanics/Molecular Mechanics (QM/MM) Methods in Computational Enzymology". *Biochemistry* **2013**, *52*, 2708–2728.

[94] Groenhof, G. *"Introduction to QM/MM Simulations"*; 2013; pp 43–66, in Monticelli, M.; Salonen, E. *Biomolecular Simulations. Methods in Molecular Biology (Methods and Protocols)*, Humana Press, Totowa.

[95] Warshel, A.; Levitt, M. "Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme". *J. Mol. Biol.* **1976**, *103*, 227–249.

[96] Lin, H.; Truhlar, D. G. "QM/MM: what have we learned, where are we, and where do we go from here?". *Theor. Chem. Acc.* **2007**, *117*, 185–199.

[97] Melo, M. C. R.; Bernardi, R. C.; Rudack, T.; Scheurer, M.; Riplinger, C.; Phillips, J. C.; Maia, J. D. C.; Rocha, G. B.; Ribeiro, J. V.; Stone, J. E.; Neese, F.; Schulten, K.; Luthey-Schulten, Z. "NAMD goes quantum: an integrative suite for hybrid simulations". *Nat. Methods* **2018**, *15*, 351–354.

[98] Sherwood, P.; de Vries, A. H.; Guest, M. F.; Schreckenbach, G.; Catlow, C. R. A.; French, S. A.; Sokol, A. A.; Bromley, S. T.; Thiel, W.; Turner, A. J.; Billeter, S.; Terstegen, F.; Thiel, S.; Kendrick, J.; Rogers, S. C.; Casci, J.; Watson, M.; King, F.; Karlsen, E.; Sjøvoll, M.; Fahmi, A.; Schäfer, A.; Lennartz, C. "QUASI: A general purpose implementation of the QM/MM approach and its application to problems in catalysis". *J. Mol. Struct.-THEOCHEM* **2003**, *632*, 1–28.

[99] Prinz, J.-H.; Keller, B.; Noé, F. "Probing molecular kinetics with Markov models: metastable states, transition pathways and spectroscopic observables". *Phys. Chem. Chem. Phys.* **2011**, *13*, 16912.

[100] Keller, B.; Hünenberger, P.; van Gunsteren, W. F. "An Analysis of the Validity of Markov State Models for Emulating the Dynamics of Classical Molecular Systems and Ensembles". *J. Chem. Theory Comput.* **2011**, *7*, 1032–1044.

[101] Schütte, C.; Sarich, M. "A critical appraisal of Markov state models". *Eur. Phys. J. Spec. Top.* **2015**, *224*, 2445–2462.

[102] Schütte, C.; Fischer, A.; Huisinga, W.; Deuflhard, P. "A Direct Approach to Conformational Dynamics Based on Hybrid Monte Carlo". *J. Comput. Phys.* **1999**, *151*, 146–168.

[103] Bacallado, S.; Chodera, J. D.; Pande, V. "Bayesian comparison of Markov models of molecular dynamics with detailed balance constraint". *J. Chem. Phys.* **2009**, *131*, 045106.

[104] Chodera, J. D.; Noé, F. "Markov state models of biomolecular conformational dynamics". *Curr. Opin. Struct. Biol.* **2014**, *25*, 135–144.

[105] Senne, M.; Trendelkamp-Schroer, B.; Mey, A. S. J. S.; Schütte, C.; Noé, F. "EMMA: A Software Package for Markov Model Building and Analysis". *J. Chem. Theory Comput.* **2012**, *8*, 2223–2238.

[106] Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. "MSM-Builder2: Modeling Conformational Dynamics on the Picosecond to Millisecond Scale". *J. Chem. Theory Comput.* **2011**, *7*, 3412–3419.

[107] Sarich, M.; Schütte, C. "Approximating selected non-dominant timescales by Markov state models". *Comm. Math. Sci.* **2012**, *10*, 1001–1013.

[108] E, W.; Vanden-Eijnden, E. "Transition-Path Theory and Path-Finding Algorithms for the Study of Rare Events". *Annu. Rev. Phys. Chem.* **2010**, *61*, 391–420.

[109] Faradjian, A. K.; Elber, R. "Computing time scales from reaction coordinates by milestoning". *J. Chem. Phys.* **2004**, *120*, 10880.

[110] Vanden-Eijnden, E.; Venturoli, M.; Ciccotti, G.; Elber, R. "On the assumptions underlying milestoning". *J. Chem. Phys.* **2008**, *129*, 174102.

[111] Vanden-Eijnden, E.; Venturoli, M. "Markovian milestoning with Voronoi tessellations". *J. Chem. Phys.* **2009**, *130*, 194101.

[112] Pérez-Hernández, G.; Paul, F.; Giorgino, T.; de Fabritiis, G.; Noé, F. "Identification of slow molecular order parameters for Markov model construction". *J. Chem. Phys.* **2013**, *139*, 015102.

[113] Krivov, S. V.; Karplus, M. "Hidden complexity of free energy surfaces for peptide (protein) folding". *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 14766–14770.

[114] Noé, F.; Fischer, S. "Transition networks for modeling the kinetics of conformational change in macromolecules". *Curr. Opin. Struct. Biol.* **2008**, *18*, 154–162.

[115] Pearson F.R.S., K. "LIII. On lines and planes of closest fit to systems of points in space". *Lond. Edinb. Dubl. Phil. Mag.* **1901**, *2*, 559–572.

[116] David, C. C.; Jacobs, D. J. *"Principal Component Analysis: A Method for Determining the Essential Dynamics of Proteins"*; 2014; pp 193–226, in Livesay, D. *Protein Dynamics. Methods in Molecular Biology (Methods and Protocols)*, Humana Press, Totowa.

[117] Schwantes, C. R.; Pande, V. S. "Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9". *J. Comp. Theory Comput.* **2013**, *9*, 2000–2009.

[118] Shao, J.; Tanner, S. W.; Thompson, N.; Cheatham, T. E. "Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms". *J. Chem. Theory Comput.* **2007**, *3*, 2312–2334.

[119] Wolf, A.; Kirschner, K. N. "Principal component and clustering analysis on molecular dynamics data of the ribosomal L11·23S subdomain". *J. Mol. Model.* **2013**, *19*, 539–549.

[120] Karpen, M. E.; Tobias, D. J.; Brooks III, C. L. "Statistical clustering techniques for the analysis of long molecular dynamics trajectories: analysis of 2.2-ns trajectories of YPGDV". *Biochemistry* **1993**, *32*, 412–420.

[121] Beyer, K.; Goldstein, J.; Ramakrishnan, R.; Shaft, U. *"When Is "Nearest Neighbor" Meaningful?"*; 1999; pp 217–235, in Beeri, C.; Buneman P.; *Database Theory – ICDT'99. ICDT 1999. Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg.

[122] Lloyd, S. "Least squares quantization in PCM". *IEEE Trans. Inf. Theory* **1982**, *28*, 129–137.

[123] Kaufmann, L.; Rousseeuw, P. *Data Analysis based on the L1-Norm and Related Methods*; 1987; pp 405–416.

[124] Ng, R. T.; Han, J. "Efficient and Effective Clustering Methods for Spatial Data Mining". In Proceeding VLDB '94 Proceedings of the 20th International Conference on Very Large Data Bases, Morgan Kaufmann Publisher, San Francisco. 1994; pp 144–155.

[125] Zhang, T.; Ramakrishnan, R.; Livny, M. "BIRCH: An Efficient Data Clustering Method for Very Large Databases". *ACM SIGMOD Record* **1996**, *25*, 103–114.

[126] Guha, S.; Rastogi, R.; Shim, K. "CURE: An Efficient Clustering Algorithm for Large Databases". *ACM SIGMOD Record* **1998**, *27*, 73–84.

[127] Dunn, J. C. "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters". *J. Cybern.* **1973**, *3*, 32–57.

[128] Dave, R. N.; Bhaswan, K. “Adaptive fuzzy c-shells clustering and detection of ellipses”. *IEEE Trans. Neural Netw.* **1992**, *3*, 643–662.

[129] Jarvis, R. A.; Patrick, E. A. “Clustering Using a Similarity Measure Based on Shared Near Neighbors”. *IEEE Trans. Comput.* **1973**, *C-22*, 1025–1034.

[130] Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”. In Proceedings of the Second International Conference of Knowledge Discovery and Data Mining, AAAI Press, Portland. 1996; pp 226–231.

[131] Ankerst, M.; Breunig, M. M.; Kriegel, H.-P.; Sander, J. “OPTICS: Ordering Points To Identify the Clustering Structure”. *ACM SIGMOD Record* **1999**, *28*, 49–60.

[132] Comaniciu, D.; Meer, P. “Mean shift: a robust approach toward feature space analysis”. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 603–619.

[133] Rodriguez, A.; Laio, A. “Clustering by fast search and find of density peaks”. *Science* **2014**, *344*, 1492–1496.

[134] Xu, D.; Tian, Y. “A Comprehensive Survey of Clustering Algorithms”. *Ann. Data Sci.* **2015**, *2*, 165–193.

[135] Lemke, O.; Keller, B. G. “Density-based cluster algorithms for the identification of core sets”. *J. Chem. Phys.* **2016**, *145*, 164104.

[136] Nagel, D.; Weber, A.; Lickert, B.; Stock, G. “Dynamical coring of Markov state models”. *J. Chem. Phys.* **2019**, *150*, 094111.

[137] Pinamonti, G.; Paul, F.; Noé, F.; Rodriguez, A.; Bussi, G. “The mechanism of RNA base fraying: Molecular dynamics simulations analyzed with core-set Markov state models”. *J. Chem. Phys.* **2019**, *150*, 154123.

[138] Kriegel, H.-P.; Kröger, P.; Sander, J.; Zimek, A. “Density-based clustering”. *WIREs Data Min. Knowl.* **2011**, *1*, 231–240.

[139] Karypis, G.; Han, E.-H.; Kumar, V. “Chameleon: A hierarchical clustering algorithm using dynamic modeling”. *IEEE Trans. Comput.* **1999**, *32*, 68–75.

[140] Marques, M. A. L.; López, X.; Varsano, D.; Castro, A.; Rubio, A. “Time-Dependent Density-Functional Approach for Biological Chromophores: The Case of the Green Fluorescent Protein. *Phys. Rev. Lett.* **2003**, *90*, 258101.

[141] Götze, J. P.; Kröner, D.; Banerjee, S.; Karasulu, B.; Thiel, W. "Carotenoids as a Shortcut for Chlorophyll Soret-to-Q Band Energy Flow". *ChemPhysChem* **2014**, *15*, 3392–3401.

[142] Yuriev, E.; Ramsland, P. A. "Latest developments in molecular docking: 2010-2011 in review". *J. Mol. Recognit.* **2013**, *26*, 215–239.

[143] Shoichet, B. K. "Virtual screening of chemical libraries". *Nature* **2004**, *432*, 862–865.

[144] de Ruyck, J.; Brysbaert, G.; Blossey, R.; Lensink, M. F. "Molecular docking as a popular tool in drug design, an in silico travel". *Adv. Appl. Bioinform. Chem.* **2016**, *9*, 1–11.

[145] Pagadala, N. S.; Syed, K.; Tuszynski, J. "Software for molecular docking: a review". *Biophys. Rev.* **2017**, *9*, 91–102.

[146] Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. "A review of protein-small molecule docking methods". *J. Comput. Aided Mol. Des.* **2002**, *16*, 151–166.

[147] Carlson, H. A.; McCammon, J. A. "Accommodating protein flexibility in computational drug design". *Mol. Pharmacol.* **2000**, *57*, 213–218.

[148] Pagano, B.; Martino, L.; Randazzo, A.; Giancola, C. "Stability and Binding Properties of a Modified Thrombin Binding Aptamer". *Biophys. J.* **2008**, *94*, 562–569.

[149] Seeliger, D.; de Groot, B. L. "Ligand docking and binding site analysis with PyMOL and Autodock/Vina". *J. Comput. Aided Mol. Des.* **2010**, *24*, 417–422.

[150] Alonso, H.; Bliznyuk, A. A.; Gready, J. E. "Combining docking and molecular dynamic simulations in drug design". *Med. Res. Rev.* **2006**, *26*, 531–568.

[151] Frenkel, D.; Smit, B. *Understanding Molecular Simulations*; Academic Press, San Diego, 2002.

[152] Verlet, L. "Computer "Experiment" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules". *Phys. Rev.* **1967**, *159*, 98–103.

[153] Hockney, R. W.; Goel, S. P.; Eastwood, J. W. "Quiet high-resolution computer models of a plasma". *J. Comput. Phys.* **1974**, *14*, 148–158.

[154] Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. "Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes". *J. Comput. Phys.* **1977**, *23*, 327–341.

[155] Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. "LINCS: A linear constraint solver for molecular simulations". *J. Comput. Chem.* **1997**, *18*, 1463–1472.

[156] Andersen, H. C. “Molecular dynamics simulations at constant pressure and/or temperature”. *J. Chem. Phys.* **1980**, *72*, 2384–2393.

[157] Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. “Molecular dynamics with coupling to an external bath”. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

[158] Bussi, G.; Donadio, D.; Parrinello, M. “Canonical sampling through velocity rescaling”. *J. Chem. Phys.* **2007**, *126*, 014101.

[159] Parrinello, M.; Rahman, A. “Polymorphic transitions in single crystals: A new molecular dynamics method”. *J. Appl. Phys.* **1981**, *52*, 7182–7190.

[160] Thiel, W. *“QM/MM Methodology: Fundamentals, Scope, and Limitations”*; 2009; pp 203–214, in Grotendorst, J.; Attig, N.; Blügel, S.; Marx, D. *Multiscale Simulation Methods in Molecular Sciences*, NIC Series, Jülich.

[161] Metz, S.; Kästner, J.; Sokol, A. A.; Keal, T. W.; Sherwood, P. “ChemShell—a modular software package for QM/MM simulations”. *WIREs Comput. Mol. Sci.* **2014**, *4*, 101–110.

[162] Pérez-Hernández, G.; Noé, F. “Hierarchical Time-Lagged Independent Component Analysis: Computing Slow Modes and Reaction Coordinates for Large Molecular Systems”. *J. Chem. Theory Comput.* **2016**, *12*, 6118–6129.

[163] Ritz, W. “Über eine neue Methode zur Lösung gewisser Variationsprobleme der mathematischen Physik.”. *J. Reine Angew. Math. (Crelle's Journal)* **1909**, *1909*, 1–61.

[164] Nüske, F.; Keller, B. G.; Pérez-Hernández, G.; Mey, A. S. J. S.; Noé, F. “Variational Approach to Molecular Kinetics”. *J. Chem. Theory Comput.* **2014**, *10*, 1739–1752.

[165] Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. “Equation of State Calculations by Fast Computing Machines”. *J. Chem. Phys.* **1953**, *21*, 1087–1092.

[166] Lemke, O.; Keller, B. G. “Common Nearest Neighbor Clustering—A Benchmark”. *Algorithms* **2018**, *11*, 19.

[167] Arthur, D.; Vassilvitskii, S. “K-means++ : the advantages of careful seeding”. In SODA ’07 Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms, Society for Industrial and Applied Mathematics, New Orleans. 2007; pp 1027–1035.

[168] Ball, G. H.; Hall, D. J. “A clustering technique for summarizing multivariate data”. *Behav. Sci.* **1967**, *12*, 153–155.

[169] Fränti, P.; Sieranoja, S. "How much can k-means be improved by using better initialization and repeats?". *Pattern Recognit.* **2019**, *93*, 95–112.

[170] Burke, K.; Wagner, L. O. "DFT in a nutshell". *Int. J. Quantum Chem.* **2013**, *113*, 96–101.

[171] Koch, W.; Holthausen, M. C. *A Chemist's Guide to Density Functional Theory*; Wiley-VCH Verlag GmbH, Weinheim, 2000.

[172] Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. "Atoms, molecules, solids, and surfaces: Applications of the generalized gradient approximation for exchange and correlation". *Phys. Rev. B* **1992**, *46*, 6671–6687.

[173] Gross, E. K. U.; Maitra, N. T. *"Introduction to TDDFT"*; 2012; pp 53–99, in Marques, M.; Maitra, N.; Nogueira F.; Gross, E.; Rubio, A. *Fundamentals of Time-Dependent Density Functional Theory. Lecture Notes in Physics*, Springer, Berlin, Heidelberg.

[174] Petersilka, M.; Gossmann, U. J.; Gross, E. K. U. "Excitation Energies from Time-Dependent Density-Functional Theory". *Phys. Rev. Lett.* **1996**, *76*, 1212–1215.

[175] Runge, E.; Gross, E. K. U. "Density-Functional Theory for Time-Dependent Systems". *Phys. Rev. Lett.* **1984**, *52*, 997–1000.

[176] Marques, M. A. L.; Ullrich, C. A.; Nogueira, F.; Rubio, A.; Burke, K.; Gross, E. K. U. *Time-Dependent Density Functional Theory*; Springer, Berlin, Heidelberg, 2006.

[177] Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. "Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function". *J. Comput. Chem.* **1998**, *19*, 1639–1662.

[178] Huey, R.; Morris, G. M.; Olson, A. J.; Goodsell, D. S. "A semiempirical free energy force field with charge-based desolvation". *J. Comput. Chem.* **2007**, *28*, 1145–1152.

[179] Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. "AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility". *J. Comput. Chem.* **2009**, *30*, 2785–2791.

[180] Horigome, D.; Satoh, H.; Itoh, N.; Mitsunaga, K.; Oonishi, I.; Nakagawa, A.; Uchida, A. "Structural Mechanism and Photoprotective Function of Water-soluble Chlorophyll-binding Protein". *J. Biol. Chem.* **2007**, *282*, 6525–6531.

[181] Satoh, H.; Uchida, A.; Nakayama, K.; Okada, M. "Water-Soluble Chlorophyll Protein in Brassicaceae Plants is a Stress-Induced Chlorophyll-Binding Protein". *Plant Cell Physiol.* **2001**, *42*, 906–911.

[182] Palm, D. M.; Agostini, A.; Tenzer, S.; Gloeckle, B. M.; Werwie, M.; Carbonera, D.; Paulsen, H. "Water-Soluble Chlorophyll Protein (WSCP) Stably Binds Two or Four Chlorophylls". *Biochemistry* **2017**, *56*, 1726–1736.

[183] Bektas, I.; Fellenberg, C.; Paulsen, H. "Water-soluble chlorophyll protein (WSCP) of Arabidopsis is expressed in the gynoecium and developing silique". *Planta* **2012**, *236*, 251–259.

[184] Kamimura, Y.; Mori, T.; Yamasaki, T.; Katoh, S. "Isolation, Properties and a Possible Function of a Water-Soluble Chlorophyll a/b-Protein from Brussels Sprouts". *Plant Cell Physiol.* **1997**, *38*, 133–138.

[185] Schmidt, K.; Fufezan, C.; Krieger-Liszkay, A.; Satoh, H.; Paulsen, H. "Recombinant Water-Soluble Chlorophyll Protein from Brassica oleracea Var. Botrys Binds Various Chlorophyll Derivatives". *Biochemistry* **2003**, *42*, 7427–7433.

[186] Satoh, H.; Nakayama, K.; Okada, M. "Molecular Cloning and Functional Expression of a Water-soluble Chlorophyll Protein, a Putative Carrier of Chlorophyll Molecules in Cauliflower". *J. Biol. Chem.* **1998**, *273*, 30568–30575.

[187] Agostini, A.; Palm, D. M.; Schmitt, F.-J.; Albertini, M.; Di Valentin, M.; Paulsen, H.; Carbonera, D. "An unusual role for the phytyl chains in the photoprotection of the chlorophylls bound to Water-Soluble Chlorophyll-binding Proteins". *Sci. Rep.* **2017**, *7*, 7504.

[188] Palm, D. M.; Agostini, A.; Pohland, A.-C.; Werwie, M.; Jaenicke, E.; Paulsen, H. "Stability of Water-Soluble Chlorophyll Protein (WSCP) Depends on Phytyl Conformation". *ACS Omega* **2019**, *4*, 7971–7979.

[189] Takahashi, S.; Yanai, H.; Nakamaru, Y.; Uchida, A.; Nakayama, K.; Satoh, H. "Molecular Cloning, Characterization and Analysis of the Intracellular Localization of a Water-Soluble Chl-Binding Protein from Brussels Sprouts (Brassica oleracea var. gemmifera)". *Plant Cell Physiol.* **2012**, *53*, 879–891.

[190] Takahashi, S.; Aizawa, K.; Nakayama, K.; Satoh, H. "Water-soluble chlorophyll-binding proteins from Arabidopsis thaliana and Raphanus sativus target the endoplasmic reticulum body". *BMC Res. Notes* **2015**, *8*, 365.

[191] Gall, A.; Berera, R.; Alexandre, M. T. A.; Pascal, A. A.; Bordes, L.; Mendes-Pinto, M. M.; Andrianambinintsoa, S.; Stoitchkova, K. V.; Marin, A.; Valkunas, L.; Horton, P.; Kennis, J. T. M.; van Grondelle, R.; Ruban, A.; Robert, B. "Molecular Adaptation of Photoprotection: Triplet States in Light-Harvesting Proteins". *Biophys. J.* **2011**, *101*, 934–942.

[192] Peterman, E. J.; Dukker, F. M.; van Grondelle, R.; van Amerongen, H. “Chlorophyll a and carotenoid triplet states in light-harvesting complex II of higher plants”. *Biophys. J.* **1995**, *69*, 2670–2678.

[193] Croce, R.; Weiss, S.; Bassi, R. “Carotenoid-binding Sites of the Major Light-harvesting Complex II of Higher Plants”. *J. Biol. Chem.* **1999**, *274*, 29613–29623.

[194] Siefermann-Harms, D. “Carotenoids in photosynthesis. I. Location in photosynthetic membranes and light-harvesting function”. *Biochim. Biophys. Acta Rev. Bioenerg.* **1985**, *811*, 325–355.

[195] Renger, G.; Pieper, J.; Theiss, C.; Trostmann, I.; Paulsen, H.; Renger, T.; Eichler, H. J.; Schmitt, F.-J. “Water soluble chlorophyll binding protein of higher plants: A most suitable model system for basic analyses of pigment–pigment and pigment–protein interactions in chlorophyll protein complexes”. *J. Plant Physiol.* **2011**, *168*, 1462–1472.

[196] Gradinaru, C. C.; van Stokkum, I. H. M.; Pascal, A. A.; van Grondelle, R.; van Amerongen, H. “Identifying the Pathways of Energy Transfer between Carotenoids and Chlorophylls in LHCII and CP29. A Multicolor, Femtosecond Pump-Probe Study”. *J. Phys. Chem. B* **2000**, *104*, 9330–9342.

[197] Strain, H. H.; Thomas, M. R.; Katz, J. J. “Spectral absorption properties of ordinary and fully deuteriated chlorophylls a and b”. *Biochim. Biophys. Acta* **1963**, *75*, 306–311.

[198] Theiss, C.; Trostmann, I.; Andree, S.; Schmitt, F. J.; Renger, T.; Eichler, H. J.; Paulsen, H.; Renger, G. “Pigment-Pigment and Pigment-Protein Interactions in Recombinant Water-Soluble Chlorophyll Proteins (WSCP) from Cauliflower”. *J. Phys. Chem. B* **2007**, *111*, 13325–13335.

[199] van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. “GROMACS: Fast, flexible, and free”. *J. Comput. Chem.* **2005**, *26*, 1701–1718.

[200] Yanai, T.; Tew, D. P.; Handy, N. C. “A new hybrid exchange–correlation functional using the Coulomb-attenuating method (CAM-B3LYP)”. *Chem. Phys. Lett.* **2004**, *393*, 51–57.

[201] Francl, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; DeFrees, D. J.; Pople, J. A. “Self-consistent molecular orbital methods. XXIII. A polarization-type basis set for second-row elements”. *J. Chem. Phys.* **1982**, *77*, 3654–3665.

[202] Hariharan, P. C.; Pople, J. A. “The influence of polarization functions on molecular orbital hydrogenation energies”. *Theor. Chim. Acta* **1973**, *28*, 213–222.

[203] Petersson, G. A.; Bennett, A.; Tensfeldt, T. G.; Al-Laham, M. A.; Shirley, W. A.; Mantzaris, J. “A complete basis set model chemistry. I. The total energies of closed-shell atoms and hydrides of the first-row elements”. *J. Chem. Phys.* **1988**, *89*, 2193–2218.

[204] Petersson, G. A.; Al-Laham, M. A. "A complete basis set model chemistry. II. Open-shell systems and the total energies of the first-row atoms". *J. Chem. Phys.* **1991**, *94*, 6081–6090.

[205] Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. "Gaussian 16 Revision B.01". 2016; Gaussian Inc. Wallingford CT.

[206] Bauernschmitt, R.; Ahlrichs, R. "Treatment of electronic excitations within the adiabatic approximation of time dependent density functional theory". *Chem. Phys. Lett.* **1996**, *256*, 454–464.

[207] Stratmann, R. E.; Scuseria, G. E.; Frisch, M. J. "An efficient implementation of time-dependent density-functional theory for the calculation of excitation energies of large molecules". *J. Chem. Phys.* **1998**, *109*, 8218–8224.

[208] Casida, M. E.; Jamorski, C.; Casida, K. C.; Salahub, D. R. "Molecular excitation energies to high-lying bound states from time-dependent densityfunctional response theory: Characterization and correction of the time-dependent local density approximation ionization threshold". *J. Chem. Phys.* **1998**, *108*, 4439–4449.

[209] Caillie, C. V.; Amos, R. D. "Geometric derivatives of excitation energies using SCF and DFT". *Chem. Phys. Lett.* **1999**, *308*, 249–255.

[210] Caillie, C. V.; Amos, R. D. "Geometric derivatives of density functional theory excitation energies using gradient-corrected functionals". *Chem. Phys. Lett.* **2000**, *317*, 159–164.

[211] Furche, F.; Ahlrichs, R. "Adiabatic time-dependent density functional methods for excited state properties". *J. Chem. Phys.* **2002**, *117*, 7433–7447.

[212] Scalmani, G.; Frisch, M. J.; Mennucci, B.; Tomasi, J.; Cammi, R.; Barone, V. "Geometries and properties of excited states in the gas phase and in solution: Theory and application of a time-

dependent density functional theory polarizable continuum model". *J. Chem. Phys.* **2006**, *124*, 094107.

[213] Kasha, M.; Rawls, H. R.; El-Bayoumi, M. A. "The exciton model in molecular spectroscopy". *Pure Appl. Chem.* **1965**, *11*, 371–392.

[214] Renger, T.; Trostmann, I.; Theiss, C.; Madjet, M. E.; Richter, M.; Paulsen, H.; Eichler, H. J.; Knorr, A.; Renger, G. "Refinement of a Structural Model of a Pigment-Protein Complex by Accurate Optical Line Shape Theory and Experiments". *J. Phys. Chem. B* **2007**, *111*, 10487–10501.

[215] Vitalini, F.; Noé, F.; Keller, B. G. "A Basis Set for Peptides for the Variational Approach to Conformational Kinetics". *J. Chem. Theory Comput.* **2015**, *11*, 3992–4004.

[216] Pande, V. S.; Beauchamp, K.; Bowman, G. R. "Everything you wanted to know about Markov State Models but were afraid to ask". *Methods* **2010**, *52*, 99–105.

[217] Lemke, O.; Keller, B. G. "CNNClustering". `https://github.com/BDGSoftware/CNNClustering`, 2017.

[218] Driggers, E. M.; Hale, S. P.; Lee, J.; Terrett, N. K. "The exploration of macrocycles for drug discovery — an underexploited structural class". *Nat. Rev. Drug Discov.* **2008**, *7*, 608–624.

[219] Horne, W. S. "Peptide and peptoid foldamers in medicinal chemistry". *Expert Opin. Drug Discov.* **2011**, *6*, 1247–1262.

[220] Marsault, E.; Peterson, M. L. "Macrocycles Are Great Cycles: Applications, Opportunities, and Challenges of Synthetic Macrocycles in Drug Discovery". *J. Med. Chem.* **2011**, *54*, 1961–2004.

[221] Mallinson, J.; Collins, I. "Macrocycles in new drug discovery". *Future Med. Chem.* **2012**, *4*, 1409–1438.

[222] Amidon, G. L.; Lee, H. J. "Absorption of Peptide and Peptidomimetic Drugs". *Annu. Rev. Pharmacol. Toxicol.* **1994**, *34*, 321–341.

[223] Rüegger, A.; Kuhn, M.; Lichti, H.; Loosli, H.-R.; Huguenin, R.; Quiquerez, C.; von Wartburg, A. "Cyclosporin A, ein immunsuppressiv wirksamer Peptidmetabolit aus Trichoderma polysporum (LINK ex PERS.) Rifai". *Helv. Chim. Acta* **1976**, *59*, 1075–1092.

[224] Wenger, R. M. "Synthesis of Cyclosporine and Analogues: Structural Requirements for Immunosuppressive Activity". *Angew. Chem. Int. Ed. Engl.* **1985**, *24*, 77–85.

[225] Handschumacher, R. E.; Harding, M. E.; Rice, J.; Drugge, R. J.; Speicher, D. W. "Cyclophilin: a specific cytosolic binding protein for cyclosporin A". *Science* **1984**, *226*, 544–547.

[226] Liu, J.; Farmer Jr., J. D.; Lane, W. S.; Friedman, J.; Weissman, I.; Schreiber, S. L. "Calcineurin is a common target of cyclophilin-cyclosporin A and FKBP-FK506 complexes". *Cell* **1991**, *66*, 807–815.

[227] Sultan, M. M.; Pande, V. S. "tICA-Metadynamics: Accelerating Metadynamics by Using Kinetically Selected Collective Variables". *J. Chem. Theory Comput.* **2017**, *13*, 2440–2447.

[228] Zhou, H.; Dong, Z.; Verkhivker, G.; Zoltowski, B. D.; Tao, P. "Allosteric mechanism of the circadian protein Vivid resolved through Markov state model and machine learning analysis". *PLOS Comput. Biol.* **2019**, *15*, e1006801.

[229] Sengupta, U.; Carballo-Pacheco, M.; Strodel, B. "Automated Markov state models for molecular dynamics simulations of aggregation and self-assembly". *J. Chem. Phys.* **2019**, *150*, 115101.

[230] Scott, W. R. P.; Hünenberger, P. H.; Tironi, I. G.; Mark, A. E.; Billeter, S. R.; Fennen, J.; Torda, A. E.; Huber, T.; Krüger, P.; van Gunsteren, W. F. "The GROMOS Biomolecular Simulation Program Package". *J. Phys. Chem. A* **1999**, *103*, 3596–3607.

[231] Schmid, N.; Christ, C. D.; Christen, M.; Eichenberger, A. P.; van Gunsteren, W. F. "Architecture, implementation and parallelisation of the GROMOS software for biomolecular simulation". *Comput. Phys. Commun.* **2012**, *183*, 890–903.

[232] Schmid, N.; Eichenberger, A. P.; Choutko, A.; Riniker, S.; Winger, M.; Mark, A. E.; van Gunsteren, W. F. "Definition and testing of the GROMOS force-field versions 54A7 and 54B7". *Eur. Biophys. J.* **2011**, *40*, 843–856.

[233] Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. *"Interaction Models for Water in Relation to Protein Hydration"*; 1981; pp 331–342, in Pullman B. *Intermolecular Forces. The Jerusalem Symposia on Quantum Chemistry and Biochemistry*, Springer, Dodrecht.

[234] Tironi, I. G.; van Gunsteren, W. F. "A molecular dynamics simulation study of chloroform". *Mol. Phys.* **1994**, *83*, 381–403.

[235] Deuflhard, P.; Weber, M. "Robust Perron cluster analysis in conformation dynamics". *Linear Algebra Appl.* **2005**, *398*, 161–184.

[236] Gonzalez, T. F. "Clustering to minimize the maximum intercluster distance". *Theor. Comput. Sci.* **1985**, *38*, 293–306.

[237] Steinley, D. "Local Optima in K-Means Clustering: What You Don't Know May Hurt You.". *Psychol. Methods* **2003**, *8*, 294–304.

[238] Sarich, M.; Noé, F.; Schütte, C. “On the Approximation Quality of Markov State Models”. *Multiscale Model. Simul.* **2010**, *8*, 1154–1177.

[239] Klippel, S.; Wieczorek, M.; Schümann, M.; Krause, E.; Marg, B.; Seidel, T.; Meyer, T.; Knapp, E.-W.; Freund, C. “Multivalent Binding of Formin-binding Protein 21 (FBP21)-Tandem-WW Domains Fosters Protein Recognition in the Pre-spliceosome”. *J. Biol. Chem.* **2011**, *286*, 38478–38487.

[240] Pabo, C. O.; Sauer, R. T. “Protein-DNA Recognition”. *Ann. Rev. Biochem.* **1984**, *53*, 293–321.

[241] Liu, D. S.; Nivón, L. G.; Richter, F.; Goldman, P. J.; Deerinck, T. J.; Yao, J. Z.; Richardson, D.; Phipps, W. S.; Ye, A. Z.; Ellisman, M. H.; Drennan, C. L.; Baker, D.; Ting, A. Y. “Computational design of a red fluorophore ligase for site-specific protein labeling in living cells”. *Proc. Natl. Acad. Sci. U.S.A.* **2014**, *111*, E4551–E4559.

[242] Szyk, A.; Deaconescu, A. M.; Piszczek, G.; Roll-Mecak, A. “Tubulin tyrosine ligase structure reveals adaptation of an ancient fold to bind and modify tubulin”. *Nat. Struct. Mol. Biol.* **2011**, *18*, 1250–1258.

[243] Ersfeld, K. “Characterization of the tubulin-tyrosine ligase”. *J. Cell Biol.* **1993**, *120*, 725–732.

[244] Griffiths, A. J. F.; Miller, J. H.; Suzuki, D. T.; Lewontin, R. C.; Gelbart, W. M. *An Introduction to Genetic Analysis*; W. H. Freeman, New York, 2000.

[245] O’Connor, M. J. “Targeting the DNA Damage Response in Cancer”. *Mol. Cell* **2015**, *60*, 547–560.

[246] Kapuscinski, J.; Darzynkiewicz, Z. “Condensation of nucleic acids by intercalating aromatic cations”. *Proc. Natl. Acad. Sci. U.S.A.* **1984**, *81*, 7368–7372.

[247] Ackerman, M.; Ben-David, S.; Brânzei, S.; Loker, D. “Weighted Clustering”. In AAAI’12 Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, AAAI Press, Toronto. 2012; pp 858–863.

[248] Mortier, J.; Dhakal, P.; Volkamer, A. “Truly Target-Focused Pharmacophore Modeling: A Novel Tool for Mapping Intermolecular Surfaces”. *Molecules* **2018**, *23*, 1959.

Selbstständigkeitserklärung

Hierdurch versichere ich, dass ich meine Dissertation mit dem Titel “Theoretical Analysis of Biomolecular Systems: Computational Simulations, Core-set Markov State Models, Clustering, Molecular Docking” selbstständig und ohne unerlaubte Hilfe angefertigt und nur die aufgeführten Hilfsmittel und Quellen verwendet habe.

Zusätzlich versichere ich, dass ich meine Dissertation nicht schon einmal in einem anderen Promotionsverfahren eingereicht habe.